piranka/Getty Images

CHAPTER 7

Inference for Means

Introduction

We began our study of data analysis in Chapter 1 by learning graphical and numerical tools for describing the distribution of a single variable and for comparing several distributions. Our study of the practice of statistical inference begins in the same way, with inference about a single distribution and comparison of two distributions. Comparing more than two distributions requires more elaborate methods, which are presented in Chapters 12 and 13.

Two important aspects of any distribution are its center and spread. If the distribution is Normal, we describe its center by the mean μ and its spread by the standard deviation σ.

In this chapter, we will meet confidence intervals and significance tests for inference about a population mean μ and the difference between two population means $\mu_1 - \mu_2$. Chapter 6 emphasized the reasoning of significance tests and confidence intervals; now we emphasize statistical practice and no longer assume that population standard deviations are known. As a result, we move away from the standard Normal sampling distribution to a new family of t distributions. The t procedures for inference about means are among the most commonly used statistical methods.

07_Moore_13387_Ch07_407-482.indd 407 16/05/16 6:56 PM

7.1 Inference for the Mean of a Population

When you complete this section, you will be able to:

- Distinguish the standard deviation of the sample mean from the standard error of the sample mean.
- Describe a level C confidence interval for the population mean in terms of an estimate and its margin of error.
- Construct a level C confidence interval for μ from a simple random sample (SRS) size n from a large population.
- Perform a one-sample t significance test and summarize the results.
- Identify when the matched pairs t procedures should be used instead of two-sample t procedures.
- Explain when t procedures can be useful for non-Normal data.

AU: PR edit okay

Both confidence intervals and tests of significance for the mean μ of a Normal population are based on the sample mean $\bar{x}$, which estimates the unknown μ. The sampling distribution of $\bar{x}$ depends on σ. This fact causes no difficulty when σ is known. When σ is unknown, however, we must estimate σ even though we are primarily interested in μ.

In this section, we meet the sampling distribution of the standardized sample mean when we use the sample standard deviation s to estimate the population standard deviation σ. This sampling distribution is then used to produce both confidence intervals and significance tests about the mean μ.

The t distributions

LOOK BACK
sampling distribution of $\bar{x}$, p. 298

Suppose that we have a simple random sample (SRS) of size n from a Normally distributed population with mean μ and standard deviation σ. The sample mean $\bar{x}$ is then Normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. When σ is not known, we estimate it with the sample standard deviation s, and then we estimate the standard deviation of $\bar{x}$ by $s/\sqrt{n}$. This quantity is called the *standard error* of the sample mean $\bar{x}$, and we denote it by $\text{SE}_{\bar{x}}$.

STANDARD ERROR

When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of the sample mean is

$$\text{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

AU/DE/PE: should these be ital, no quotes? Cf pp. 80, 132, 165

The term "standard error" is sometimes used for the actual standard deviation of a statistic. The estimated value is then called the "estimated standard error." In this book, we will use the term "standard error" only when the

standard deviation of a statistic is estimated from the data. The term has this meaning in the output of many statistical computer packages and in research reports that apply statistical methods.

In the previous chapter, the standardized sample mean, or one-sample z statistic,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

is the basis for inference about μ when σ is known. This statistic has the standard Normal distribution $N(0, 1)$. However, when we substitute the standard error $s/\sqrt{n}$ for the standard deviation of $\bar{x}$, the statistic does *not* have a Normal distribution. It has a distribution that is new to us, called a t *distribution*.

AU: Please check. Since term is italicized here, okay to reverse on the "t"?

THE *t* DISTRIBUTIONS

Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population. Then the **one-sample *t* statistic**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the ***t* distribution** with $n - 1$ **degrees of freedom**.

degrees of freedom, p. 40

A particular t distribution is specified by giving the *degrees of freedom*. We use $t(k)$ to stand for the t distribution with k degrees of freedom. The degrees of freedom for this t statistic come from the sample standard deviation s in the denominator of t. We showed earlier that s has $n - 1$ degrees of freedom. Thus, there is a different t distribution for each sample size. There are also other t statistics with different degrees of freedom, some of which we will meet later in this chapter.

The t distributions were discovered in 1908 by William S. Gosset. Gosset was a statistician employed by the Guinness brewing company, which prohibited its employees from publishing their discoveries that were brewing related. In this case, the company let him publish under the pen name "Student" using an example that did not involve brewing. The t distribution is often called "Student's t" in his honor.

The density curves of the $t(k)$ distributions are similar in shape to the standard Normal curve. That is, they are symmetric about 0 and are bell-shaped. Figure 7.1 compares the density curves of the standard Normal distribution and the t distributions with 5 and 10 degrees of freedom. The similarity in shape is apparent, as is the fact that the t distributions have more probability in the tails and less in the center.

In reference to the standardized sample mean, this greater spread is due to the extra variability caused by substituting the random variable s for the fixed parameter σ. In Figure 7.1, we see that as the degrees of freedom k increase, the $t(k)$ density gets closer to the $N(0, 1)$ curve. This reflects the fact that s will be closer to σ (more precise) as the sample size increases.

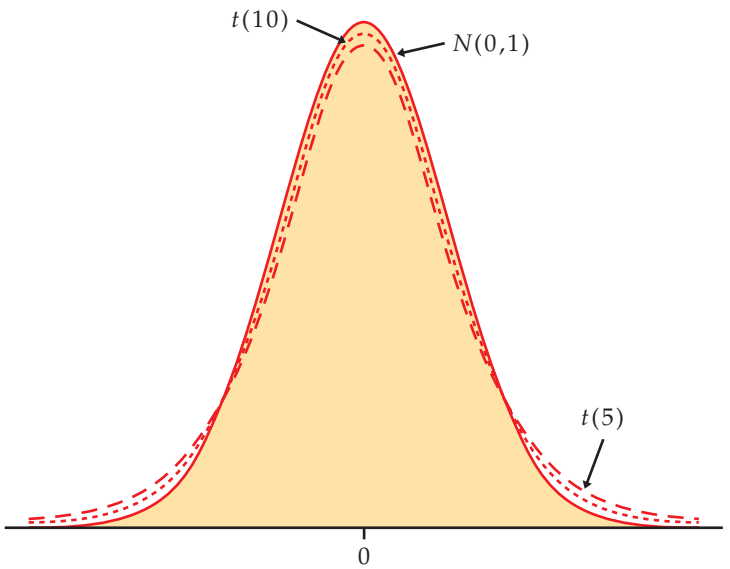


FIGURE 7.1 Density curves for the standard Normal, $t(10)$, and $t(5)$ distributions. All are symmetric with center 0. The t distributions have more probability in the tails than the standard Normal distribution.

USE YOUR KNOWLEDGE

7.1 One-bedroom apartment rates. You randomly choose 16 unfurnished one-bedroom apartments from a large number of advertisements in your local newspaper. You calculate that their mean monthly rent is \$766 and their standard deviation is \$180.

(a) What is the standard error of the mean?

(b) What are the degrees of freedom for a one-sample t statistic?

7.2 Changing the sample size. Refer to the previous exercise. Suppose that instead of an SRS of 16, you sampled 25 advertisements.

(a) Would you expect the standard error of the mean to be larger or smaller in this case? Explain your answer.

(b) State why you can't be certain that the standard error for this new SRS will be larger or smaller.

With the t distributions to help us, we can now analyze a sample from a Normal population with unknown σ or a large sample from a non-Normal population with unknown σ. Table D in the back of the book gives critical values t^* for the t distributions. For convenience, we have labeled the table entries both by the value of p needed for significance tests and by the confidence level C (in percent) required for confidence intervals. The standard Normal critical values are in the bottom row of entries and labeled z^*. As in the case of the Normal table (Table A), computer software often makes Table D unnecessary.

The one-sample t confidence interval

LOOK BACK
z confidence interval, p. 349

The one-sample t confidence interval is similar in both reasoning and computational detail to the z confidence interval of Chapter 6. There, the margin of error for the population mean was $z^*\sigma/\sqrt{n}$. When σ is unknown, we replace it with its estimate s and switch from z^* to t^*. This means that the margin of error for the population mean when we use the data to estimate σ is $t^*s/\sqrt{n}$.

THE ONE-SAMPLE *t* CONFIDENCE INTERVAL

Suppose that an SRS of size n is drawn from a population having unknown mean μ. A level C **confidence interval** for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. The quantity

$$t^* \frac{s}{\sqrt{n}}$$

is the **margin of error**. The confidence level is exactly C when the population distribution is Normal and is approximately correct for large n in other cases.

EXAMPLE 7.1

Watching traditional television. The Nielsen Company is a global information and media company and one of the leading suppliers of media information. In their annual Total Audience Report, the Nielsen Company states that adults age 18 to 24 years old average 18.5 hours per week watching traditional television.[1] Does this average seem reasonable for college students? They tend to watch a lot of television, but given their unusual schedules, they may be more likely to binge-watch or stream episodes after they air. To investigate, let's construct a 95% confidence interval for the average time (hours per week) spent watching traditional television among full-time U.S. college students. We draw the following SRS of size 8 from this population:

3.0 16.5 10.5 40.5 5.5 33.5 0.0 6.5

The sample mean is

$$\bar{x} = \frac{3.0 + 16.5 + \cdots + 6.5}{8} = 14.5$$

and the standard deviation is

$$s = \sqrt{\frac{(3.0-14.5)^2 + (16.5-14.5)^2 + \cdots + (6.5-14.5)^2}{8-1}} = 14.854$$

with degrees of freedom $n - 1 = 7$. The standard error is

$$\text{SE}_{\bar{x}} = s/\sqrt{n} = 14.854/\sqrt{8} = 5.252$$

From Table D, we find $t^* = 2.365$. The 95% confidence interval is

$$\begin{aligned} \bar{x} \pm t^* \frac{s}{\sqrt{n}} &= 14.5 \pm 2.365 \frac{14.854}{\sqrt{8}} \\ &= 14.5 \pm (2.365)(5.252) \\ &= 14.5 \pm 12.421 \\ &= (2.08, 26.92) \end{aligned}$$

We are 95% confident that among U.S. college students the average time spent watching traditional television is between 2.1 and 26.9 hours per week.

df = 7

t^*	1.895	2.365	2.517
C	0.90	0.95	0.96

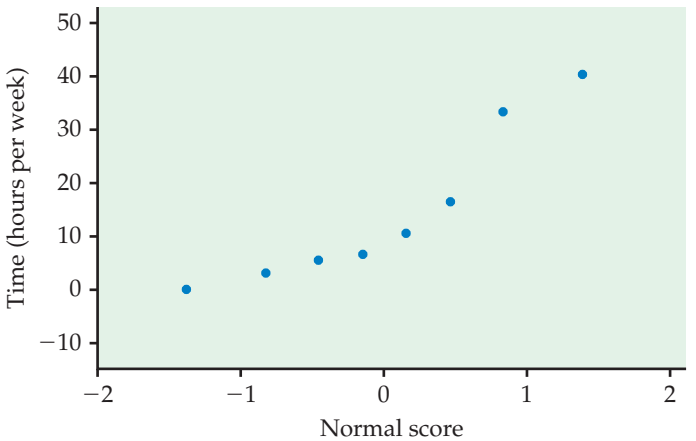


FIGURE 7.2 Normal quantile plot of data, Example 7.1.

In this example, we have given the actual interval (2.1, 26.9) hours per week as our answer. Sometimes, we prefer to report the mean and margin of error: the mean time is 14.5 hours per week with a margin of error of 12.4 hours per week. This is a large margin of error in relation to the estimated mean. In Section 7.3, we will return to this example and discuss determining an appropriate sample size for a desired margin of error such as ±5 hours a week.

AU: add xref page?

Valid interpretation of the t confidence interval in Example 7.1 rests on assumptions that appear reasonable here. First, we assume that our random sample is an SRS from the U.S. population of college students. Second, we assume that the distribution of watching times is Normal. Figure 7.2 shows the Normal quantile plot. With only eight observations, this assumption cannot be effectively checked. In fact, because a watching time cannot be negative, we might expect this distribution to be skewed to the right. With these data, however, there are no extreme outliers to suggest a severe departure from Normality.

USE YOUR KNOWLEDGE

7.3 More on apartment rents. Recall Exercise 7.1 (page 410). Construct a 95% confidence interval for the mean monthly rent of all advertised one-bedroom apartments.

7.4 Finding critical t^*-values. What critical value t^* from Table D should be used to construct

(a) a 95% confidence interval when $n = 25$?

(b) a 99% confidence interval when $n = 11$?

(c) a 90% confidence interval when $n = 61$?

LOOK BACK
four steps of significance test, p. 370

The one-sample t test

Significance tests using the standard error are also very similar to the z test that we studied in the last chapter. We still carry out the four steps common to all significance tests, but because we use s in place of σ, we use a t distribution to find the P-value.

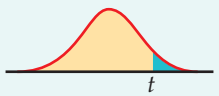
THE ONE-SAMPLE *t* TEST

Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test the hypothesis H_0: $\mu = \mu_0$ based on an SRS of size n, compute the **one-sample *t* statistic**

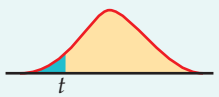
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In terms of a random variable T having the $t(n - 1)$ distribution, the P-value for a test of H_0 against

H_a: $\mu > \mu_0$ is $P(T \geq t)$



H_a: $\mu < \mu_0$ is $P(T \leq t)$



H_a: $\mu \neq \mu_0$ is $2P(T \geq |t|)$

These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

EXAMPLE 7.2

Significance test for watching traditional television. We want to test whether the average time that U.S. college students spend watching traditional television differs from the reported overall U.S. average of 18- to 24-year-olds at the 0.05 significance level. Specifically, we want to test

$$H_0: \mu = 18.5$$
$$H_a: \mu \neq 18.5$$

Recall that $n = 8$, $\bar{x} = 14.5$, and $s = 14.854$. The t test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{14.5 - 18.5}{14.854/\sqrt{8}}$$
$$= -0.762$$

This means that the sample mean $\bar{x} = 14.5$ is slightly more than 0.75 standard deviations below the null hypothesized value $\mu = 18.5$. Because the degrees of freedom are $n - 1 = 7$, this t statistic has the $t(7)$ distribution. Figure 7.3 shows that the P-value is $2P(T \geq 0.762)$, where T has the $t(7)$ distribution. From Table D, we see that $P(T \geq 0.711) = 0.25$ and $P(T \geq 0.896) = 0.20$.

Therefore, we conclude that the P-value is between $2 \times 0.20 = 0.40$ and $2 \times 0.25 = 0.50$. Software gives the exact value as $P = 0.4711$. These data are compatible with a mean of 18.5 hours per week. Under H_0, a difference this large or larger would occur about half the time simply due to chance. There is not enough evidence to reject the null hypothesis at the 0.05 level.

df = 7

p	0.25	0.20
t^*	0.711	0.896

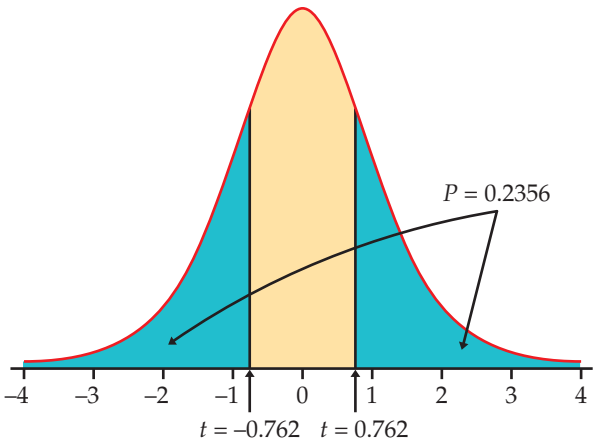


FIGURE 7.3 Sketch of the *P*-value calculation, Example 7.2.

In this example, we tested the null hypothesis $\mu = 18.5$ hours per week against the two-sided alternative $\mu \neq 18.5$ hours per week because we had no prior suspicion that the average among college students would be larger or smaller. If we had suspected that the average would be smaller (for example, expected more streaming of shows), we would have used a one-sided test.

EXAMPLE 7.3

TVTIME

One-sided test for watching traditional television. For the problem described in the previous example, we want to test whether the U.S. college student average is smaller than the overall U.S. population average. Here we test

$$H_0: \mu = 18.5$$

versus

$$H_a: \mu < 18.5$$

The t test statistic does not change: $t = -0.762$. As Figure 7.4 illustrates, however, the P-value is now $P(T \leq -0.762)$, half of the value in the previous example. From Table D, we can determine that $0.20 < P < 0.25$; software gives the exact value as $P = 0.2356$. Again, there is not enough evidence to reject the null hypothesis in favor of the alternative at the 0.05 significance level.

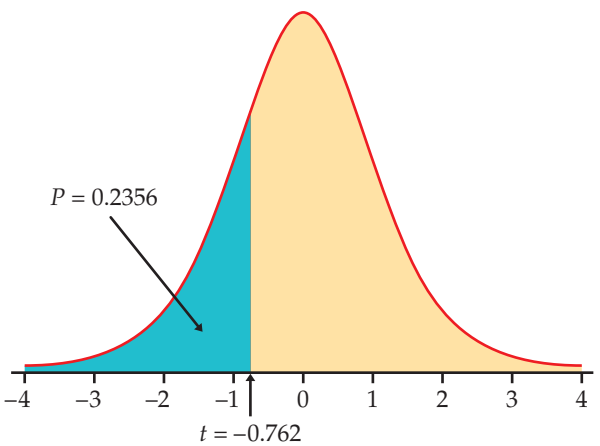


FIGURE 7.4 Sketch of the *P*-value calculation, Example 7.3.

For the watching-television example, our conclusion did not depend on the choice between a one-sided and a two-sided test. Sometimes, however, this choice *will* affect the conclusion, so this choice needs to be made prior to analysis. If in doubt, always use a two-sided test. *It is wrong to examine the data first and then decide to do a one-sided test in the direction indicated by the data.* Often, a significant result for a two-sided test can be used to justify a one-sided test for *another* sample from the same population.

USE YOUR KNOWLEDGE

7.5 Significance test using the *t* distribution. A test of a null hypothesis versus a two-sided alternative gives $t = 2.148$.

(a) The sample size is 23. Is the test result significant at the 5% level? Explain how you obtained your answer.

(b) The sample size is 9. Is the test result significant at the 5% level? Explain how you obtained your answer.

(c) Sketch the two t distributions to illustrate your answers.

7.6 Significance test for apartment rents. Refer to Exercise 7.1 (page 410). Does this SRS give good reason to believe that the mean rent of all advertised one-bedroom apartments is greater than \$700? State the hypotheses, find the t statistic and its P-value, and state your conclusion.

For small data sets, such as the one in Example 7.1 (page 411), it is easy to perform the computations for confidence intervals and significance tests with an ordinary calculator. For larger data sets, however, we prefer to use software or a statistical calculator.

EXAMPLE 7.4

Stock portfolio diversification? An investor with a stock portfolio worth several hundred thousand dollars sued his broker and brokerage firm because lack of diversification in his portfolio led to poor performance. Table 7.1 gives the rates of return for the 39 months that the account was managed by the broker.[2]

Figure 7.5 gives a histogram for these data, and Figure 7.6 gives the Normal quantile plot. There are no outliers and the distribution shows no strong skewness. We are reasonably confident that the distribution of

TABLE 7.1 Monthly Rates of Return on a Portfolio (%)

−8.36	1.63	−2.27	−2.93	−2.70	−2.93	−9.14	−2.64
6.82	−2.35	−3.58	6.13	7.00	−15.25	−8.66	−1.03
−9.16	−1.25	−1.22	−10.27	−5.11	−0.80	−1.44	1.28
−0.65	4.34	12.22	−7.21	−0.09	7.34	5.04	−7.24
−2.14	−1.01	−1.41	12.03	−2.56	4.33	2.35	

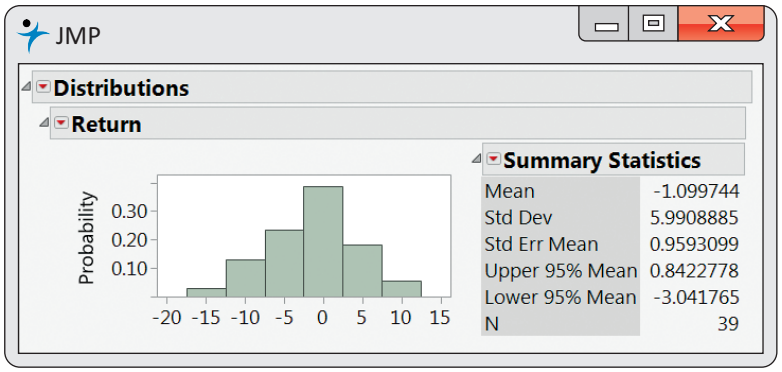


FIGURE 7.5 Histogram of monthly rates of return for a stock portfolio, Example 7.4.

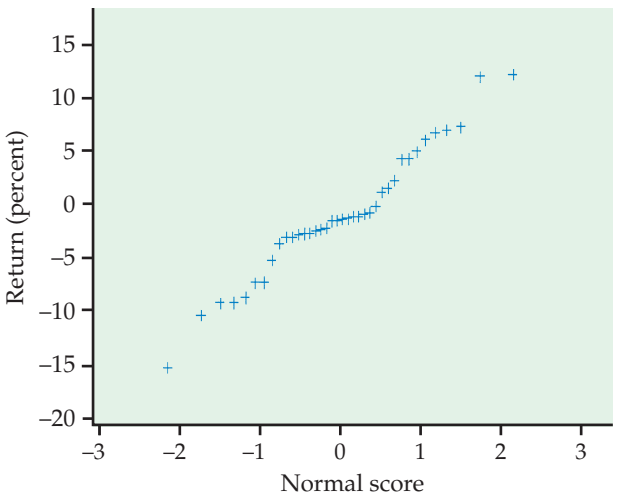


FIGURE 7.6 Normal quantile plot, Example 7.4.

AU: "for" removed before "Example". Okay as adjusted?

$\bar{x}$ is approximately Normal, and we proceed with our inference based on Normal theory.

The arbitration panel compared these returns with the average of the Standard & Poor's 500 stock index for the same period. Consider the 39 monthly returns as a random sample from the population of monthly returns the brokerage firm would generate if it managed the account forever. Are these returns compatible with a population mean of $\mu = 0.95\%$, the S&P 500 average? Our hypotheses are

$$H_0: \mu = 0.95$$
$$H_a: \mu \neq 0.95$$

Minitab and SPSS outputs appear in Figure 7.7. Output from other software will be similar.

Here is one way to report the conclusion: the mean monthly return on investment for this client's account was $\bar{x} = -1.1\%$. This is significantly worse than the performance of the S&P 500 stock index for the same period ($t = -2.14$, df $= 38$, $P = 0.039$).

AU: throughout, make "output" plural when more than one screenshot from software is provided?

Minitab

One-Sample T: Return

```
Test of μ = 0.95 vs ≠ 0.95

Variable   N    Mean  StDev  SE Mean        95% CI          T      P
Return    39  -1.100  5.991    0.959  (-3.042, 0.842)  -2.14  0.039
```

IBM SPSS Statistics Viewer

One-Sample Test

	Test Value = 0.95					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Return	-2.137	38	.039	-2.04974	-3.9918	-.1077

FIGURE 7.7 Minitab and SPSS output, Example 7.4.

The hypothesis test in Example 7.4 leads us to conclude that the mean return on the client's account differs from that of the S&P 500 stock index. Now let's assess the return on the client's account with a confidence interval.

EXAMPLE 7.5

Estimating the mean monthly return. The mean monthly return on the client's portfolio was $\bar{x} = -1.1\%$, and the standard deviation was $s = 5.99\%$. Figure 7.7 gives Minitab output, and Figure 7.8 gives JMP and Excel outputs

FIGURE 7.8 Excel and JMP output, Example 7.5.

AU: Okay as adjusted here to correspond with figure.

Excel

	A	B
1	*Return*	
2		
3	Mean	-1.09974359
4	Standard Error	0.95930991
5	Standard Deviation	5.99088847
6	Sample Variance	35.8907447
7	Count	39
8	Confidence Level(95.0%)	1.94202137

AU: Should we transpose here?

JMP

Distributions

Return

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	-1.09974	-3.04176	0.842278	0.950
Std Dev	5.990888	4.896029	7.720927	0.950

for a 95% confidence interval for the population mean μ. Note that Excel gives the margin of error next to the label "Confidence Level(95.0%)" rather than the actual confidence interval. We see that the 95% confidence interval is $(-3.04, 0.84)$, or (from Excel) -1.0997 ± 1.9420.

Because the S&P 500 return, 0.95%, falls outside this interval, we know that μ differs significantly from 0.95% at the $\alpha = 0.05$ level. Example 7.4 gave the actual P-value as $P = 0.039$.

The confidence interval suggests that the broker's management of this account had a long-term mean somewhere between a loss of 3.04% and a gain of 0.84% per month. We are interested, not in the actual mean, but in the difference between the performance of the client's portfolio and that of the diversified S&P 500 stock index.

EXAMPLE 7.6

Estimating the difference from a standard. Following the analysis accepted by the arbitration panel, we are considering the S&P 500 monthly average return as a constant standard. (It is easy to envision scenarios where we would want to treat this type of quantity as random.) The difference between the mean of the investor's account and the S&P 500 is $\bar{x} - \mu = -1.10 - 0.95 = -2.05\%$. In Example 7.5, we found that the 95% confidence interval for the investor's account was $(-3.04, 0.84)$.

To obtain the corresponding interval for the difference, subtract 0.95 from each of the endpoints. The resulting interval is $(-3.04 - 0.95, 0.84 - 0.95)$, or $(-3.99, -0.11)$. We conclude with 95% confidence that the underperformance was between -3.99 % and -0.11%. This interval is presented in the SPSS output of Figure 7.7. This estimate helps to set the compensation owed the investor.

The assumption that these 39 monthly returns represent an SRS from the population of monthly returns is certainly questionable. If the monthly S&P 500 returns were available, an alternative analysis would be to compare the average difference between each monthly return for this account and for the S&P 500. This method of analysis is discussed next.

USE YOUR KNOWLEDGE

7.7 Using software to obtain a confidence interval. In Example 7.1 (page 411), we calculated the 95% confidence interval for the U.S. college student average of hours per month spent watching traditional television. Use software to compute this interval and verify that you obtain the same interval.

7.8 Using software to perform a significance test. In Example 7.2 (page 413), we tested whether the average time that U.S. college students spend watching traditional television differs from the reported overall U.S. average of 18- to 24-year-olds at the 0.05 significance level. Use software to perform this test and obtain the exact P-value.

Matched pairs *t* procedures

LOOK BACK
confounding, p. 150
matched pairs design, p. 182

The watching-television problem of Example 7.1 (page 411) concerns only a single population. We know that comparative studies are usually preferred to single-sample investigations because of the protection they offer against confounding. For that reason, inference about a parameter of a single distribution is less common than comparative inference.

One common comparative design, however, makes use of single-sample procedures. In a matched pairs study, subjects are matched in pairs, and their outcomes are compared within each matched pair. For example, an experiment to compare two smartphone packages might use pairs of subjects who are the same age, sex, and income level. The experimenter could toss a coin to assign the two packages to the two subjects in each pair. The idea is that matched subjects are more similar than unmatched subjects, so comparing outcomes within each pair is more efficient (smaller σ).

Matched pairs are also common when randomization is not possible. For example, one situation calling for matched pairs is when observations are taken on the same subjects under two different conditions or before and after some intervention. Here is an example.

EXAMPLE 7.7

GEPARTS

The effect of altering a software parameter. The MeasureMind® 3D MultiSensor metrology software is used by various companies to measure complex machine parts. As part of a technical review of the software, researchers at GE Healthcare discovered that unchecking one software option reduced measurement time by 10%. This time reduction would help the company's productivity provided the option has no impact on the measurement outcome. To investigate this, the researchers measured 51 parts using the software both with and without this option checked.[3] The experimenters tossed a fair coin to decide which measurement (with or without the option) to take first.

Table 7.2 gives the measurements (in microns) for the first 20 parts. For analysis, we subtract the measurement with the option on from the measurement with the option off. These differences form a single sample and appear in the "Diff" columns for each part.

TABLE 7.2 Parts Measurements Using Optical Software

Part	OptionOn	OptionOff	Diff	Part	OptionOn	OptionOff	Diff
1	118.63	119.01	0.38	11	119.03	118.66	−0.37
2	117.34	118.51	1.17	12	118.74	118.88	0.14
3	119.30	119.50	0.20	13	117.96	118.23	0.27
4	119.46	118.65	−0.81	14	118.40	118.96	0.56
5	118.12	118.06	−0.06	15	118.06	118.28	0.22
6	117.78	118.04	0.26	16	118.69	117.46	−1.23
7	119.29	119.25	−0.04	17	118.20	118.25	0.05
8	120.26	118.84	−1.42	18	119.54	120.26	0.72
9	118.42	117.78	−0.64	19	118.28	120.26	1.98
10	119.49	119.66	0.17	20	119.13	119.15	0.02

To assess whether there is a difference between the measurements with and without this option, we test

$$H_0: \mu = 0$$
$$H_a: \mu \neq 0$$

Here, μ is the mean difference for the entire population of parts. The null hypothesis says that there is no difference, and H_a says that there is a difference, but does not specify a direction.

The 51 differences have

$$\bar{x} = 0.0504 \quad \text{and} \quad s = 0.6943$$

Figure 7.9 shows a histogram of the differences. It is reasonably symmetric with no outliers, so we can comfortably use the one-sample t procedures. *Remember to always check assumptions before proceeding with statistical inference.*

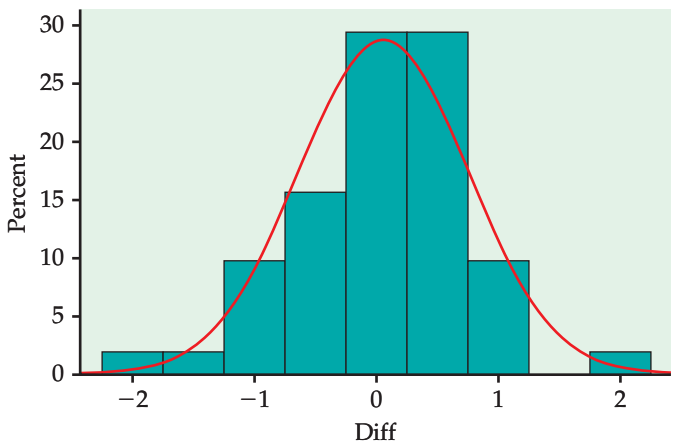


FIGURE 7.9 Histogram of differences in times, Example 7.7.

The one-sample t statistic is

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{0.0504}{0.6943/\sqrt{51}}$$
$$= 0.52$$

The P-value is found from the $t(50)$ distribution. Remember that the degrees of freedom are 1 less than the sample size.

Table D shows that 0.52 lies to the left of the first column entry. This means the P-value is greater than $2(0.25) = 0.50$. Software gives the exact value $P = 0.6054$. There is little evidence to suggest this option has an impact on the measurements. When reporting results, it is usual to omit the details of routine statistical procedures; our test would be reported in the form: "The difference in measurements was not statistically significant ($t = 0.52$, df $= 50$, $P = 0.61$)."

This result, however, does not fully address the goal of this study. *A lack of statistical significance does not prove the null hypothesis is true.* If that were the case, we would simply design poor experiments whenever we wanted to prove

equivalence testing

the null hypothesis. The more appropriate method of inference in this setting is to consider **equivalence testing**. With this approach, we try to prove that the mean difference is within some acceptable region around 0. We can actually perform this test using a confidence interval.

EXAMPLE 7.8

GEPARTS

df = 50

t^*	1.676	2.009
C	90%	95%

Are the two means equivalent? Suppose the GE Healthcare researchers state that a mean difference less than 0.20 micron is not important. To see if the data support a mean difference within 0.00 ± 0.20 micron, we construct a 90% confidence interval for the mean difference.

The standard error is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.6943}{\sqrt{51}} = 0.0972$$

so the margin of error is

$$m = t^* \times SE_{\bar{x}} = (1.676)(0.0972) = 0.1629$$

where the critical value $t^* = 1.676$ comes from Table D using 50 degrees of freedom. The confidence interval is

$$\begin{aligned} \bar{x} \pm m &= 0.0504 \pm 0.1629 \\ &= (-0.112, 0.2133) \end{aligned}$$

This interval is *not* entirely within the 0.00 ± 0.20 micron region that the researchers state is not important. Thus, we *cannot* conclude at the 5% signficance level that the two means are equivalent. Because the observed mean difference is close to zero and well within the "equivalent region," the company may want to consider a larger study to improve precision.

ONE SAMPLE TEST OF EQUIVALENCE

Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test, at significance level α, if μ is within a range of equivalency to μ_0, specified by the interval $\mu_0 \pm \delta$:

1. Compute the confidence interval with $C = 1 - 2\alpha$.
2. Compare this interval with the range of equivalency.

If the confidence interval falls entirely within $\mu_0 \pm \delta$, conclude that μ is equivalent to μ_0. If the confidence interval is outside the equivalency range or contains values both within and outside the range, conclude the μ is not equivalent to μ_0.

One can also use statistical software to perform an equivalence test. Figure 7.10 shows the Minitab output for Example 7.8. It is common to visually present the test using the confidence interval and the user-specified upper and lower equivalence limits.

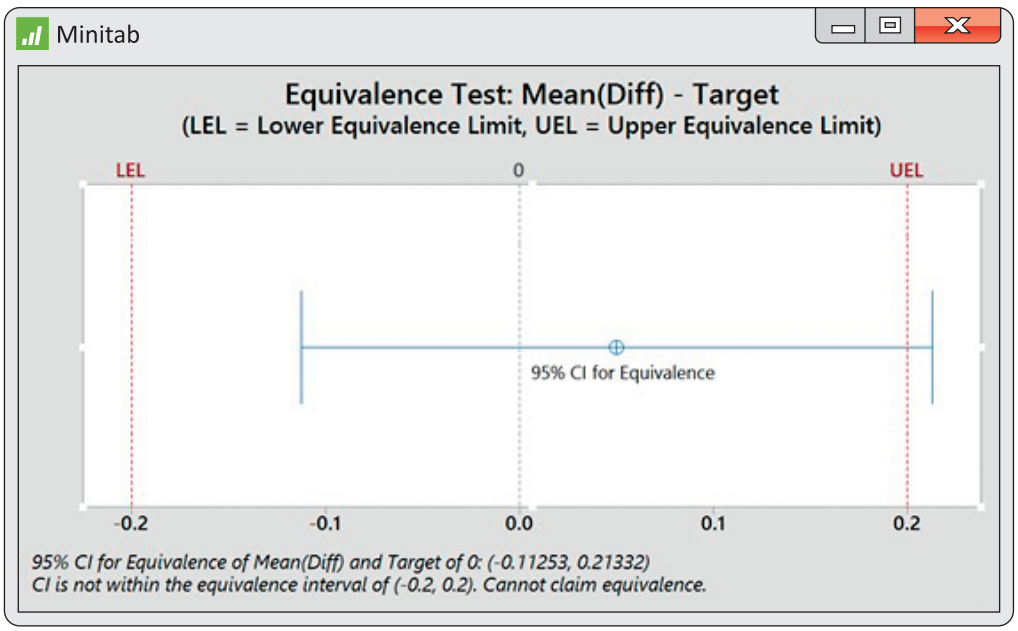


FIGURE 7.10 Minitab output of the equivalence test, Example 7.8.

USE YOUR KNOWLEDGE

AU: note edit

7.9 Female wolf spider mate preferences. As part of a study on factors affecting mate choice, researchers exposed 18 premature female wolf spiders twice a day until maturity to iPod videos of three courting males with average size tufts. Once mature, each female spider was exposed to two videos, one involving a male with large tufts and the other involving a male with small tufts. The number of receptivity displays by the female toward each male was recorded.[4] Explain why a paired t-test is appropriate in this setting.

7.10 Oil-free deep fryer. Researchers at Purdue University are developing an oil-free deep fryer that will produce fried food faster, healthier, and safer than hot oil.[5] As part of this development, they ask food experts to compare foods made with hot oil and their oil-free fryer. Consider the following table comparing the taste of hash browns. Each hash brown was rated on a 0 to 100 scale, with 100 being the highest rating. For each expert, a coin was tossed to see which type of hash brown was tasted first.

	Expert				
	1	2	3	4	5
Hot oil:	78	84	62	73	63
Oil free:	75	85	67	75	66

Is there a difference in taste? State the appropriate hypotheses, and carry out a matched pairs t test using $\alpha = 0.05$.

7.11 95% confidence interval for the difference in taste. To a restaurant owner, the real question is how much difference there is in taste. Use the data to give a 95% confidence interval for the mean difference in taste scores between oil-free and hot-oil frying.

Robustness of the *t* procedures

The matched pairs *t* procedures and test of equivalence use one-sample *t* confidence intervals and significance tests for differences. They are, therefore, based on an assumption that the population of differences has a Normal distribution. For the histogram of the 51 differences in Example 7.7 shown in Figure 7.9 (page 420), the data appear to be slightly skewed. Does this slight non-Normality suggest that we should not use the *t* procedures for these data?

All inference procedures are based on some conditions, such as Normality. Procedures that are not strongly affected by violations of a condition are called *robust*. Robust procedures are very useful in statistical practice because they can be used over a wide range of conditions with good performance.

ROBUST PROCEDURES

A statistical inference procedure is called **robust** if the required probability calculations are insensitive to violations of the assumptions made.

AU/PUB: Please check here. Okay with two Look Back arrows here?

resistant measure, p. 30

The assumption that the population is Normal rules out outliers, so the presence of outliers shows that this assumption is not valid. The *t* procedures are not robust against outliers because $\bar{x}$ and s are not resistant to outliers.

Fortunately, the *t* procedures are quite robust against non-Normality of the population except in the case of outliers or strong skewness. Larger samples improve the accuracy of *P*-values and critical values from the *t* distributions when the population is not Normal. This is true for two reasons:

central limit theorem, p. 298

law of large numbers, p. 250

1. The sampling distribution of the sample mean $\bar{x}$ from a large sample is close to Normal (that's the central limit theorem). Normality of the individual observations is of little concern when the sample is large.
2. As the sample size n grows, the sample standard deviation s will be an accurate estimate of σ whether or not the population has a Normal distribution. This fact is closely related to the law of large numbers.

To convince yourself of this fact, use the *t Statistic* applet to study the sampling distribution of the one-sample *t* statistic. From one of three population distributions, 10,000 SRSs of a user-specified sample size n are generated, and a histogram of the *t* statistics is constructed. You have the option to compare this estimated sampling distribution with the $t(n - 1)$ distribution. When the population distribution is Normal, the sampling distribution of the *t* statistic is always *t* distributed. For the other two population distributions, you should see that as n increases, the histogram of *t* statistics looks more like the $t(n - 1)$ distribution.

To assess whether the *t* procedures can be used in practice, a Normal quantile plot, stemplot, or boxplot is a good tool to check for skewness and outliers. For most purposes, the one-sample *t* procedures can be safely used when $n \geq 15$ unless an outlier or clearly marked skewness is present.

Except in the case of small samples, the assumption that the data are an SRS from the population of interest is more crucial than the assumption that the population distribution is Normal. Here are practical guidelines for inference on a single mean:[6]

- *Sample size less than 15:* Use t procedures if the data are close to Normal. If the data are clearly non-Normal or if outliers are present, do not use t.
- *Sample size at least 15 and less than 40:* The t procedures can be used except in the presence of outliers or strong skewness.
- *Large samples:* The t procedures can be used even for clearly skewed distributions when the sample is large, roughly $n \geq 40$.

For the measurement data in Example 7.7 (page 419), there is only slight skewness and no outliers. With $n = 51$ observations, we should feel comfortable that the t procedures give approximately correct results.

USE YOUR KNOWLEDGE

7.12 *t* procedures for time to start a business? Consider the data from Exercise 1.43 (page 29) but with Suriname removed. Would you be comfortable applying the t procedures in this case? Explain your answer.

7.13 *t* procedures for ticket prices? Consider the data on StubHub! ticket prices presented in Figure 1.32 (page 69). Would you be comfortable applying the t procedures in this case? In explaining your answer, recall that these t procedures focus on the mean μ.

BEYOND THE BASICS

The Bootstrap

Confidence intervals are based on sampling distributions. In this section, we have used the fact that the sampling distribution of $\bar{x}$ is $N(\mu, \sigma/\sqrt{n})$ when the data are an SRS from an $N(\mu, \sigma)$ population. If the data are not Normal, the central limit theorem tells us that this sampling distribution is still a reasonable approximation as long as the distribution of the data is not strongly skewed and there are no outliers. Even a fair amount of skewness can be tolerated when the sample size is large.

What if the population does not appear to be Normal and we have only a small sample? Then we do not know what the sampling distribution of $\bar{x}$ looks like. The **bootstrap** is a procedure for approximating sampling distributions when theory cannot tell us their shape.[7]

bootstrap

The basic idea is to act as if our sample were the population. We take many samples from it. Each of these is called a **resample**. We calculate the mean $\bar{x}$ for each resample. We get different results from different resamples because we sample *with replacement*. Thus, an observation in the original sample can appear more than once in a resample. We treat the resulting distribution of $\bar{x}$s as if it were the sampling distribution and use it to perform inference. If we want a 95% confidence interval, for example, we could use the middle 95% of this resample distribution.

resample

EXAMPLE 7.9

TVTIME

A bootstrap confidence interval. Consider the eight time measurements (in hours per week) spent watching traditional television in Example 7.1 (page 411):

3.0 16.5 10.5 40.5 5.5 33.5 0.0 6.5

We defended the use of the one-sided t confidence interval for an earlier analysis. Let's now compare those results with the confidence interval constructed using the bootstrap.

We decide to collect the $\bar{x}$'s from 1000 resamples of size $n = 8$. We use software to do this very quickly. One resample was

5.5 6.5 5.5 40.5 16.5 33.5 10.5 6.5

with $\bar{x} = 15.6251$. The middle 95% of our 1000 $\bar{x}$'s runs from 7.0 to 25.0. We repeat the procedure and get the interval (6.6, 25.1).

The two bootstrap intervals are relatively close to each other and are more narrow than the one-sample t confidence interval (2.1, 26.9). This suggests that the standard t interval is likely a little wider than it needs to be for this data set.

The bootstrap is practical only when you can use a computer to take 1000 or more resamples quickly. It is an example of how the use of fast and easy computing is changing the way we do statistics. More details about the bootstrap can be found in Chapter 16.

SECTION 7.1 SUMMARY

- Significance tests and confidence intervals for the mean μ of a Normal population are based on the sample mean $\bar{x}$ of an SRS. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample is large.
- The **standard error** of the sample mean is

$$\text{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- The standardized sample mean, or **one-sample z statistic**,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has the $N(0, 1)$ distribution. If the standard deviation $\sigma/\sqrt{n}$ of $\bar{x}$ is replaced by the **standard error** $s/\sqrt{n}$, the **one-sample t statistic**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the ***t* distribution** with $n - 1$ degrees of freedom.

- There is a t distribution for every positive **degrees of freedom k**. All are symmetric distributions similar in shape to Normal distributions. The $t(k)$ distribution approaches the $N(0, 1)$ distribution as k increases.
- A level C **confidence interval for the mean** μ of a Normal population is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. The quantity

$$t^*\frac{s}{\sqrt{n}}$$

is the **margin of error**.

- Significance tests for H_0: $\mu = \mu_0$ are based on the t statistic. P-values or fixed significance levels are computed from the $t(n-1)$ distribution.
- A matched pairs analysis is needed when subjects or experimental units are matched in pairs or when there are two measurements on each individual or experimental unit and the question of interest concerns the difference between the two measurements.
- The one-sample procedures are used to analyze **matched pairs** data by first taking the differences within the matched pairs to produce a single sample.
- One-sample **equivalence testing** assesses whether a population mean μ is practically different from a hypothesized mean μ_0. This test requires a threshold δ, which represents the largest difference between μ and μ_0 such that the means are considered equivalent.
- The t procedures are relatively **robust** against non-Normal populations. The t procedures are useful for non-Normal data when $15 \le n < 40$ unless the data show outliers or strong skewness. When $n \ge 40$, the t procedures can be used even for clearly skewed distributions.

SECTION 7.1 EXERCISES

For Exercises 7.1 and 7.2, see page 410; for Exercises 7.3 and 7.4, see page 412; for Exercises 7.5 and 7.6, see page 415; for Exercises 7.7 and 7.8, see page 418; for Exercises 7.9 through 7.11, see pages 422–423; and for Exercises 7.12 and 7.13, see page 424.

7.14 What is wrong? In each of the following situations, identify what is wrong and then either explain why it is wrong or change the wording of the statement to make it true.

(a) As the degrees of freedom k decrease, the t distribution density curve gets closer to the $N(0,1)$ curve.

(b) The standard error of the sample mean is s^2/n.

(c) A researcher wants to test H_0: $\bar{x} = 30$ versus the one-sided alternative H_a: $\bar{x} < 30$.

(d) The 95% margin of error for the mean μ of a Normal population with unknown σ is the same for all SRS of size n.

7.15 Finding the critical value t^*. What critical value t^* from Table D should be used to calculate the margin of error for a confidence interval for the mean of the population in each of the following situations?

(a) A 95% confidence interval based on $n = 15$ observations.

(b) A 95% confidence interval from an SRS of 28 observations.

(c) A 90% confidence interval from a sample of size 28.

(d) These cases illustrate how the size of the margin of error depends upon the confidence level and the sample size. Summarize these relationships.

7.16 Distribution of the t statistic. Assume a sample size of $n = 24$. Draw a picture of the distribution of the t statistic under the null hypothesis. Use Table D and your picture to illustrate the values of the test statistic that would lead to rejection of the null hypothesis at the 5% level for a two-sided alternative.

7.17 More on the distribution of the t statistic. Repeat the previous exercise for the two situations where the alternative is one-sided.

7.18 One-sided versus two-sided P-values. Computer software reports $\bar{x} = 11.2$ and $P = 0.075$ for a t test of H_0: $\mu = 0$ versus H_a: $\mu \neq 0$. Based on prior knowledge, you justified testing the alternative H_a: $\mu > 0$. What is the P-value for your significance test?

7.19 More on one-sided versus two-sided *P*-values. Suppose that computer software reports $\bar{x} = -11.2$ and $P = 0.075$ for a t test of H_0: $\mu = 0$ versus H_a: $\mu \neq 0$. Would this change your P-value for the alternative hypothesis in the previous exercise? Use a sketch of the distribution of the test statistic under the null hypothesis to illustrate and explain your answer.

7.20 A one-sample *t* test. The one-sample t statistic for testing

$$H_0: \mu = 8$$
$$H_a: \mu > 8$$

from a sample of $n = 22$ observations has the value $t = 2.24$.

(a) What are the degrees of freedom for this statistic?

(b) Give the two critical values t^* from Table D that bracket t.

(c) Between what two values does the P-value of the test fall?

(d) Is the value $t = 2.24$ significant at the 5% level? Is it significant at the 1% level?

(e) If you have software available, find the exact P-value.

7.21 Another one-sample *t* test. The one-sample t statistic for testing

$$H_0: \mu = 40$$
$$H_a: \mu \neq 40$$

from a sample of $n = 13$ observations has the value $t = 2.78$.

(a) What are the degrees of freedom for t?

(b) Locate the two critical values t^* from Table D that bracket t.

(c) Between what two values does the P-value of the test fall?

(d) Is the value $t = 2.78$ statistically significant at the 5% level? At the 1% level?

(e) If you have software available, find the exact P-value.

7.22 A final one-sample *t* test. The one-sample t statistic for testing

$$H_0: \mu = 20$$
$$H_a: \mu < 20$$

based on $n = 9$ observations has the value $t = -1.85$.

(a) What are the degrees of freedom for this statistic?

(b) Between what two values does the P-value of the test fall?

(c) If you have software available, find the exact P-value.

7.23 Two-sided to one-sided *P*-value. Most software gives P-values for two-sided alternatives. Explain why you cannot always divide these P-values by 2 to obtain P-values for one-sided alternatives.

7.24 Business bankruptcies in Canada. Business bankruptcies in Canada are monitored by the Office of the Superintendent of Bankruptcy Canada (OSB).[8] Included in each report are the assets and liabilities the company declared at the time of the bankruptcy filing. A study is based on a random sample of 75 reports from the current year. The average debt (liabilities minus assets) is $92,172 with a standard deviation of $111,538.

(a) Construct a 95% one-sample t confidence interval for the average debt of these companies at the time of filing.

(b) Because the sample standard deviation is larger than the sample mean, this debt distribution is skewed. Provide a defense for using the t confidence interval in this case.

7.25 Fuel economy. Although the Environmental Protection Agency (EPA) establishes the tests to determine the fuel economy of new cars, it often does not perform them. Instead, the test protocols are given to the car companies, and the companies perform the tests themselves. To keep the industry honest, the EPA runs some spot checks each year. Recently, the EPA announced that Hyundai and Kia must lower their fuel economy estimates for many of their models.[9] Here are some city miles per gallon (mpg) values for one of the models the EPA investigated: MILEAGE

28.0	25.7	25.8	28.0	28.5	29.8	30.2	30.4
26.9	28.3	29.8	27.2	26.7	27.7	29.5	28.0

Give a 95% confidence interval for μ, the mean city mpg for this model.

7.26 Testing the sticker information. Refer to the previous exercise. The vehicle sticker information for this model stated a city average of 30 mpg. Are these mpg values consistent with the vehicle sticker? Perform a significance test using the 0.05 significance level. Be sure to specify the hypotheses, the test statistic, the P-value, and your conclusion. MILEAGE

7.27 UberX driver earnings. On its blog, Uber posted a scatterplot using a sample of several thousand drivers in New York City. The plot shows each driver's average net earnings per hour versus the number of hours worked.[10] Here is a sample of earnings (dollars) for 27 drivers working 40 hours a week. UBERX

26.25	33.51	43.91	31.91	31.78	43.37	36.66	31.69	31.25
46.86	35.44	40.30	30.93	37.80	42.44	43.80	49.64	36.79
34.10	37.54	30.93	38.40	37.83	21.73	41.62	26.25	33.51

(a) Do you think it is appropriate to use the t methods of this section to compute a 95% confidence interval for the average earnings per hour of New York City UberX drivers working 40 hours a week? Generate a plot to support your answer.

(b) Report the 95% confidence interval for μ, the average earnings per hour of New York City UberX drivers working 40 hours a week, as an estimate and margin of error.

(c) Report the 95% confidence interval for the average annual earnings of New York City UberX drivers working 40 hours a week.

(d) According to Uber, the median wage of an UberX driver working at least 40 hours in New York City is $90,766. Can these data be used to assess this claim? Explain your answer.

7.28 Number of friends on Facebook. To mark Facebook's 10th birthday, Pew Research surveyed people using Facebook to see what they like and dislike about the site. The survey found that among adult Facebook users, the average number of friends is 338. This distribution takes only integer values, so it is certainly not Normal. It is also highly skewed to the right with a median of 200 friends.[11] Consider the following SRS of $n = 30$ Facebook users from your large university. FACEFR

107	246	289	177	155	101	80	461	336	78
463	264	827	180	221	1065	79	691	70	921
126	672	296	60	11	227	84	787	18	82

(a) Are these data also heavily skewed? Use graphical methods to examine the distribution. Write a short summary of your findings.

(b) Do you think it is appropriate to use the t methods of this section to compute a 95% confidence interval for the mean number of friends that Facebook users at your large university have? Explain why or why not.

(c) Compute the sample mean and standard deviation, the standard error of the mean, and the margin of error for 95% confidence.

(d) Report the 95% confidence interval for μ, the average number of friends for Facebook users at your large university.

7.29 Alcohol content in beer. In February 2013, two California residents filed a class-action lawsuit against Anheuser-Busch, alleging the company was watering down beers to boost profits.[12] They argued that because water was being added, the true alcohol content of the beer by volume is less than the advertised amount. For example, they alleged that Budweiser beer has an alcohol content by volume of 4.7% instead of the stated 5%. CNN, NPR, and a local St. Louis news team picked up on this suit and hired independent labs to test samples of Budweiser beer and find the alcohol content. Below is a summary of these tests each done on a single can. BUD

4.94 5.00 4.99

(a) Even though we have a very small sample, test the null hypothesis that the alcohol content is 4.7% by volume. Do the data provide evidence against the claim of the two residents?

(b) Construct a 95% confidence interval for the true alcohol content in Budweiser.

(c) U.S. government standards require that the true alcohol content in all cans and bottles be within $\pm 0.3\%$ of the advertised level. Do these tests provide strong evidence that this is the case for Budweiser beer? Explain your answer.

7.30 Using the Internet on a computer. The Nielsen Company reported that U.S. residents aged 18 to 24 years spend an average of 32.5 hours per month using the Internet on a computer.[13] You wonder if this it true for students at your large university because so many students use their smartphone to access the Internet. You collect an SRS of $n = 75$ students and obtain $\bar{x} = 28.5$ hours with $s = 23.1$ hours.

AU: plural?

(a) Report the 95% confidence interval for μ, the average number of hours per month that students at your university use the Internet on a computer.

(b) Use this interval to test whether the average time for students at your university is different from the average reported by Nielsen. Use the 5% significance level. Summarize your results.

7.31 Rudeness and its effect on onlookers. Many believe that an uncivil environment has a negative effect on people. A pair of researchers performed a series of experiments to test whether witnessing rudeness and disrespect affects task performance.[14] In one study, 34 participants met in small groups and witnessed the group organizer being rude to a "participant" who showed up late for the group meeting. After the exchange, each participant performed an individual brainstorming task in which he or she was asked to produce as many uses for a brick as possible in five minutes. The mean number of uses was 7.88 with a standard deviation of 2.35.

(a) Suppose that prior research has shown that the average number of uses a person can produce in five minutes under normal conditions is 10. Given that the researchers hypothesize that witnessing this rudeness will decrease performance, state the appropriate null and alternative hypotheses.

(b) Carry out the significance test using a significance level of 0.05. Give the *P*-value and state your conclusion.

7.32 Fuel efficiency *t* test. Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the miles per gallon were recorded each time the gas tank was filled, and the computer was then reset.[15] Here are the mpg values for a random sample of 20 of these records: MPG

41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3

(a) Describe the distribution using graphical methods. Is it appropriate to analyze these data using methods based on Normal distributions? Explain why or why not.

(b) Find the mean, standard deviation, standard error, and margin of error for 95% confidence.

(c) Report the 95% confidence interval for μ, the mean miles per gallon for this vehicle based on these data.

7.33 Tree diameter confidence interval. A study of 584 longleaf pine trees in the Wade Tract in Thomas County, Georgia, is described in Example 6.1 (page 342). For each tree in the tract, the researchers measured the diameter at breast height (DBH). This is the diameter of the tree at a height of 4.5 feet, and the units are centimeters (cm). Only trees with DBH greater than 1.5 cm were sampled. Here are the diameters of a random sample of 40 of these trees: PINES

10.5	13.3	26.0	18.3	52.2	9.2	26.1	17.6	40.5	31.8
47.2	11.4	2.7	69.3	44.4	16.9	35.7	5.4	44.2	2.2
4.3	7.8	38.1	2.2	11.4	51.5	4.9	39.7	32.6	51.8
43.6	2.3	44.6	31.5	40.3	22.3	43.3	37.5	29.1	27.9

(a) Use a histogram or stemplot and a boxplot to examine the distribution of DBHs. Include a Normal quantile plot if you have the necessary software. Write a careful description of the distribution.

(b) Is it appropriate to use the methods of this section to find a 95% confidence interval for the mean DBH of all trees in the Wade Tract? Explain why or why not.

(c) Report the mean with the margin of error and the confidence interval. Write a short summary describing the meaning of the confidence interval.

(d) Do you think these results would apply to other similar trees in the same area? Give reasons for your answer.

7.34 Nutritional intake among Canadian high-performance male athletes. Recall Exercise 6.74 (page 382). For one part of the study, $n = 114$ male athletes from eight Canadian sports centers were surveyed. Their average caloric intake was 3077.0 kilocalories per day (kcal/d) with a standard deviation of 987.0. The recommended amount is 3421.7. Is there evidence that Canadian high-performance male athletes are deficient in their caloric intake?

(a) State the appropriate H_0 and H_a to test this.

(b) Carry out the test, give the *P*-value, and state your conclusion.

(c) Construct a 95% confidence interval for the average deficiency in caloric intake.

7.35 Average number of Instagram posts. LocoWise provides social media analytics to companies and marketing agencies through a variety of online tools. One tool is the Instagram Analyzer, which allows a user to compare a profile with 2500 other Instagram profiles. Recently, it reported that the 2500 profiles it monitors averaged 2.55 posts per day, with a minimum value of 0 posts and a maximum value of 95 posts.[16]

(a) A common estimator of the standard deviation when provided the range R is $s = R/6$. Compute this estimate of s for these data.

(b) Construct the 95% confidence interval for the average number of Instagram posts per day.

(c) These data are clearly skewed and possibly have a few outliers. Do you think it is appropriate to use the *t* procedures? Explain your answer.

7.36 Stress levels in parents of children with ADHD. In a study of parents who have children with attention-deficit/hyperactivity disorder (ADHD), parents were asked to rate their overall stress level using the Parental Stress Scale (PSS).[17] This scale has 18 items that contain statements regarding both positive and negative aspects of parenthood. Respondents are asked to rate their agreement with each statement using a five-point scale (1 = strongly disagree to 5 = strongly agree). The scores are summed such that a higher score indicates greater stress. The mean rating for the 50 parents in the study was reported as 52.98 with a standard deviation of 10.34.

(a) Do you think that these data are approximately Normally distributed? Explain why or why not.

(b) Is it appropriate to use the methods of this section to compute a 90% confidence interval? Explain why or why not.

(c) Find the 90% margin of error and the corresponding confidence interval. Write a sentence explaining the interval and the meaning of the 90% confidence level.

(d) To recruit parents for the study, the researchers visited a psychiatric outpatient service in Rohtak, India, and selected 50 consecutive families who met the inclusion and exclusion criteria. To what extent do you think the results can be generalized to all parents with children who have ADHD in India or in other locations around the world?

7.37 Are the parents feeling extreme stress? Refer to the previous exercise. The researchers considered a score greater than 45 to represent extreme stress. Is there evidence that the average stress level for the parents in this study is above this level? Perform a test of significance using $\alpha = 0.10$ and summarize your results.

7.38 Food intake and weight gain. If we increase our food intake, we generally gain weight. Nutrition scientists can calculate the amount of weight gain that would be associated with a given increase in calories. In one study, 16 nonobese adults, aged 25 to 36 years, were fed 1000 calories per day in excess of the calories needed to maintain a stable body weight. The subjects maintained this diet for eight weeks, so they consumed a total of 56,000 extra calories.[18] According to theory, 3500 extra calories will translate into a weight gain of 1 pound. Therefore, we expect each of these subjects to gain 56,000/3500 = 16 pounds (lb). Here are the weights before and after the eight-week period, expressed in kilograms (kg): WTGAIN

Subject	1	2	3	4	5	6	7	8
Weight before	55.7	54.9	59.6	62.3	74.2	75.6	70.7	53.3
Weight after	61.7	58.8	66.0	66.2	79.0	82.3	74.3	59.3
Subject	9	10	11	12	13	14	15	16
Weight before	73.3	63.4	68.1	73.7	91.7	55.9	61.7	57.8
Weight after	79.1	66.0	73.4	76.9	93.1	63.0	68.2	60.3

(a) For each subject, subtract the weight before from the weight after to determine the weight change.

(b) Find the mean and the standard deviation for the weight change.

(c) Calculate the standard error and the margin of error for 95% confidence. Report the 95% confidence interval for weight change in a sentence that explains the meaning of the 95%.

(d) Convert the mean weight gain in kilograms to mean weight gain in pounds. Because there are 2.2 kg per pound, multiply the value in kilograms by 2.2 to obtain pounds. Do the same for the standard deviation and the confidence interval.

(e) Test the null hypothesis that the mean weight gain is 16 lb. Be sure to specify the null and alternative hypotheses, the test statistic with degrees of freedom, and the *P*-value. What do you conclude?

(f) Write a short paragraph explaining your results.

7.39 Food intake and NEAT. Nonexercise activity thermogenesis (NEAT) provides a partial explanation for the results you found in the previous analysis. NEAT is energy burned by fidgeting, maintenance of posture, spontaneous muscle contraction, and other activities of daily living. In the study of the previous exercise, the 16 subjects increased their NEAT by 328 calories per day, on average, in response to the additional food intake. The standard deviation was 256.

(a) Test the null hypothesis that there was no change in NEAT versus the two-sided alternative. Summarize the results of the test and give your conclusion.

(b) Find a 95% confidence interval for the change in NEAT. Discuss the additional information provided by the confidence interval that is not evident from the results of the significance test.

7.40 Potential insurance fraud? Insurance adjusters are concerned about the high estimates they are receiving from Jocko's Garage. To see if the estimates are unreasonably high, each of 10 damaged cars was taken to Jocko's and to another garage and the estimates (in dollars) were recorded. Here are the results: JOCKO

Car	1	2	3	4	5
Jocko's	1410	1550	1250	1300	900
Other	1250	1300	1250	1200	950
Car	6	7	8	9	10
Jocko's	1520	1750	3600	2250	2840
Other	1575	1600	3380	2125	2600

(a) For each car, subtract the estimate of the other garage from Jocko's estimate. Find the mean and the standard deviation for this difference.

(b) Test the null hypothesis that there is no difference between the estimates of the two garages. Be sure to specify the null and alternative hypotheses, the test statistic with degrees of freedom, and the *P*-value. What do you conclude using the 0.05 significance level?

(c) Construct a 95% confidence interval for the difference in estimates.

(d) The insurance company is considering seeking repayment from 1000 claims filed with Jocko's last year. Using your answer to part (c), what repayment would you recommend the insurance company seek? Explain your answer.

7.41 Fuel efficiency comparison *t* test. Refer to Exercise 7.32. In addition to the computer calculating

miles per gallon, the driver also recorded the miles per gallon by dividing the miles driven by the number of gallons at fill-up. The driver wants to determine if these calculations are different. MPGDIFF

Fill-up	1	2	3	4	5	6	7	8	9	10
Computer	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
Driver	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0	42.8	39.2
Fill-up	11	12	13	14	15	16	17	18	19	20
Computer	43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3
Driver	38.8	44.5	45.4	45.3	45.7	34.2	35.2	39.8	44.9	47.5

(a) State the appropriate H_0 and H_a.

(b) Carry out the test using a significance level of 0.05. Give the *P*-value, and then interpret the result.

7.42 Counts of picks in a one-pound bag. A guitar supply company must maintain strict oversight on the number of picks they package for sale to customers. Their current advertisement specifies between 900 and 1000 picks in every bag. An SRS of 36 one-pound bags of picks was collected as part of a quality improvement effort within the company. The number of picks in each bag is shown in the following table. PICKS

924	925	967	909	959	937	970	936	952
919	965	921	913	886	956	962	916	945
957	912	961	950	923	935	969	916	952
917	977	940	924	957	920	986	895	923

(a) Create (i) a histogram or stemplot, (ii) a boxplot, and (iii) a Normal quantile plot of these counts. Write a careful description of the distribution. Make sure to note any outliers, and comment on the skewness and Normality of the data.

(b) Based on your observations in part (a), is it appropriate to analyze these data using the *t* procedures? Briefly explain your response.

(c) Find the mean, the standard deviation, and the standard error of the mean for this sample.

(d) Calculate the 90% confidence interval for the mean number of picks in a one-pound bag.

7.43 Significance test for the average number of picks. Refer to the previous exercise. PICKS

(a) Do these data provide evidence that the average number of picks in a one-pound bag is greater than 925? Using a significance level of 5%, state your hypotheses, the *P*-value, and your conclusions.

(b) Do these data provide evidence that the average number of picks in a one-pound bag is greater than 935? Using a significance level of 5%, state your hypotheses, the *P*-value, and your conclusion.

(c) Explain the relationship between your conclusions in parts (a) and (b) and the 90% confidence interval calculated in the previous problem.

7.44 A customer satisfaction survey. Many organizations are doing surveys to determine the satisfaction of their customers. Attitudes toward various aspects of campus life were the subject of one such study conducted at Purdue University. Each item was rated on a 1 to 5 scale, with 5 being the highest rating. The average response of 1568 first-year students to "Feeling welcomed at Purdue" was 3.83 with a standard deviation of 1.10. Assuming that the respondents are an SRS, give a 90% confidence interval for the mean of all first-year students.

7.45 Comparing operators of a DXA machine. Dual-energy X-ray absorptiometry (DXA) is a technique for measuring bone health. One of the most common measures is total body bone mineral content (TBBMC). A highly skilled operator is required to take the measurements. Recently, a new DXA machine was purchased by a research lab, and two operators were trained to take the measurements. TBBMC for eight subjects was measured by both operators.[19] The units are grams (g). A comparison of the means for the two operators provides a check on the training they received and allows us to determine if one of the operators is producing measurements that are consistently higher than the other. Here are the data: TBBMC

	Subject							
Operator	1	2	3	4	5	6	7	8
1	1.328	1.342	1.075	1.228	0.939	1.004	1.178	1.286
2	1.323	1.322	1.073	1.233	0.934	1.019	1.184	1.304

(a) Take the difference between the TBBMC recorded for Operator 1 and the TBBMC for Operator 2. Describe the distribution of these differences. Is it appropriate to analyze these data using the *t* methods? Explain why or why not.

(b) Use a significance test to examine the null hypothesis that the two operators have the same mean. Be sure to give the test statistic with its degrees of freedom, the *P*-value, and your conclusion.

(c) The sample here is rather small, so we may not have much power to detect differences of interest. Use a 95% confidence interval to provide a range of differences that are compatible with these data.

(d) The eight subjects used for this comparison were not a random sample. In fact, they were friends of the researchers whose ages and weights were similar to those of the types of people who would be measured with

this DXA machine. Comment on the appropriateness of this procedure for selecting a sample, and discuss any consequences regarding the interpretation of the significance-testing and confidence interval results.

7.46 Equivalence of paper and computer-based questionnaires. Computers are commonly being used to complete questionnaires because of the increased efficiency of data collection and reduction in coding errors. Studies, however, have shown that questionnaire format can influence responses, especially for items of a sensitive nature.[20] Consider the small study below comparing paper and computer survey formats of a self-report measure of mental health. Each participant completed both forms on adjacent days with the order determined by a flip of a coin. EQUIV

Subject	Paper	Computer	Diff	Subject	Paper	Computer	Diff
1	5	2	3	11	6	5	1
2	4	3	1	12	5	5	0
3	4	4	0	13	3	7	−4
4	7	8	−1	14	3	6	−3
5	4	5	−1	15	4	4	0
6	6	7	−1	16	2	3	−1
7	4	3	1	17	7	10	−3
8	6	8	−2	18	8	7	1
9	6	5	1	19	4	6	−2
10	2	3	−1	20	6	8	−2

(a) Explain to someone unfamiliar with statistics why this experiment is a matched pairs design.

(b) The measure involves 10 items and produces a whole number score ranging between 0 and 20. Do you think it is appropriate to use the t procedures on the difference in survey scores? Explain your answer.

(c) Perform an equivalency test at the 0.05 level using the limits ± 0.5 and state your conclusion.

7.47 Assessment of a foreign-language institute. The National Endowment for the Humanities sponsors summer institutes to improve the skills of high school teachers of foreign languages. One such institute hosted 20 French teachers for four weeks. At the beginning of the period, the teachers were given the Modern Language Association's listening test of understanding of spoken French. After four weeks of immersion in French in and out of class, the listening test was given again. (The actual French spoken in the two tests was different, so that simply taking the first test should not improve the score on the second test.) The maximum possible score on the test is 36.[21] Here are the data: SUMLANG

Teacher	Pretest	Posttest	Gain	Teacher	Pretest	Posttest	Gain
1	32	34	2	11	30	36	6
2	31	31	0	12	20	26	6
3	29	35	6	13	24	27	3
4	10	16	6	14	24	24	0
5	30	33	3	15	31	32	1
6	33	36	3	16	30	31	1
7	22	24	2	17	15	15	0
8	25	28	3	18	32	34	2
9	32	26	−6	19	23	26	3
10	20	26	6	20	23	26	3

To analyze these data, we first subtract the pretest score from the posttest score to obtain the improvement for each teacher. These 20 differences form a single sample. They appear in the "Gain" columns. The first teacher, for example, improved from 32 to 34, so the gain is $34 - 32 = 2$.

(a) State appropriate null and alternative hypotheses for examining the question of whether or not the course improves French spoken-language skills.

(b) Describe the gain data. Use numerical and graphical summaries.

(c) Perform the significance test. Give the test statistic, the degrees of freedom, and the P-value. Summarize your conclusion.

(d) Give a 95% confidence interval for the mean improvement.

7.2 Comparing Two Means

When you complete this section, you will be able to:

- Describe a level C confidence interval for the difference between two population means in terms of an estimate and its margin of error.
- Construct a level C confidence interval for the difference between two population means $\mu_1 - \mu_2$ from two SRSs of size n_1 and n_2, respectively.
- Perform a two-sample t significance test and summarize the results.
- Explain when the t procedures can be useful for non-Normal data.

A psychologist wants to compare male and female college students' impressions of personality based on selected Facebook pages. A nutritionist is interested in the effect of increased calcium on blood pressure. A bank wants to know which of two incentive plans will most increase the use of its debit cards. Two-sample problems such as these are among the most common situations encountered in statistical practice.

TWO-SAMPLE PROBLEMS

- The goal of inference is to compare the means of the response variable in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.

LOOK BACK

randomized comparative experiment, p. 177

side-by-side boxplots, p. 37

A two-sample problem can arise from a randomized comparative experiment that randomly divides the subjects into two groups and exposes each group to a different treatment. A two-sample problem can also arise when comparing random samples separately selected from two populations. Unlike the matched pairs designs studied earlier, there is no matching of the units in the two samples, and the two samples may be of different sizes. As a result, inference procedures for two-sample data differ from those for matched pairs.

We can present two-sample data graphically by a back-to-back stemplot (for small samples) or by side-by-side boxplots (for larger samples). Now we will apply the ideas of formal inference in this setting. When both population distributions are symmetric, and especially when they are at least approximately Normal, a comparison of the mean responses in the two populations is most often the goal of inference.

We have two independent samples, from two distinct populations (such as subjects given the latest Apple iPhone and those given the latest Samsung Galaxy smartphone). The same response variable—say, battery life—is measured for both samples. We will call the variable x_1 in the first population and x_2 in the second because the variable may have different distributions in the two populations. Here is the notation that we will use to describe the two populations:

AU: Please check correction

Population	Variable	Mean	Standard deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

We want to compare the two population means, either by giving a confidence interval for $\mu_1 - \mu_2$ or by testing the hypothesis of no difference, H_0: $\mu_1 = \mu_2$.

Inference is based on two independent SRSs, one from each population. Here is the notation that describes the samples:

Population	Sample size	Sample mean	Sample standard deviation
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

Throughout this section, the subscripts 1 and 2 show the population to which a parameter or a sample statistic refers.

The two-sample *z* statistic

The natural estimator of the difference $\mu_1 - \mu_2$ is the difference between the sample means, $\bar{x}_1 - \bar{x}_2$. If we are to base inference on this statistic, we must know its sampling distribution. Here are some facts from our study of probability:

AU/PUB: Please check here. Okay with two Look Back arrows here?

LOOK BACK
addition rule for means, p. 254

- The mean of the difference $\bar{x}_1 - \bar{x}_2$ is the difference between the means $\mu_1 - \mu_2$. This follows from the addition rule for means and the fact that the mean of any $\bar{x}$ is the same as the mean μ of the population.

LOOK BACK
addition rule for variances, p. 258
linear combination of Normal random variables p. 304

- The variance of the difference $\bar{x}_1 - \bar{x}_2$ is the sum of their variances, which is

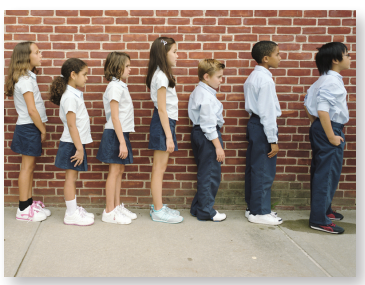
$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

This follows from the addition rule for variances. Because the samples are independent, their sample means $\bar{x}_1$ and $\bar{x}_2$ are independent random variables.

- If the two population distributions are both Normal, then the distribution of $\bar{x}_1 - \bar{x}_2$ is also Normal. This is true because each sample mean alone is Normally distributed and because a difference between independent Normal random variables is also Normal.

We now know the sampling distribution of $\bar{x}_1 - \bar{x}_2$ when both populations are Normally distributed. The mean and variance of this distribution can be expressed in terms of the parameters of the two populations.

EXAMPLE 7.10


Robert Warren/Getty Images

Heights of 10-year-old girls and boys. A fourth-grade class has 12 girls and 8 boys. The children's heights are recorded on their 10th birthdays. What is the chance that the girls are taller than the boys? Of course, it is very unlikely that all the girls are taller than all the boys. We translate the question into the following: what is the probability that the mean height of the girls is greater than the mean height of the boys?

Based on information from the National Health and Nutrition Examination Survey, we assume that the heights (in inches) of 10-year-old girls are $N(56.9, 2.8)$ and the heights of 10-year-old boys are $N(56.0, 3.5)$.[22] The heights of the students in our class are assumed to be random samples from these populations. The two distributions are shown in Figure 7.11(a).

The difference $\bar{x}_1 - \bar{x}_2$ between the female and male mean heights varies in different random samples. The sampling distribution has mean

$$\mu_1 - \mu_2 = 56.9 - 56.0 = 0.9 \text{ inches}$$

and variance

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{2.8^2}{12} + \frac{3.5^2}{8} = 2.18$$

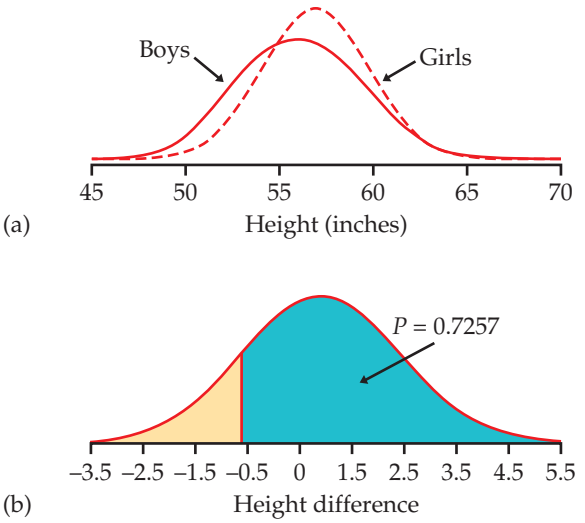


FIGURE 7.11 Distributions, Example 7.10. (a) Distributions of heights of 10-year-old boys and girls. (b) Distribution of the difference between mean heights of 12 girls and 8 boys.

The standard deviation of the difference in sample means is, therefore, $\sqrt{2.18} = 1.48$ inches.

If the heights vary Normally, the difference in sample means is also Normally distributed. The distribution of the difference in heights is shown in Figure 7.11(b). We standardize $\bar{x}_1 - \bar{x}_2$ by subtracting its mean (0.9) and dividing by its standard deviation (1.48). Therefore, the probability that the girls, on average, are taller than the boys is

$$P(\bar{x}_1 - \bar{x}_2 > 0) = P\left(\frac{(\bar{x}_1 - \bar{x}_2) - 0.9}{1.48} > \frac{0 - 0.9}{1.48}\right)$$
$$= P(Z > -0.61) = 0.7257$$

Even though the population mean height of 10-year-old girls is greater than the population mean height of 10-year-old boys, the probability that the sample mean of the girls is greater than the sample mean of the boys in our class is only 73%. *Large samples are needed to see the effects of small differences.*

As Example 7.10 reminds us, any Normal random variable has the $N(0, 1)$ distribution when standardized. We have arrived at a new z statistic.

TWO-SAMPLE z STATISTIC

Suppose that $\bar{x}_1$ is the mean of an SRS of size n_1 drawn from an $N(\mu_1, \sigma_1)$ population and that $\bar{x}_2$ is the mean of an independent SRS of size n_2 drawn from an $N(\mu_2, \sigma_2)$ population. Then the **two-sample z statistic**

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

has the standard Normal $N(0, 1)$ sampling distribution.

In the unlikely event that both population standard deviations are known, the two-sample z statistic is the basis for inference about $\mu_1 - \mu_2$. Exact z procedures are seldom used, however, because σ_1 and σ_2 are rarely known. In Chapter 6, we discussed the one-sample z procedures in order to introduce the ideas of inference. Here we move directly to the more useful t procedures.

The two-sample *t* procedures

Suppose now that the population standard deviations σ_1 and σ_2 are not known. We estimate them by the sample standard deviations s_1 and s_2 from our two samples. Following the pattern of the one-sample case, we substitute the standard errors for the standard deviations used in the two-sample z statistic. The result is the *two-sample t statistic:*

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

Unfortunately, this statistic does *not* have a t distribution. A t distribution replaces the $N(0, 1)$ distribution only when a single standard deviation (σ) in a z statistic is replaced by its sample standard deviation (s). In this case, we replace two standard deviations (σ_1 and σ_2) by their estimates (s_1 and s_2), which does not produce a statistic having a t distribution.

df approximation

Nonetheless, we can approximate the distribution of the two-sample t statistic by using the $t(k)$ distribution with an **approximation for the degrees of freedom *k***. We use these approximations to find approximate values of t^* for confidence intervals and to find approximate P-values for significance tests. Here are two approximations:

Satterthwaite approximation

1. Use an approximation known as the **Satterthwaite approximation** for the value of k. It is calculated from the data and, in general, will not be a whole number.

2. Use k equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

In practice, the choice of approximation rarely makes a difference in our conclusion. Most statistical software uses the first option to approximate the $t(k)$ distribution for two-sample problems unless the user requests another method. Use of this approximation without software is a bit complicated; we give the details later in this section (see page 447).

If you are not using software, the second approximation is preferred. This approximation is appealing because it is conservative.[23] Margins of error for the level C confidence intervals are a bit larger than they need to be, so the true confidence level is larger than C. For significance testing, the P-values are a bit larger; thus, for tests at a fixed significance level, we are a little less likely to reject H_0 when it is true.

The two-sample *t* confidence interval

We now apply the basic ideas about t procedures to the problem of comparing two means when the standard deviations are unknown. We start with confidence intervals.

THE TWO-SAMPLE *t* CONFIDENCE INTERVAL

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. The **confidence interval for $\mu_1 - \mu_2$** given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

has confidence level at least C no matter what the population standard deviations may be. The quantity

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

is the **margin of error**. Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^*. The value of the degrees of freedom k is approximated by software, or we use the smaller of $n_1 - 1$ and $n_2 - 1$. Similarly, we can use either software or the conservative approach with Table D to approximate the value of t^*.

EXAMPLE 7.11

RICHARD HUTCHINGS/Science Source/Getty Images

Directed reading activities assessment. An educator believes that new directed reading activities in the classroom will help elementary school pupils improve some aspects of their reading ability. She arranges for a third-grade class of 21 students to take part in these activities for an eight-week period. A control classroom of 23 third-graders follows the same curriculum without the activities. At the end of the eight weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve. The data appear in Table 7.3.[24]

DRP

The design of the study in Example 7.11 is not ideal. Random assignment of students was not possible in a school environment, so existing third-grade classes were used. The effect of the reading programs is, therefore,

TABLE 7.3 DRP Scores for Third-Graders

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

confounding, p. 150

confounded with any other differences between the two classes. The classes were chosen to be as similar as possible—for example, in terms of the social and economic status of the students. Extensive pretesting showed that the two classes were, on the average, quite similar in reading ability at the beginning of the experiment. To avoid the effect of two different teachers, the researcher herself taught reading in both classes during the eight-week period of the experiment. Therefore, we can be somewhat confident that the two-sample test is detecting the effect of the treatment and not some other difference between the classes. This example is typical of many situations in which an experiment is carried out but randomization is not possible.

EXAMPLE 7.12

DRP

Computing an approximate 95% confidence interval for the difference in means. First examine the data:

Control		Treatment
970	1	
860	2	4
773	3	3
8632221	4	3334699
5543	5	23467789
20	6	127
	7	1
5	8	

The back-to-back stemplot suggests that there is a mild outlier in the control group but no deviation from Normality serious enough to forbid use of t procedures. Separate Normal quantile plots for both groups (Figure 7.12) confirm that both distributions are approximately Normal. The scores of

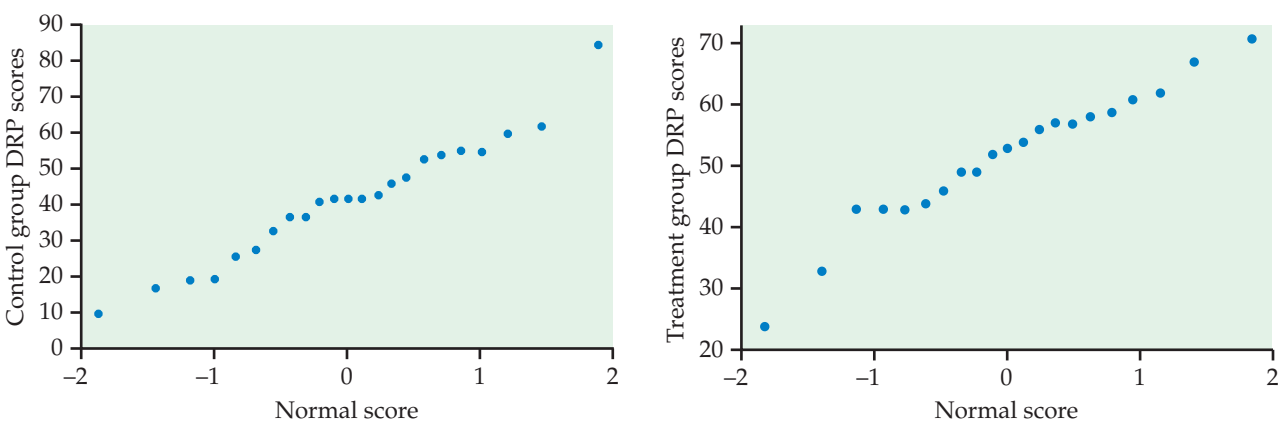


FIGURE 7.12 Normal quantile plots of the DRP scores in Table 7.3.

the treatment group appear to be somewhat higher than those of the control group. The summary statistics are

Group	n	$\bar{x}$	s
Treatment	21	51.48	11.01
Control	23	41.52	17.15

To describe the size of the treatment effect, let's construct a confidence interval for the difference between the treatment group and the control group means. The interval is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (51.48 - 41.52) \pm t^* \sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}}$$

$$= 9.96 \pm 4.31t^*$$

df = 20

t^*	1.725	2.086	2.197
C	0.90	0.95	0.96

The second degrees of freedom approximation uses the $t(20)$ distribution.

Table D gives $t^* = 2.086$. With this approximation, we have

$$9.96 \pm (4.31 \times 2.086) = 9.96 \pm 8.99 = (1.0, 18.9)$$

We estimate the mean improvement to be about 10 points, with a margin of error of almost 9 points. Unfortunately, the data do not allow a very precise estimate of the size of the average improvement.

USE YOUR KNOWLEDGE

7.48 Two-sample *t* confidence interval. Suppose a study similar to Example 7.11 was performed using two second-grade classes. Assume the summary statistics are $\bar{x}_1 = 46.32$, $\bar{x}_2 = 32.85$, $s_1 = 11.53$, $s_2 = 15.33$, $n_1 = 26$, and $n_2 = 24$. Find a 95% confidence interval for the difference between the treatment (Group 1) and the control (Group 2) means using the second approximation for degrees of freedom. Also write a one-sentence summary of what this confidence interval says about the difference in means.

7.49 Smaller sample sizes. Refer to the previous exercise. Suppose instead that the two classes are smaller, so the summary statistics are $\bar{x}_1 = 46.32$, $\bar{x}_2 = 32.85$, $s_1 = 11.53$, $s_2 = 15.33$, $n_1 = 16$, and $n_2 = 14$. Find a 95% confidence interval for the difference using the second approximation for degrees of freedom. Compare this interval with the one in the previous exercise and discuss the impact smaller sample sizes have on a confidence interval.

The two-sample *t* significance test

The same ideas that we used for the two-sample t confidence interval also apply to *two-sample t significance tests*. We can use either software or the conservative approach with Table D to approximate the P-value.

THE TWO-SAMPLE *t* SIGNIFICANCE TEST

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. To test the hypothesis H_0: $\mu_1 - \mu_2 = \Delta_0$, compute the **two-sample *t* statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use *P*-values or critical values for the $t(k)$ distribution, where the degrees of freedom k either are approximated by software or are the smaller of $n_1 - 1$ and $n_2 - 1$.

EXAMPLE 7.13

DRP

Is there an improvement? For the DRP study described in Example 7.11 (page 437), we hope to show that the treatment (Group 1) performs better than the control (Group 2). For a formal significance test, the hypotheses are

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 > \mu_2$$

The two-sample t test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$= \frac{51.48 - 41.52}{\sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}}}$$
$$= 2.31$$

The *P*-value for the one-sided test is $P(T \geq 2.31)$. For the second approximation, the degrees of freedom k are equal to the smaller of

$$n_1 - 1 = 21 - 1 = 20 \quad \text{and} \quad n_2 - 1 = 23 - 1 = 22$$

Comparing 2.31 with the entries in Table D for 20 degrees of freedom, we see that *P* lies between 0.01 and 0.02.

The data strongly suggest that directed reading activity improves the DRP score ($t = 2.31$, df = 20, $0.01 < P < 0.02$).

df = 20

p	0.02	0.01
t^*	2.197	2.528

USE YOUR KNOWLEDGE

7.50 A two-sample *t* significance test. Refer to Exercise 7.48 (page 439). Perform a significance test at the 0.05 level to assess whether the average improvement is five points versus the alternative that it is greater than five points. Write a one-sentence conclusion.

7.51 Interpreting the confidence interval. Refer to the previous exercise and Exercise 7.48. Can the confidence interval in Exercise 7.48 be used to determine whether the significance test of the previous exercise rejects or does not reject the null hypothesis? Explain your answer.

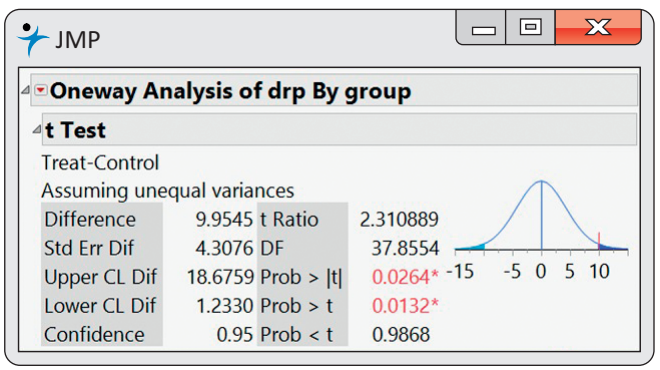
Most statistical software requires the raw data for analysis. A few, like Minitab, will also perform a t test on data in summarized form (such as the summary statistics table in Example 7.12, pages 438–439). It is always preferable to work with the raw data because one can also examine the data through plots such as the back-to-back stemplot and those in Figure 7.12.

EXAMPLE 7.14

DRP

Using software. Figure 7.13 shows JMP and Minitab output for the comparison of DRP scores. Both outputs include the 95% confidence interval and the significance test that the means are equal. JMP reports the difference as the mean of treatment minus the mean of control, while Minitab reports the difference in the opposite order.

AU: plural? outputs



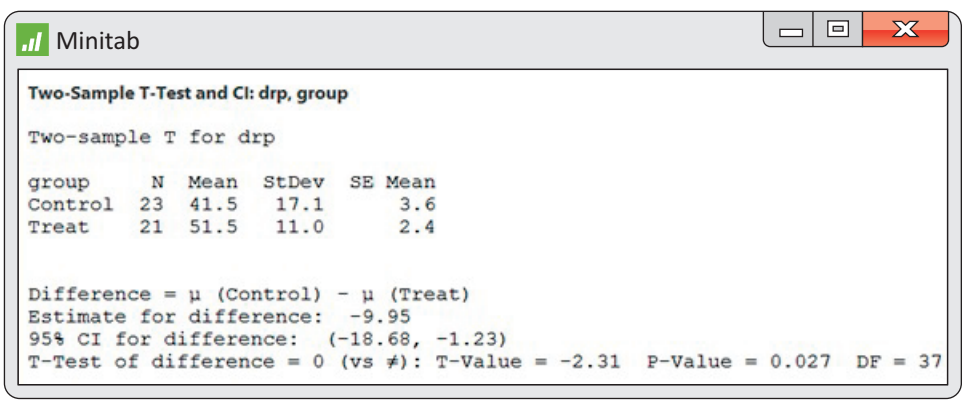


FIGURE 7.13 JMP and Minitab output, Example 7.14.

Recall the confidence interval (treatment minus control) is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (51.48 - 41.52) \pm t^* \sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}}$$
$$= 9.96 \pm 4.31t^*$$

From the JMP output, we see that the degrees of freedom under the first approximation are 37.9. Using these degrees of freedom, the interval is (1.2, 18.7). This interval, as expected, is more narrow than the confidence interval in Example 7.12 (pages 438–439), which uses the conservative approach. The difference, however, is pretty small.

For the significance test, the *P*-value for the one-sided significance test is $P(T \geq 2.31)$. JMP gives the approximate *P*-value as 0.0132, again using 37.9 as the degrees of freedom.

Minitab also uses the first degrees of freedom approximation but rounds the degrees of freedom down to the nearest integer ($37.9 \rightarrow 37$). As a result, the margin of error is slightly wider than that of JMP and the *P*-value of the significance test is slightly larger.

In order to get a confidence interval as part of the Minitab output, the two-sided alternative was considered. If your software gives you the *P*-value for only the two-sided alternative, $2P(T \geq |t|)$, you need to divide the reported value by 2 after checking that the means differ in the direction specified by the alternative hypothesis.

Robustness of the two-sample procedures

The two-sample *t* procedures are more robust than the one-sample *t* methods. When the sizes of the two samples are equal and the distributions of the two populations being compared have similar shapes, probability values from the *t* table are quite accurate for a broad range of distributions when the sample sizes are as small as $n_1 = n_2 = 5$.[25] When the two population distributions have different shapes, larger samples are needed.

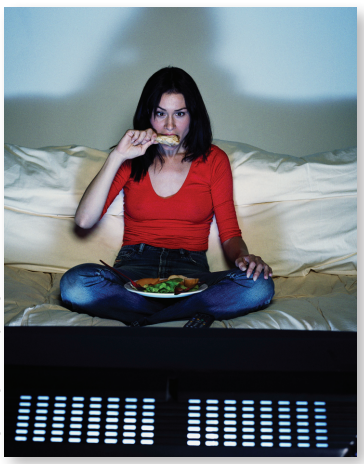
The guidelines for the use of one-sample *t* procedures can be adapted to two-sample procedures by replacing "sample size" with the "sum of the sample sizes" $n_1 + n_2$. Specifically,

- *If $n_1 + n_2$ is less than 15:* Use *t* procedures if the data are close to Normal. If the data in either sample are clearly non-Normal or if outliers are present, do not use *t*.
- *If $n_1 + n_2$ is at least 15 and less than 40:* The *t* procedures can be used except in the presence of outliers or strong skewness.
- *Large samples:* The *t* procedures can be used even for clearly skewed distributions when the sample is large, roughly $n_1 + n_2 \geq 40$.

These guidelines are rather conservative, especially when the two samples are of equal size. *In planning a two-sample study, choose equal sample sizes if you can.* The two-sample *t* procedures are most robust against non-Normality in this case, and the conservative probability values are most accurate.

Here is an example with large sample sizes that are almost equal. Even if the distributions are not Normal, we are confident that the sample means will be approximately Normal. The two-sample t test is very robust in this case.

EXAMPLE 7.15



Serge Krouglikoff/Getty Images

Timing of food intake and weight loss. There is emerging evidence of a relationship between timing of feeding and weight regulation. In one study, researchers followed 402 obese or overweight individuals through a 20-week weight-loss treatment.[26] To investigate the timing of food intake, participants were grouped into early eaters and late eaters, based on the timing of their main meal. Here are the summary statistics of their weight loss over the 20 weeks, in kilograms (kg):

Group	n	$\bar{x}$	s
Early eater	202	9.9	5.8
Late eater	200	7.7	6.1

The early eaters lost more weight on average. Can we conclude that these two groups are not the same? Or is this observed difference merely what we could expect to see given the variation among participants?

While other evidence suggests that early eaters should lose more weight, the researchers did not specify a direction for the difference. Thus, the hypotheses are

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

Because the samples are large, we can confidently use the t procedures even though we lack the detailed data and so cannot verify the Normality condition.

The two-sample t statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$= \frac{9.9 - 7.7}{\sqrt{\frac{5.8^2}{202} + \frac{6.1^2}{200}}}$$
$$= 3.71$$

df = 100

p	0.0005
t^*	3.390

The conservative approach finds the P-value by comparing 3.71 to critical values for the $t(199)$ distribution because the smaller sample has 200 observations. Because Table D does not contain a row for 199 degrees of freedom, we will be even more conservative and use the first row in the table with degrees of freedom less than 199. This means we'll use the $t(100)$ distribution to compute the P-value.

Our calculated value of t is larger than the $p = 0.0005$ entry in the table. We must double the table tail area p because the alternative is two-sided, so we conclude that the P-value is less than 0.001. The data give conclusive evidence that early eaters lost more weight, on average, than late eaters ($t = 3.71$, df = 100, $P < 0.001$).

In this example the exact P-value is very small because $t = 3.71$ says that the observed difference in means is over 3.5 standard errors above the hypothesized difference of zero ($\mu_1 = \mu_2$). In this study, the researchers also compared energy intake and energy expenditure between late and early eaters. Despite the observed weight loss difference of 2.2 kg, no significant differences in these variables were found.

In this and other examples, we can choose which population to label 1 and which to label 2. After inspecting the data, we chose early eaters as Population 1 because this choice makes the t statistic a positive number. This avoids any possible confusion from reporting a negative value for t. *Choosing the population labels is* ***not*** *the same as choosing a one-sided alternative after looking at the data.* Choosing hypotheses after seeing a result in the data is a violation of sound statistical practice.

Inference for small samples

Small samples require special care. We do not have enough observations to examine the distribution shapes, and only extreme outliers stand out. The power of significance tests tends to be low, and the margins of error of confidence intervals tend to be large. Despite these difficulties, we can often draw important conclusions from studies with small sample sizes. If the size of an effect is very large, it should still be evident even if the n's are small.

EXAMPLE 7.16

EATER

Timing of food intake. In the setting of Example 7.15, let's consider a much smaller study that collects weight loss data from only five participants in each eating group. Also, given the results of this past example, we choose the one-sided alternative. The data are

Group	Weight loss (kg)				
Early eater	6.3	15.1	9.4	16.8	10.2
Late eater	7.8	0.2	1.5	11.5	4.6

First, examine the distributions with a back-to-back stemplot (the data are rounded to the nearest integer).

Early		Late
	0	02
96	0	58
0	1	2
75	1	

While there is variation among weight losses within each group, there is also a noticeable separation. The early-eaters group contains four of the five largest losses, and the late-eaters group contains four of the five smallest losses. A significance test can confirm whether this pattern can arise just by chance or if the early-eaters group has a higher mean. We test

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 > \mu_2$$

The average weight loss is higher in the early-eater group ($t = 2.28$, df $= 7.96$, $P = 0.0262$). The difference in sample means is 6.44 kg.

Figure 7.14 gives outputs for this analysis from several software packages. Although the formats differ, the basic information is the same. All report the sample sizes, the sample means and standard deviations (or variances), the t statistic, and its P-value. All agree that the P-value is small, though some give more detail than others. Software often labels the groups in alphabetical order. Always check the means first and report the statistic (you may need to change the sign) in an appropriate way. Be sure to also mention the size of the effect you observed, such as "The mean weight loss for the early eaters was 6.44 kg higher than for the late eaters."

plural?

FIGURE 7.14 SAS, Excel, JMP, and SPSS output, Example 7.16.

SAS

The TTEST Procedure

Variable: loss

Grp	N	Mean	StdDev	StdErr	Minimum	Maximum
Early	5	11.5600	4.3062	1.9258	6.3000	16.8000
Late	5	5.1200	4.6224	2.0672	0.2000	11.5000
Diff (1-2)		6.4400	4.4671	2.8252		

Grp	Method	Mean	95%	CL Mean	Std Dev	95%	CL Std Dev
Early		11.5600	6.2132	16.9068	4.3062	2.5800	12.3740
Late		5.1200	−0.6195	10.8595	4.6224	2.7695	13.2829
Diff (1–2)	Pooled	6.4400	−0.0750	12.9550	4.4671	3.0173	8.5579
Diff (1–2)	Satterthwaite	6.4400	−0.0807	12.9607			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	8	2.28	0.0521
Satterthwaite	Unequal	7.9601	2.28	0.0523

Excel

	A	B	C
1	**t-Test: Two-Sample Assuming Unequal Variances**		
2			
3		*Early*	*Late*
4	Mean	11.56	5.12
5	Variance	18.543	21.367
6	Observations	5	5
7	Hypothesized Mean Difference	0	
8	Df	8	
9	t Stat	2.27944966	
10	P(T<=t) one-tail	0.026058036	
11	t Critical one-tail	1.859548033	
12	P(T<=t) two-tail	0.052116073	
13	t Critical two-tail	2.306004133	

FIGURE 7.14 *Continued*

JMP

Oneway Analysis of WTLOSS By Group

Means and Std Deviations

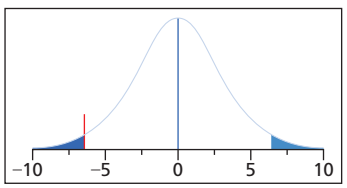
Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Early	5	11.5600	4.30616	1.9258	6.213	16.907
Late	5	5.1200	4.62245	2.0672	−0.620	10.860

t Test

Late-early

Assuming unequal variances

Difference	−6.440	t Ratio	−2.27945
Std Err Dif	2.825	DF	7.960145
Upper CL Dif	0.081	Prob > |t|	0.0523
Lower CL Dif	−12.961	Prob > t	0.9739
Confidence	0.95	Prob < t	0.0261*



IBM SPSS Statistics Viewer

Group Statistics

	grp	N	Mean	Std. Deviation	Std. Error Mean
Loss	Early	5	11.560	4.3062	1.9258
	Late	5	5.120	4.6224	2.0672

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Loss	Equal variances assumed	2.279	8	.052	6.4400	2.8252	−.0750	12.9550
	Equal variances not assumed	2.279	7.960	.052	6.4400	2.8252	−.0807	12.9607

There are two other things to notice in the outputs. First, SAS and SPSS only give results for the two-sided alternative. To get the *P*-value for the one-sided alternative, we must first check the mean difference to make sure it is in the proper direction. If it is, we divide the given *P*-value by 2. Also, SAS and SPSS report the results of *two t* procedures: a special procedure that assumes

that the two population variances are equal and the general two-sample procedure that we have just studied. We don't recommend the "equal-variances" procedures, but we describe them later, in the section on pooled two-sample t procedures.

Software approximation for the degrees of freedom

We noted earlier that the two-sample t statistic does not have a t distribution. Moreover, the distribution changes as the unknown population standard deviations σ_1 and σ_2 change. However, the distribution can be approximated by a t distribution with degrees of freedom given by

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

This is the approximation used by most statistical software. It is quite accurate when both sample sizes n_1 and n_2 are 5 or larger.

EXAMPLE 7.17

Degrees of freedom for directed reading assessment. For the DRP study of Example 7.11 (page 437), the following table summarizes the data:

Group	n	$\bar{x}$	s
1	21	51.48	11.01
2	23	41.52	17.15

For greatest accuracy, we will use critical points from the t distribution with degrees of freedom given by the preceding equation:

$$\text{df} = \frac{\left(\frac{11.01^2}{21} + \frac{17.15^2}{23}\right)^2}{\frac{1}{20}\left(\frac{11.01^2}{21}\right)^2 + \frac{1}{22}\left(\frac{17.15^2}{23}\right)^2}$$

$$= \frac{344.486}{9.099} = 37.86$$

This is the value that we reported in Example 7.14 (pages 441–442), where we gave the results produced by software.

The number df given by the preceding approximation is always at least as large as the smaller of $n_1 - 1$ and $n_2 - 1$. On the other hand, the number df is never larger than the sum $n_1 + n_2 - 2$ of the two individual degrees of freedom. The number df is generally not a whole number. There is a t distribution with any positive degrees of freedom, even though Table D contains entries only for whole-number degrees of freedom. When the number df is small

and is not a whole number, interpolation between entries in Table D may be needed to obtain an accurate critical value or *P*-value. Because of this and the need to calculate df, we do not recommend regular use of this approximation if a computer is not doing the arithmetic. With a computer, however, the more accurate procedures are painless.

USE YOUR KNOWLEDGE

7.52 Calculating the degrees of freedom. Assume that $s_1 = 5$, $s_2 = 8$, $n_1 = 25$, and $n_2 = 32$. Find the approximate degrees of freedom.

The pooled two-sample *t* procedures

There is one situation in which a t statistic for comparing two means has exactly a t distribution. This is when the two Normal population distributions have the *same* standard deviation. As we've done with other t statistics, we will first develop the z statistic and then, from it, the t statistic. In this case, notice that we need to substitute only a single standard error when we go from the z to the t statistic. This is why the resulting t statistic has a t distribution.

Call the common—and still unknown—standard deviation of both populations σ. Both sample variances s_1^2 and s_2^2 estimate σ^2. The best way to combine these two estimates is to average them with weights equal to their degrees of freedom. This gives more weight to the sample variance from the larger sample, which is reasonable. The resulting estimator of σ^2 is

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

pooled estimator of σ^2

This is called the **pooled estimator of σ^2** because it combines the information in both samples.

When both populations have variance σ^2, the addition rule for variances says that $\bar{x}_1 - \bar{x}_2$ has variance equal to the *sum* of the individual variances, which is

$$\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

The standardized difference between means in this equal-variance case is, therefore,

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

This is a special two-sample z statistic for the case in which the populations have the same σ. Replacing the unknown σ by the estimate s_p gives a t statistic. The degrees of freedom are $n_1 + n_2 - 2$, the sum of the degrees of freedom of the two sample variances. This t statistic is the basis of the pooled two-sample t inference procedures.

THE POOLED TWO-SAMPLE *t* PROCEDURES

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. Suppose also that the two populations have the same standard deviation. A level C **confidence interval for $\mu_1 - \mu_2$** is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here, $t^*$is the value for the $t(n_1 + n_2 - 2)$ density curve with area C between $-t^*$ and t^*. The quantity

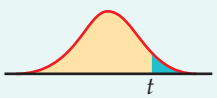
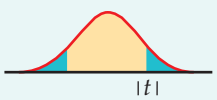
$$t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is the **margin of error**.

To test the hypothesis H_0: $\mu_1 - \mu_2 = \Delta_0$, compute the **pooled two-sample *t* statistic**

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

In terms of a random variable T having the $t(n_1 + n_2 - 2)$ distribution, the P-value for a test of H_0 against

H_a: $\mu_1 - \mu_2 > \Delta_0$	is	$P(T \geq t)$	
H_a: $\mu_1 - \mu_2 < \Delta_0$	is	$P(T \leq t)$	
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	is	$2P(T \geq \lvert t \rvert)$	

EXAMPLE 7.18

AU: Please check. You had a blank note here in your corrections. Did you mean to add anything?

Calcium and blood pressure. Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure, but such observational studies do not establish causation. Animal experiments, however, showed that calcium supplements do reduce blood pressure in rats, justifying an experiment with human subjects. A randomized comparative experiment gave one group of 10 black men a calcium supplement for 12 weeks. The control group of 11 black men received a placebo that appeared identical. (In fact, a block design with black and white men as the blocks was used. We will look only at the results for blacks because

TABLE 7.4 Seated Systolic Blood Pressure (mm Hg)

Calcium Group			Placebo Group		
Begin	End	Decrease	Begin	End	Decrease
107	100	7	123	124	−1
110	114	−4	109	97	12
123	105	18	112	113	−1
129	112	17	102	105	−3
112	115	−3	98	95	3
111	116	−5	114	119	−5
107	106	1	119	114	5
112	102	10	114	112	2
136	125	11	110	121	−11
102	104	−2	117	118	−1
			130	133	−3

BP_CA

the earlier survey suggested that calcium is more effective for blacks.) The experiment was double-blind. Table 7.4 gives the seated systolic (heart contracted) blood pressure for all subjects at the beginning and end of the 12-week period, in millimeters of mercury (mm Hg). Because the researchers were interested in decreasing blood pressure, Table 7.4 also shows the decrease for each subject. An increase appears as a negative entry.[27]

As usual, we first examine the data. To compare the effects of the two treatments, take the response variable to be the amount of the decrease in blood pressure. Inspection of the data reveals that there are no outliers. Side-by-side boxplots and Normal quantile plots (Figures 7.15 and 7.16) give a more detailed picture. The calcium group has a somewhat short left tail, but there are no severe departures from Normality that will prevent use of t procedures. To examine the question of the researchers who collected these data, we perform a significance test.

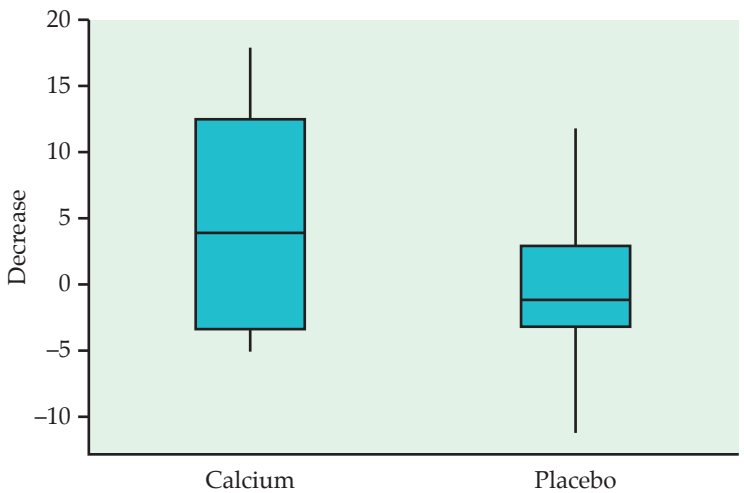


FIGURE 7.15 Side-by-side boxplots of the decrease in blood pressure from Table 7.4.

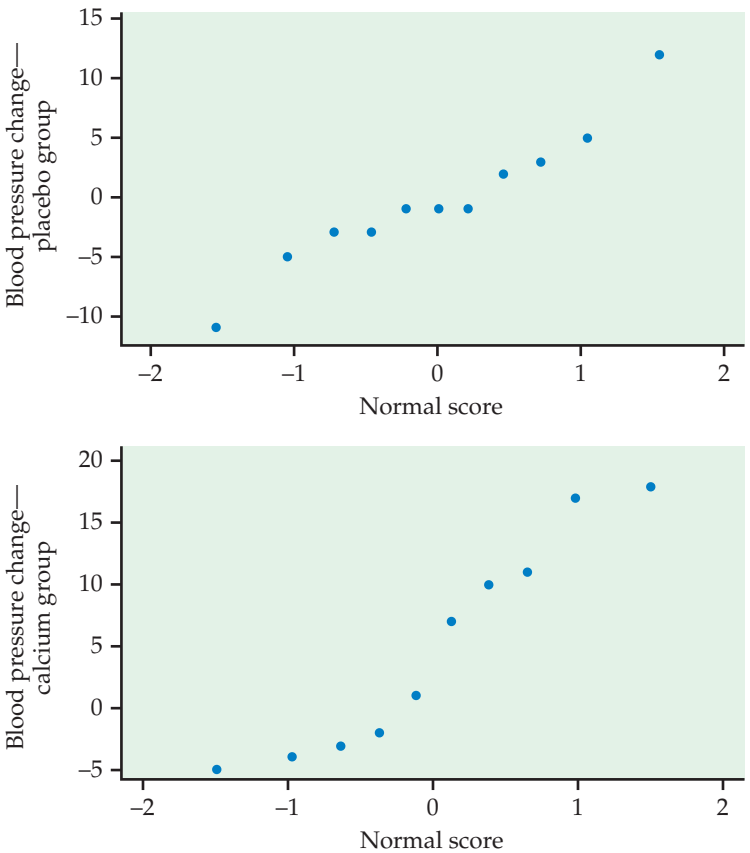


FIGURE 7.16 Normal quantile plots of the change in blood pressure from Table 7.4.

EXAMPLE 7.19

BP_CA

Does increased calcium reduce blood pressure? Take Group 1 to be the calcium group and Group 2 to be the placebo group. The evidence that calcium lowers blood pressure more than a placebo is assessed by testing

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 > \mu_2$$

Here are the summary statistics for the decrease in blood pressure:

Group	Treatment	n	$\bar{x}$	s
1	Calcium	10	5.000	8.743
2	Placebo	11	−0.273	5.901

The calcium group shows a drop in blood pressure, and the placebo group has a small increase. The sample standard deviations do not rule out equal population standard deviations. A difference this large will often arise by chance in samples this small. We are willing to assume equal population standard deviations. The pooled sample variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
$$= \frac{(10 - 1)8.743^2 + (11 - 1)5.901^2}{10 + 11 - 2} = 54.536$$

so that

$$s_p = \sqrt{54.536} = 7.385$$

The pooled two-sample t statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{5.000 - (-0.273)}{7.385\sqrt{\frac{1}{10} + \frac{1}{11}}}$$

$$= \frac{5.273}{3.227} = 1.634$$

The P-value is $P(T \geq 1.634)$, where T has the $t(19)$ distribution.

From Table D, we can see that P falls between the $\alpha = 0.10$ and $\alpha = 0.05$ levels. Statistical software gives the exact value $P = 0.059$. The experiment found evidence that calcium reduces blood pressure, but the evidence falls a bit short of the traditional 5% and 1% levels.

df = 19

p	0.10	0.05
t^*	1.328	1.729

Sample size strongly influences the P-value of a test. An effect that fails to be significant at a specified level α in a small sample can be significant in a larger sample. In the light of the rather small samples in Example 7.19, the evidence for some effect of calcium on blood pressure is rather good. The published account of the study combined these results for blacks with the results for whites and adjusted for pretest differences among the subjects. Using this more detailed analysis, the researchers were able to report a P-value of 0.008.

Of course, a P-value is almost never the last part of a statistical analysis. To make a judgment regarding the size of the effect of calcium on blood pressure, we need a confidence interval.

EXAMPLE 7.20

BP_CA

How different are the calcium and placebo groups? We estimate that the effect of calcium supplementation is the difference between the sample means of the calcium and the placebo groups, $\bar{x}_1 - \bar{x}_2 = 5.273$ mm Hg. A 90% confidence interval for $\mu_1 - \mu_2$ uses the critical value $t^* = 1.729$ from the $t(19)$ distribution. The interval is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = [5.000 - (-0.273)] \pm (1.729)(7.385)\sqrt{\frac{1}{10} + \frac{1}{11}}$$

$$= 5.273 \pm 5.579$$

We are 90% confident that the difference in means is in the interval $(-0.306, 10.852)$. The calcium treatment reduced blood pressure by about 5.3 mm Hg more than a placebo on the average, but the margin of error for this estimate is 5.6 mm Hg.

The pooled two-sample t procedures are anchored in statistical theory and so have long been the standard version of the two-sample t in textbooks. *But they require the assumption that the two unknown population*

standard deviations are equal. We discuss methods to assess this condition in Chapter 12.

The pooled t procedures are therefore a bit risky. They are reasonably robust against both non-Normality and unequal standard deviations when the sample sizes are nearly the same. When the samples are quite different in size, the pooled t procedures become sensitive to unequal standard deviations and should be used with caution unless the samples are large. Unequal standard deviations are quite common. In particular, it is not unusual for the spread of data to increase when the center of the data increases. We recommend regular use of the unpooled t procedures because most software automates the Satterthwaite approximation.

USE YOUR KNOWLEDGE

7.53 Timing of food intake revisited. Figure 7.14 (pages 445–446) gives the outputs from four software packages for comparing the weight loss of two groups with different eating schedules. Some of the software reports both pooled and unpooled analyses. Which outputs give the pooled results? What are the pooled t and its P-value?

7.54 Equal sample sizes. The software outputs in Figure 7.14 give the *same value* for the pooled and unpooled t statistics. Do some simple algebra to show that this is always true when the two sample sizes n_1 and n_2 are the same. In other cases, the two t statistics usually differ.

SECTION 7.2 SUMMARY

- Significance tests and confidence intervals for the difference between the means μ_1 and μ_2 of two Normal populations are based on the difference $\bar{x}_1 - \bar{x}_2$ between the sample means from two independent SRSs. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample sizes are large.

- When independent SRSs of sizes n_1 and n_2 are drawn from two Normal populations with parameters μ_1, σ_1 and μ_2, σ_2 the **two-sample z statistic**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

has the $N(0, 1)$ distribution.

- The **two-sample t statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

does *not* have a t distribution. However, good approximations are available.

- **Conservative inference procedures** for comparing μ_1 and μ_2 are obtained from the two-sample t statistic by using the $t(k)$ distribution with degrees of freedom k equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

- **More accurate probability values** can be obtained by estimating the degrees of freedom from the data. This is the usual procedure for statistical software.
- An approximate level C **confidence interval** for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^*, where k is computed from the data by software or is the smaller of $n_1 - 1$ and $n_2 - 1$. The quantity

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

is the **margin of error**.

- Significance tests for H_0: $\mu_1 - \mu_2 = \Delta_0$ use the **two-sample *t* statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The P-value is approximated using the $t(k)$ distribution where k is estimated from the data using software or is the smaller of $n_1 - 1$ and $n_2 - 1$.

- The guidelines for practical use of two-sample t procedures are similar to those for one-sample t procedures. Equal sample sizes are recommended.
- If we can assume that the two populations have equal variances, **pooled two-sample *t* procedures** can be used. These are based on the **pooled estimator**

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

of the unknown common variance and the $t(n_1 + n_2 - 2)$ distribution. We do not recommend this procedure for regular use.

AU: x-ref page corrected here.

SECTION 7.2 EXERCISES

For Exercises 7.48 and 7.49, see page 439; for Exercises 7.50 and 7.51, see pages 440–441; for Exercise 7.52, see page 448; and for Exercises 7.53 and 7.54, see page 453.

In exercises that call for two-sample t procedures, you may use either of the two approximations for the degrees of freedom that we have discussed: the value given by your software or the smaller of $n_1 - 1$ and $n_2 - 1$. Be sure to state clearly which approximation you have used.

7.55 What is wrong? In each of the following situations, identify what is wrong and then either explain why it is wrong or change the wording of the statement to make it true.

(a) A researcher wants to test H_0: $\bar{x}_1 = \bar{x}_2$ versus the two-sided alternative H_a: $\bar{x}_1 \neq \bar{x}_2$.

(b) A study recorded the IQ scores of 100 college freshmen. The scores of the 56 males in the study were compared with the scores of all 100 freshmen using the two-sample methods of this section.

(c) A two-sample t statistic gave a P-value of 0.94. From this, we can reject the null hypothesis with 90% confidence.

(d) A researcher is interested in testing the one-sided alternative H_a: $\mu_1 < \mu_2$. The significance test gave $t = 2.15$. Because the P-value for the two-sided alternative is 0.036, he concluded that his P-value was 0.018.

7.56 Basic concepts. For each of the following, answer the question and give a short explanation of your reasoning.

(a) A 95% confidence interval for the difference between two means is reported as (0.8, 2.3). What can you conclude about the results of a significance test of the null hypothesis that the population means are equal versus the two-sided alternative?

(b) Will larger samples generally give a larger or smaller margin of error for the difference between two sample means?

7.57 More basic concepts. For each of the following, answer the question and give a short explanation of your reasoning.

(a) A significance test for comparing two means gave $t = -1.97$ with 10 degrees of freedom. Can you reject the null hypothesis that the μ's are equal versus the two-sided alternative at the 5% significance level?

(b) Answer part (a) for the one-sided alternative that the difference between means is negative.

7.58 Physical demands of women's rugby seven matches. Rugby sevens is rapidly growing in popularity and will be included in the 2016 Olympics. Matches are played on a full rugby field and consist of two seven-minute halves. Each team also consists of seven players. To better understand the demands of women's rugby sevens, a group of researchers compared the physical qualities of elite players from the Canadian National team with a university squad. The following table summarizes some of these qualities:[28]

	Elite ($n = 16$)		University ($n = 13$)	
Quality	$\bar{x}$	s	$\bar{x}$	s
Sprint speed (km/hr)	27.3	0.7	26.0	1.5
Peak heart rate (bpm)	192.0	6.0	193.0	6.0
Intermittent recovery test (m)	1160	191	781	129

Carry out the significance tests using $\alpha = 0.05$. Report the test statistic with the degrees of freedom and the P-value. Write a short summary of your conclusion.

7.59 Noise levels in fitness classes. Fitness classes often have very loud music that could affect hearing. One study collected noise levels (decibels) in both high-intensity and low-intensity fitness classes across eight commercial gyms in Sydney, Australia.[29] NOISE

(a) Create a histogram or Normal quantile plot for the high-intensity classes. Do the same for the low-intensity classes. Are the distributions reasonably Normal? Summarize the distributions in words.

(b) Test the equality of means using a two-sided alternative hypothesis and significance level $\alpha = 0.05$.

(c) Are the t procedures appropriate given your observations in part (a)? Explain your answer.

(d) Remove the one low decibel reading for the low-intensity group and redo the significance test. How does this outlier affect the results?

(e) Do you think the results of the significance test from part (b) or (d) should be reported? Explain your answer.

7.60 Noise levels in fitness classes, continued. Refer to the previous exercise. In most countries, the workplace noise standard is 85 db (over eight hours). For every 3 dB increase above that, the amount of exposure time is halved. This means that the exposure time for a dB level of 91 is two hours and for a dB level of 94 it is one hour. NOISE

(a) Construct a 95% confidence interval for the mean dB level in high-intensity classes.

(b) Using the interval in part (a), construct a 95% confidence interval for the number of one-hour classes per day an instructor can teach before possibly risking hearing loss. (*Hint*: This is a linear transformation.)

(c) Repeat parts (a) and (b) for low-intensity classes.

(d) Explain how one might use these intervals to determine the staff size of a new gym.

7.61 When is 30/31 days not equal to a month? Time can be expressed on different levels of scale; days, weeks, months, and years. Can the scale provided influence perception of time? For example, if you placed an order over the phone, would it make a difference if you were told the package would arrive in four weeks or one month? To investigate this, two researchers asked a group of 267 college students to imagine their car needed major repairs and would have to stay at the shop. Depending on the group he or she was randomized to, the student was either told it would take one month or 30/31 days. Each student was then asked to give best- and worst-case estimates of when the car would be ready. The interval between these two estimates (in days) was the response. Here are the results:[30]

Group	n	$\bar{x}$	s
30/31 days	177	20.4	14.3
One month	90	24.8	13.9

(a) Given that the interval cannot be less than 0, the distributions are likely skewed. Comment on the appropriateness of using the t procedures.

(b) Test that the average interval is the same for the two groups using the $\alpha = 0.05$ significance level. Report the test statistic, the degrees of freedom, and the *P*-value. Give a short summary of your conclusion.

7.62 When is 52 weeks not equal to a year? Refer to the previous exercise. The researchers also had 60 marketing students read an announcement about a construction project. The expected duration was either one year or 52 weeks. Each student was then asked to state the earliest and latest completion date.

Group	n	$\bar{x}$	s
52 weeks	30	84.1	55.8
1 year	30	139.6	73.1

Test that the average interval is the same for the two groups using the $\alpha = 0.05$ significance level. Report the test statistic, the degrees of freedom, and the *P*-value. Give a short summary of your conclusion.

7.63 Trustworthiness and eye color. Why do we naturally tend to trust some strangers more than others? One group of researchers decided to study the relationship between eye color and trustworthiness.[31] In their experiment, the researchers took photographs of 80 students (20 males with brown eyes, 20 males with blue eyes, 20 females with brown eyes, and 20 females with blue eyes), each seated in front of a white background looking directly at the camera with a neutral expression. These photos were cropped so the eyes were horizontal and at the same height in the photo and so the neckline was visible. They then recruited 105 participants to judge the trustworthiness of each student photo. This was done using a 10-point scale, where 1 meant very untrustworthy and 10 very trustworthy. The 80 scores from each participant were then converted to z -scores, and the average z -score of each photo (across all 105 participants) was used for the analysis. Here is a summary of the results:

Eye color	n	$\bar{x}$	s
Brown	40	0.55	1.68
Blue	40	−0.38	1.53

Can we conclude from these data that brown-eyed students appear more trustworthy compared to their blue-eyed counterparts? Test the hypothesis that the average scores for the two groups are the same.

7.64 Facebook use in college. Because of Facebook's rapid rise in popularity among college students, there is a great deal of interest in the relationship between Facebook use and academic performance. One study collected information on $n = 1839$ undergraduate students to look at the relationships among frequency of Facebook use, participation in Facebook activities, time spent preparing for class, and overall GPA.[32]

Students reported preparing for class an average of 706 minutes per week with a standard deviation of 526 minutes. Students also reported spending an average of 106 minutes per day on Facebook with a standard deviation of 93 minutes; 8% of the students reported spending no time on Facebook.

(a) Construct a 95% confidence interval for the average number of minutes per week a student prepares for class.

(b) Construct a 95% confidence interval for the average number of minutes per week a student spends on Facebook. (*Hint:* Be sure to convert from minutes per day to minutes per week.)

(c) Explain why you might expect the population distributions of these two variables to be highly skewed to the right. Do you think this fact makes your confidence intervals invalid? Explain your answer.

7.65 Possible biases? Refer to the previous exercise. The researcher surveyed students at a four-year, public university in the northeastern United States ($N = 3866$). Each student was emailed a link to the survey hosted on SurveyMonkey.com. The researcher also states:

> *For the students who did not participate immediately, two additional reminders were sent, one week apart. Participants were offered a chance to enter a drawing to win one of 90 $10 Amazon.com gift cards as incentive. A total of 1839 surveys were completed for an overall response rate of 48%.*

Discuss how these factors influence your interpretation of the results of this survey.

7.66 Comparing means. Refer to Exercise 7.64. Suppose that you wanted to compare the average minutes per week spent on Facebook with the average minutes per week spent preparing for class.

(a) Provide an estimate of this difference.

(b) Explain why it is incorrect to use the two-sample *t* test to see if the means differ.

7.67 Sadness and spending. The "misery is not miserly" phenomenon refers to a person's spending judgment going haywire when the person is sad. In a study, 31 young adults were given $10 and randomly assigned to either a sad or a neutral group. The participants in the sad group watched a video about the death of a boy's mentor (from *The Champ*), and those in the neutral group watched a video on the Great Barrier Reef. After the video, each participant was offered the chance to trade $0.50 increments of the $10 for an insulated water bottle.[33] Here are the data: SADNESS

Group	Purchase price ($)								
Neutral	0.00	2.00	0.00	1.00	0.50	0.00	0.50		
	2.00	1.00	0.00	0.00	0.00	0.00	1.00		
Sad	3.00	4.00	0.50	1.00	2.50	2.00	1.50	0.00	1.00
	1.50	1.50	2.50	4.00	3.00	3.50	1.00	3.50	

(a) Examine each group's prices graphically. Is use of the t procedures appropriate for these data? Carefully explain your answer.

(b) Make a table with the sample size, mean, and standard deviation for each of the two groups.

(c) State appropriate null and alternative hypotheses for comparing these two groups.

(d) Perform the significance test at the $\alpha = 0.05$ level, making sure to report the test statistic, degrees of freedom, and P-value. What is your conclusion?

(e) Construct a 95% confidence interval for the mean difference in purchase price between the two groups.

AU: Please note adjustment here.

7.68 Diet and mood. Researchers were interested in comparing the long-term psychological effects of being on a high-carbohydrate, low-fat (LF) diet versus a high-fat, low-carbohydrate (LC) diet.[34] A total of 106 overweight and obese participants were randomly assigned to one of these two energy-restricted diets. At 52 weeks, 32 LC dieters and 33 LF dieters remained. Mood was assessed using a total mood disturbance score (TMDS), where a lower score is associated with a less negative mood. A summary of these results follows:

Group	n	$\bar{x}$	s
LC	32	47.3	28.3
LF	33	19.3	25.8

(a) Is there a difference in the TMDS at Week 52? Test the null hypothesis that the dieters' average mood in the two groups is the same. Use a significance level of 0.05.

(b) Critics of this study focus on the specific LC diet (that it, the science) and the dropout rate. Explain why the dropout rate is important to consider when drawing conclusions from this study.

7.69 Drive-thru customer service. *QSRMagazine.com* assessed 1855 drive-thru visits at quick-service restaurants.[35] One benchmark assessed was customer service. Responses ranged from "Rude (1)" to "Very Friendly (5)." The following table breaks down the responses according to two of the chains studied. DRVTHRU

	Rating				
Chain	1	2	3	4	5
Taco Bell	0	5	41	143	119
McDonald's	1	22	55	139	100

(a) A researcher decides to compare the average rating of McDonald's and Taco Bell. Comment on the appropriateness of using the average rating for these data.

(b) Assuming an average of these ratings makes sense, comment on the use of the t procedures for these data.

(c) Report the means and standard deviations of the ratings for each chain separately.

(d) Test whether the two chains, on average, have the same customer satisfaction. Use a two-sided alternative hypothesis and a significance level of 5%.

7.70 Comparison of two web page designs. You want to compare the daily number of hits for two different website designs for your indie rock band. You assign the next 30 days to either Design A or Design B, 15 days to each.

(a) Would you use a one-sided or a two-sided significance test for this problem? Explain your choice.

(b) If you use Table D to find the critical value, what are the degrees of freedom using the second approximation?

(c) If you perform the significance test using $\alpha = 0.05$, how large (positive or negative) must the t statistic be to reject the null hypothesis that the two designs result in the same average hits?

7.71 Comparison of dietary composition. Refer to Example 7.15 (page 443). That study also broke down the dietary composition of the main meal. The following table summarizes the total fats, protein, and carbohydrates in the main meal (g) for the two groups:

	Early eaters (n = 202)		Late eaters (n = 200)	
	$\bar{x}$	s	$\bar{x}$	s
Fats	23.1	12.5	21.4	8.2
Protein	27.6	8.6	25.7	6.8
Carbohydrates	64.1	21.0	63.5	20.8

(a) Is it appropriate to use the two-sample t procedures that we studied in this section to analyze these data for group differences? Give reasons for your answer.

(b) Describe appropriate null and alternative hypotheses for comparing the two groups in terms of fats consumed.

(c) Carry out the significance test using $\alpha = 0.05$. Report the test statistic with the degrees of freedom and the P-value. Write a short summary of your conclusion.

(d) Find a 95% confidence interval for the difference between the two means. Compare the information given by the interval with the information given by the significance test.

7.72 More on dietary composition. Refer to the previous exercise. Repeat parts (b) through (d) for protein and for carbohydrates. Combining these results with the results of Exercise 7.71, write a short summary of your findings.

7.73 Change in portion size. A study of food portion sizes reported that over a 17-year period, the average size of a soft drink consumed by Americans aged two years and older increased from 13.1 ounces (oz) to 19.9 oz. The authors state that the difference is statistically significant with $P < 0.01$.[36] Explain what additional information you would need to compute a confidence interval for the increase, and outline the procedure that you would use for the computations. Do you think that a confidence interval would provide useful additional information? Explain why or why not.

7.74 Beverage consumption. The results in the previous exercise were based on two national surveys with a very large number of individuals. Here is a study that also looked at beverage consumption, but the sample sizes were much smaller. One part of this study compared 20 children who were 7 to 10 years old with 5 children who were 11 to 13.[37] The younger children consumed an average of 8.2 oz of sweetened drinks per day, while the older ones averaged 14.5 oz. The standard deviations were 10.7 oz and 8.2 oz, respectively.

(a) Do you think that it is reasonable to assume that these data are Normally distributed? Explain why or why not. (*Hint:* Think about the 68–95–99.7 rule.)

(b) Using the methods in this section, test the null hypothesis that the two groups of children consume equal amounts of sweetened drinks versus the two-sided alternative. Report all details of the significance-testing procedure with your conclusion.

(c) Give a 95% confidence interval for the difference in means.

(d) Do you think that the analyses performed in parts (b) and (c) are appropriate for these data? Explain why or why not.

(e) The children in this study were all participants in an intervention study at the Cornell Summer Day Camp at Cornell University. To what extent do you think that these results apply to other groups of children?

7.75 Study design is important! Recall Exercise 7.70 (page 457). You are concerned that day of the week may affect the number of hits. So to compare the two web page designs, you choose two successive weeks in the middle of a month. You flip a coin to assign one Monday to the first design and the other Monday to the second. You repeat this for each of the seven days of the week. You now have seven hit amounts for each design. It is *incorrect* to use the two-sample t test to see if the mean hits differ for the two designs. Carefully explain why.

7.76 New hybrid tablet and laptop? The purchasing department has suggested your company switch to a new hybrid tablet and laptop. As CEO, you want data to be assured that employees will like these new hybrids over the old laptops. You designate the next 16 employees needing a new laptop to participate in an experiment in which eight will be randomly assigned to receive the standard laptop and the remainder will receive the new hybrid tablet and laptop. After a month of use, these employees will express their satisfaction with their new computers by responding to the statement "I like my new computer" on a scale from 1 to 5, where 1 represents "strongly disagree," 2 is "disagree," 3 is "neutral," 4 is "agree," and 5 is "strongly agree."

(a) The employees with the hybrid computers have an average satisfaction score of 4.3 with standard deviation 0.7. The employees with the standard laptops have an average of 3.7 with standard deviation 1.5. Give a 95% confidence interval for the difference in the mean satisfaction scores for all employees.

(b) Would you reject the null hypothesis that the mean satisfaction for the two types of computers is the same versus the two-sided alternative at significance level 0.05? Use your confidence interval to answer this question. Explain why you do not need to calculate the test statistic.

7.77 Why randomize? Refer to the previous exercise. A coworker suggested that you give the new hybrid computers to the next eight employees who need new computers and the standard laptop to the following eight. Explain why your randomized design is better.

7.78 Does ad placement matter? Corporate advertising tries to enhance the image of the corporation. A study compared two ads from two sources, the *Wall Street Journal* and the *National Enquirer*. Subjects were asked to pretend that their company was considering a major investment in Performax, the fictitious sportswear firm in the ads. Each subject was asked to respond to the question "How trustworthy was the source in the sportswear company ad for Performax?" on a 7-point scale. Higher values indicated more trustworthiness.[38] Here is a summary of the results:

Ad source	n	$\bar{x}$	s
Wall Street Journal	66	4.77	1.50
National Enquirer	61	2.43	1.64

(a) Compare the two sources of ads using a t test. Be sure to state your null and alternative hypotheses, the test statistic with degrees of freedom, the P-value, and your conclusion.

(b) Give a 95% confidence interval for the difference.

(c) Write a short paragraph summarizing the results of your analyses.

AU: Check x-ref

7.79 Size of trees in the northern and southern halves. The study of 584 longleaf pine trees in the Wade Tract in Thomas County, Georgia, had several purposes. Are trees in one part of the tract more or less like trees in any other part of the tract or are there differences? In Example 6.1 (page 342), we examined how the trees were distributed in the tract and found that the pattern was not random. In this exercise, we will examine the sizes of the trees. In Exercise 7.33 (page 429), we analyzed the sizes, measured as diameter at breast height (DBH), for a random sample of 40 trees. Here, we divide the tract into northern and southern halves and take random samples of 30 trees from each half. Here are the diameters in centimeters (cm) of the sampled trees: NSPINES

	27.8	14.5	39.1	3.2	58.8	55.5	25.0	5.4	19.0	30.6
North	15.1	3.6	28.4	15.0	2.2	14.2	44.2	25.7	11.2	46.8
	36.9	54.1	10.2	2.5	13.8	43.5	13.8	39.7	6.4	4.8
	44.4	26.1	50.4	23.3	39.5	51.0	48.1	47.2	40.3	37.4
South	36.8	21.7	35.7	32.0	40.4	12.8	5.6	44.3	52.9	38.0
	2.6	44.6	45.5	29.1	18.7	7.0	43.8	28.3	36.9	51.6

(a) Use a back-to-back stemplot and side-by-side boxplots to examine the data graphically. Describe the patterns in the data.

(b) Is it appropriate to use the methods of this section to compare the mean DBH of the trees in the north half of the tract with the mean DBH of the trees in the south half? Give reasons for your answer.

(c) What are appropriate null and alternative hypotheses for comparing the two samples of tree DBHs? Give reasons for your choices.

(d) Perform the significance test. Report the test statistic, the degrees of freedom, and the *P*-value. Summarize your conclusion.

(e) Find a 95% confidence interval for the difference in mean DBHs. Explain how this interval provides additional information about this problem.

7.80 Size of trees in the eastern and western halves. Refer to the previous exercise. The Wade Tract can also be divided into eastern and western halves. Here are the DBHs of 30 randomly selected longleaf pine trees from each half: EWPINES

	23.5	43.5	6.6	11.5	17.2	38.7	2.3	31.5	10.5	23.7
East	13.8	5.2	31.5	22.1	6.7	2.6	6.3	51.1	5.4	9.0
	43.0	8.7	22.8	2.9	22.3	43.8	48.1	46.5	39.8	10.9
	17.2	44.6	44.1	35.5	51.0	21.6	44.1	11.2	36.0	42.1
West	3.2	25.5	36.5	39.0	25.9	20.8	3.2	57.7	43.3	58.0
	21.7	35.6	30.9	40.6	30.7	35.6	18.2	2.9	20.4	11.4

Using the questions in the previous exercise, analyze these data.

7.81 Sales of a small appliance across months. A market research firm supplies manufacturers with estimates of the retail sales of their products from samples of retail stores. Marketing managers are prone to look at the estimate and ignore sampling error. Suppose that an SRS of 60 stores this month shows mean sales of 53 units of a small appliance, with standard deviation 12 units. During the same month last year, an SRS of 58 stores gave mean sales of 50 units, with standard deviation 10 units. An increase from 50 to 53 is a rise of 6%. The marketing manager is happy because sales are up 6%.

(a) Use the two-sample *t* procedure to give a 95% confidence interval for the difference in mean number of units sold at all retail stores.

(b) Explain in language that the manager can understand why he cannot be certain that sales rose by 6%, and that in fact sales may even have dropped.

7.82 An improper significance test. A friend has performed a significance test of the null hypothesis that two means are equal. His report states that the null hypothesis is rejected in favor of the alternative that the first mean is larger than the second. In a presentation on his work, he notes that the first sample mean was larger than the second mean and this is why he chose this particular one-sided alternative.

(a) Explain what is wrong with your friend's procedure and why.

(b) Suppose that he reported $t = 1.93$ with a *P*-value of 0.06. What is the correct *P*-value that he should report?

7.83 Breast-feeding versus baby formula. A study of iron deficiency among infants compared samples of infants following different feeding regimens. One group contained breast-fed infants, while the infants in another group were fed a standard baby formula without any iron supplements. Here are summary results on blood hemoglobin levels at 12 months of age:[39]

Group	n	$\bar{x}$	s
Breast-fed	23	13.3	1.7
Formula	19	12.4	1.8

(a) Is there significant evidence that the mean hemoglobin level is higher among breast-fed babies? State H_0 and H_a and carry out a *t* test. Give the *P*-value. What is your conclusion?

(b) Give a 95% confidence interval for the mean difference in hemoglobin level between the two populations of infants.

(c) State the assumptions that your procedures in parts (a) and (b) require in order to be valid.

7.84 Revisiting the sadness and spending study. In Exercise 7.67 (page 456), the purchase price of a water bottle was analyzed using the two-sample t procedures that do not assume equal standard deviations. Compare the means using a significance test and find the 95% confidence interval for the difference using the pooled methods. How do the results compare with those you obtained in Exercise 7.67? SADNESS

7.85 Revisiting the diet and mood study. In Exercise 7.68 (page 457), the total mood disturbance score means were compared using the two-sample t procedures that do not assume equal standard deviations. Compare the means using a significance test and find the 95% confidence interval for the difference using the pooled methods. How do the results compare with those you obtained in Exercise 7.68?

7.86 Revisiting dietary composition. In Exercise 7.71 (page 457), the total amount of fats was analyzed using the two-sample t procedures that do not assume equal standard deviations. Compare the means using a significance test and find the 95% confidence interval for the difference using the pooled methods. How do the results compare with those you obtained in Exercise 7.71?

7.87 Revisiting the size of trees. Refer to the Wade Tract DBH data in Exercise 7.79 (page 459), where we compared a sample of trees from the northern half of the tract with a sample from the southern half. Because the standard deviations for the two samples are quite close, it is reasonable to analyze these data using the pooled procedures. Perform the significance test and find the 95% confidence interval for the difference in means using these methods. Summarize your results and compare them with what you found in Exercise 7.79. NSPINES

7.88 Revisiting the food-timing study. Example 7.15 (page 443) gives summary statistics for weight loss in early eaters and late eaters. The two sample standard deviations are quite similar, so we may be willing to assume equal population standard deviations. Calculate the pooled t test statistic and its degrees of freedom from the summary statistics. Use Table D to assess significance. How do your results compare with the unpooled analysis in the example?

7.89 Computing the degrees of freedom. Use the Wade Tract data in Exercise 7.79 to calculate the software approximation to the degrees of freedom using the formula on page 447. Verify your calculation with software. NSPINES

7.90 Again computing the degrees of freedom. Use the Wade Tract data in Exercise 7.80 (page 459) to calculate the software approximation to the degrees of freedom using the formula on page 447. Verify your calculation with software. EWPINES

7.91 Revisiting the small-sample example. Refer to Example 7.16 (page 444). This is a case where the sample sizes are quite small. With only five observations per group, we have very little information to make a judgment about whether the population standard deviations are equal. The potential gain from pooling is large when the sample sizes are small. Assume that we will perform a two-sided test using the 5% significance level. EATER

(a) Find the critical value for the unpooled t test statistic that does not assume equal variances. Use the minimum of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom.

(b) Find the critical value for the pooled t test statistic.

(c) How does comparing these critical values show an advantage of the pooled test?

7.92 Two-sample test of equivalence. In Section 7.1 (page 421), we were introduced to the one-sample test of equivalence. Using those same concepts, describe how to perform a two-sample test of equivalence.

7.3 Additional Topics on Inference

When you complete this section, you will be able to:

- Compute the sample size n needed for a desired margin of error for a mean μ.
- Define the power of a significance test.
- Calculate the power of the one-sample t-test to detect an alternative for a given sample size n.
- Determine the sample size necessary to have adequate power to detect a scaled difference in means of size δ.
- Identify alternative strategies of inference for non-Normal populations.

In this section, we discuss two topics that are related to the procedures we have learned for inference about population means. First, we focus on planning a study—in particular, choosing the sample size. *A wise user of statistics does not plan for inference without at the same time planning data collection.* The second topic introduces us to various inference methods for non-Normal populations. These would be used when our populations are clearly non-Normal and we do not think that the sample size is large enough to rely on the robustness of the t procedures.

Choosing the sample size

We describe sample size procedures for both confidence intervals and significance tests. For anyone planning to design a study, a general understanding of these procedures is necessary. While the actual formulas are a bit technical, statistical software now makes it trivial to get sample size results.

AU: Check edits

Sample size for confidence intervals We can arrange to have both high confidence and a small margin of error by choosing an appropriate sample size. Let's first focus on the one-sample t confidence interval. Its margin of error is

$$m = t^*\text{SE}_{\bar{x}} = t^*\frac{s}{\sqrt{n}}$$

Besides the confidence level C and sample size n, this margin of error depends on the sample standard deviation s. Because we don't know the value of s until we collect the data, we guess a value to use in the calculations. Because s is our estimate of the population standard deviation σ, this value can also be considered our guess of the population standard deviation.

We will call this guessed value s^*. We typically guess at this value using results from a pilot study or from similar published studies. *It is always better to use a value of the standard deviation that is a little larger than what is expected.* This may result in a sample size that is a little larger than needed, but it helps avoid the situation where the resulting margin of error is larger than desired.

Given an estimate for s and the desired margin of error m, we can find the sample size by plugging everything into the margin of error formula and solving for n. The one complication, however, is that t^* depends not only on the confidence level C, but also on the sample size n. Here are the details.

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR FOR A MEAN μ

The level C confidence interval for a mean μ will have an expected margin of error less than or equal to a specified value m when the sample size is such that

$$m \geq t^*s^*/\sqrt{n}$$

Here t^* is the critical value for confidence level C with $n - 1$ degrees of freedom, and s^* is the guessed value for the population standard deviation.

Finding the smallest sample size n that satisfies this requirement can be done using the following iterative search:

1. Get an initial sample size by replacing t^* with z^*. Compute $n = (z^*s^*/m)^2$ and round up to the nearest integer.
2. Use this sample size to obtain t^*, and check if $m \geq t^*s^*/\sqrt{n}$.
3. If the requirement is satisfied, then this n is the needed sample size. If the requirement is not satisfied, increase n by 1 and return to Step 2.

Notice that this method makes no reference to the size of the *population*. It is the size of the *sample* that determines the margin of error. The size of the population does not influence the sample size we need as long as the population is much larger than the sample. Here is an example.

EXAMPLE 7.21

Planning a survey of college students. In Example 7.1 (page 411), we calculated a 95% confidence interval for the mean hours per week a college student watches traditional television. The margin of error based on an SRS of $n = 8$ students was 12.42 hours. Suppose that a new study is being planned and the goal is to have a margin of error of five hours. How many students need to be sampled?

The sample standard deviation in Example 7.1 is $s = 14.854$ hours. To be conservative, we'll guess that the population standard deviation is 17.5 hours.

1. To compute an initial n, we replace t^* with z^*. This results in

$$n = \left(\frac{z^*s^*}{m}\right)^2 = \left[\frac{1.96(17.5)}{5}\right]^2 = 47.06$$

Round up to get $n = 48$.

2. We now check to see if this sample size satisfies the requirement when we switch back to t^*. For $n = 48$, we have $n - 1 = 47$ degrees of freedom and $t^* = 2.011$. Using this value, the expected margin of error is

$$2.011(17.5)/\sqrt{48} = 5.08$$

This is larger than $m = 5$, so the requirement is not satisfied.

3. The following table summarizes these calculations for some larger values of n.

n	$t^*s^*/\sqrt{n}$
49	5.03
50	4.97
51	4.92

The requirement is first satisfied when $n = 50$. Thus, we need to sample at least $n = 50$ students for the expected margin of error to be no more than five hours.

Figure 7.17 shows the Minitab input window used to do these calculations. Because the default confidence level is 95%, only the desired margin of error m and the estimate for s need to be entered.

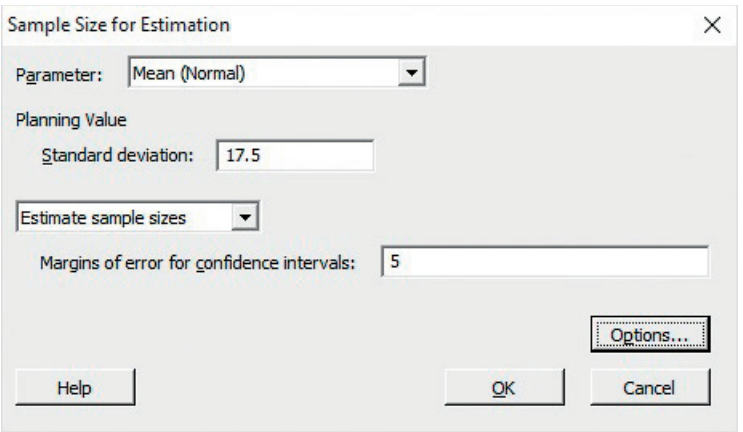


FIGURE 7.17 Minitab input window used to compute the sample size for the desired margin of error, Example 7.21.

AU: insert "from"?

Note that the $n = 50$ refers to the *expected* margin of error being no more than five hours. This does not guarantee that the margin of error for the collected sample will be less than five hours. That is because the sample standard deviation s varies sample to sample and these calculations are treating it as a fixed quantity. More advanced sample size procedures ask you to also specify the probability of obtaining a margin of error less than the desired value. For our approach, this probability is roughly 50%. For a probability closer to 100%, the sample size will need to be larger. For example, suppose we wanted this probability to be roughly 80%. In SAS, we'd perform these calculations using the command

```
proc power;
  onesamplemeans CI=t stddev=17.5 halfwidth=5 probwidth=0.80 ntotal=.;
run;
```

The needed sample size increases from $n = 50$ to $n = 57$.

Unfortunately, the actual number of usable observations is often less than that planned at the beginning of a study. This is particularly true of data collected in surveys or studies that involve a time commitment from the participants. Careful study designers often assume a nonresponse rate or dropout rate that specifies what proportion of the originally planned sample will fail to provide data. We use this information to calculate the sample size to be used at the start of the study. For example, if, in the preceding survey, we expect only 40% of those students to respond, we would need to start with a sample size of $2.5 \times 50 = 125$ to obtain usable information from 50 students.

nonresponse, p. 196

These sample size calculations also do not account for collection costs. In practice, taking observations costs time and money. There are times when the required sample size may be impossibly expensive. In those situations, one might consider a larger margin of error and/or a lower confidence level to be acceptable.

For the two-sample t confidence interval, the margin of error is

$$m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

A similar type of iterative search can be used to determine the sample sizes n_1 and n_2, but now we need to guess both standard deviations and decide on an estimate for the degrees of freedom. We suggest taking the conservative approach and using the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of

freedom. Another approach is to consider the standard deviations and sample sizes are equal, so the margin of error is

$$m = t^* \sqrt{\frac{2s^2}{n}}$$

AU: Please note edits here.

and use degrees of freedom $2(n - 1)$. That is the approach most statistical software takes.

EXAMPLE 7.22

Planning a new blood pressure study. In Example 7.20 (page 452), we calculated a 90% confidence interval for the mean difference in blood pressure. The 90% margin of error was roughly 5.6 mm Hg. Suppose that a new study is being planned and the desired margin of error at 90% confidence is 2.8 mm Hg. How many subjects per group do we need?

The pooled sample standard deviation in Example 7.20 is 7.385. To be a bit conservative, we'll guess that the two population standard deviations are both 8.0. To compute an initial n, we replace t^* with z^*. This results in

$$n = \left(\frac{\sqrt{2}z^*s^*}{m}\right)^2 = \left[\frac{\sqrt{2}(1.645)(8)}{2.8}\right]^2 = 44.2$$

We round up to get $n = 45$. The following table summarizes the margin of error for this and some larger values of n.

n	$t^*s^*/\sqrt{2/n}$
45	2.834
46	2.801
47	2.770

The requirement is first satisfied when $n = 47$. In SAS, we'd perform these calculations using the command

```
proc power;
 twosamplemeans CI=diff alpha=0.1 stddev=8 halfwidth=2.8
                probwidth=0.50 npergroup=.;
run;
```

This sample size is roughly 4.5 times the sample size used in Example 7.20. This researcher may not be able to recruit this large a sample. If so, we should consider a larger margin of error.

USE YOUR KNOWLEDGE

7.93 Starting salaries. In a recent survey by the National Association of Colleges and Employers, the average starting salary for college graduate with a computer and information sciences degree was reported to be \$62,194.[40] You are planning to do a survey of starting salaries for recent computer science majors from your university. Using an estimated standard deviation of \$11,605, what sample size do you need to have a margin of error equal to \$5000 with 95% confidence?

7.94 Changes in sample size. Suppose that, in the setting of the previous exercise, you have the resources to contact 35 recent graduates. If all respond, will your margin of error be larger or smaller than $5000? What if only 50% respond? Verify your answers by performing the calculations.

The power of the one-sample *t* test The power of a statistical test measures its ability to detect deviations from the null hypothesis. In practice, we carry out the test in the hope of showing that the null hypothesis is false, so high power is important. Power calculations are a way to assess whether or not a sample size is sufficiently large to answer the research question.

The power of the one-sample t test against a specific alternative value of the population mean μ is the probability that the test will reject the null hypothesis when this alternative is true. To calculate the power, we assume a fixed level of significance, usually $\alpha = 0.05$.

AU: Please check x-ref page

LOOK BACK

power calculation, p. 392

Calculation of the exact power of the t test takes into account the estimation of σ by s and requires a new distribution. We will describe that calculation when discussing the power of the two-sample t test. Fortunately, an approximate calculation that is based on assuming that σ is known is almost always adequate for planning a study in the one-sample case. This calculation is very much like that for the z test, presented in Section 6.4. The steps are

1. Write the event, in terms of $\bar{x}$, that the test rejects H_0.
2. Find the probability of this event when the population mean has the alternative value.

Here is an example.

EXAMPLE 7.23

Is the sample size large enough? Recall Example 7.2 (page 413) on the average time that U.S. college students spend watching traditional television. The sample mean of $n = 8$ students was four hours lower than the U.S. average of 18- to 24-year-olds but not found significantly different. Suppose a new study is being planned using a sample size of $n = 50$ students. Does this study have adequate power when the population mean is four hours less than the U.S. average?

We wish to compute the power of the t test for

$$H_0: \mu = 18.5$$
$$H_a: \mu < 18.5$$

against the alternative that $\mu = 18.5 - 4 = 14.5$ when $n = 50$. This gives us most of the information we need to compute the power. The other important piece is a rough guess of the size of σ. In planning a large study, a pilot study is often run for this and other purposes. In this case, we can use the standard deviation from the earlier survey. Similar to Example 7.21 (page 462), we will round up and use $\sigma = 17.5$ and $s = 17.5$ in the approximate calculation.

Step 1. The t test with 50 observations rejects H_0 at the 5% significance level if the t statistic

$$t = \frac{\bar{x} - 18.5}{s/\sqrt{50}}$$

is less than the lower 5% point of $t(49)$, which is -1.677. Taking $s = 17.5$, the event that the test rejects H_0 is, therefore,

$$t = \frac{\bar{x} - 18.5}{17.5/\sqrt{50}} \le -1.677$$

$$\bar{x} \le 18.5 - 1.677\frac{17.5}{\sqrt{50}}$$

$$\bar{x} \le 14.35$$

Step 2. The power is the probability that $\bar{x} \le 14.351$ when $\mu = 14.5$. Taking $\sigma = 17.5$, we find this probability by standardizing $\bar{x}$:

$$P(\bar{x} \le 14.35 \text{ when } \mu = 14.5) = P\left(\frac{\bar{x} - 14.5}{17.5/\sqrt{50}} \le \frac{14.35 - 14.5}{17.5/\sqrt{50}}\right)$$

$$= P(Z \le -0.061)$$

$$= 0.4761$$

A mean value of 14.5 hours per week will produce significance at the 5% level in only 47.6% of all possible samples. Figure 7.18 shows Minitab output for the exact power calculation. It is about 48% and is represented by a dot on the power curve at a difference of -4. This curve is very informative. For many studies, 80% in consider the standard value for desirable power. We see that with a sample size of 50, the power is greater than 80% only for reductions larger than 6.25 hours per week. If we want to detect a reduction of only four hours, we definitely need to increase the sample size.

AU: should this be "is considered"?

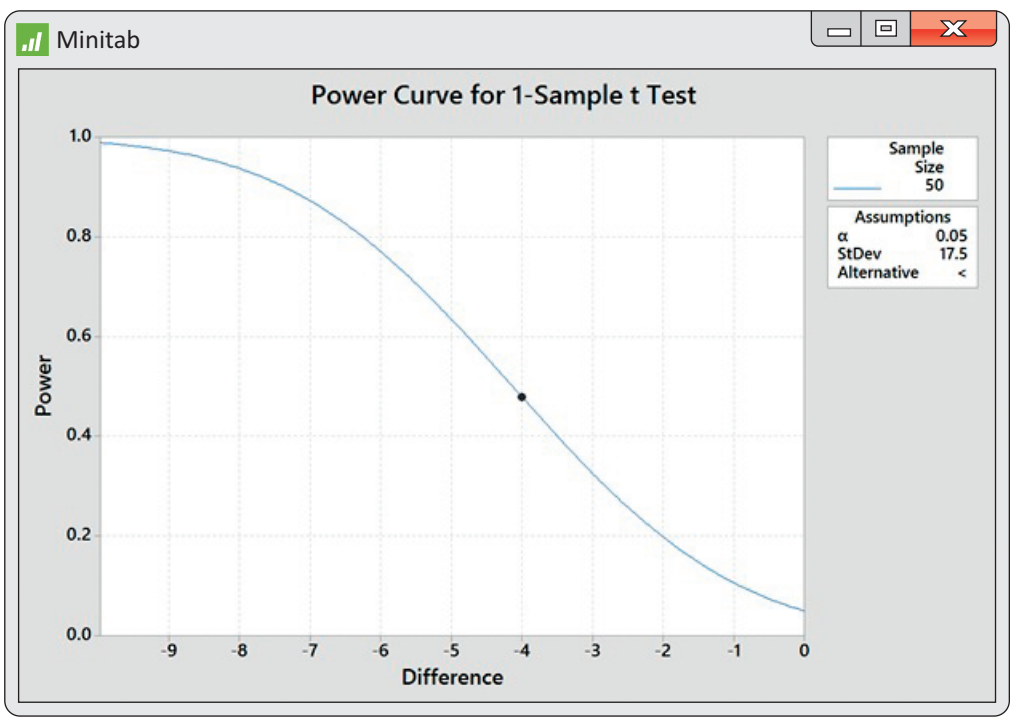


FIGURE 7.18 Minitab output (a power curve) for the one-sample power calculation, Example 7.23.

Power calculations are used in planning studies to ensure that we have a reasonable chance of detecting effects of interest. They give us some guidance in selecting a sample size. In making these calculations, we need assumptions about the standard deviation and the alternative of interest. In our example, we assumed that the standard deviation would be 17.5, but in practice, we are hoping that the value will be somewhere around this value. Similarly, we have used a somewhat arbitrary alternative of 14.5. This is a guess based on the results of the previous study. *Beware of putting too much trust in fine details of the results of these calculations.* They serve as a guide, not a mandate.

USE YOUR KNOWLEDGE

7.95 Power for other values of μ. If you repeat the calculation in Example 7.23 for values of μ that are smaller than 14.5, would you expect the power to be higher or lower than 0.4761? Why?

7.96 Another power calculation. Verify your answer to the previous exercise by doing the calculation for the alternative $\mu = 12$ hours per week.

The power of the two-sample *t* test The two-sample t test is one of the most used statistical procedures. Unfortunately, because of inadequate planning, users frequently fail to find evidence for the effects that they believe to be present. This is often the result of an inadequate sample size. Power calculations, performed prior to running the experiment, will help avoid this occurrence.

We just learned how to approximate the power of the one-sample t test. The basic idea is the same for the two-sample case, but we will describe the exact method rather than an approximation again. The exact power calculation involves a new distribution, the **noncentral *t* distribution**. This calculation is not practical by hand but is easy with software that calculates probabilities for this distribution.

noncentral *t* distribution

We consider only the common case where the null hypothesis is "no difference," $\mu_1 - \mu_2 = 0$. We illustrate the calculation for the pooled two-sample t test. A simple modification is needed when we do not pool. The unknown parameters in the pooled t setting are μ_1, μ_2, and a single common standard deviation σ. To find the power for the pooled two-sample t test, follow these steps.

Step 1. Specify these quantities:

(a) An alternative value for $\mu_1 - \mu_2$ that you consider important to detect.

(b) The sample sizes, n_1 and n_2.

(c) A fixed significance level α, often $\alpha = 0.05$.

(d) An estimate of the standard deviation σ from a pilot study or previous studies under similar conditions.

Step 2. Find the degrees of freedom $\text{df} = n_1 + n_2 - 2$ and the value of t^* that will lead to rejecting H_0 at your chosen level α.

noncentrality parameter

Step 3. Calculate the **noncentrality parameter**

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Step 4. The power is the probability that a noncentral t random variable with degrees of freedom df and noncentrality parameter δ will be greater than t^*. Use software to calculate this probability. In SAS, the command is `1 - PROBT(tstar, df, delta)`. In R the commmand is `1-pt(tstar, df, delta)`. If you do not have software that can perform this calculation, you can approximate the power as the probability that a standard Normal random variable is greater than $t^* - \delta$, that is, $P(Z > t^* - \delta)$. Use Table A or software for standard Normal probabilities.

Note that the denominator in the noncentrality parameter,

$$\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is our guess at the standard error for the difference in the sample means. Therefore, if we wanted to assess a possible study in terms of the margin of error for the estimated difference, we would examine t^* times this quantity.

If we do not assume that the standard deviations are equal, we need to guess both standard deviations and then combine these to get an estimate of the standard error:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This guess is then used in the denominator of the noncentrality parameter. Use the conservative value, the smaller of $n_1 - 1$ and $n_2 - 1$, for the degrees of freedom.

EXAMPLE 7.24

Planning a new study of calcium versus placebo groups. In Example 7.19 (page 451), we examined the effect of calcium on blood pressure by comparing the means of a treatment group and a placebo group using a pooled two-sample t test. The P-value was 0.059, failing to achieve the usual standard of 0.05 for statistical significance. Suppose that we wanted to plan a new study that would provide convincing evidence—say, at the 0.01 level—with high probability. Let's examine a study design with 45 subjects in each group ($n_1 = n_2 = 45$) to see if this meets our goals.

Step 1. Based on our previous results, we choose $\mu_1 - \mu_2 = 5$ as an alternative that we would like to be able to detect with $\alpha = 0.01$. For σ we use 7.4, our pooled estimate from Example 7.19.

Step 2. The degrees of freedom are $n_1 + n_2 - 2 = 88$, which leads to $t^* = 2.37$ for the significance test.

Step 3. The noncentrality parameter is

$$\delta = \frac{5}{7.4\sqrt{\frac{1}{45} + \frac{1}{45}}} = \frac{5}{1.56} = 3.21$$

Step 4. Software gives the power as 0.7965, or 80%. The Normal approximation gives 0.7983, a very accurate result.

With this choice of sample sizes, we are just barely below 80% power. If we judge this to be large enough power, we can proceed to the recruitment of our samples.

With software it is often very easy to examine the effects of variations in a study design. For example, Figure 7.19 shows the JMP power calculator for the two-sample t test. You input values for α, σ, $n_1 + n_2$, and δ (Step 1) and it computes the power (Steps 2–4). Figure 7.19 shows the results of the calculations for Example 7.24. The JMP calculator only considers the two-sided alternative so to get the power for a one-sided alternative, the significance level must be input as 2α. Most other software, such as Minitab, provides the option to choose the alternative.

USE YOUR KNOWLEDGE

7.97 Power and the choice of alternative. If you were to repeat the calculation in Example 7.24 for the two-sided alternative, would the power increase or decrease? Explain your answer.

7.98 Power and the standard deviation. If the true population standard deviation were 8 instead of the 7.4 hypothesized in Example 7.24, would the power increase or decrease? Explain.

7.99 Power and statistical software. Refer to the two previous exercises. Use statistical software to compute the exact power of each scenario.

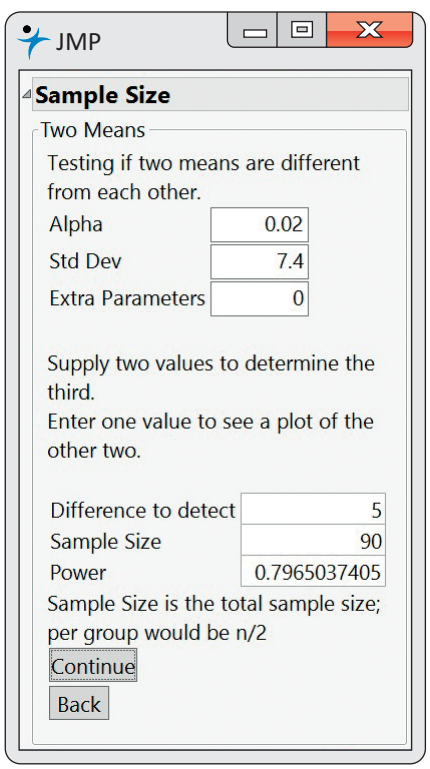


FIGURE 7.19 JMP input/output window for the two-sample power calculation, Example 7.24.

Inference for non-Normal populations

We have not discussed how to do inference about the mean of a clearly non-Normal distribution based on a small sample. If you face this problem, you should consult an expert. Three general strategies are available:

- In some cases, a distribution other than a Normal distribution describes the data well. There are many non-Normal models for data, and inference procedures for these models are available.

- Because skewness is the chief barrier to the use of t procedures on data without outliers, you can attempt to transform skewed data so that the distribution is symmetric and as close to Normal as possible. Confidence levels and P-values from the t procedures applied to the transformed data will be quite accurate for even moderate sample sizes. Methods are generally available for transforming the results back to the original scale.

distribution-free procedures
nonparametric procedures

- Use a **distribution-free** inference procedure. Such procedures do not assume that the population distribution has any specific form, such as Normal. Distribution-free procedures are often called **nonparametric procedures**. Chapter 15 discusses several of these procedures.

Each of these strategies can be effective, but each quickly carries us beyond the basic practice of statistics. We emphasize procedures based on Normal distributions because they are the most common in practice, because their robustness makes them widely useful, and (most important) because we are first of all concerned with understanding the principles of inference. Therefore, we will not discuss procedures for non-Normal continuous distributions. We will be content with illustrating by example the use of a transformation and of a simple distribution-free procedure.

COMP: move this all down one line

LOOK BACK
log transformation, p. 91

Transforming data When the distribution of a variable is skewed, it often happens that a simple transformation results in a variable whose distribution is symmetric and even close to Normal. The most common transformation is the logarithm, or log. The logarithm tends to pull in the right tail of a distribution. For example, the data 2, 3, 4, 20 show an outlier in the right tail. Their common logarithms 0.30, 0.48, 0.60, 1.30 are much less skewed. Taking logarithms is a possible remedy for right-skewness. Instead of analyzing values of the original variable X, we compute their logarithms and analyze the values of $\log X$. Here is an example of this approach.

EXAMPLE 7.25

Length of audio files on an iPod. Table 7.5 presents data on the length (in seconds) of audio files found on an iPod. There was a total of 10,003 audio

Justin Sullivan/Staff/Getty Images

SONGS

TABLE 7.5 Length (in Seconds) of Audio Files Sampled from an iPod

240	316	259	46	871	411	1366
233	520	239	259	535	213	492
315	696	181	357	130	373	245
305	188	398	140	252	331	47
309	245	69	293	160	245	184
326	612	474	171	498	484	271
207	169	171	180	269	297	266
1847						

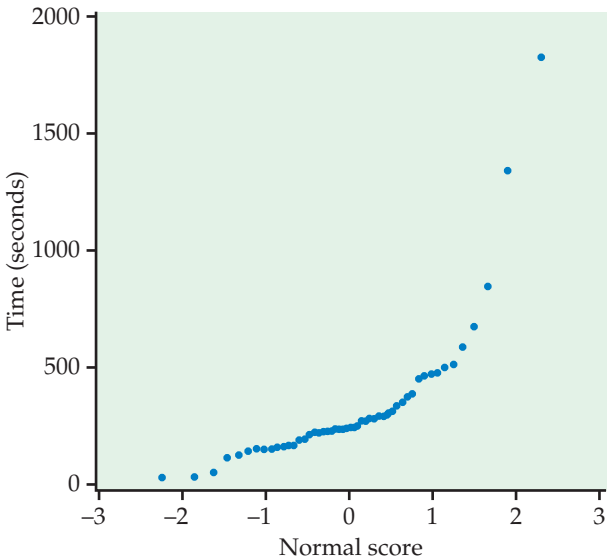


FIGURE 7.20 Normal quantile plot of audio file length, Example 7.25. This sort of pattern occurs when a distribution is skewed to the right.

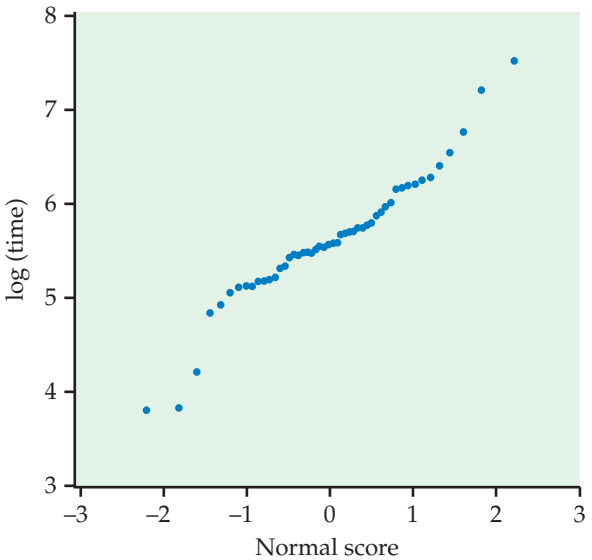


FIGURE 7.21 Normal quantile plot of the logarithms of the audio file lengths, Example 7.25. This distribution appears approximately Normal.

files, and 50 files were randomly selected using the "shuffle songs" command.[41] We would like to give a confidence interval for the average audio file length μ for this iPod.

A Normal quantile plot of the audio data from Table 7.5 (Figure 7.20) shows that the distribution is skewed to the right. Because there are no extreme outliers, the sample mean of the 50 observations will nonetheless have an approximately Normal sampling distribution. The t procedures could be used for approximate inference. For more exact inference, we will transform the data so that the distribution is more nearly Normal. Figure 7.21 is a Normal quantile plot of the natural logarithms of the time measurements. The transformed data are very close to Normal, so t procedures will give quite exact results.

The application of the t procedures to the transformed data is straightforward. Call the original length values from Table 7.5 the variable X. The transformed data are values of $X_{new} = \log X$. In most software packages, it is an easy task to transform data in this way and then analyze the new variable.

EXAMPLE 7.26

Software output of audio length data. Analysis of the natural log of the length values in Minitab produces the following output:

N	Mean	StDev	SE Mean	95.0% C.I.
50	5.6315	0.6840	0.0967	(5.4371, 5.8259)

For comparison, the 95% t confidence interval for the original mean μ is found from the original data as follows:

N	Mean	StDev	SE Mean	95.0% C.I.
50	354.1	307.9	43.6	(266.6, 441.6)

The advantage of analyzing transformed data is that use of procedures based on the Normal distributions is better justified and the results are more exact. The disadvantage is that a confidence interval for the mean μ in the original scale (in our example, seconds) cannot be easily recovered from the confidence interval for the mean of the logs. One approach based on the lognormal distribution[42] results in an interval of (285.5, 435.5), which is narrower and slightly asymmetric compared with the t interval.

Use of a distribution-free procedure Perhaps the most straightforward way to cope with non-Normal data is to use a *distribution-free,* or *nonparametric,* procedure. As the name indicates, these procedures do not require the population distribution to have any specific form, such as Normal. Distribution-free significance tests are quite simple and are available in most statistical software packages.

Distribution-free tests have two drawbacks. First, they are generally less powerful than tests designed for use with a specific distribution, such as the t test. Second, we must often modify the statement of the hypotheses in order to use a distribution-free test. A distribution-free test concerning the center of a distribution, for example, is usually stated in terms of the *median* rather than the mean. This is sensible when the distribution may be skewed. But the distribution-free test does not ask the same question (Has the mean changed?) that the t test does.

sign test

The simplest distribution-free test, and one of the most useful, is the **sign test**. The test gets its name from the fact that we look only at the signs of the differences, not their actual values. The following example illustrates this test.

EXAMPLE 7.27

GEPARTS

The effect of altering a software parameter. Example 7.7 (page 419) describes an experiment to compare the measurements obtained from two software algorithms. In that example, we used the matched pairs t test on these data, despite some skewness, which makes the P-value only roughly correct. The sign test is based on the following simple observation: of the 51 parts measured, 29 had a larger measurement with the option off and 22 had a larger measurement with the option on.

To perform a significance test based on these counts, let p be the probability that a randomly chosen part would have a larger measurement with the option turned on. The null hypothesis of "no effect" says that these two measurements are just repeat measurements, so the measurement with the option off is equally likely to be larger or smaller than the measurement with the option on. Therefore, we want to test

$$H_0: p = 1/2$$
$$H_a: p \neq 1/2$$

AU: Check x-ref page

binomial distribution, p. 312

The 51 parts are independent trials, so the number that had larger measurements with the option off has the binomial distribution $B(51, 1/2)$ if H_0 is true. The P-value for the observed count 29 is, therefore, $2P(X \geq 29)$, where X has the $B(51, 1/2)$ distribution. You can compute this probability with software or the Normal approximation to the binomial:

$$\begin{aligned} 2P(X \geq 29) &= 2P\left(Z \geq \frac{29 - 25.5}{\sqrt{12.75}}\right) \\ &= 2P\,(Z \geq 0.98) \\ &= 2(0.1635) \\ &= 0.3270 \end{aligned}$$

As in Example 7.7, there is not strong evidence that the two measurements are different.

There are several varieties of sign test, all based on counts and the binomial distribution. The sign test for matched pairs is the most useful. The null hypothesis of "no effect" is then always H_0: $p = 1/2$. The alternative can be one-sided in either direction or two-sided, depending on the type of change we are considering.

SIGN TEST FOR MATCHED PAIRS

Ignore pairs with difference 0; the number of trials n is the count of the remaining pairs. The test statistic is the count X of pairs with a positive difference. P-values for X are based on the binomial $B(n, 1/2)$ distribution.

The matched pairs t test in Example 7.7 tested the hypothesis that the mean of the distribution of differences is 0. The sign test in Example 7.27 is, in fact, testing the hypothesis that the *median* of the differences is 0. If p is the probability that a difference is positive, then $p = 1/2$ when the median is 0. This is true because the median of the distribution is the point with probability 1/2 lying to its right. As Figure 7.22 illustrates, $p > 1/2$ when the median is greater than 0, again because the probability to the right of the median is always 1/2. The sign test of H_0: $p = 1/2$ against H_a: $p > 1/2$ is a test of

$$H_0\text{: population median} = 0$$
$$H_a\text{: population median} > 0$$

The sign test in Example 7.27 makes no use of the actual scores—it just counts how many parts had a larger measurement with the option off. Any parts that did not have different measurements would be ignored altogether. Because the sign test uses so little of the available information, it is much less powerful than the t test when the population is close to Normal. Chapter 15 describes other distribution-free tests that are more powerful than the sign test.

USE YOUR KNOWLEDGE

7.100 Sign test for the oil-free frying comparison. Exercise 7.10 (page 422) gives data on the taste of hash browns made using a hot-oil fryer and an oil-free fryer. Is there evidence that the medians are different? State the hypotheses, carry out the sign test, and report your conclusion.

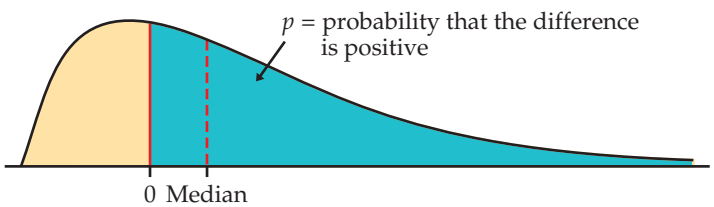


FIGURE 7.22 Why the sign test tests the median difference: when the median is greater than 0, the probability p of a positive difference is greater than 1/2, and vice versa.

SECTION 7.3 SUMMARY

- The **sample size** required to obtain a confidence interval with an expected margin of error no larger than m for a population mean satisfies the constraint

$$m \geq t^* s^* / \sqrt{n}$$

where t^* is the critical value for the desired level of confidence with $n - 1$ degrees of freedom, and s^* is the guessed value for the population standard deviation.

- The sample sizes necessary for a two-sample confidence interval can be obtained using a similar constraint, but guesses of both standard deviations and an estimate for the degrees of freedom are required. We suggest using the smaller of $n_1 - 1$ and $n_2 - 1$ for degrees of freedom.

- The **power** of the one-sample t test can be calculated like that of the z test, using an approximate value for both σ and s.

- The **power** of the two-sample t test is found by first finding the critical value for the significance test, the degrees of freedom, and the **noncentrality parameter** for the alternative of interest. These are used to calculate the power from a **noncentral *t* distribution**. A Normal approximation works quite well. Calculating margins of error for various study designs and conditions is an alternative procedure for evaluating designs.

- The **sign test** is a **distribution-free test** because it uses probability calculations that are correct for a wide range of population distributions.

- The sign test for "no treatment effect" in matched pairs counts the number of positive differences. The P-value is computed from the $B(n, 1/2)$ distribution, where n is the number of non-0 differences. The sign test is less powerful than the t test in cases where use of the t test is justified.

SECTION 7.3 EXERCISES

For Exercise 7.93 and 7.94, see pages 464–465; for Exercises 7.95 and 7.96, see page 467; for Exercises 7.97 through 7.99, see page 469; and for Exercise 7.100, see page 473.

7.101 What is wrong? In each of the following situations, identify what is wrong, and then either explain why it is wrong or change the wording of the statement to make it true.

(a) To reduce the margin of error in half, the sample size needs to be doubled.

(b) The sign test for matched pairs is more powerful than the paired t test when the differences are close to Normal.

(c) When testing H_0: $\mu = 10$ versus the two-sided alternative, the power at $\mu = 3$ is larger than at $\mu = 17$.

(d) Increasing sample size increases the power for all alternatives and decreases the probability of a Type I error.

7.102 Apartment rental rates. You hope to rent an unfurnished one-bedroom apartment in Dallas next year. You call a friend who lives there and ask him to give you an estimate of the mean monthly rate. Having taken a statistics course recently, the friend asks about the desired margin of error and confidence level for this estimate. He also tells you that the standard deviation of monthly rents for one-bedrooms is about $300.

(a) For 95% confidence and a margin of error of $150, how many apartments should the friend randomly sample from the local newspaper?

(b) Suppose that you want the margin of error to be no more than $50. How many apartments should the friend sample?

(c) Why is the sample size in part (b) not just nine times larger than the sample size in part (a)?

7.103 More on apartment rental rates. Refer to the previous exercise. Will the 95% confidence interval include approximately 95% of the rents of all unfurnished one-bedroom apartments in this area? Explain why or why not.

7.104 Average hours per week on the Internet. The *Student Monitor* surveys 1200 undergraduates from 100 colleges semiannually to understand trends among college students.[43] Recently, the *Student Monitor* reported that the average amount of time spent per week on the Internet was 19.0 hours. You suspect that this amount is far too small for your campus and plan a survey.

(a) You feel that a reasonable estimate of the standard deviation is 10.0 hours. What sample size is needed so that the expected margin of error of your estimate is not larger than one hour for 95% confidence?

(b) The distribution of times is likely to be heavily skewed to the right. Do you think that this skewness will invalidate the use of the t confidence interval in this case? Explain your answer.

7.105 Average hours per week listening to the radio. Refer to the previous exercise. The *Student Monitor* also reported that the average amount of time listening to the radio was 11.5 hours.

(a) Given an estimated standard deviation of 5.2 hours, what sample size is needed so that the expected margin of error of your estimate is not larger than one hour for 95% confidence?

(b) If your survey is going to ask about Internet use and radio use, which of the two calculated sample sizes should you use? Explain your answer.

7.106 Accuracy of a laboratory scale. To assess the accuracy of a laboratory scale, a standard weight known to weigh 10 grams is weighed repeatedly. The scale readings are Normally distributed with unknown mean (this mean is 10 grams if the scale has no bias). The standard deviation of the scale readings in the past has been 0.0013 gram.

(a) The weight is measured five times. The mean result is 10.0009 grams. Give a 98% confidence interval for the mean of repeated measurements of the weight.

(b) How many measurements must be averaged to get an expected margin of error no more than 0.001 with 98% confidence?

7.107 Accuracy of a laboratory scale, continued. Refer to the previous exercise. Suppose that instead of a confidence interval, the researchers want to perform a test (with $\alpha = 0.05$) that the scale is unbiased ($\mu = 10$).

(a) What sample size n is necessary to have at least 90% power when the alternative mean is $\mu = 10.001$?

(b) Suppose they can only perform a maximum of $n = 10$ measurements. Based on your answer in part (a), will the power be more or less than 90%? Explain your answer.

(c) Verify your answer in part (b), by computing the power when $n = 10$.

7.108 Sample size calculations. You are designing a study to test the null hypothesis that $\mu = 0$ versus the alternative that μ is positive. Assume that σ is 20. Suppose that it would be important to be able to detect the alternative $\mu = 4$. What sample size is needed to detect this alternative with power of at least 0.80?

7.109 Power of the comparison of DXA machine operators. Suppose that the bone researchers in Exercise 7.45 (page 431) want to be able to detect an alternative mean difference of 0.002. Find the power for this alternative for a sample size of 20 patients. Make sure to explain the reasoning of your choice of standard deviation in these calculations.

7.110 Determining the sample size. Consider Example 7.23 (page 465). What is the minimum sample size needed for the power to be greater than 80% when $\mu = 14.5$?

7.111 Changing the significance level. In Example 7.24 (page 468), we assessed the power of a new study of calcium on blood pressure assuming $n_1 = n_2 = 45$ subjects. The power was based on $\alpha = 0.01$. Suppose that we wanted to use $\alpha = 0.05$ instead.

(a) Would the power increase or decrease? Explain your answer in terms someone unfamiliar with power calculations can understand.

(b) Verify your answer by computing the power.

7.112 Planning a study to compare tree size. In Exercise 7.79 (page 459), DBH data for longleaf pine trees in two parts of the Wade Tract are compared. Suppose that you are planning a similar study in which you will measure the diameters of longleaf pine trees. Based on Exercise 7.79, you are willing to assume that the standard deviation for both halves is 20 cm. Suppose that a difference in mean DBH of 10 cm or more would be important to detect. You will use a t statistic and a two-sided alternative for the comparison.

(a) Find the power if you randomly sample 20 trees from each area to be compared.

(b) Repeat the calculations for 60 trees in each sample.

(c) If you had to choose between the 20 and 60 trees per sample, which would you choose? Give reasons for your answer.

7.113 More on planning a study to compare tree size. Refer to the previous exercise. Find the two standard deviations from Exercise 7.79. Do the same for the data in Exercise 7.80 (page 459), which is a similar setting. These are somewhat smaller than the assumed value that you used in the previous exercise. Explain why it is generally a better idea to assume a standard deviation that is larger than you expect than one that is smaller. Repeat the power calculations for some other reasonable values of σ and comment on the impact of the size of σ for planning the new study.

7.114 Planning a study to compare ad placement. Refer to Exercise 7.78 (page 458), where we compared trustworthiness ratings for ads from two different publications. Suppose that you are planning a similar study using two different publications that are not expected to show the differences seen when comparing the *Wall Street Journal* with the *National Enquirer*. You would like to detect a difference of 1.5 points using a two-sided significance test with a 5% level of significance. Based on Exercise 7.78, it is reasonable to use 1.6 as the value of the common standard deviation for planning purposes.

(a) What is the power if you use sample sizes similar to those used in the previous study—for example, 65 for each publication?

(b) Repeat the calculations for 100 in each group.

(c) What sample size would you recommend for the new study?

7.115 Sign test for potential insurance fraud. The differences in the repair estimates in Exercise 7.40 (page 430) can also be analyzed using a sign test. Set up the appropriate null and alternative hypotheses, carry out the test, and summarize the results. How do these results compare with those that you obtained in Exercise 7.40? JOCKO

7.116 Sign test for the comparison of operators. The differences in the TBBMC measures in Exercise 7.45 (pages 431–432) can also be analyzed using a sign test. Set up the appropriate null and alternative hypotheses, carry out the test, and summarize the results. How do these results compare with those that you obtained in Exercise 7.45? TBBMC

7.117 Sign test for fuel efficiency comparison. Use the sign test to assess whether the computer calculates a higher mpg than the driver in Exercise 7.41 (pages 430–431). State the hypotheses, give the *P*-value using the binomial table (Table C), and report your conclusion. MPGDIFF

7.118 Insulation study. A manufacturer of electric motors tests insulation at a high temperature (250° C) and records the number of hours until the insulation fails.[44] The data for five specimens are

446 326 372 377 310

The small sample size makes judgment from the data difficult, but engineering experience suggests that the logarithm of the failure time will have a Normal distribution. Take the logarithms of the five observations and use t procedures to give a 90% confidence interval for the mean of the log failure time for insulation of this type. INSULAT

CHAPTER 7 EXERCISES

7.119 LSAT scores. The scores of four senior roommates on the Law School Admission Test (LSAT) are

153 162 166 133

Find the mean, the standard deviation, and the standard error of the mean. Is it appropriate to calculate a confidence interval based on these data? Explain why or why not. LSAT

7.120 Converting a two-sided *P*-value. You use statistical software to perform a significance test of the null hypothesis that two means are equal. The software reports a *P*-value for the two-sided alternative. Your alternative is that the first mean is greater than the second mean.

(a) The software reports $t = 1.85$ with a P-value of 0.075. Would you reject H_0 at $\alpha = 0.05$? Explain your answer.

(b) The software reports $t = -1.85$ with a P-value of 0.075. Would you reject H_0 at $\alpha = 0.05$? Explain your answer.

7.121 Degrees of freedom and *t.** As the degrees of freedom increase, the t distributions get closer and closer to the z ($N(0, 1)$) distribution. One way to see this is to look at how the value of t^* for a 95% confidence interval changes with the degrees of freedom.

(a) Make a plot with degrees of freedom from 10 to 100 by 10 on the x axis and t^* on the y axis. Also draw a horizontal line on the plot corresponding to the value of $z^* = 1.96$.

(b) Summarize the main features of the plot.

(c) Describe how this plot would change if you considered a 90% confidence interval.

7.122 Sample size and margin of error. The margin of error for a confidence interval for μ depends on the confidence level, the sample standard deviation s, and the sample size. Fix the confidence level at 95% and the sample standard deviation at $s = 1$ to examine the effect of the sample size. Find the margin of error for sample sizes of 11 to 101 by 10s—that is, let $n = 11, 21, 31, \ldots, 101$. Plot the margins of error versus the sample size and summarize the relationship.

7.123 Which design? The following situations all require inference about a mean or means. Identify each as (1) a single sample, (2) matched pairs, or (3) two independent samples. Explain your answers.

(a) Your customers are college students. You are interested in comparing the interest in a new product that you are developing between those students who live in the dorms and those who live elsewhere.

(b) Your customers are college students. You are interested in finding out which of two new product labels is more appealing.

(c) Your customers are college students. You are interested in assessing their interest in a new product.

7.124 Which design? The following situations all require inference about a mean or means. Identify each as (1) a single sample, (2) matched pairs, or (3) two independent samples. Explain your answers.

(a) You want to estimate the average age of your store's customers.

(b) You do an SRS survey of your customers every year. One of the questions on the survey asks about customer satisfaction on a seven-point scale with the response 1 indicating "very dissatisfied" and 7 indicating "very satisfied." You want to see if the mean customer satisfaction has improved from last year.

(c) You ask an SRS of customers their opinions on each of two new floor plans for your store.

7.125 Number of critical food violations. The results of a major city's restaurant inspections are available through its online newspaper.[45] Critical food violations are those that put patrons at risk of getting sick and must immediately be corrected by the restaurant. An SRS of $n = 200$ inspections from the more than 16,000 inspections since January 2012 were collected, resulting in $\bar{x} = 0.995$ violations and $s = 1.822$ violations.

(a) Test the hypothesis that the average number of critical violations is less than 1.5 using a significance level of 0.05. State the two hypotheses, the test statistic, and P-value.

(b) Construct a 95% confidence interval for the average number of critical violations and summarize your result.

(c) Which of the two summaries (significance test versus confidence interval) do you find more helpful in this case? Explain your answer.

(d) These data are integers ranging from 0 to 10. The data are also skewed to the right, with 79% of the values either a 0 or a 1. Given this information, do you think use of the t procedures is appropriate? Explain your answer.

7.126 Two-sample t test versus matched pairs t test. Consider the following data set. The data were actually collected in pairs, and each row represents a pair. PAIRED

Group 1	Group 2
48.86	48.88
50.60	52.63
51.02	52.55
47.99	50.94
54.20	53.02
50.66	50.66
45.91	47.78
48.79	48.44
47.76	48.92
51.13	51.63

(a) Suppose that we ignore the fact that the data were collected in pairs and mistakenly treat this as a two-sample problem. Compute the sample mean and variance for each group. Then compute the two-sample t statistic, degrees of freedom, and P-value for the two-sided alternative.

(b) Now analyze the data in the proper way. Compute the sample mean and variance of the differences. Then compute the t statistic, degrees of freedom, and P-value.

(c) Describe the differences in the two test results.

7.127 Two-sample t test versus matched pairs t test, continued. Refer to the previous exercise. Perhaps an easier way to see the major difference in the two analysis approaches for these data is by computing 95% confidence intervals for the mean difference.

(a) Compute the 95% confidence interval using the two-sample t confidence interval.

(b) Compute the 95% confidence interval using the matched pairs t confidence interval.

(c) Compare the estimates (that is, the centers of the intervals) and margins of error. What is the major difference between the two approaches for these data?

7.128 Average service time. Recall the drive-thru study in Exercise 7.69 (page 457). Another benchmark that was measured was the service time. A summary of the results (in seconds) for two of the chains is shown below.

Chain	n	$\bar{x}$	s
Taco Bell	308	158.03	33.8
McDonald's	317	189.49	41.3

(a) Is there a difference in the average service time between these two chains? Test the null hypothesis that the chains' average service time is the same. Use a significance level of 0.05.

(b) Construct a 95% confidence interval for the difference in average service time.

(c) Lex plans to go to Taco Bell and Sam to McDonald's. Does the interval in part (b) contain the difference in their service times that they're likely to encounter? Explain your answer.

7.129 Interracial friendships in college. A study utilized the random roommate assignment process of a small college to investigate the interracial mix of friends among students in college.[46] As part of this study, the researchers looked at 238 white students who were randomly assigned a roommate in their first year and recorded the proportion of their friends (not including the first-year roommate) who were black. The following table summarizes the results, broken down by roommate race, for the middle of the first and third years of college.

Middle of First Year			
Randomly assigned	n	$\bar{x}$	s
Black roommate	41	0.085	0.134
White roommate	197	0.063	0.112

Middle of Third Year			
Randomly assigned	n	$\bar{x}$	s
Black roommate	41	0.146	0.243
White roommate	197	0.062	0.154

(a) Proportions are not Normally distributed. Explain why it may still be appropriate to use the t procedures for these data.

(b) For each year, state the null and alternative hypotheses for comparing these two groups.

(c) For each year, perform the significance test at the $\alpha = 0.05$ level, making sure to report the test statistic, degrees of freedom, and P-value.

(d) Write a one-paragraph summary of your conclusions from these two tests.

7.130 Interracial friendships in college, continued. Refer to the previous exercise. For each year, construct a 95% confidence interval for the difference in means $\mu_1 - \mu_2$ and describe how these intervals can be used to test the null hypotheses in part (b) of the previous exercise.

7.131 Alcohol consumption and body composition. Individuals who consume large amounts of alcohol do not use the calories from this source as efficiently as calories from other sources. One study examined the effects of moderate alcohol consumption on body composition and the intake of other foods. Fourteen subjects participated in a crossover design where they either drank wine for the first six weeks and then abstained for the next six weeks or vice versa.[47] During the period when they drank wine, the subjects, on average, lost 0.4 kilogram (kg) of body weight; when they did not drink wine, they lost an average of 1.1 kg. The standard deviation of the difference between the weight lost under these two conditions is 8.6 kg. During the wine period, they consumed an average of 2589 calories; with no wine, the mean consumption was 2575. The standard deviation of the difference was 210.

(a) Compute the differences in means and the standard errors for comparing body weight and caloric intake under the two experimental conditions.

(b) A report of the study indicated that there were no significant differences in these two outcome measures. Verify this result for each measure, giving the test statistic, degrees of freedom, and the P-value.

(c) One concern with studies such as this, with a small number of subjects, is that there may not be sufficient power to detect differences that are potentially important. Address this question by computing 95% confidence intervals for the two measures and discuss the information provided by the intervals.

(d) Here are some other characteristics of the study. The study periods lasted for six weeks. All subjects were males between the ages of 21 and 50 years who weighed between 68 and 91 kg. They were all from the same city. During the wine period, subjects were told to consume two 135-milliliter (ml) servings of red wine per day and no other alcohol. The entire six-week supply was given to each subject at the beginning of the period. During the other period, subjects were instructed to refrain from any use of alcohol. All subjects reported that they complied with these instructions except for three subjects, who said that they drank no more than three to four 12-ounce bottles of beer during the no-alcohol period. Discuss how these factors could influence the interpretation of the results.

7.132 The wine makes the meal? In one study, 39 diners were given a free glass of cabernet sauvignon wine to accompany a French meal.[48] Although the wine was identical, half the bottle labels claimed the wine was from California and the other half claimed it was from North Dakota. The following table summarizes the grams of entrée and wine consumed during the meal.

	Wine label	n	Mean	St. Dev
Entrée	California	24	499.8	87.2
	North Dakota	15	439.0	89.2
Wine	California	24	100.8	23.3
	North Dakota	15	110.4	9.0

Did the patrons who thought that the wine was from California consume more? Analyze the data and write a report summarizing your work. Be sure to include details regarding the statistical methods you used, your assumptions, and your conclusions.

7.133 Can mockingbirds learn to identify specific humans? A central question in urban ecology is why some animals adapt well to the presence of humans and others do not. The following results summarize part of a study of the northern mockingbird (*Mimus polyglottos*) that took place on a campus of a large university.[49] For four consecutive days, the same human approached a nest and stood 1 meter away for 30 seconds, placing his or her hand on the rim of the nest. On the fifth day, a new person did the same thing. Each day, the distance of the human from the nest when the bird flushed was recorded. This was repeated for 24 nests. The human intruder varied his or her appearance (that is, wore different clothes) over the four days. We report results for only Days 1, 4, and 5 here. The response variable is flush distance measured in meters.

Day	Mean	s
1	6.1	4.9
4	15.1	7.3
5	4.9	5.3

(a) Explain why this should be treated as a matched design.

(b) Unfortunately, the research article does not provide the standard error of the difference, only the standard error of the mean flush distance for each day. However, we can use the general addition rule for variances (page 258) to approximate it. If we assume that the correlation between the flush distance at Day 1 and Day 4 for each nest is $\rho = 0.40$, what is the standard deviation for the difference in distance?

AU: check x-ref page

(c) Using your result in part (b), test the hypothesis that there is no difference in the flush distance across these two days. Use a significance level of 0.05.

(d) Repeat parts (b) and (c) but now compare Day 1 and Day 5, assuming a correlation between flush distances for each nest of $\rho = 0.30$.

(e) Write a brief summary of your conclusions.

7.134 Sign test for assessment of a foreign-language institute. Use the sign test to assess whether the summer institute of Exercise 7.47 (page 432) improves French listening skills. State the hypotheses, give the P-value using the binomial table (Table C), and report your conclusion. SUMLANG

7.135 Study design information. Refer to Exercise 7.132. In this study, diners were seated alone or in groups of two, three, four, and, in one case, nine (for a total of $n = 16$ tables). Also, each table, not each patron, was randomly assigned a particular wine label. Does this information alter how you might do the analysis in the previous problem? Explain your answer.

7.136 Analysis of tree size using the complete data set. The data used in Exercises 7.33 (page 429), 7.79, and 7.80 (page 459) were obtained by taking SRSs from the 584 longleaf pine trees that were measured in the Wade Tract. The entire data set is given in the WADE data set. Find the 95% confidence interval for the mean DBH using the entire data set, and compare this interval with the one that you calculated in Exercise 7.33. Write a report about these data. Include comments on the effect of the sample size on the margin of error, the distribution of the data, the appropriateness of the Normality-based methods for this problem, and the generalizability of the results to other similar stands of longleaf pine or other kinds of trees in this area of the United States and other areas. WADE

AU: Check edits

7.137 Can snobby salespeople boost retail sales? Researchers asked 180 women to read a hypothetical shopping experience where they entered a luxury store (e.g., Louis Vuitton, Gucci, Burberry) and ask a salesperson for directions to the items they seek. For half the women, the salesperson was condescending while doing this. The other half were directed in a neutral manner. After reading the experience, participants were asked various questions, including what price they were willing to pay (in dollars) for a particular product from the brand.[50] Here is a summary of the results.

Chain	n	$\bar{x}$	s
Condescending	90	4.44	3.98
Neutral	90	3.95	2.88

Were the participants who were treated rudely willing to pay more for the product? Analyze the data, and write a report summarizing your work. Be sure to include details regarding the statistical methods you used, your assumptions, and your conclusions.

7.138 A comparison of female high school students. A study was performed to determine the prevalence of the female athlete triad (low energy availability, menstrual dysfunction, and low bone mineral density) in high school students.[51] A total of 80 high school athletes and 80 sedentary students were assessed. The following table summarizes several measured characteristics:

	Athletes		Sedentary	
Characteristic	$\bar{x}$	s	$\bar{x}$	s
Body fat (%)	25.61	5.54	32.51	8.05
Body mass index	21.60	2.46	26.41	2.73
Calcium deficit (mg)	297.13	516.63	580.54	372.77
Glasses of milk/day	2.21	1.46	1.82	1.24

(a) For each of the characteristics, test the hypothesis that the means are the same in the two groups. Use a significance level of 0.05 for each test.

(b) Write a short report summarizing your results.

7.139 More on snobby salespeople. Refer to Exercise 7.137. Researchers also asked a different 180 women to read the same hypothetical shopping experience, but now they entered a mass market (e.g., Gap, American Eagle, H&M). Here are those results (in dollars) for the two conditions:

Chain	n	$\bar{x}$	s
Condescending	90	2.90	3.28
Neutral	90	2.98	3.24

Were the participants who were treated rudely willing to pay more for the product? Analyze the data, and write a report summarizing your work. Be sure to include details regarding the statistical methods you used, your assumptions, and your conclusions. Also compare these results with the ones from Exercise 7.137.

7.140 Transforming the response. Refer to Exercises 7.137 and 7.139. The researchers state that they took the natural log of the willingness to pay variable in order to "normalize the distribution" prior to analysis. Thus, their test results are based on log dollar measurements. For the t procedures used in the previous two exercises, do you feel this transformation is necessary? Explain your answer.

7.141 Competitive prices? A retailer entered into an exclusive agreement with a supplier who guaranteed to provide all products at competitive prices. The retailer eventually began to purchase supplies from other vendors who offered better prices. The original supplier filed a legal action claiming violation of the agreement. In defense, the retailer had an audit performed on a random sample of invoices. For each audited invoice, all purchases made from other suppliers were examined and the prices were compared with those offered by the original supplier. For each invoice, the percent of purchases for which the alternate supplier offered a lower price than the original supplier was recorded.[52] Here are the data:

0	100	0	100	33	34	100	48	78	100	77	100	38
68	100	79	100	100	100	100	100	100	89	100	100	

Report the average of the percents with a 95% margin of error. Do the sample invoices suggest that the original supplier's prices are not competitive on the average? COMPETE

7.142 Weight-loss programs. In a study of the effectiveness of weight-loss programs, 47 subjects who were at least 20% overweight took part in a group support program for 10 weeks. Private weighings determined each subject's weight at the beginning of the program and six months after the program's end. The matched pairs t test was used to assess the significance of the average weight loss. The paper reporting the study said, "The subjects lost a significant amount of weight over time, $t(46) = 4.68$, $p < 0.01$." It is common to report the results of statistical tests in this abbreviated style.[53]

(a) Why was the matched pairs statistic appropriate?

(b) Explain to someone who knows no statistics but is interested in weight-loss programs what the practical conclusion is.

(c) The paper follows the tradition of reporting significance only at fixed levels such as $\alpha = 0.01$. In fact, the results are more significant than "$p < 0.01$" suggests. What can you say about the P-value of the t test?

7.143 Do women perform better in school? Some research suggests that women perform better than men in school, but men score higher on standardized tests. Table 1.3 (page 26) presents data on a measure of school performance, grade point average (GPA), and a standardized test, IQ, for 78 seventh-grade students. Do these data lend further support to the previously found gender differences? Give graphical displays of the data and describe the distributions. Use significance tests and confidence intervals to examine this question, and prepare a short report summarizing your findings. GRADES

AU: Check x-ref page

7.144 Self-concept and school performance. Refer to the previous exercise. Although self-concept

in this study was measured on a scale with values in the data set ranging from 20 to 80, many prefer to think of this kind of variable as having only two possible values: low self-concept or high self-concept. Find the median of the self-concept scores in Table 1.3, and define those students with scores at or below the median to be low-self-concept students and those with scores above the median to be high-self-concept students. Do high-self-concept students have GPAs that differ from those of low-self-concept students? What about IQ? Prepare a report addressing these questions. Be sure to include graphical and numerical summaries and confidence intervals, and state clearly the details of significance tests. GRADES

7.145 Behavior of pet owners. On the morning of March 5, 1996, a train with 14 tankers of propane derailed near the center of the small Wisconsin town of Weyauwega. Six of the tankers were ruptured and burning when the 1700 residents were ordered to evacuate the town. Researchers study disasters like this so that effective relief efforts can be designed for future disasters. About half the households with pets did not evacuate all their pets. A study conducted after the derailment focused on problems associated with retrieval of the pets after the evacuation and characteristics of the pet owners. One of the scales measured "commitment to adult animals," and the people who evacuated all or some of their pets were compared with those who did not evacuate any of their pets. Higher scores indicate that the pet owner is more likely to take actions that benefit the pet.[54] Here are the data summaries:

Group	n	$\bar{x}$	s
Evacuated all or some pets	116	7.95	3.62
Did not evacuate any pets	125	6.26	3.56

Analyze the data and prepare a short report describing the results.

7.146 Sample size calculation. Example 7.10 (page 434) tells us that the mean height of 10-year-old girls is $N(56.9, 2.8)$ and for boys it is $N(56.0, 3.5)$. The null hypothesis that the mean heights of 10-year-old boys and girls are equal is clearly false. The difference in mean heights is $56.9 - 56.0 = 0.9$ inch. Small differences such as this can require large sample sizes to detect. To simplify our calculations, let's assume that the standard deviations are the same—say, $\sigma = 3.2$—and that we will measure the heights of an equal number of girls and boys. How many would we need to measure to have a 90% chance of detecting the (true) alternative hypothesis?

7.147 Different methods of teaching reading. In the READ data set, the response variable Post3 is to be compared for three methods of teaching reading. The Basal method is the standard, or control, method, and the two new methods are DRTA and Strat. We can use the methods of this chapter to compare Basal with DRTA and Basal with Strat. Note that to make comparisons among three treatments it is more appropriate to use the procedures that we will learn in Chapter 12. READ

(a) Is the mean reading score with the DRTA method higher than that for the Basal method? Perform an analysis to answer this question, and summarize your results.

(b) Answer part (a) for the Strat method in place of DRTA.

7.148 Designing a new stress management survey. Refer to Exercise 6.17 (page 358). Suppose you want to draw a new SRS of millenials such that the expected margin of error with 99% confidence is 0.2 points. What sample size do you need?

7.149 Conditions for inference. Suppose that your state contains 85 school corporations and each corporation reports its expenditures per pupil. Is it proper to apply the one-sample t method to these data to give a 95% confidence interval for the average expenditure per pupil? Explain your answer.

AU: Check x-ref page