Example 1

How to Get a Better Tan

Picture the Scenario

Statistics students were asked to design an experiment about a topic of choice, conduct the experiment, and then analyze the data. One student, Allison, decided to compare tanning methods without exposure to the sun to avoid skin cancer risk. She investigated two treatments—a bronze tanning lotion applied twice over a two-day period, or a tanning studio where the person is exposed to ultraviolet (UV) light. We'll refer to these treatments as "tanning lotion" and "tanning studio."

The tanning lotion is much less expensive, but Allison predicted that the tanning studio would give a better tan. To investigate this hypothesis, she recruited five untanned female friends to participate in an experiment. Another student in the class used a random number generator to pick three of the friends to use the tanning lotion. The other two friends used the tanning studio. After three days, Allison evaluated the tans produced. She was blinded to the treatment allocation, not knowing which participants used which tanning method. Allison ranked the friends in terms of the quality of their tans. The ranks went from 1 to 5, with 1 = most natural looking and 5 = least natural-looking.

Questions to Explore

- Once Allison ranked the five tanned participants, how could she summarize the evidence in favor of one treatment over the other?
- How can Allison find a P-value to determine if one treatment is significantly better than the other?

Thinking Ahead

You learned in Sections 10.2 and 10.3 how to compare means for two treatments using *t* tests. The tests assume a *normal* distribution for a quantitative response variable. The *t* tests are *robust*, usually working well even when population distributions are *not* normal. An exception is when the distribution is skewed, the sample sizes are small, and the alternative hypothesis is one-sided.

To use the *t* test, suppose Allison created a quantitative variable by assigning a score between 0 and 10 for each girl to describe the quality of tan. With such small sample sizes (only 2 and 3), she would not be able to assess whether quality of tan is approximately normal. Moreover, her prediction that the studio gives a better tan than the lotion was one-sided. In any case, Allison found it easier to rank the participants than to create a quantitative variable. For these reasons, then, it's not appropriate for her to use a *t* test to compare the tanning methods.

Example 2

Tanning Studio Versus Tanning Lotion

Picture the Scenario

Example 1 describes Allison's experiment to determine whether a tanning lotion or a tanning studio produced a better tan. Table 15.1 showed the possible rankings for five tans. Table 15.2 showed the sampling distribution of the difference between the sample mean ranks, presuming the null hypothesis is true that the tanning treatments have identical effects. For Allison's actual experiment, the ranks were (2, 4, 5) for the three using the tanning lotion and (1, 3) for the two using the tanning studio.

Questions to Explore

a. Find and interpret the P-value for comparing the treatments, using the one-sided alternative hypothesis that the tanning studio gives a better tan than the tanning lotion. That is, H_a states that the expected mean

Think It Through

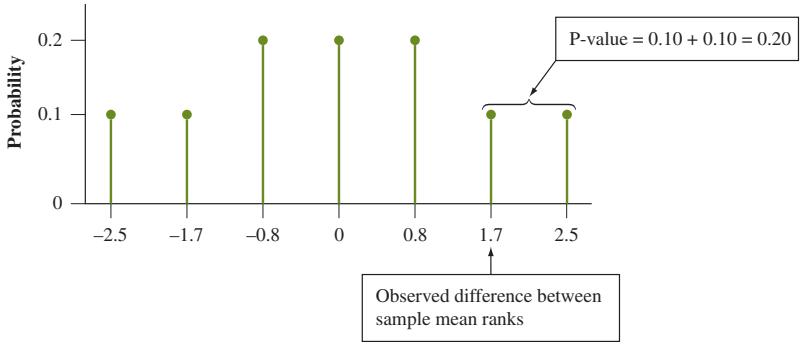
a. For the observed sample, the mean ranks are $(2 + 4 + 5)/3 = 3.7$ for the tanning lotion and $(1 + 3)/2 = 2$ for the tanning studio. The test statistic is the difference between the sample mean ranks, $3.7 - 2 = 1.7$. The right tail of the sampling distribution in Figure 15.1 has the large positive differences, for which the ranks tended to be better (lower) with the tanning studio. For the one-sided H_a, the P-value is the probability,

$$\text{P-value} = \text{P(difference between sample mean ranks at least as large as observed).}$$

That is, under the presumption that H_0 is true,

$$\text{P-value} = \text{P(difference between sample mean ranks} \geq 1.7).$$

From Table 15.2 (shown again in the margin) or Figure 15.1 (reproduced below), the probability of a sample mean difference of 1.7 or even larger is $1/10 + 1/10 = 2/10 = 0.20$. This is the P-value. It is not very close to 0. Although there is some evidence that the tanning studio gives a better tan (it *did* have a lower sample mean rank), the evidence is not strong. If the treatments had identical effects, the probability would be 0.20 of getting a sample like we observed or even more extreme.



b. In this experiment, suppose the tanning studio gave the two most natural-looking tans. The ranks would then be (1, 2) for the tanning studio and (3, 4, 5) for the tanning lotion. The difference of sample means then equals $4 - 1.5 = 2.5$. It is the most extreme possible sample, and (from Table 15.2 or Figure 15.1) its tail probability is 0.10. This is the smallest possible one-sided P-value.

Insight

With sample sizes of only 2 for one treatment and 3 for the other treatment, it's not possible to get a very small P-value. If Allison wanted to make a decision using a 0.05 significance level, she would never be able to get strong enough evidence to reject the null hypothesis. To get informative results, she'd need to conduct an experiment with larger sample sizes.

Try Exercises 15.1 and 15.2

Example 3

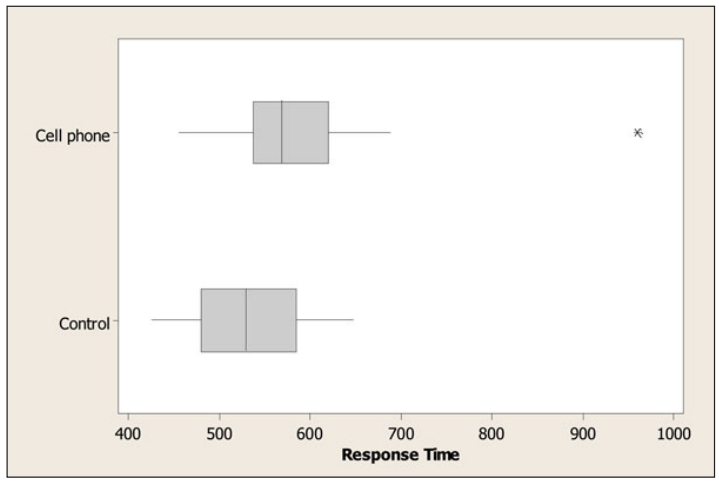
Driving Reaction Times

Picture the Scenario

Example 9 in Chapter 10 discussed an experiment investigating whether or not cell phone use impairs drivers' reaction times. A sample of 64 college students was randomly assigned to a cell phone group or a control group, 32 to each. On a machine that simulated driving situations, participants were instructed to press a "brake button" as soon as possible when they detected a red light. The control group listened to a radio broadcast or to books-on-tape while they performed the simulated driving. The cell phone group carried out a conversation on a cell phone with someone in a separate room.

A subject's reaction time observation is defined to be his or her response time to the red lights (in milliseconds), averaged over all the trials. Figure 15.2 shows box plots of the data for the two groups. Here's some of the data showing the four smallest observations and the four largest observations for each treatment.

Cell phone:	456	468	482	501	. . . . .	672	679	688	960
Control:	426	436	444	449	. . . . .	626	626	642	648



▲ **Figure 15.2** Box Plots of Response Times for Cell Phone Study. Question Does either box plot show any irregularities that suggest it's safer to use a nonparametric test than a two-sample *t* test?

The t inferences for comparing the treatment means assume normal population distributions. The box plots do not show any substantial skew, but there is an extreme outlier for the cell phone group. One subject in that group had a very slow mean reaction time, 960 milliseconds.

Questions to Explore

a. Explain how to find the ranks for the Wilcoxon test by showing which of the 64 observations get ranks 1, 2, 63, and 64.

b. Table 15.5 shows the SPSS output for conducting the Wilcoxon test. Report and interpret the mean ranks.

Table 15.5 SPSS Output for Wilcoxon Test with Data from Cell Phone Study

```
Ranks
                group          N     Mean Rank    Sum of Ranks
TIME            Control        32      27.03         865.00
                Cell phone     32      37.97        1215.00
                Total          64
Test Statistics
                              TIME
Wilcoxon W                  865.000
Z                            −2.350
Asymp. Sig. (2-tailed).019
```

c. Report the test statistic and the P-value for the two-sided Wilcoxon test. Interpret.

Think It Through

a. Let's look at the smallest and largest observations for each group that were shown above:

Cell phone:	456	468	482	501	.....	672	679	688	960
Control:	426	436	444	449	.....	626	626	642	648

We give rank 1 to the *smallest* reaction time, so the value 426 gets rank 1. The second smallest observation is 436, which gets rank 2. The largest of the 64 reaction times, which was 960, gets rank 64. The next largest observation, 688, gets rank 63.

b. Table 15.5 reports mean ranks of 27.03 for the control group and 37.97 for the cell phone group. The smaller mean for the control group suggests that that group tends to have smaller ranks, and thus faster reaction times.

c. The z test statistic takes the difference between the sample mean ranks and divides it by a standard error. Table 15.5 reports $z = -2.35$. The P-value of 0.019, reported as "Asymp. Sig. (2-tailed)," is the two-tail probability for the two-sided H_a. It shows strong evidence against the null hypothesis that the distribution of reaction time is identical for the two treatments. Specifically, the sample mean ranks suggest that reaction times tend to be slower for those using cell phones.

Insight

The observation of 960 would get rank 64 if it were *any* number larger than 688 (the second largest value). So, *the Wilcoxon test is not affected by an outlier.* No matter how far the largest observation falls from the next largest, it still gets the same rank. Likewise, no matter how far the smallest observation is below the next smallest, it still gets the rank of 1.

Try Exercise 15.4

Example 4

Difference Between Median Reaction Times

Picture the Scenario

Example 3 used the Wilcoxon test to compare reaction time distributions in a simulated driving experiment for subjects using cell phones and for a control group. The MINITAB output in Table 15.6 shows results of comparing the distributions using medians. (It uses the Greek letter name *eta*, which is η, to denote the median.)

Table 15.6 MINITAB Output for Comparing Medians for Cell Phone Group and Control Group

```
                N       Median
Cell phone      32      569.00
Control         32      530.00
Point estimate for ETA1-ETA2 is 44.50
95.1 Percent CI for ETA1-ETA2 is (8.99, 79.01)
Test of ETA1 = ETA2 vs. ETA1 not = ETA2 is significant at 0.0184
```

ETA is MINITAB notation for the median

P-value

Questions to Explore

a. Report the sample medians and the point estimate of the difference between the population medians.

b. Report and interpret the 95% confidence interval for the difference between the population medians.

Think It Through

a. The median reaction times were 569 milliseconds for the cell phone group and 530 milliseconds for the control group. For the cell phone group, for example, half of the reaction times were smaller than 569 milliseconds and half were larger than 569. Table 15.6 reports a point estimate of the difference between the population medians for the two groups of 44.5 milliseconds. (*Note*: This is not the same as the difference between the two sample medians, which is an alternative estimate.)

b. Table 15.6 reports that the 95% confidence interval for the difference between the population medians is (9, 79). Since zero is not contained in the 95% interval, this interval supports that the median reaction times are not the same for the cell phone and control groups. We infer from the interval that the population median reaction time for the cell phone group is between 9 milliseconds and 79 milliseconds larger than for the control group. This inference agrees with the conclusion of the Wilcoxon test that the reaction time distributions differ for the two groups (P-value = 0.02).

Insight

Example 9 in Chapter 10 estimated the difference between the population *means* to be 51.5, with a 95% confidence interval of (12, 91). However, those results were influenced by the outlier of 960 for the cell phone group. When estimation focuses on medians rather than means, outliers do not influence the analysis. *The lack of an influence of outliers is an advantage of the analysis reported here for the medians.*

Try Exercise 15.5

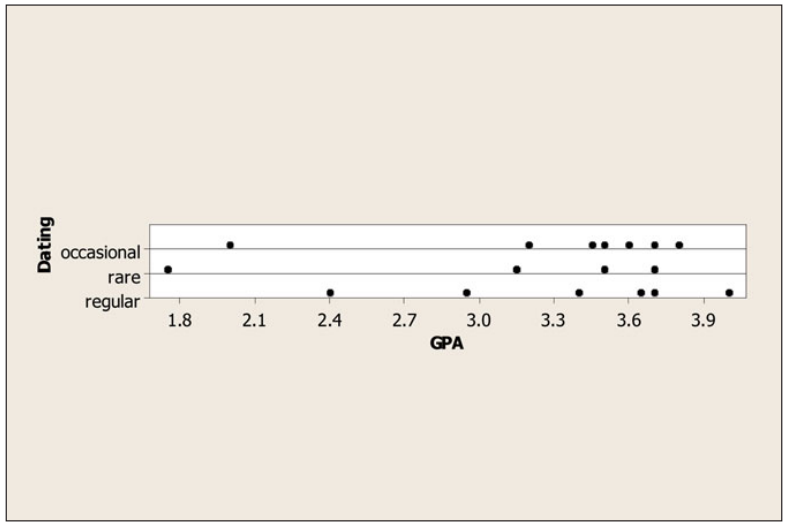
Example 5

Frequent Dating and College GPA

Picture the Scenario

Tim decided to study whether dating was associated with college GPA. He wondered whether students who date a lot tend to have poorer GPAs. He asked the 17 students in the class to anonymously fill out a short questionnaire in which they were asked to give their college GPA (0 to 4 scale) and to indicate whether during their college careers they had dated regularly, occasionally, or rarely.

Figure 15.4 shows the dot plots that Tim constructed of the GPA data for the three dating groups. Since the dot plots showed evidence of severe skew to the left and since the sample size was small in each group, he felt safer analyzing the data with the Kruskal-Wallis test than with the ordinary ANOVA *F* test.



▲ **Figure 15.4** Dot Plots of GPA by Dating Group. **Question** Why might we be nervous about using the ordinary ANOVA *F* test to compare mean GPA for the three dating groups?

Table 15.7 shows the data, with the college GPA values ordered from smallest to largest for each dating group. The table in the margin shows the combined sample of 17 observations from the three groups and their ranks. Table 15.7 also shows these ranks as well as the mean rank for each group.

Table 15.7 College GPA by Dating Group

Dating Group	GPA Observations	Ranks	Mean Rank
Rare	1.75, 3.15, 3.50, 3.68	1, 5, 9.5, 13	7.1
Occasional	2.00, 3.20, 3.44, 3.50, 3.60, 3.71, 3.80	2, 6, 8, 9.5, 11, 15, 16	9.6
Regular	2.40, 2.95, 3.40, 3.67, 3.70, 4.00	3, 4, 7, 12, 14, 17	9.5

Question to Explore

Table 15.8 shows MINITAB output for the Kruskal-Wallis test. MINITAB denotes the chi-squared test statistic by H. Interpret these results.

Table 15.8 Results of Kruskal-Wallis Test for Data in Table 15.7

Kruskal-Wallis Test: GPA versus Dating

```
Dating          N     Median  AVE. Rank
rare            4      3.325      7.1
occasional      7      3.500      9.6
regular         6      3.535      9.5
H = 0.72     DF = 2   P = 0.696 (adjusted for ties)
```

Think It Through

If H_0: identical population distributions for the three groups were true, the Kruskal-Wallis test statistic would have an approximate chi-squared distribution with $df = 2$. Table 15.8 reports that the test statistic is H = 0.72. The P-value is the right-tail probability above 0.72. Table 15.8 reports this as 0.696, about 0.7. It is plausible that GPA is independent of dating group. Table 15.8 shows that the sample median GPAs are not very different, and since the sample sizes are small, these sample medians do not give much evidence against H_0.

Insight

If the P-value had been small, to find out which pairs of groups significantly differ, we could follow up the Kruskal-Wallis test by a Wilcoxon test to compare each pair of dating groups. Or, we could find a confidence interval for the difference between the population medians for each pair.

Try Exercise 15.8

Example 6

Time Browsing the Internet or Watching TV

Picture the Scenario

Which do most students spend more time doing—browsing the Internet or watching TV? Let's consider the students surveyed at the University of Georgia whose responses are in the Georgia Student Survey data file. The results for the first three students in the data file (in minutes per day) were

Student	Internet	TV
1	60	120
2	20	120
3	60	90

All three spent more time watching TV. For the entire sample, 35 students spent more time watching TV and 19 students spent more time browsing the Internet. (The analysis ignores the 5 students who reported the same time for each.)

Question to Explore

Let p denote the population proportion who spent more time watching TV. Find the test statistic and P-value for the sign test of H_0: $p = 0.50$ against H_a: $p \neq 0.50$. Interpret.

Think It Through

Here, $n = 35 + 19 = 54$. The sample proportion who spent more time watching TV was $35/54 = 0.648$. For testing that $p = 0.50$, the *se* of the sample proportion is

$$se = \sqrt{(0.50)(0.50)/n} = \sqrt{(0.50)(0.50)/54} = 0.068.$$

The test statistic is

$$z = (\hat{p} - 0.50)/se = (0.648 - 0.50)/0.068 = 2.18.$$

From the normal distribution table (Table A or software), the two-sided P-value is 0.03. This provides considerable evidence that most students spend more time watching TV than browsing the Internet. The conclusion must be tempered by the fact that the data resulted from a convenience sample (students in a class for a statistics course) rather than a random sample of all college students.

Insight

The sign test uses merely the information about *which* response is higher and *how many* responses are higher, not the quantitative information about *how much* higher. This is a disadvantage compared to the corresponding parametric test, the matched pairs t test of Section 10.4, which analyzes the mean of the differences between the two responses. The sign test is most appropriate when we can order the responses but do not have quantitative information, such as in the next example.

Try Exercise 15.10

Example 7

Crossover Experiment Comparing Tanning Methods

Picture the Scenario

When Allison told another student in the class (Megan) about her planned experiment to compare tanning methods, Megan decided to do a separate tanning experiment. She used a crossover design for a different sample of five untanned female friends. The results of her experiment were that the tanning studio gave a better tan than the tanning lotion for four of the five participants.

Question to Explore

Find and interpret the P-value for testing that the population proportion p of participants for whom the tanning studio gives a better tan than the tanning lotion equals 0.50. Use the alternative hypothesis that this population proportion is larger than 0.50, because Megan predicted that the tanning studio would give better tans.

Think It Through

The null hypothesis is $H_0: p = 0.50$. For $H_a: p > 0.50$, the P-value is the probability of the observed sample outcome or an even larger one. The sample size ($n = 5$) was small, so we use the binomial distribution rather than its normal approximation to find the P-value.

If $p = 0.50$, from the margin Recall box the binomial probability that $x = 4$ of the $n = 5$ participants would get better tans with the tanning studio is

$$P(4) = \frac{5!}{4!(5-4)!}(0.50)^4(0.50)^1 = 0.156.$$

The more extreme result that all five participants would get better tans with the tanning studio has probability $P(5) = (0.50)^5 = 0.031$. The P-value is the right-tail probability of the observed result and the more extreme one, that is, $0.156 + 0.031 = 0.187$. See the margin figure. In summary, the evidence is not strong that more participants get a better tan from the tanning studio than the tanning lotion.

Insight

Megan would instead use the two-sided alternative, $H_a: p \neq 0.50$, if she did not make a prior prediction about which tanning method would be better. The P-value would then be $2(0.19) = 0.38$. With only $n = 5$ observations, the smallest possible two-sided P-value would be $2(0.031) = 0.06$, which occurs when $x = 0$ or when $x = 5$.

Try Exercise 15.11

Example 8

GRE Test Scores

Picture the Scenario

If you want to attend graduate school, taking the Graduate Record Examination (GRE) is usually a requirement. Many graduate schools consider GRE scores for admittance, to qualify for financial aid, to determine fellowships and grants, and for other program research or teaching assignments. The GRE includes three sections designed to test verbal, quantitative, and writing skills.

The verbal and quantitative sections are each scored between 200 and 800. The analytical writing portion of the GRE is given a score between 0 and 6 in half point increments.

In our example, three students volunteered for a study to determine if taking a two-day workshop on GRE preparation improved their GRE analytical writing score from a previous score. Note: The original data was larger ($n = 12$) but a small sample is used in this example to make it easier to explain. The results are shown in the following table:

	Subject		
	1	2	3
Before	2.5	4	1.5
After	3	3.5	3

A negative difference (after − before) represents a lower score. A positive difference represents an improved score. Let's test the null hypothesis that the two-day GRE workshop has no effect, in the sense that the population median gained score is 0, against the alternative hypothesis that the population median gained score is positive.

Observations that have a tie score use a decimal point ranking and are ranked the same. The next ranking includes all of the previous rankings.

Questions to Explore

a. For ranks applied to the absolute values of the differences, find the rank sum for the differences that were positive.

b. Consider each of the possible ways that positive and negative signs could be assigned to these three differences. For each case, find the rank sum for the positive differences. Create the sampling distribution of this rank sum that applies if the workshop truly has no effect.

c. Find the P-value for the Wilcoxon signed-ranks test, using the sampling distribution of the rank sum created in part b.

Subject	Before	After	Difference	Absolute Value	Rank of Absolute Value	Signed Rank
1	2.5	3	0.5	0.5	1.5	1.5
2	4	3.5	−0.5	0.5	1.5	−1.5
3	1.5	3	1.5	1.5	3	3

b. For the difference values of the three subjects, 0.5, − 0.5, and 1.5, Table 15.9 shows all the possible ways the differences could have been positive or negative. For each sample, this table also shows the sum of ranks for the positive differences. The observed data are Sample 1, which had a rank sum of 4.5.

Table 15.9 Possible Samples with Absolute Difference Values of Sample

Subject	1	2	3	4	5	6	7	8	Rank of Absolute Value
1	0.5	0.5	−0.5	0.5	−0.5	0.5	−0.5	−0.5	1.5
2	−0.5	−0.5	−0.5	0.5	−0.5	0.5	0.5	0.5	1.5
3	1.5	−1.5	1.5	1.5	−1.5	−1.5	1.5	−1.5	3
Sum of Ranks for Positive Differences									
	4.5	1.5	3	6	0	3	4.5	1.5	

If the workshop has no effect, then the eight possible samples in Table 15.9 are equally likely. The table that follows summarizes the sampling distribution of the rank sum for the positive differences, presuming no effect of the workshop. For example, the rank sum was 4.5 for two of the eight samples, so its probability is 2/8.

Sampling Distribution of Rank Sum for the Positive Differences

Rank Sum	Probability
0	1/8
1.5	2/8
3	2/8
4.5	2/8
6	1/8

c. The larger the sum of ranks for the positive differences, the greater the evidence that the workshop has a positive effect. So, the P-value is the probability that this sum of ranks is at least as large as observed. Since three of the eight possible samples had a rank sum for the positive differences of at least 4.5 (the observed value), the P-value is 3/8 = 0.375.

Insight

Suppose we instead used the sign test. We then observe that two of the three differences are positive. For the alternative hypothesis that the workshop has a positive effect, the P-value is the probability that at least two of the three differences are positive, when the chance is 0.50 that any particular difference is positive. Using the binomial distribution, you can find that the P-value is 0.50 (Exercise 15.12).

The sign test ignores the fact that the two positive differences are larger than the negative difference. The Wilcoxon signed-ranks test uses this information. By taking this extra information into account, its P-value of 0.375 is smaller than the P-value of 0.50 from the sign test. However, the P-value is still not small. With only three observations, the one-sided P-value can be no smaller than one-eighth, which is the P-value for the largest possible value (which is 6) for the rank sum of positive differences.

The MINITAB output for the complete data set is shown below. The test of medians between the two groups results in a statistically significant P-value of 0.018.

Wilcoxon Signed Rank CI: Before, After

	N	Estimated Median	Achieved Confidence	Confidence Interval Lower	Upper
Before	12	3.00	94.5	2.25	3.75
After	12	4.00	94.5	3.00	4.75

Wilcoxon Signed Rank Test: Diff

Test of median = 0.000000 versus median not = 0.000000

	N	Test	Statistic	P	Median
Diff	12	11	60.0	0.018	1.000

Try Exercise 15.12