CHAPTER

27

Nonparametric Tests

Blickwinkel/Alamy

IN THIS CHAPTER WE COVER...

- Comparing two samples: the Wilcoxon rank sum test
- Matched pairs: the Wilcoxon signed rank test
- Comparing several samples: the Kruskal-Wallis test

The most commonly used methods of inference about the means of quantitative response variables assume that the variables in question have Normal distributions in the population or populations from which we draw our data. In practice, of course, no distribution is exactly Normal. Fortunately, our usual methods of inference about population means (the one-sample and two-sample t procedures and analysis of variance) are quite **robust.** That is, the results of inference are not very sensitive to moderate lack of Normality, especially when the samples are reasonably large. Practical guidelines for taking advantage of the robustness of these methods appear in Chapters 17, 18, and 24.

What can we do if plots suggest that the data are clearly not Normal, especially when we have only a few observations? This is not a simple question. Here are the basic options:

robustness

outliers

1. If lack of Normality is due to outliers, it may be legitimate to remove **outliers** if you have reason to think that they do not come from the same population as the other observations. Equipment failure that produced a bad measurement, for example, entitles you to remove the outlier and analyze the remaining data. *But if an outlier appears to be "real data," you should not arbitrarily remove it.* The discussion topic in Chapter 2 (page 55) addresses in more depth how to recognize and treat outliers.

CAUTION

transforming data

2. Sometimes we can **transform** our data so that their distribution is more nearly Normal. Transformations such as the logarithm that shorten the long tail of right-skewed distributions are particularly helpful. We used the logarithm transformation in Example 4.2 (page 96).

other distributions

3. In some settings, **other standard distributions** replace the Normal distributions as models for the overall pattern in the population. The lifetimes in service of equipment and the survival times of cancer patients after treatment usually have right-skewed distributions. Statistical studies in these areas use families of right-skewed distributions rather than Normal distributions. There are inference procedures for the parameters of these distributions that replace the *t* procedures.

bootstrap methods
permutation tests

4. Modern **bootstrap methods** and **permutation tests** use heavy computing to avoid requiring Normality or any other specific form of sampling distribution. We recommend these methods unless the sample is so small that it may not represent the population well. For an introduction, see the Companion Chapter 16 of the somewhat more advanced text *Introduction to the Practice of Statistics*, available online at `www.macmillanlearning.com/ips9e`.

nonparametric methods

5. Finally, there are other **nonparametric methods,** which do not assume any specific form for the distribution of the population. Unlike bootstrap and permutation methods, common nonparametric methods do not make use of the actual values of the observations.

This chapter concerns one type of nonparametric procedure: tests that can replace the *t* tests and one-way analysis of variance *when the Normality conditions for those tests are either not met or questionable*. The most useful nonparametric tests are **rank tests** based on the rank (place in order) of each observation in the set of all the data.

rank tests

Figure 27.1 presents an outline of the standard tests (based on Normal distributions) and the rank tests that compete with them. These rank tests require that the population or populations have *continuous distributions*. That is, each distribution must be described by a *density curve* (Chapter 9, page 225) that allows observations to take any value within some interval of outcomes. The Normal curves are one shape of density curve. Rank tests allow curves of any shape.

The rank tests we will study concern the *center* of a population or populations. When a population has at least roughly a Normal distribution, we describe its center by the mean. The "Normal tests" in Figure 27.1 all test hypotheses about population means. When distributions are strongly skewed, we often prefer the median to the mean as a measure of center. In their simplest form, the hypotheses for rank tests just replace mean with median.

We begin by describing the most common rank test, the one for comparing two samples. In this setting we also explain ideas common to all rank tests: the

Setting	Normal test	Rank test
One sample	One-sample *t* test Chapter 17	Wilcoxon signed rank test
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample *t* test Chapter 18	Wilcoxon rank sum test (Mann-Whitney test)
Several independent samples	One-way ANOVA *F* test Chapter 24	Kruskal-Wallis test

▶ FIGURE 27.1
Comparison of tests based on Normal distributions with rank tests for similar settings.

big idea of using ranks, the conditions required by rank tests, the nature of the hypotheses tested, and the difference between exact distributions for use with small samples and Normal approximations for use with larger samples.

COMPARING TWO SAMPLES: THE WILCOXON RANK SUM TEST

Two-sample problems (see Chapter 18) are among the most common in statistics. The most useful nonparametric significance test compares two distributions. Here is an example of this setting.

EXAMPLE 27.1 Weeds among the corn

STATE: Does the presence of small numbers of weeds reduce the yield of corn? Lamb's-quarter is a common weed in corn fields. A researcher planted corn at the same rate in 8 small plots of ground, then weeded the corn rows by hand to allow no weeds in 4 randomly selected plots and exactly 3 lamb's-quarter plants per meter of row in the other 4 plots. Here are the yields of corn (bushels per acre) in each of the plots:[1]

0 weeds per meter	166.7	172.2	165.0	176.9
3 weeds per meter	158.6	176.4	153.1	156.0

Blickwinkel/Alamy

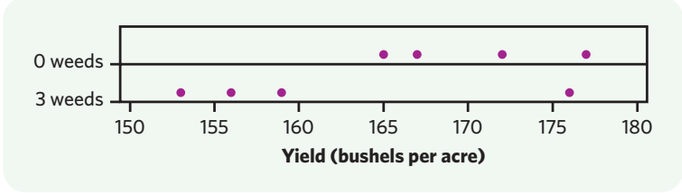
PLAN: Make a graph to compare the two sets of yields. Test the null hypothesis that there is no difference against the one-sided alternative that yields are higher when no weeds are present.

SOLVE (first steps): A dotplot (Figure 27.2) suggests that yields may be higher when there are no weeds. There is one outlier; because it is correct data, we cannot remove it. The samples are too small to rely on the robustness of the two-sample t test. We will now develop a test that does not require Normality in order to be valid.

First, rank all 8 observations together. To do this, arrange them in order from smallest to largest:

153.1 156.0 158.6 **165.0** **166.7** **172.2** 176.4 **176.9**

The boldface entries in the list are the yields with no weeds present. We see that four of the five highest yields come from that group, suggesting that yields are higher with no weeds. The idea of rank tests is to look just at position in this



◀ FIGURE 27.2
Dotplot of corn yields from plots with no weeds and with 3 weeds per meter of row.

ordered list. To do this, replace each observation with its order, from 1 (smallest) to 8 (largest). These numbers are the *ranks:*

Yield	153.1	156.0	158.6	**165.0**	**166.7**	**172.2**	176.4	**176.9**
Rank	1	2	3	**4**	**5**	**6**	7	**8**

RANKS

To rank observations, first arrange them in order from smallest to largest. The **rank** of each observation is its position in this ordered list, starting with rank 1 for the smallest observation.

Moving from the original observations to their ranks retains only the ordering of the observations and makes no other use of their numerical values. Working with ranks allows us to dispense with specific conditions on the shape of the distribution, such as Normality.

If the presence of weeds reduces corn yields, we expect the ranks of the yields from plots without weeds to be larger as a group than the ranks from plots with weeds. Let's compare the *sums* of the ranks from the two treatments:

Treatment	Sum of ranks
No weeds	23
Weeds	13

These sums measure how much the ranks of the weed-free plots as a group exceed those of the weedy plots. In fact, the sum of the ranks from 1 to 8 is equal to 36, so it is enough to report the sum for one of the two groups. If the sum of the ranks for the weed-free group is 23, the ranks for the other group must add to 13, because $23 + 13 = 36$. If the weeds had no effect, we would expect the sum of the ranks in each group to be 18 (half of 36). Here are the facts we need in a more general form that takes account of the fact that our two samples need not be the same size.

THE WILCOXON RANK SUM TEST

Draw an SRS of size n_1 from one population and draw an independent SRS of size n_2 from a second population. There are N observations in all, where $N = n_1 + n_2$. Rank all N observations. The sum W of the ranks for the first sample is the **Wilcoxon rank sum statistic.** If the two populations have the same continuous distribution, then W has mean

$$\mu_W = \frac{n_1(N+1)}{2}$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

The **Wilcoxon rank sum test** rejects the hypothesis that the two populations have identical distributions when the rank sum W is far from its mean.

In the corn yield study of Example 27.1, we want to test the hypotheses

H_0: no difference in distribution of yields
H_a: yields are systematically higher in weed-free plots

Our test statistic is the rank sum $W = 23$ for the weed-free plots.

EXAMPLE 27.2 Weeds among the corn: inference

SOLVE: First, note that the conditions for the Wilcoxon test are met: The data come from a randomized comparative experiment, and the yield of corn in bushels per acre has a continuous distribution.

There are $N = 8$ observations in all, with $n_1 = 4$ and $n_2 = 4$. The sum of ranks for the weed-free plots has mean

$$\mu_W = \frac{n_1(N+1)}{2} = \frac{(4)(9)}{2} = 18$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}} = \sqrt{\frac{(4)(4)(9)}{12}} = \sqrt{12} = 3.464$$

Although the observed rank sum $W = 23$ is higher than the mean, it is only about 1.4 standard deviations higher. We now suspect that the data do not give strong evidence that yields are higher in the population of weed-free corn.

The P-value for our one-sided alternative is $P(W \geq 23)$, the probability that W is at least as large as the value for our data when H_0 is true. Software tells us that this probability is $P = 0.1$.

CONCLUDE: The data provide some evidence ($P = 0.1$) that corn yields are lower when weeds are present. There are only 4 observations in each group, so even quite large effects can fail to reach the levels of significance usually considered convincing, such as $P < 0.05$. A larger experiment might clarify the effect of weeds on corn yield.

APPLY YOUR KNOWLEDGE

27.1 Daily activity and obesity. Our lead example for the two-sample t procedures in Chapter 18 concerned a study comparing the level of physical activity of lean and mildly obese people who don't exercise. Here are the minutes per day that the subjects spent standing or walking over a 10-day period:

Lean subjects		Obese subjects	
511.100	543.388	260.244	416.531
607.925	677.188	464.756	358.650
319.212	555.656	367.138	267.344
584.644	374.831	413.667	410.631
578.869	504.700	347.375	426.356

The data are a bit irregular but not distinctly non-Normal. Let's use the Wilcoxon test for comparison with the two-sample t test.

(a) Find the median minutes spent standing or walking for each group. Which group appears more active?

(b) Arrange all 20 observations in order and find the ranks.

(c) Take W to be the sum of the ranks for the lean group. What is the value of W? If the null hypothesis (no difference between the groups) is true, what are the mean and standard deviation of W?

(d) Does comparing W with the mean and standard deviation suggest that the lean subjects are more active than the obese subjects?

27.2 Immune response. NOD receptors are involved in the immune response to bacterial infections. Researchers investigated the impact of a deleterious mutation ($Bid^{-/-}$) on NOD-mediated immune signaling in mice. Here are the immune responses to a NOD trigger in wild-type and $Bid^{-/-}$ mice, as assessed by the serum concentration of a specific protein marker (in picograms per milliliter, pg/ml):[2]

Wild-type	17.3	22.8	30.8	38.0	55.1
$Bid^{-/-}$	45.9	27.1	25.7	27.7	38.8

(a) The sample sizes are small. We may prefer a nonparametric test to compare the groups. Find the Wilcoxon rank sum W for the wild-type group, along with its mean and standard deviation under the null hypothesis (no difference between groups).

(b) Do you think that W is far enough from the mean under H_0 to suggest that there may be a difference between the groups?

The Normal approximation for *W* To calculate the P-value $P(W \geq 23)$ for Example 27.2, we need to know the sampling distribution of the rank sum W when the null hypothesis is true. This distribution depends on the two sample sizes n_1 and n_2. Tables of such values are therefore unwieldy, though you can find them in handbooks of statistical tables. Most statistical software will give you P-values, as well as carry out the ranking and calculate W. However, many software packages give only approximate P-values. You must learn what your software offers.

With or without software, P-values for the Wilcoxon test are often based on the fact that **the rank sum statistic W becomes approximately Normal as the two sample sizes increase.** We can then form yet another z statistic by standardizing W:

$$z = \frac{W - \mu_W}{\sigma_W}$$
$$= \frac{W - n_1(N+1)/2}{\sqrt{n_1 n_2 (N+1)/12}}$$

Use standard Normal probability calculations to find P-values for this statistic. Because W takes only whole-number values, an idea called the *continuity correction* improves the accuracy of the approximation.

CONTINUITY CORRECTION

To apply the **continuity correction** in a Normal approximation for a variable that takes only whole-number values, act as if each whole number occupies the entire interval from 0.5 below the number to 0.5 above it.

EXAMPLE 27.3 Weeds among the corn: Normal approximation

The standardized rank sum statistic W in our corn yield example is

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{23 - 18}{3.464} = 1.44$$

We expect W to be larger when the alternative hypothesis is true, so the approximate P-value is (from Table B)

$$P(Z \geq 1.44) = 0.0749$$

We can improve this approximation by using the continuity correction. To do this, act as if the whole number 23 occupies the entire interval from 22.5 to 23.5. Calculate the P-value $P(W \geq 23)$ as $P(W \geq 22.5)$, because the value 23 is included in the range whose probability we want. Here is the calculation:

$$\begin{aligned} P(W \geq 22.5) &= P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{22.5 - 18}{3.464}\right) \\ &= P(Z \geq 1.30) \\ &= 0.0968 \end{aligned}$$

using Table B (we get $P = 0.0970$ using technology for a more accurate Normal computation). If you do not use the exact distribution of W (from software or tables), you should always use the continuity correction in calculating P-values.

APPLY YOUR KNOWLEDGE

27.3 Daily activity and obesity, continued. In Exercise 27.1 you found the Wilcoxon rank sum W and its mean and standard deviation. We want to test the null hypothesis that the two groups don't differ in activity against the alternative hypothesis that the lean subjects spend more time standing and walking.

(a) What is the probability expression for the P-value of W if we use the continuity correction?

(b) Find the P-value. What do you conclude?

27.4 Immune response, continued. Use your values of W, μ_W, and σ_W from Exercise 27.2 to see whether the Bid$^{-/-}$ mutation has an effect on immune response.

(a) The two-sided P-value is $2P(W \leq ?)$. Using the continuity correction, what number replaces the ? in this probability?

(b) Find the P-value. What do you conclude about the effect of the Bid$^{-/-}$ mutation?

27.5 Fruit fly reproduction. Reproduction has a high physiological cost. Researchers hypothesized that there is a trade-off between longevity and reproduction. They designed an experiment comparing the reproductive output of wild-type fruit flies and mutants with a longer reproductive cycle and higher longevity. Here are the number of eggs produced per female when the fruit flies were fed a protein-rich diet:[3]

Wild-type	63.80	86.45	82.60	73.35	85.50
Mutant	40.35	67.95	48.50	64.25	60.35

Do the data provide evidence that the mutants invest less in reproduction (as measured by the number of eggs produced per female) than the wild-type fruit flies? Follow the four-step process as illustrated in Examples 27.1 and 27.2.

Using technology For samples as small as those in the corn yield study of Example 27.1, we prefer software that gives the exact *P*-value for the Wilcoxon test rather than the Normal approximation. Neither the Excel spreadsheet nor the TI-83 calculator has menu entries for rank tests. Some statistical programs offer only the Normal approximation for the Wilcoxon test. Other statistical programs carry out the **Mann-Whitney test** rather than the Wilcoxon test. The two tests always have the same *P*-value, because the two test statistics are related by simple algebra.

Mann-Whitney test

EXAMPLE 27.4 Weeds among the corn: software output

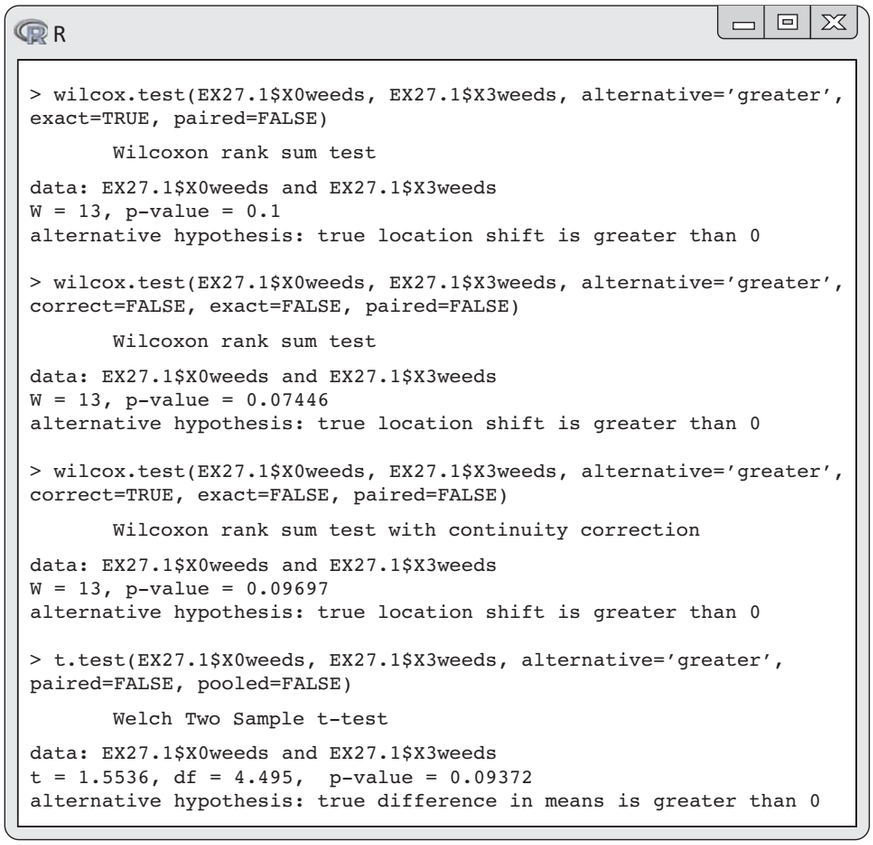
Figure 27.3 displays output from the software program R for the corn yield data. The top three queries reflect the three methods available to obtain the *P*-value of the Wilcoxon sum rank test: exact, Normal, and Normal with a continuity correction, respectively. The exact *P*-value is $P = 0.1$. The Normal approximation with continuity correction, $P = 0.097$ in Example 27.3 and in the third R query, is quite accurate.

The last query in Figure 27.3 is the two-sample *t* test from Chapter 18, which does not assume that the two populations have the same standard deviation. It gives $P = 0.0937$, close to the Wilcoxon *P*-value. Because the *t* test is quite robust, it is somewhat unusual for *P*-values derived from *t* and those from *W* to differ greatly.

APPLY YOUR KNOWLEDGE

27.6 Immune response: software. Use your software to repeat the Wilcoxon test you did in Exercise 27.4. By comparing the results, state how your software finds *P*-values for *W*: exact distribution, Normal approximation with continuity correction, or Normal approximation without continuity correction.

27.7 Daily activity and obesity: software. Use your software to carry out the one-sided Wilcoxon rank sum test that you did by hand in Exercise 27.3. Use the exact distribution if your software will do it. Compare the software result with your result in Exercise 27.3.



```
R

> wilcox.test(EX27.1$X0weeds, EX27.1$X3weeds, alternative='greater',
exact=TRUE, paired=FALSE)

        Wilcoxon rank sum test

data: EX27.1$X0weeds and EX27.1$X3weeds
W = 13, p-value = 0.1
alternative hypothesis: true location shift is greater than 0

> wilcox.test(EX27.1$X0weeds, EX27.1$X3weeds, alternative='greater',
correct=FALSE, exact=FALSE, paired=FALSE)

        Wilcoxon rank sum test

data: EX27.1$X0weeds and EX27.1$X3weeds
W = 13, p-value = 0.07446
alternative hypothesis: true location shift is greater than 0

> wilcox.test(EX27.1$X0weeds, EX27.1$X3weeds, alternative='greater',
correct=TRUE, exact=FALSE, paired=FALSE)

        Wilcoxon rank sum test with continuity correction

data: EX27.1$X0weeds and EX27.1$X3weeds
W = 13, p-value = 0.09697
alternative hypothesis: true location shift is greater than 0

> t.test(EX27.1$X0weeds, EX27.1$X3weeds, alternative='greater',
paired=FALSE, pooled=FALSE)

        Welch Two Sample t-test

data: EX27.1$X0weeds and EX27.1$X3weeds
t = 1.5536, df = 4.495,  p-value = 0.09372
alternative hypothesis: true difference in means is greater than 0
```

◀ FIGURE 27.3
Output from *R* for the data in Example 27.1. The output compares the three methods for computing the *P*-value of the Wilcoxon rank sum test and the *P*-value obtained with a two-sample *t* test.

27.8 Weeds among the corn. The corn yield study of Example 27.1 also examined yields in four plots having 9 lamb's-quarter plants per meter of row. The yields (bushels per acre) in these plots were

162.8 142.4 162.7 162.4

There is a clear outlier, but rechecking the results found that this is the correct yield for this plot. The outlier makes us hesitant to use t procedures because $\bar{x}$ and s are not resistant to outliers.

(a) Is there evidence that the presence of 9 weeds per meter reduces corn yields when compared with weed-free corn? Use the Wilcoxon rank sum test with the data above and part of the data from Example 27.1 to answer this question.

(b) Compare the results from (a) with those from the two-sample t test for these data.

(c) Now remove the low outlier 142.4 from the data with 9 weeds per meter. Repeat both the Wilcoxon and t analyses. By how much did the outlier reduce the mean yield in its group? By how much did it increase the standard deviation? Did it have a practically important impact on your conclusions?

What hypotheses does Wilcoxon test? Our null hypothesis is that weeds do not affect yield. The alternative hypothesis is that yields are lower when weeds are present. If we are willing to assume that yields are Normally distributed, or if

we have reasonably large samples, we can use the two-sample t test for means. Our hypotheses then have the form

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

When the distributions may not be Normal, we might restate the hypotheses in terms of population medians rather than means:

$$H_0: \text{population median}_1 = \text{population median}_2$$
$$H_a: \text{population median}_1 > \text{population median}_2$$

The Wilcoxon rank sum test provides a test of these median-based hypotheses, but only if an additional condition is met: Both populations must have distributions of *the same shape*. That is, the density curve for corn yields with 3 weeds per meter looks exactly like that for yields with no weeds except that it may slide to a different location on the scale of yields.

The same-shape condition is too strict to be reasonable in practice. Fortunately, the Wilcoxon test also applies in a more useful setting. It compares *any two* continuous distributions, whether or not they have the same shape, by testing hypotheses that we can state in words as

H_0: the two distributions are the same

H_a: one has values that are systematically larger

A more exact statement of the "systematically larger" alternative hypothesis is a bit tricky, so we won't try to give it here.[4] The R output of Figure 27.3 states the hypotheses in terms of location shift. These hypotheses really are "nonparametric," because they do not involve any specific parameter such as the mean or median. If the two distributions do have the same shape, the general hypotheses are reduced to comparing medians. *Many texts and computer outputs state the hypotheses in terms of medians, sometimes ignoring the same-shape condition.* We recommend that you express the hypotheses in words rather than symbols. "Yields are systematically higher in weed-free plots" is easy to understand and is a good statement of the effect that the Wilcoxon test looks for.

Why don't we discuss the confidence intervals for the difference in population medians that some statistical programs offer? These intervals require the unrealistic same-shape condition. The more general "systematically larger" hypothesis does not involve a specific parameter, so there is no accompanying confidence interval.

APPLY YOUR KNOWLEDGE

27.9 Daily activity and obesity: hypotheses. We could use either two-sample t or the Wilcoxon rank sum to test the null hypothesis that lean and mildly obese people don't differ in the time they spend standing and walking against the alternative hypothesis that lean people generally spend more time in these activities. Explain carefully what H_0 and H_a are for t and for W.

27.10 Immune response: hypotheses. We are interested in whether the deleterious mutation Bid$^{-/-}$ changes the serum concentration of a protein marker of NOD-mediated immune signaling in mice "on average."

(a) State null and alternative hypotheses in terms of population means. What test would we typically use for these hypotheses? What conditions does this test require?

(b) State null and alternative hypotheses in terms of population medians. What test would we typically use for these hypotheses? What conditions does this test require?

Dealing with ties in rank tests We have chosen our examples and exercises to this point rather carefully: They all involve data in which *no two values are the same*. This allowed us to rank all the values. In practice, however, we often find observations tied at the same value. What shall we do? The usual practice is to *assign all tied values the* **average** *of the ranks they occupy*. Here is an example with 6 observations:

average ranks

Observation	153	155	158	158	161	164
Rank	1	2	3.5	3.5	5	6

The tied observations occupy the third and fourth places in the ordered list, so they share rank 3.5.

The exact distribution for the Wilcoxon rank sum W applies only to data without ties. Moreover, the standard deviation σ_W must be adjusted if ties are present. The Normal approximation can be used after the standard deviation is adjusted. Statistical software will detect ties, make the necessary adjustment, and switch to the Normal approximation. *In practice, software is required to use rank tests when the data contain tied values.*

Some data have many ties because the scale of measurement has only a few values. Rank tests are often used for such data. Here is an example.

EXAMPLE 27.5 Food safety at fairs

STATE: Food sold at outdoor fairs and festivals may be less safe than food sold in restaurants because it is prepared in temporary locations and often by volunteer help. What do people who attend fairs think about the safety of the food served? One study asked this question of people at a number of fairs in the Midwest: "How often do you think people become sick because of food they consume prepared at outdoor fairs and festivals?" The possible responses were

1 = very rarely
2 = once in a while
3 = often
4 = more often than not
5 = always

In all, 303 people answered the question. Of these, 196 were women and 107 were men.[5] We suspect that women are more concerned than men about food safety. Is there good evidence for this conclusion?

PLAN: Do data analysis to understand the difference between women and men. Because the 1-to-5 ratings are in fact rankings rather than actual quantitative data, we should use a statistical procedure involving ranks. Check the conditions required by the Wilcoxon test. If the conditions are met, use the Wilcoxon test for the hypotheses

H_0: men and women do not differ in their responses

H_a: women give systematically higher responses than men

SOLVE: Here are the data, presented as a two-way table of counts:

	Response					
	1	2	3	4	5	Total
Female	13	108	50	23	2	196
Male	22	57	22	5	1	107
Total	35	165	72	28	3	303

Comparing row percents shows that the women in the sample do tend to give higher responses (showing more concern):

	Response					
	1	2	3	4	5	Total
Percent of females	6.6	55.1	25.5	11.7	1.0	100
Percent of males	20.6	53.3	20.6	4.7	1.0	100

Are these differences between women and men statistically significant?

The most important condition for inference is that the subjects be a *random sample* of people who attend fairs, at least in the Midwest. The researcher visited 11 different fairs. She stood near the entrance and stopped every 25th adult who passed. Because no personal choice was involved in choosing the subjects, we can reasonably treat the data as coming from a random sample. (As usual, there was some nonresponse, which could create bias.) The Wilcoxon test also requires that responses have *continuous distributions*. We think that the subjects' opinions about how often people become sick from food at fairs really do form a continuous distribution. The questionnaire asks them to round off their opinions to the nearest value in the five-point scale. So we are willing to use the Wilcoxon test.

Because the responses can take only five values, there are many ties. All 35 people who chose "very rarely" are tied at 1, and all 165 who chose "once in a while" are tied at 2. Figure 27.4 gives output from Minitab. The Wilcoxon (reported as Mann-Whitney) test for the one-sided alternative that women are more concerned about food safety at fairs is highly significant ($P = 0.0004$, adjusted for tie).

CONCLUDE: There is very strong evidence ($P = 0.0004$) that women are more concerned than men about the safety of food served at fairs.

Minitab

Mann-Whitney Test: F, M

```
     N   Median
F  196   2.0000
M  107   2.0000

Point estimate for ETA1-ETA2 is 0.0000
W = 31996.5
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0.0012
The test is significant at 0.0004 (adjusted for ties)
```

◀ FIGURE 27.4
Output from Minitab for the data of Example 27.5.

Because the sample sizes are so large in Example 27.5, a *t* test would have given similar results. However, the *t* statistic treats the response values 1 through 5 as meaningful numbers. In particular, the possible responses are treated as though they are equally spaced. The difference between "very rarely" and "once in a while" is considered the same as the difference between "once in a while" and "often." This may not make sense. The rank test, on the other hand, uses only the order of the responses, not their actual values. The responses are arranged in order from least to most concerned about safety, so the rank test makes sense. *Statisticians often avoid using t procedures when there is not a fully meaningful scale of measurement.*

Because we have a two-way table, we might have applied the chi-square test (Chapter 22), which asks if there is a significant relationship *of any kind* between gender and response. The chi-square test ignores the ordering of the responses and so doesn't tell us whether women are *more* concerned than men about the safety of the food served. This question depends on the ordering of responses from least concerned to most concerned.

APPLY YOUR KNOWLEDGE

Software is required to adequately carry out the Wilcoxon rank sum test in the presence of ties. All of the following exercises concern data with ties.

27.11 Do birds learn to time their breeding? Blue titmice eat caterpillars but do they adjust their breeding based on the peak caterpillar supply in the previous year? Researchers randomly assigned 7 bird pairs to have the natural caterpillar supply supplemented when the birds were feeding their young and another 6 pairs to serve as a control group relying only on the natural food supply. The next year, they measured how many days after the caterpillar peak the birds produced their nestlings.[6] Here are the data (days after the caterpillar peak):


Hugh Clark/Frank Lane Picture Agency/CORBIS

Control	4.6	2.3	7.7	6.0	4.6	−1.2	
Supplemented	15.5	11.3	5.4	16.5	11.3	11.4	7.7

The null hypothesis is no difference in timing; the alternative hypothesis is that the supplemented birds miss the peak by more days because they don't adjust their breeding date.

(a) There are three sets of ties, at 4.6, 7.7, and 11.3. Arrange the observations in order and assign average ranks to each tied observation.

(b) Take W to be the rank sum for the supplemented group. What is the value of W?

(c) Use software to find the P-value of the Wilcoxon test, and conclude.

27.12 Mycorrhizal colonies and plant nutrition. Mycorrhizal fungi are present in the roots of many plants. In this symbiotic relationship, the plant supplies nutrition to the fungus and the fungus helps the plant absorb nutrients from the soil. An experiment compared two genetic types of tomato plants, a wild-type variety that is susceptible to mycorrhizal colonies and a mutant that is not. In particular, researchers expect that, in the absence of fertilizer, plants with mycorrhizal fungi would better be able to absorb what little nitrogen might be available in the soil. Here are plant nitrogen contents at harvest time (in percent of whole dry mass):[7]

Wild-type	1.21	1.57	1.30	1.19	1.23	1.29
Mutant	1.23	1.05	1.28	1.04	1.25	1.21

(a) Arrange the observations in order and find their ranks.

(b) Take W to be the rank sum for wild-type tomato plants. What is the value of W?

(c) Use software: Does W provide significant evidence that the nitrogen contents of the wild-type tomato plants with mycorrhizal fungi are systematically larger than those in mutants without mycorrhizal fungi?

27.13 Good smells and purchasing behavior. Many aspects of human behavior are influenced by our environment. One study of this phenomenon took place in a small pizzeria in France on two Saturday evenings in May. On one of these evenings, a relaxing lavender odor was spread through the restaurant. The two evenings were comparable in many ways (weather, customer count, and so on), so we are willing to regard the data as independent SRSs from spring Saturday evenings at this restaurant. The authors say, "Therefore at this stage it would be impossible to generalize the results to other restaurants." Here are the amounts spent (in euros) by customers:[8]

No odor									
15.9	18.5	15.9	18.5	18.5	21.9	15.9	15.9	15.9	15.9
15.9	18.5	18.5	18.5	20.5	18.5	18.5	15.9	15.9	15.9
18.5	18.5	15.9	18.5	15.9	18.5	15.9	25.5	12.9	15.9
Lavender odor									
21.9	18.5	22.3	21.9	18.5	24.9	18.5	22.5	21.5	21.9
21.5	18.5	25.5	18.5	18.5	21.9	18.5	18.5	24.9	21.9
25.9	21.9	18.5	18.5	22.8	18.5	21.9	20.7	21.9	22.5

Examine the data and comment on departures from Normality. Is there significant evidence that the lavender odor encourages customers to spend more? Follow the four-step process.

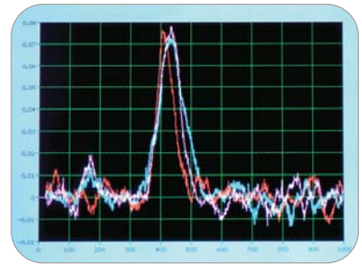
MATCHED PAIRS: THE WILCOXON SIGNED RANK TEST

We use the one-sample t procedures for inference about the mean of one population or for inference about the mean difference in a matched pairs setting. The matched pairs setting is more important, because good studies are generally comparative. We will now look at a rank test for this setting.

EXAMPLE 27.6 Visual receptive fields

STATE: Neurons in the visual cortex fire an increased number of action potentials ("spikes") when a specific area of the retina is visually stimulated. This area is called the neuron's receptive field. A neuron in the primary visual cortex is recorded while a monkey first looks at a dot on a screen (spontaneous activity, SA) and then looks at a pattern of random dots (response, R). Here is the neural activity in number of spikes per second for 9 different recordings of that neuron:[9]

Recording	1	2	3	4	5	6	7	8	9
SA	2.5	7.5	10.0	0.0	12.5	2.5	0.0	2.5	17.5
R	16.7	20.0	23.3	16.7	56.7	0.0	26.7	36.7	10.0
Difference (R – SA)	14.2	12.5	13.3	16.7	44.2	−2.5	26.7	34.2	−7.5



Donald Pye/Alamy

The neuron's receptive field is activated if the neuron's response is systematically higher following presentation of the visual pattern (R) than just before (SA). Is this neuron's receptive field activated by the visual pattern presented?

PLAN: We would like to test the hypotheses

H_0: neural activity has the same distribution for both SA and R

H_a: neural activity is systematically higher for R

SOLVE (first steps): Because this is a matched pairs design, we base our inference on the differences. The matched pairs t test gives $t = 3.085$ with one-sided P-value $P = 0.007$. We cannot reliably assess Normality from so few observations. We would therefore like to use a rank test.

Positive differences in Example 27.6 indicate that the neuron's activity was stronger during the stimulus presentation (R). If neural activity is generally higher for the response to stimulus, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction. We therefore compare the **absolute values** of the differences, that is, their magnitudes without a sign. Here they are, with boldface indicating the positive values:

absolute value

14.2 **12.5** **13.3** **16.7** **44.2** 2.5 **26.7** **34.2** 7.5

Arrange these in increasing order and assign ranks, keeping track of which values were originally positive. Tied values receive the average of their ranks. If there are zero differences, discard them before ranking.

Absolute value	2.5	7.5	**12.5**	**13.3**	**14.2**	**16.7**	**26.7**	**34.2**	**44.2**
Rank	1	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**

The test statistic is the sum of the ranks of the positive differences. (We could equally well use the sum of the ranks of the negative differences.) This is the *Wilcoxon signed rank statistic*. Its value here is $W^+ = 42$.

THE WILCOXON SIGNED RANK TEST FOR MATCHED PAIRS

Draw an SRS of size n from a population for a matched pairs study and take the differences in responses within pairs. Rank the absolute values of these differences. The sum W^+ of the ranks for the positive differences is the **Wilcoxon signed rank statistic.** If the distribution of the responses is not affected by the different treatments within pairs, then W^+ has mean

$$\mu_{W^+} = \frac{n(n+1)}{4}$$

and standard deviation

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The **Wilcoxon signed rank test** rejects the hypothesis that there are no systematic differences within pairs when the rank sum W^+ is far from its mean.

EXAMPLE 27.7 Visual receptive fields, continued

SOLVE: In the neural receptive field study of Example 27.6, $n = 9$. If the null hypothesis (no systematic effect of the visual stimulus) is true, the mean of the signed rank statistic is

$$\mu_{W^+} = \frac{n(n+1)}{4} = \frac{(9)(10)}{4} = 22.5$$

The standard deviation of W^+ under the null hypothesis is

$$\begin{aligned}\sigma_{W^+} &= \sqrt{\frac{n(n+1)(2n+1)}{24}} \\ &= \sqrt{\frac{(9)(10)(19)}{24}} \\ &= \sqrt{71.25} = 8.441\end{aligned}$$

The observed value $W^+ = 42$ is much larger than the mean. We now expect that the data are statistically significant.

The P-value for our one-sided alternative is $P(W^+ \geq 42)$, calculated using the distribution of W^+ when the null hypothesis is true. Software gives a P-value of $P = 0.0098$.

CONCLUDE: The data give strong evidence ($P < 0.01$) that neural activity is higher during the response to the visual stimulus (R) than just before (SA). This neuron's receptive field is activated by the visual pattern presented.

APPLY YOUR KNOWLEDGE

27.14 Sitting versus squatting. Squatting for defecation continues to be the traditional position in most of Asia and Africa. In contrast, sitting on toilet seats has proliferated in Western counties since the 19th century with the development of the sewage system. This difference, along with differences in diet, might explain the higher rate of hemorrhoids and constipation in Western countries. A study asked if body position affects defecation in humans. Researchers recruited 6 healthy adult volunteers and took X-rays of their intestinal system to measure the anorectal angle (in degrees). Note that, as with a water hose, an angle of 180 degrees should provide the least amount of flow disruption. For each subject, an X-ray was taken once in a sitting position and once in a squatting position. Here are the findings:[10]

Borderlands/Alamy

Subject	Sitting	Squatting
1	109	122
2	73	127
3	117	141
4	77	103
5	98	121
6	125	142

(a) Find the differences within pairs, arrange them in order, and rank the absolute values. What is the signed rank statistic W^+?

(b) If the null hypothesis (no difference in anorectal angle) is true, what are the mean and standard deviation of W^+? Does comparing W^+ with this mean lead to a tentative conclusion?

27.15 Floral scents and learning. We hear that listening to Mozart improves students' performance on tests. In the EESEE case study "Floral Scents and Learning," investigators asked whether pleasant odors have a similar effect. Twenty-one subjects worked a paper-and-pencil maze while wearing a mask. The mask was either unscented or carried a floral scent. The response variable is the subjects' average time on three trials per condition. Each subject worked the maze with both masks, in a random order. The randomization is important because subjects tend to improve their times as they work a maze repeatedly. Table 27.1 gives the subjects' average times with each mask. Does the scent improve performance (that is,

TABLE 27.1 Average time (seconds) to complete a maze

Subject	Unscented	Scented	Difference	Subject	Unscented	Scented	Difference
1	30.60	37.97	−7.37	12	58.93	83.50	−24.57
2	48.43	51.57	−3.14	13	54.47	38.30	16.17
3	60.77	56.67	4.10	14	43.53	51.37	−7.84
4	36.07	40.47	−4.40	15	37.93	29.33	8.60
5	68.47	49.00	19.47	16	43.50	54.27	−10.77
6	32.43	43.23	−10.80	17	87.70	62.73	24.97
7	43.70	44.57	−0.87	18	53.53	58.00	−4.47
8	37.10	28.40	8.70	19	64.30	52.40	11.90
9	31.17	28.23	2.94	20	47.37	53.63	−6.26
10	51.23	68.47	−17.24	21	53.67	47.00	6.67
11	65.40	51.10	14.30				

shorten the time needed to complete the maze)? A matched pairs t test works well and gives $P = 0.3652$. Let's compare the Wilcoxon signed rank test.

(a) What are the ranks for the absolute values of the differences in Table 27.1? What is the value of W^+?

(b) What would be the mean and standard deviation of W^+ if the null hypothesis (scent makes no difference) were true? Compare W^+ with this mean (in standard deviation units) to reach a tentative conclusion about significance.

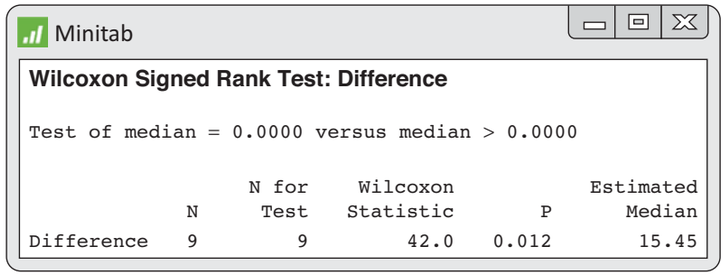
The Normal approximation for W^+ The distribution of the signed rank statistic when the null hypothesis (no difference) is true becomes approximately Normal as the sample size becomes large. We can then use Normal probability calculations (with the continuity correction) to obtain approximate P-values for W^+. Let's see how this works in the receptive field example, even though $n = 9$ is not a large sample.

EXAMPLE 27.8 **Visual receptive field: Normal approximation**

For $n = 9$ observations, we saw in Example 27.7 that $\mu_{W^+} = 22.5$ and that $\sigma_{W^+} = 8.441$. We observed $W^+ = 42$, so the one-sided P-value is $P(W^+ \geq 42)$. The continuity correction calculates this as $P(W^+ \geq 41.5)$, treating the value $W^+ = 42$ as occupying the interval from 41.5 to 42.5. We find the Normal approximation for the P-value either from software or by standardizing and using the standard Normal table:

$$P(W^+ \geq 41.5) = P\left(\frac{W^+ - 22.5}{8.441} \geq \frac{41.5 - 22.5}{8.441}\right)$$
$$= P(Z \geq 2.25)$$
$$= 0.0122$$

Figure 27.5 displays the output of four statistical programs. Minitab uses the Normal approximation and agrees with our calculation $P = 0.0122$. SPSS also uses the Normal approximation but does not perform the continuity correction and gets a slightly different two-sided P-value. CrunchIt! uses the exact distribution of W^+ and finds that the exact one-sided P-value for the Wilcoxon signed rank test is $P = 0.0098$, as we reported in Example 27.7. The Normal approximation is quite close to this. R gives us the option to compute the P-value with either method. A matched pairs t test (not shown) finds $P = 0.007$, but its validity depends on the distribution of the difference in neural activity. All tests tell us that there is strong evidence that this neuron's receptive field is activated by the visual pattern presented.

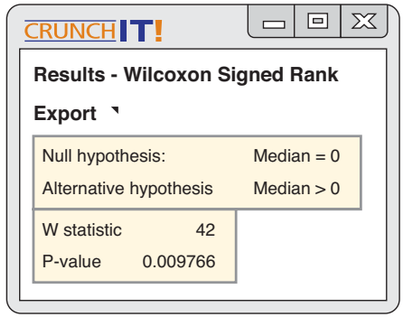


Minitab

Wilcoxon Signed Rank Test: Difference

Test of median = 0.0000 versus median > 0.0000

	N	N for Test	Wilcoxon Statistic	P	Estimated Median
Difference	9	9	42.0	0.012	15.45

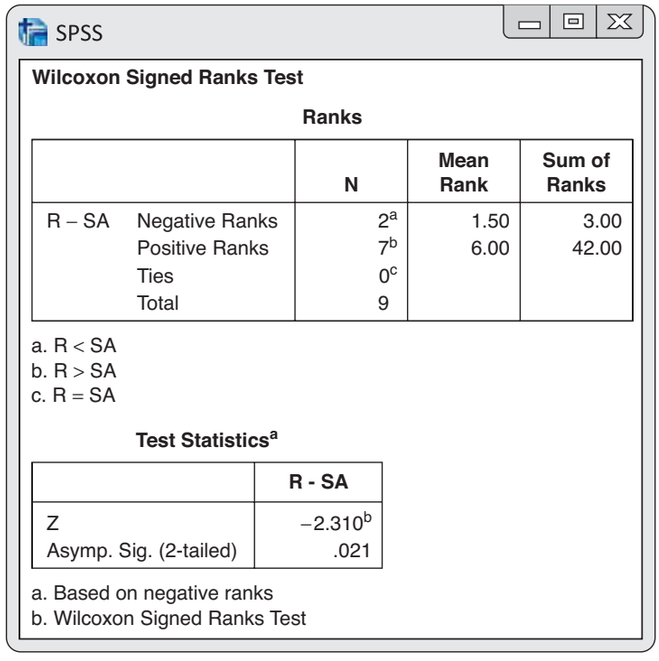


CRUNCHIT!

Results - Wilcoxon Signed Rank

Export

Null hypothesis:	Median = 0
Alternative hypothesis	Median > 0
W statistic	42
P-value	0.009766



SPSS

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
R – SA	Negative Ranks	2[a]	1.50	3.00
	Positive Ranks	7[b]	6.00	42.00
	Ties	0[c]		
	Total	9		

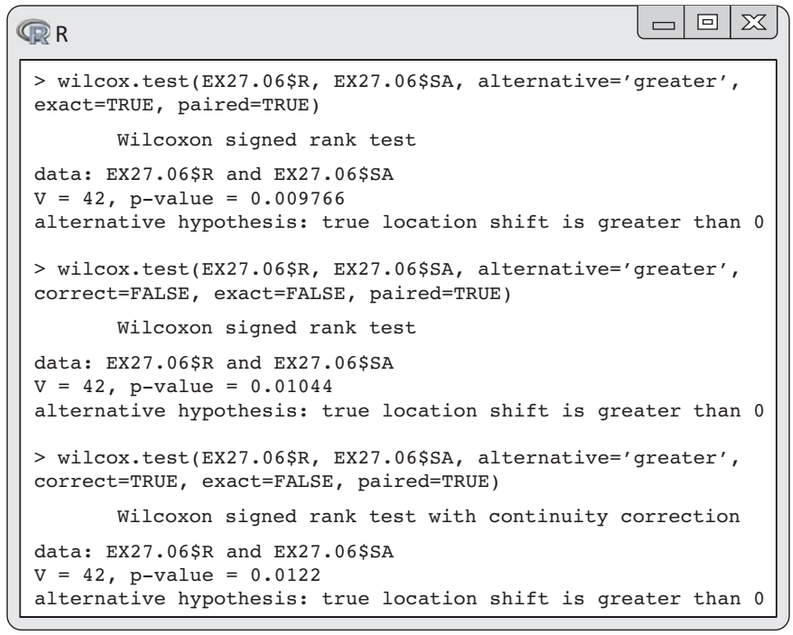
a. R < SA
b. R > SA
c. R = SA

Test Statistics[a]

	R - SA
Z	−2.310[b]
Asymp. Sig. (2-tailed)	.021

a. Based on negative ranks
b. Wilcoxon Signed Ranks Test

◀ FIGURE 27.5
Output from Minitab, CrunchIt!, SPSS, and R for the neural activity data of Example 27.6. CrunchIt! uses the exact method for calculating P, whereas Minitab and SPSS use a Normal approximation (one with and the other without a continuity correction). R gives the option of selecting the computation method (all three are included for comparison).



▶ **FIGURE 27.5**
(Continued)

APPLY YOUR KNOWLEDGE

27.16 Sitting versus squatting: Normal approximation. Continue your work from Exercise 27.14. Use the Normal approximation with continuity correction to find the *P*-value for the signed rank test against a two-sided alternative. What do you conclude?

27.17 Sitting versus squatting: W^+ versus *t*. Find the two-sided *P*-value for the matched pairs *t* test applied to the data in Exercise 27.14. The smaller *P*-value of *t* relative to W^+ means that *t* gives stronger evidence of the effect of sitting position. The *t* test takes advantage of assuming that the data are Normal, a considerable advantage for these very small samples. However, if you are not sure whether the data really are Normal, the nonparametric test is your only option.

27.18 Floral scents and learning: Normal approximation. Use the Normal approximation with continuity correction to find the *P*-value for the test in Exercise 27.15. Does the Wilcoxon signed rank test lead to essentially the same result as the $P = 0.3652$ for the *t* test?

27.19 Ancient air. Amber can preserve little samples of the past. Here are the percents of nitrogen found in the gas bubbles of 9 specimens of amber from the late Cretaceous era (75 to 95 million years ago):[11]

63.4 65.0 64.4 63.3 54.8 64.5 60.8 49.1 51.0

We wonder if ancient air differs significantly from the present atmosphere, which is 78.1% nitrogen.

(a) Graph the data and comment on skewness and outliers. A rank test is appropriate.

(b) We would like to test hypotheses about the median percent of nitrogen in ancient air (the population):

$$H_0: \text{median} = 78.1$$
$$H_a: \text{median} \neq 78.1$$

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 78.1. (This is the one-sample version of the test.) What do you conclude?

Dealing with ties in the signed rank test Ties among the absolute differences are handled by assigning average ranks. A tie *within* a pair creates a difference of zero. Because these are neither positive nor negative, we drop such pairs from our sample. Ties within pairs simply reduce the number of observations, but ties among the absolute differences complicate finding a *P*-value. There is no longer a usable exact distribution for the signed rank statistic W^+, and the standard deviation σ_{W^+} must be adjusted for the ties before we can use the Normal approximation. Software will do this. Here is an example.

EXAMPLE 27.9 Treatment for amyloidosis

STATE: Chronic rheumatoid arthritis can lead to secondary amyloidosis, a disabling and potentially fatal disease due to the buildup of amyloid proteins in specific organs. A Spanish study examined the efficacy and safety of anti–tumor necrosis factor in treating secondary amyloidosis affecting the kidneys.

Creatinine is a normal, nonprotein waste product of muscular function and is filtered out by the kidneys. The serum creatinine level is therefore a good marker of kidney function. In healthy individuals it remains fairly constant. Here are the serum creatinine levels (in milligrams per deciliters, mg/dl) before and after treatment of 14 female patients:[12]

Before	2.7	2.4	4.0	2.9	0.9	1.4	1.2	1.5	0.8	0.8	1.1	0.7	2.0	5.9
After	2.2	2.0	4.2	2.3	0.9	1.7	1.4	1.2	1.0	1.0	1.2	0.9	1.5	4.6
Diff.	−0.5	−0.4	0.2	−0.6	0.0	0.3	0.2	−0.3	0.2	0.2	0.1	0.2	−0.5	−1.3

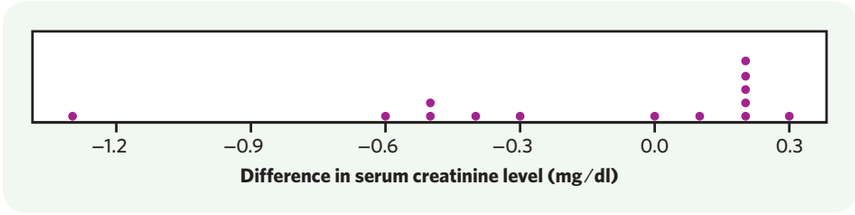
Negative differences represent a decrease in serum creatinine level. Based on this sample, can we conclude that female patients given the anti-tumor necrosis factor will experience a change in their serum creatinine level?

PLAN: We would like to test the hypotheses that in women given the anti–tumor necrosis factor,

H_0: serum creatinine levels are the same before and after treatment

H_a: serum creatinine levels are systematically different after treatment

SOLVE: A dotplot of the differences (Figure 27.6) shows a left-skew with one very low outlier. We will use the Wilcoxon signed rank test.



▶ **FIGURE 27.6**
Dotplot of the differences in serum creatinine level before and after treatment, for Example 27.9.

Minitab

Wilcoxon Signed Rank Test: diff

```
Test of median = 0.000000 versus median not = 0.000000

                  N for    Wilcoxon            Estimated
             N     Test   Statistic       P      Median
diff        14       13        28.5   0.249     -0.1500
```

Paired T-Test and CI: after, before

```
Paired T for after – before

               N      Mean    StDev   SE Mean
after         14     1.864    1.173     0.313
before        14     2.021    1.481     0.396
Difference    14    –0.157    0.460     0.123

T-Test of mean difference = 0 (vs not = 0): T-value = –1.28 P-value = 0.224
```

▲ **FIGURE 27.7** Output from Minitab for the serum creatinine data of Example 27.9. Because there are ties, a Normal approximation must be used for the Wilcoxon signed rank test. The results of a paired *t* test are provided for comparison.

Figure 27 displays the Minitab output for the serum creatinine data. The first query is the Wilcoxon test with its statistic $W^+ = 28.5$ and a two-sided P-value $P = 0.249$. The second query is the output of a matched pairs t test, for which $P = 0.224$. The two P-values are somewhat similar and the practical conclusion is the same. The nonparametric test is more reliable for samples with outliers.

CONCLUDE: These data give no evidence for a systematic change in serum creatinine level following treatment. Kidney function appears stable. ■

Let's see where the value $W^+ = 28.5$ came from. One difference was equal to zero and thus was omitted. The absolute values of the remaining differences, with boldface indicating those that were negative, are

0.5 **0.4** 0.2 **0.6** 0.3 0.2 **0.3** 0.2 0.2 0.1 0.2 **0.5** **1.3**

Arrange these in increasing order and assign ranks, keeping track of which values were originally negative. Tied values receive the average of their ranks.

Abs. value	0.1	0.2	0.2	0.2	0.2	0.2	**0.3**	0.3	**0.4**	**0.5**	**0.5**	**0.6**	**1.3**
Rank	1	4	4	4	4	4	**7.5**	7.5	9	**10.5**	**10.5**	12	13

The Wilcoxon signed rank statistic is the sum $W^+ = 28.5$ of the ranks of the positive differences. (We could equally well use the sum of the ranks of the negative differences.)

APPLY YOUR KNOWLEDGE

27.20 **Does nature heal best?** Table 25.2 (page 644) gives data on the healing rate (micrometers per hour) of the skin of newts under two conditions. This is a matched pairs design, with the body's natural electric field for one limb (control) and half the natural value for another limb of the same newt (experimental). We want to know if the healing rates are systematically different under the two conditions. You decide to use a rank test.

(a) There are several ties among the absolute differences. Find the ranks and give the value of the signed rank statistic W^+.

(b) Use software to find the P-value. Give a conclusion. Be sure to include a description of what the data show in addition to the test results.

27.21 **Sweetening colas.** Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before storage minus sweetness after storage) found by 10 tasters for one new cola recipe:

2.0 0.4 0.7 2.0 −0.4 2.2 −1.3 1.2 1.1 2.3

Are these data good evidence that the cola lost sweetness?

(a) These data are the differences from a matched pairs design. State hypotheses in terms of the median difference in the population of all tasters, carry out a test, and give your conclusion.

(b) Software gives $P = 0.0123$ for the one-sample t test for these data. How does this compare with your result from (a)? What are the hypotheses for the t test? What conditions must be met for the t test? What conditions must be met for the Wilcoxon test?

COMPARING SEVERAL SAMPLES: THE KRUSKAL-WALLIS TEST

We have now considered alternatives to the paired-sample and two-sample t tests for comparing the magnitude of responses to two treatments. To compare mean responses for more than two treatments, we use one-way analysis of variance (ANOVA) if the distributions of the responses to each treatment are at least roughly Normal and have similar spreads. What can we do when these distribution requirements are violated?

EXAMPLE 27.10 Weeds among the corn

STATE: Lamb's-quarter is a common weed that interferes with the growth of corn. A researcher planted corn at the same rate in 16 small plots of ground, then randomly assigned the plots to four groups. He weeded the plots by hand to allow a fixed number of lamb's-quarter plants to grow in each meter of corn row. These numbers were 0, 1, 3, and 9 in the four groups of plots. No other weeds were

allowed to grow, and all plots received identical treatment except for the weeds. Here are the yields of corn (bushels per acre) in each of the plots:[13]

Weeds per meter	Corn yield	Weeds per meter	Corn yield	Weeds per meter	Corn yield	Weeds per meter	Corn yield
0	166.7	1	166.2	3	158.6	9	162.8
0	172.2	1	157.3	3	176.4	9	142.4
0	165.0	1	166.7	3	153.1	9	162.7
0	176.9	1	161.1	3	156.0	9	162.4

Do yields change as the number of weeds changes?

PLAN: Do data analysis to see how the yields change. Test the null hypothesis "no difference in the distribution of yields" against the alternative that the groups do differ.

SOLVE (first steps): The summary statistics are

Weeds	*n*	Median	Mean	Std. dev.
0	4	169.45	170.200	5.422
1	4	163.65	162.825	4.469
3	4	157.30	161.025	10.493
9	4	162.55	157.575	10.118

The mean yields do go down as more weeds are added. ANOVA tests whether the differences are statistically significant. Can we safely use ANOVA? Outliers are present in the yields for 3 and 9 weeds per meter. The outliers explain the differences between the means and the medians. They are the correct yields for their plots, so we cannot remove them. Moreover, the sample standard deviations do not quite satisfy our rule of thumb for ANOVA that the largest should not exceed twice the smallest. We may prefer to use a nonparametric test.

Hypotheses and conditions for the Kruskal-Wallis test The ANOVA *F* test concerns the means of the several populations represented by our samples. For Example 27.10 the ANOVA hypotheses are

$$H_0: \mu_0 = \mu_1 = \mu_3 = \mu_9$$
$$H_a: \text{not all four means are equal}$$

For example, μ_0 is the mean yield in the population of all corn planted under the conditions of the experiment with no weeds present. The data should consist of four independent random samples from the four populations, all Normally distributed with the same standard deviation.

The *Kruskal-Wallis test* is a rank test that can replace the ANOVA *F* test. The condition about data production (independent random samples from each

population) remains important, but we can relax the Normality condition. We assume only that the response has a continuous distribution in each population. The hypotheses tested in our example are

H_0: yields have the same distribution in all groups

H_a: yields are systematically higher in some groups than in others

If all the population distributions have the same shape (Normal or not), these hypotheses take a simpler form. The null hypothesis is that all four populations have the same *median* yield. The alternative hypothesis is that not all four median yields are equal. The different standard deviations suggest that the four distributions in Example 27.10 do *not* all have the same shape.

The Kruskal-Wallis test statistic Recall the analysis of variance idea: We write the total observed variation in the responses as the sum of two parts, one measuring variation among the groups (sum of squares for groups, SSG) and one measuring variation among individual observations within the same group (sum of squares for error, SSE). The ANOVA F test rejects the null hypothesis that the mean responses are equal in all groups if SSG is large relative to SSE.

The idea of the Kruskal-Wallis rank test is to rank all the responses from all groups together and then apply one-way ANOVA to the ranks rather than to the original observations. If there are N observations in all, the ranks are always the whole numbers from 1 to N. The total sum of squares for the ranks is therefore a fixed number no matter what the data are. So we do not need to look at both SSG and SSE. Although it isn't obvious without some unpleasant algebra, the Kruskal-Wallis test statistic is essentially just SSG for the ranks. We give the formula, but you should rely on software to do the arithmetic. When SSG is large, that is evidence that the groups differ.

THE KRUSKAL-WALLIS TEST

Draw independent SRSs of sizes $n_1, n_2, \ldots, n_k$ from k populations. There are N observations in all. Rank all N observations, and let R_i be the sum of the ranks for the ith sample. The **Kruskal-Wallis statistic** is

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

When the sample sizes n_i are large and all k populations have the same continuous distribution, H has approximately the chi-square distribution with $k - 1$ degrees of freedom.

The **Kruskal-Wallis test** rejects the null hypothesis that all populations have the same distribution when H is large.

We now see that, like the Wilcoxon rank sum statistic, the Kruskal-Wallis statistic is based on the sums of the ranks for the groups we are comparing. The more different these sums are, the stronger is the evidence that responses are systematically larger in some groups than in others.

The exact distribution of the Kruskal-Wallis statistic H under the null hypothesis depends on all the sample sizes n_1 to n_k, so tables are awkward. The calculation of the exact distribution is so time-consuming for all but the smallest problems that even most statistical software uses the chi-square approximation to obtain P-values. As usual, there is no usable exact distribution when there are ties among the responses. We again assign average ranks to tied observations.

EXAMPLE 27.11 Weeds among the corn, continued

4 STEP

SOLVE (inference): In Example 27.10 there are $k = 4$ populations and $N = 16$ observations. The sample sizes are equal, $n_i = 4$. The 16 observations arranged in increasing order, with their ranks, are

Yield	142.4	153.1	156.0	157.3	158.6	161.1	162.4	162.7
Rank	1	2	3	4	5	6	7	8
Yield	162.8	165.0	166.2	166.7	166.7	172.2	176.4	176.9
Rank	9	10	11	12.5	12.5	14	15	16

There is one pair of tied observations. The ranks for each of the four treatments are

Weeds	Ranks				Sum of ranks
0	10	12.5	14	16	52.5
1	4	6	11	12.5	33.5
3	2	3	5	15	25.0
9	1	7	8	9	25.0

The Kruskal-Wallis statistic is therefore

$$\begin{aligned} H &= \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \\ &= \frac{12}{(16)(17)} \left(\frac{52.5^2}{4} + \frac{33.5^2}{4} + \frac{25^2}{4} + \frac{25^2}{4} \right) - (3)(17) \\ &= \frac{12}{272}(1282.125) - 51 \\ &= 5.56 \end{aligned}$$

Referring to the table of chi-square critical points (Table D) with df = 3, we see that the P-value lies in the interval $0.10 < P < 0.15$.

CONCLUDE: This small experiment does not provide convincing evidence that weeds decrease corn yield.

Figure 27.8 displays the Minitab output for both the Kruskal-Wallis test and the ANOVA test. Minitab finds that $H = 5.56$ and gives $P = 0.135$. Minitab also gives the results of an adjustment that makes the chi-square approximation more accurate when there are ties. For these data, the adjustment has no

Minitab

Kruskal-Wallis Test: Yield versus Weeds

```
Kruskal-Wallis Test on Yield

Weeds       N     Median     Ave Rank         Z
0           4      169.5         13.1      2.24
1           4      163.6          8.4     -0.06
3           4      157.3          6.3     -1.09
9           4      162.6          6.3     -1.09
Overall    16                     8.5

H = 5.56  DF = 3  P = 0.135
H = 5.57  DF = 3  P = 0.134 (adjusted for ties)

* NOTE * One or more small samples
```

One-way ANOVA: Yield versus Weeds

```
Analysis of Variance for Yield
Source     DF         SS        MS        F        P
Weeds       3      340.7     113.6     1.73    0.213
Error      12      785.5      65.5
Total      15     1126.2
                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level       N       Mean    StDev ---+--------+--------+--------+----
0           4     170.20     5.42              (-------*-------)
1           4     162.82     4.47       (-------*-------)
3           4     161.03    10.49     (-------*-------)
9           4     157.57    10.12    (-------*-------)
                                   ---+--------+--------+--------+----
Pooled StDev =     8.09             150       160       170       180
```

◀ FIGURE 27.8
Minitab output for the corn yield data of Example 27.10. For comparison, both the Kruskal-Wallis test and the one-way ANOVA are shown.

practical effect. It would be important if there were many ties. A very lengthy computer calculation shows that the exact P-value is $P = 0.1299$. The chi-square approximation is quite accurate.

The ANOVA F test gives $F = 1.73$ with $P = 0.213$. Although the practical conclusion is the same, ANOVA and Kruskal-Wallis do not agree closely in this example. The rank test is more reliable for these small samples with outliers.

APPLY YOUR KNOWLEDGE

27.22 Which color attracts beetles best? To detect the presence of harmful insects in farm fields, we can put up boards covered with a sticky material and examine the insects trapped on the boards. Experimenters placed six boards of each of four colors at random locations in a field of oats and measured the number of cereal leaf beetles trapped. Here are the data:[14]

Color	Beetles trapped					
Blue	16	11	20	21	14	7
Green	37	32	20	29	37	32
White	21	12	14	17	13	20
Yellow	45	59	48	46	38	47

The samples are small. If you have no reasons to believe that the data are Normally distributed (say, from similar studies with large sample sizes), it would be safer to apply a nonparametric test.

(a) Find the median number of beetles trapped by boards of each color. Which colors appear more effective?

(b) What hypotheses does ANOVA test? What hypotheses does Kruskal-Wallis test?

(c) What are k, the n_i, and N? Arrange the counts in order and assign ranks. Be careful about ties.

(d) Calculate the Kruskal-Wallis statistic H. How many degrees of freedom should you use for the chi-square approximation to its null distribution? Use the chi-square table to give an approximate P-value. What does the test lead you to conclude?

27.23 Logging in the rain forest: species richness. Table 24.2 (page 616) contains data comparing the number of trees and number of tree species in plots of land in a tropical rain forest that had never been logged with similar plots nearby that had been logged 1 year earlier and 8 years earlier. The third response variable is species richness, the number of tree species divided by the number of trees. There are low outliers in the data, and a histogram of the ANOVA residuals shows outliers as well. Because of lack of Normality and small samples, we may prefer the Kruskal-Wallis test.

(a) Make a graph to compare the distributions of richness for the three groups of plots. Also give the median richness for the three groups.

(b) Use the Kruskal-Wallis test to compare the distributions of richness. State hypotheses, the test statistic and its P-value, and your conclusions.

27.24 More rain for California? Exercise 24.29 (page 629) describes an experiment that examines the effect on plant biomass in plots of California grassland randomly assigned to receive added water in the winter, added water in the spring, or no added water. The experiment continued for several years. Here are data for 2004 (mass in grams per square meter):

Winter	Spring	Control
254.6453	517.6650	178.9988
233.8155	342.2825	205.5165
253.4506	270.5785	242.6795
228.5882	212.5324	231.7639
158.6675	213.9879	134.9847
212.3232	240.1927	212.4862

The sample sizes are small and the data contain some possible outliers. We will apply a nonparametric test.

(a) Examine the data. Show that the conditions for ANOVA (page 618) are not met. What appear to be the effects of extra water in winter or spring?

(b) What hypotheses does ANOVA test? What hypotheses does Kruskal-Wallis test?

(c) What are k, the n_i, and N? Arrange the data in order and assign ranks.

(d) Calculate the Kruskal-Wallis statistic H. How many degrees of freedom should you use for the chi-square approximation to its null distribution? Use the chi-square table to give an approximate P-value. What does the test lead you to conclude?

CHAPTER 27 SUMMARY

- **Nonparametric tests** do not require any specific form for the distributions of the populations from which our samples come.
- **Rank tests** are nonparametric tests based on the **ranks** of observations, their positions in the list ordered from smallest (rank 1) to largest. Tied observations receive the average of their ranks. Use rank tests when the data come from random samples or randomized comparative experiments and the populations have continuous distributions.
- The **Wilcoxon rank sum test** compares two distributions to assess whether one has systematically larger values than the other. The Wilcoxon test is based on the **Wilcoxon rank sum statistic *W*,** which is the sum of the ranks of one of the samples. The Wilcoxon test can replace the **two-sample *t* test.** Software may perform the **Mann-Whitney test,** another form of the Wilcoxon test.
- ***P*-values** for the rank sum test are based on the sampling distribution of the rank sum statistic W when the null hypothesis (no difference in distributions) is true. You can find P-values from special tables, software, or a Normal approximation (with continuity correction).
- The **Wilcoxon signed rank test** applies to matched pairs studies. It tests the null hypothesis that there is no systematic difference within pairs against alternatives that assert a systematic difference (either one-sided or two-sided).
- The test is based on the **Wilcoxon signed rank statistic W^+,** which is the sum of the ranks of the positive (or negative) differences when we rank the absolute values of the differences. The **matched pairs *t* test** is an alternative test in this setting.
- ***P*-values** for the signed rank test are based on the sampling distribution of W^+ when the null hypothesis is true. You can find P-values from special tables, software, or a Normal approximation (with continuity correction).
- The **Kruskal-Wallis test** compares several populations on the basis of independent random samples from each population. This is the **one-way analysis of variance** setting.
- The null hypothesis for the Kruskal-Wallis test is that the distribution of the response variable is the same in all the populations. The alternative hypothesis is that responses are systematically larger in some populations than in others.
- The **Kruskal-Wallis statistic *H*** can be viewed in two ways. It is essentially the result of applying one-way ANOVA to the ranks of the observations. It is also a comparison of the sums of the ranks for the several samples.
- When the sample sizes are not too small and the null hypothesis is true, the Kruskal-Wallis test statistic for comparing k populations has approximately the chi-square distribution with $k - 1$ degrees of freedom. We use this approximate distribution to obtain P-values.

STATISTICS IN SUMMARY

Here are the most important skills you should have acquired from reading this chapter.

A. RANKS

1. Assign ranks to a moderate number of observations. Use average ranks if there are ties among the observations.
2. From the ranks, calculate the rank sums when the observations come from two or several samples.

B. RANK TEST STATISTICS

1. Determine which of the rank sum tests is appropriate in a specific problem setting.
2. Calculate the Wilcoxon rank sum statistic W from ranks for two samples, the Wilcoxon signed rank sum statistic W^+ for matched pairs, and the Kruskal-Wallis statistic H for two or more samples.
3. State the hypotheses tested by each of these statistics in specific problem settings.
4. Determine when it is appropriate to state the hypotheses for W and H in terms of population medians.

C. RANK TESTS

1. Use software to carry out any of the rank tests. Combine the test with data description and give a clear statement of findings in specific problem settings.
2. Use the Normal approximation with continuity correction to find approximate P-values for W and W^+. Use a table of chi-square critical values to approximate the P-value of H.

THIS CHAPTER IN CONTEXT

In Chapter 17 we described the one-sample t procedures used to perform statistical inference for one population mean μ. This was followed in Chapter 18 by the two-sample t procedures used to perform statistical inference for the difference $\mu_1 - \mu_2$ between the means of two distinct populations. In Chapter 24 we described the one-way analysis of variance F test, a method used to compare the means $\mu_1, \mu_2, \ldots, \mu_k$ of several populations while avoiding the problems associated with multiple comparisons.

The t procedures and ANOVA are used when the response variable is quantitative, and they assume that the underlying populations are Normally distributed. So, checking the conditions for inference typically requires that we examine the data for evidence of skew or outliers, as seen in Chapters 1 and 2. The t procedures and ANOVA, fortunately, are robust with respect to this assumption of Normality when the sample sizes are large enough.

In this chapter we examine nonparametric inference procedures equivalent to the one-sample and matched pairs t tests, the two-sample t test, and the ANOVA F test. Respectively, these are the Wilcoxon signed rank test, the Wilcoxon rank sum test (also called the Mann-Whitney test), and the Kruskal-Wallis test. These procedures rely on the ranking of the data rather than the actual numerical values. Normality of the response variable is not required. Therefore, these nonparametric procedures are great alternatives when we cannot be sure that the response variable is truly continuous or Normally distributed, and the samples sizes are too small to rely on the robustness of the t and F methods.

CHECK YOUR SKILLS

27.25 A study of blood pressure and age compares the blood pressures of men in three age groups: less than 30 years, 30 to 55 years, and over 55 years. To analyze the data you would use the

(a) Wilcoxon rank sum test.

(b) Wilcoxon signed rank test.

(c) Kruskal-Wallis test.

27.26 The previous study also compares the blood pressures of men to those of women for the 30 to 55 age group. To analyze the data you would use the

(a) Wilcoxon rank sum test.

(b) Wilcoxon signed rank test.

(c) Kruskal-Wallis test.

27.27 You interview 75 students in their freshman year and again in their senior year. Each interview includes information about height and weight, allowing the computation of each student's body mass index (BMI). To assess whether there has been a significant change from freshman to senior year in their BMIs, you would use the

(a) Wilcoxon rank sum test.
(b) Wilcoxon signed rank test.
(c) Kruskal-Wallis test.

27.28 When some plants are attacked by leaf-eating insects, they release chemical compounds that repel the insects. Here are data on emissions of one compound by 6 plants attacked by leaf bugs and by 6 plants in an undamaged control group:

Control group	14.4	15.2	12.6	11.9	5.1	8.0
Attacked group	10.6	15.3	25.2	19.8	17.1	14.6

The rank sum W for the control group is
(a) 21. (b) 26. (c) 52.

27.29 If there is no difference in emissions between the attacked group and the control group, the mean of W in the previous exercise is
(a) 39. (b) 78. (c) 6.2.

27.30 Suppose that the 12 observations in Exercise 27.28 were

Control group	14.4	15.2	12.6	11.9	5.1	8.0
Attacked group	12.6	15.3	25.2	19.8	17.1	14.4

The rank sum for the control group would then be
(a) 21. (b) 25. (c) 26.

27.31 Interview 10 young married couples, wife and husband separately. One question asks how important the attractiveness of their spouse is to them on a scale of 1 to 10. Here are the responses:

	Couple									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Husband	7	7	7	3	9	5	10	6	6	7
Wife	4	2	5	2	2	2	4	7	1	5

The Wilcoxon signed rank statistic W^+ (based on husband's score minus wife's score) is
(a) 51. (b) 53.5. (c) 54.

27.32 If husbands and wives don't differ in how important the attractiveness of their spouse is, the mean of W^+ in the previous exercise is
(a) 27.5. (b) 55. (c) 105.

27.33 Suppose that the responses in Exercise 27.31 were

	Couple									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Husband	7	7	7	3	9	5	10	6	6	5
Wife	4	2	5	3	2	2	4	7	1	5

The Wilcoxon signed rank statistic W^+ (based on husband's score minus wife's score) is now
(a) 35. (b) 36. (c) 52.

27.34 You compare the LDL ("bad") cholesterol levels of 4 college freshmen, 5 sophomores, 6 juniors, and 7 seniors. If the four LDL cholesterol distributions are the same, the Kruskal-Wallis statistic H has approximately a chi-square distribution. The degrees of freedom are
(a) 3. (b) 4. (c) 18.

CHAPTER 27 EXERCISES

One of the rank tests discussed in this chapter is appropriate for each of the following exercises. Follow the **Plan, Solve,** *and* **Conclude** *parts of the four-step process in your answers. It may be helpful to restate in your own words the* **State** *information given in the exercise.*

27.35 Effectiveness of subliminal messages. A "subliminal" message is below our threshold of awareness but may nonetheless influence us. Can subliminal messages help students learn math? Table 27.2 gives the pretest and posttest scores for two groups of students taking a program to improve their basic mathematics skills.[15] In addition to the math program, all students received a daily subliminal message, flashed on a screen too rapidly to be consciously read. The treatment group of 10 students (chosen at random) was exposed to "Each day I am getting better in math." The control group of 8 students was exposed to a neutral message, "People are walking on the street." Is there evidence of significant improvement within each group? Did the treatment group show significantly greater improvement than the control group?

TABLE 27.2 Mathematics skills scores before and after a subliminal message

Treatment group		Control group	
Before	After	Before	After
18	24	18	29
18	25	24	29
21	33	20	24
18	29	18	26
18	33	24	38
20	36	22	27
23	34	15	22
23	36	19	31
21	34		
17	27		

27.36 Diet and reproduction in fruit flies. Exercise 27.5 introduced an experiment about fruit fly reproduction. The researchers also studied how diet composition affects their reproductive output. Wild-type fruit flies were assigned to receive varying amounts of yeast supplements (in milligrams, mg, per day), a protein-rich food, in their diet for two weeks, after which the number of eggs produced per female was assessed. Here are the results:

0 mg	7.55	6.10	9.25	5.40	3.75
1 mg	10.75	18.15	18.70	16.95	11.15
3 mg	18.35	39.25	44.60	42.80	24.20
7 mg	40.35	67.95	48.50	64.25	60.35

Does reproductive output differ based on diet richness?

27.37 Fighting cancer. Lymphocytes (white blood cells) play an important role in defending our bodies against tumors and infections. Can lymphocytes be genetically modified to recognize and destroy cancer cells? In one study of this idea, modified cells were infused into 11 patients with metastatic melanoma (serious skin cancer) that had not responded to existing treatments. Here are data for an "ELISA" test for the presence of cells that trigger an immune response, in counts per 100,000 cells before and after infusion.[16] High counts suggest that infusion had a beneficial effect.

Patient	1	2	3	4	5	6	7	8	9	10	11
Before	14	0	1	0	0	0	0	20	1	6	0
After	41	7	1	215	20	700	13	530	35	92	108

What do you conclude about the effect of infusing modified cells on the ELISA count?

27.38 Canine compulsive disorder. Obsessive compulsive disorder (OCD) is a human anxiety disorder with a poorly understood origin. Canine compulsive disorder (CCD) is an analogous behavioral disorder affecting dogs that may provide an animal model for OCD. Using magnetic resonance imaging, researchers compared the brain structures of Doberman dogs with and without CCD. Here are standardized fractional anisotropy values in the corpus callosum, indicative of fiber density in this white-matter structure connecting the cerebral hemispheres:[17]

Control	0.47	0.46	0.38	0.27	0.26	0.27	0.24	0.19
CCD	0.73	0.62	0.61	0.49	0.45	0.37	0.3	0.32

Is there significant evidence that fractional anisotropy values in the corpus callosum are systematically larger in dogs with CCD or in those without CCD?

27.39 Nematodes and plant growth. A botanist prepares 16 identical planting pots and then introduces different numbers of nematodes (microscopic worms) into the pots. A tomato seedling is transplanted into each pot. Here are data on the increase in height of the seedlings (in centimeters) 16 days after planting:[18]

Nematodes	Seedling growth			
0	10.8	9.1	13.5	9.2
1,000	11.1	11.1	8.2	11.3
5,000	5.4	4.6	7.4	5.0
10,000	5.8	5.3	3.2	7.5

Do nematodes in soil affect plant growth?

27.40 Evolution in bacteria. Can bacteria evolve a preference for the pH of their environment? An evolutionary biologist took lines of *E. coli* bacteria grown and kept at neutral pH 7.2 and grew them for 2000 generations (about 300 days) at a stressful acidic pH of 5.5. The ancestor bacteria (kept frozen all that time) and the "acid-evolved" bacteria were then grown together at pH 5.5 to compute a relative fitness score. (A score of 1 indicates equal growth, a larger score indicates that the evolved bacteria grow better than their ancestors.) A control group consisting of some of the original bacteria was grown for 2000 generations in a neutral pH, and a relative fitness score (relative to the ancestor line) was later obtained at pH 5.5. Here are the scores for the acid-evolved and the control treatments from 6 replications of the experiment:[19]

Acid-evolved	1.24	1.22	1.23	1.24	1.18	1.09
Control	1.10	1.02	0.99	1.04	1.08	1.12

Do the data provide evidence that *E. coli* can evolve adaptations to an acidic pH over the course of 2000 generations? That is, is there evidence that the acid-evolved bacteria have higher relative fitness scores than control bacteria grown at a neutral pH?

27.41 A treatment for progeria. In Example 17.4 (page 431) we examined the effect of the drug lonafarnib on pulse wave velocity (PWV) in children diagnosed with progeria, a rare genetic condition that produces rapid aging in children. PWV is the standard measure of vascular stiffness, an important factor in cardiovascular health. Table 17.1 (page 431) gives the PWV values (in meters per second, m/s) at the beginning ("Untreated") and at the end ("Treated") of the two-year study for each child. Use a nonparametric method to determine whether PWV in children with progeria is significantly lower after treatment with lonafanib.

27.42 Vaccine protection against SIV. A research team investigated the protective effect of 4 vaccine variants against the simian immunodeficiency virus (SIV) in rhesus monkeys. V2-specific antibodies provide an indirect assessment of protection against SIV infection. Here are V2-specific antibody productions measured by surface plasmon resonance (in response units) for monkeys assigned to a vaccine variant or to a sham (fake) vaccine:[20]

Vaccine 1	Vaccine 2	Vaccine 3	Vaccine 4	Sham
100.1	83.5	171.6	196.1	11.4
53.9	73.0	118.9	154.3	4.8
25.3	32.3	111.1	119.4	6.3
17.9	14.5	103.4	120.2	4.0
22.1	15.3	93.3	77.9	0.1
14.0	1.3	88.6	65.5	1.7
4.3	0.6	77.0	74.0	1.3
10.9	0.6	60.0	68.6	1.3

(a) Examine the 5 samples. Are the data appropriate for ANOVA?

(b) Use the Kruskal-Wallis test to determine whether the data provide evidence that V2-specific antibody production is systematically larger with some treatments than with others.

27.43 Cicadas as fertilizer? Every 17 years, swarms of cicadas emerge from the ground in the eastern United States, live for about six weeks, then die. There are so many cicadas that their dead bodies can serve as fertilizer. In an experiment, a researcher added cicadas under some plants in a natural plot of bellflowers on the forest floor, leaving other plants undisturbed. Here are the seed masses (in milligrams) produced by 39 cicada-fertilized plants and by 33 undisturbed (control) plants:[21]

Cicada plants				Control plants			
0.237	0.277	0.241	0.142	0.212	0.188	0.263	0.253
0.109	0.209	0.238	0.277	0.261	0.265	0.135	0.170
0.261	0.227	0.171	0.235	0.203	0.241	0.257	0.155
0.276	0.234	0.255	0.296	0.215	0.285	0.198	0.266
0.239	0.266	0.296	0.217	0.178	0.244	0.190	0.212
0.238	0.210	0.295	0.193	0.290	0.253	0.249	0.253
0.218	0.263	0.305	0.257	0.268	0.190	0.196	0.220
0.351	0.245	0.226	0.276	0.246	0.145	0.247	0.140
0.317	0.310	0.223	0.229	0.241			
0.192	0.201	0.211					

Do the data show that dead cicadas increase seed mass?

27.44 More on visual receptive fields. Example 27.6 describes a neuron in the visual cortex stimulated by the presentation of a visual pattern on a screen. By using 3D glasses that introduce a visual disparity between the two retinas, that pattern can be made to appear as if it is floating in front of or behind the screen. Some neurons in the primary visual cortex specialize in the recognition of such 3D clues. Here is the response of the neuron from Example 27.6 to a visual pattern with varying amounts of disparity presented in random order:

Disparity	Neural response (in spikes per second)								
−0.30	13.3	3.3	46.7	0.0	13.3	60.0	0.0	53.3	6.7
−0.15	13.3	6.7	6.7	0.0	6.7	6.7	10.0	13.3	0.0
0.00	16.7	20.0	23.3	16.7	56.7	0.0	26.7	36.7	10.0
0.15	86.7	3.3	46.7	86.7	50.0	56.7	80.0	73.3	106.7
0.30	53.3	23.3	0.0	13.3	0.0	23.3	0.0	0.0	13.3
0.60	23.3	40.0	20.0	0.0	23.3	6.7	6.7	23.3	6.7

Neurons specializing in 3D perception from retinal disparity respond differentially to patterns presented with different amounts of disparity. Is this such a neuron? What does it respond to best? (A positive disparity corresponds to the impression of a far object.)

27.45 Smoking during pregnancy. Cigarette labels warn pregnant women against smoking. Does nicotine actually reach the fetus, crossing the protective placental barrier? Researchers selected consecutive pregnant women delivering at an Egyptian hospital and categorized them as active smokers, passive smokers, or nonsmokers. They then analyzed the newborns' meconium for cotinine content, the metabolized form of nicotine. Meconium is a newborn's first stool right after birth and is a good biological marker for fetal exposure to drugs or other chemical agents. Here are the meconium cotinine levels (in nanograms per milliliter, ng/ml) for the 3 groups:[22]

Active smokers	490	418	405	328	700	292	295	272	240	232
Passive smokers	254	219	287	257	271	282	148	273	350	293
Nonsmokers	158	163	153	207	211	159	199	187	200	213

(a) Examine the 3 samples. What are the overall shapes of the distributions? Are there outliers? What are the sample standard deviations? Explain why ANOVA cannot be used safely on these data.

(b) Use a nonparametric method to determine whether meconium cotinine levels systematically change with the mother's smoking status.

27.46 Nanomedicine. Researchers examined a new treatment for advanced ovarian cancer in a mouse model. They created a nanoparticle-based delivery system for a suicide gene therapy to be delivered directly to the tumor cells. The grafted tumors were injected either with the new treatment or with only some buffer solution to serve as a comparison. The data below give the increase in size of the tumor after two weeks in 20 mice. A 1 represents no change, and a 2 represents a doubling in volume of the tumor.[23]

Buffer solution									
9.1	8.1	7.8	7.0	6.8	5.4	5.4	4.1	3.8	3.3
Nanoparticle-delivered gene therapy									
4.1	3.5	2.1	2.1	1.8	1.8	1.4	1.2	1.1	1.1

Do the data indicate that the nanoparticle-based delivery system is more effective than a placebo in slowing ovarian tumor growth in mice?

27.47 Ink toxicity. The National Toxicology Program evaluates the toxicity of chemicals found in manufacturing, in consumer products, or in the environment after disposal. Toxicity is assessed through a battery of tests. Here are some results from a study of the toxicity of black newsprint ink in 7-week-old female rats. The rats' fur was locally clipped twice a week for 13 weeks. One group of rats received a dermal application of ink right after each clipping, and a control group of rats was left untreated. Here are the body weights (in grams) of female rats at the beginning of the study and at the end of the 13 weeks:[24]

Control group		Treatment group	
Week 0	Week 13	Week 0	Week 13
111.2	191.6	107.3	187.0
105.4	191.2	116.7	189.5
110.8	210.7	112.2	179.2
105.6	185.2	103.4	172.2
106.1	195.0	113.2	178.7
104.4	188.3	110.6	180.9
114.0	188.4	110.6	188.3
115.1	195.6	100.5	188.9
109.2	204.6	106.3	183.1
111.3	195.7	112.5	184.5

First, verify that the two experimental groups are not significantly different at the beginning of the study. Then examine the data for evidence that ink application impairs growth in female rats between 7 and 20 weeks of age. That is, do rats treated with ink gain less weight than control rats overall?

27.48 Food safety at restaurants. Example 27.5 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. The full data file is available on the companion website as the file *ex27-48.dat*. The variable "srest" contains responses to the same question asked about food served in restaurants. The variable "gender" contains F if the respondent is a woman, M if he is a man. We saw that women are more concerned than men about the safety of food served at fairs. Is this also true for restaurants?

How does the meeting of large rivers influence the diversity of fish? A study of the Amazon and 13 of its major tributaries concentrated on electric fish, which are common in South America. The researchers trawled in more than 1000 locations in the Amazon above and below where each tributary enters the Amazon and in the lower parts of the tributaries themselves. In all, they found 43 species of electric fish. These distinctive fish can "stand in" for fish in general, which are too numerous to count easily. The researchers concluded that the number of fish species increases when a tributary joins the Amazon but that the effect is local: There is no steady increase in diversity as we move downstream. Table 27.3 gives the estimated numbers of electric fish species in the Amazon upstream and downstream from each tributary and in the tributaries themselves just before they flow into the Amazon.[25] *The researchers used nonparametric tests to assess the statistical significance of their results. Exercises 27.49 to 27.51 quote conclusions from the study.*

27.49 Downstream versus upstream. "We identified a significant positive effect of tributaries on Amazon mainstream species richness in two respects. First, we found that sample stations downstream of each tributary contained more species than did their respective upstream stations." Do a test to confirm the statistical significance of this effect and report your conclusion.

TABLE 27.3 Electric fish species in the Amazon

	Species counts		
Tributary	Upstream	Tributary	Downstream
Iça	14	23	19
Jutaí	11	15	18
Juruá	8	13	8
Japurá	9	16	11
Coari	5	7	7
Purus	10	23	16
Manacapuru	5	8	6
Negro	23	26	24
Madeira	29	24	30
Trombetas	19	20	16
Tapajós	16	5	20
Xingu	25	24	21
Tocantins	10	12	12

27.50 Tributary versus upstream. "Second, we found that species richness within tributaries exceeded that within their adjacent upstream mainstream stations." Again, do a test to confirm significance and report your finding.

27.51 Tributary versus downstream. Species richness "was comparable between tributaries and their adjacent downstream mainstream stations." Verify this conclusion by comparing tributary and downstream species counts.

NOTES AND DATA SOURCES

1. Data provided by Samuel Phillips, Purdue University.
2. G. Yeretssian et al., "Is BID required for NOD signalling?" *Nature*, 474 (2011), pp. 96–99, doi:10.1038/nature11367.
3. F. H. Simmons, "Physiology of the trade-off between fecundity and survival in *Drosophila melanogaster*, as revealed through dietary manipulation," MS thesis, University of California at Irvine, 1996.
4. The precise meaning of "yields are systematically larger in plots with no weeds" is that for every fixed value a, the probability that the yield with no weeds is larger than a is at least as great as the same probability for the yield with weeds.

5. H. Chern Boo, "Consumers' perceptions and concerns about safety and healthfulness of food served at fairs and festivals," MS thesis, Purdue University, 1997.
6. From a graph in F. Grieco, A. J. van Noordwijk, and M. E. Visser, "Evidence for the effect of learning on timing of reproduction in blue tits," *Science*, 296 (2002), pp. 136–138.
7. We thank David LeBauer of the University of California at Irvine for providing the data.
8. N. Guéeguen and C. Petr, "Odors and consumer behavior in a restaurant," *Journal of Hospitality Management*, 25 (2006), pp. 335–339. We thank Nicolas Guéguen for providing the data.
9. B. Stricanne, "Etudes d'intégration multisensorielles dans la voie visuelle occipito-pariétale du primate," PhD thesis, Université Paris VI, 1996.
10. R. Sakakibara et al., "Influence of body position on defecation in humans," *Lower Urinary Tract Symptoms*, 2 (2010), pp. 16–21, doi:10.1111/j.1757-5672.2009.00057.x.
11. R. A. Berner and G. P. Landis, "Gas bubbles in fossil amber as possible indicators of the major gas composition of ancient air," *Science*, 239 (1988), pp. 1406–1409.
12. A. Fernández-Nebro et al., "Treatment of rheumatic inflammatory disease in 25 patients with secondary amyloidosis using tumor necrosis factor alpha antagonists," *American Journal of Medicine*, 118 (2005), pp. 552–556.
13. See Note 1.
14. Modified from M. C. Wilson and R. E. Shade, "Relative attractiveness of various Luminescent colors to the cereal leaf beetle and the meadow spittlebug," *Journal of Economic Entomology*, 60 (1967), pp. 578–580.
15. Data provided by Warren Page, New York City Technical College, from a study done by John Hudesman.
16. R. A. Morgan et al., "Cancer regression in patients after transfer of genetically engineered lymphocytes," *Science*, 314 (2006), pp. 126–129. The data appear in the Online Supplementary Material.
17. N. Ogata et al., "Brain structural abnormalities in Doberman pinschers with canine compulsive disorder," *Progress in Neuro-psychopharmacology and Biological Psychiatry*, 45 (2013), pp. 1–6, doi:10.1016/j.pnpbp.2013.04.002.
18. Data provided by Matthew Moore.
19. We thank Brad Hughes of the University of California at Irvine for providing the data.
20. D. H. Barouch et al., "Vaccine protection against acquisition of neutralization-resistant SIV challenges in rhesus monkeys," *Nature*, 482 (2012), pp. 89–93, doi:10.1038/nature10766.
21. L. H. Yang, "Periodical cicadas as resource pulses in North American forests," *Science*, 306 (2004), pp. 1565–1567. The data are simulated Normal values that match the means and standard deviations reported in this article.
22. N. A. Sherif et al., "Detection of cotinine in neonate meconium as a marker for nicotine exposure in utero," *Eastern Mediterranean Health Journal*, 10 (2004), pp. 96–105.
23. Y.-H. Huang et al., "Nanoparticle-delivered suicide gene therapy effectively reduces ovarian tumor burden in mice," *Cancer Research*, 69 (2009), pp. 6184–6191.
24. *Black Newsprint Inks*, National Toxicology Program Toxicity Report Series No. 17, `ntp.niehs.nih.gov`.
25. C. Cox Fernandes, J. Podos, and J. G. Lundberg, "Amazonian ecology: tributaries enhance the diversity of electric fishes," *Science*, 305 (2004), pp. 1960–1962.