CHAPTER 14

Nonparametric Tests

Introduction

The most commonly used methods for 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 for 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. Some practical guidelines for taking advantage of the robustness of these methods appear in Chapter 7.

robustness

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:

outliers

1. If there are extreme **outliers** in a small data set, any inference method may be suspect. An outlier is an observation that may not come from the same population as the others. To decide what to do, you must find the cause of the outlier. Equipment failure that produced a bad measurement, for example, entitles you to remove the outlier and analyze the remaining data. If the outlier appears to be "real data," it is risky to draw any conclusion from just a few observations. This is the advice we gave to the child development researcher in Example 2.19 (page 163).

transforming data

2. Sometimes we can **transform** our data so that their distribution is more nearly normal. Transformations such as the logarithm that pull in the long tail of right-skewed distributions are particularly helpful. We discussed transformations in detail in Section 2.6.

other standard distributions

3. In some settings, **other standard distributions** replace the normal distributions as models for the overall pattern in the population. We mentioned in Section 5.2 (page 400) that the Weibull distributions are common models for the lifetimes in service of equipment in statistical studies of reliability. There are inference procedures for the parameters of these distributions that replace the *t* procedures when we use specific nonnormal models.

nonparametric methods

4. Finally, there are inference procedures that do not require any specific form for the distribution of the population. These are called **nonparametric methods.** The *sign test* (page 509) is an example of a nonparametric test.

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 not met. 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 14.1 presents an outline of the standard tests (based on normal distributions) and the rank tests that compete with them. All of these tests

Setting	Normal test	Rank test
One sample	One-sample *t* test	Wilcoxon signed rank test
	Section 7.1	Section 14.2
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample *t* test	Wilcoxon rank sum test
	Section 7.2	Section 14.1
Several independent samples	One-way ANOVA *F* test	Kruskal-Wallis test
	Chapter 12	Section 14.3

FIGURE 14.1 Comparison of tests based on normal distributions with nonparametric tests for similar settings.

continuous distribution

require that the population or populations have **continuous distributions.** That is, each distribution must be described by a density curve that allows observations to take any value in 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 14.1 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 simplest form, the hypotheses for rank tests just replace mean by median.

We devote a section of this chapter to each of the rank procedures. Section 14.1, which discusses the most common of these tests, also contains general information about rank tests. The kind of assumptions required, the nature of the hypotheses tested, the big idea of using ranks, and the contrast between exact distributions for use with small samples and approximations for use with larger samples are common to all rank tests. Sections 14.2 and 14.3 more briefly describe other rank tests.

14.1 The Wilcoxon Rank Sum Test

Two-sample problems (see Section 7.2) are among the most common in statistics. The most useful nonparametric significance test compares two distributions. Here is an example of this setting.

EXAMPLE 14.1

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]

Weeds per meter	Yield (bu/acre)			
0	166.7	172.2	165.0	176.9
3	158.6	176.4	153.1	156.0

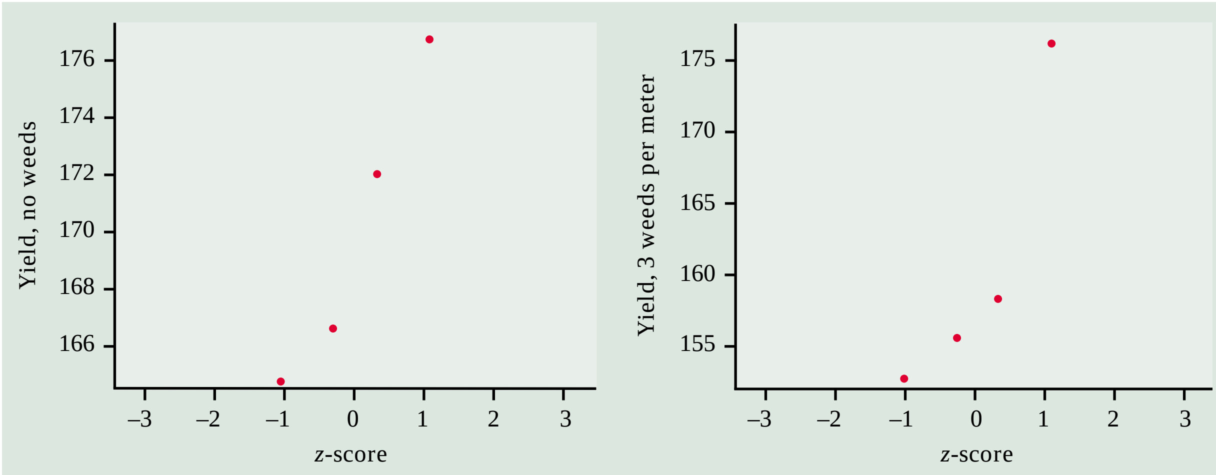

FIGURE 14.2 Normal quantile plots of corn yields from plots with no weeds (left) and with 3 weeds per meter of row (right).

Normal quantile plots (Figure 14.2) suggest that the data may be right-skewed. The samples are too small to assess normality adequately or to rely on the robustness of the two-sample *t* test. We may prefer to use a test that does not require normality.

The rank transformation

We 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 ordered list. To do this, replace each observation by 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 is a transformation of the data, like moving from the observations to their logarithms. The rank transformation 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 assumptions about the shape of the distribution, such as normality.

The Wilcoxon rank sum test

If the presence of weeds reduces corn yields, we expect the ranks of the yields from plots with weeds to be smaller as a group than the ranks from plots without weeds. We might 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 always 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 have no effect, we would expect the sum of the ranks in either 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.*

*This test was invented by Frank Wilcoxon (1892–1965) in 1945. Wilcoxon was a chemist who met statistical problems in his work at the research laboratories of American Cyanimid Company.

In the corn yield study of Example 14.1, we want to test

H_0: no difference in distribution of yields

against the one-sided alternative

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 14.2

In Example 14.1, $n_1 = 4$, $n_2 = 4$, and there are $N = 8$ observations in all. 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.

To calculate the P-value $P(W \geq 23)$, 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 are therefore a bit 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, some software gives only approximate P-values. You must learn what your software offers.

EXAMPLE 14.3

Figure 14.3 shows the output from software that calculates the exact sampling distribution of W. We see that the sum of the ranks in the weed-free group is $W = 23$, with P-value $P = 0.10$ against the one-sided alternative that weed-free plots have higher yields. There is some evidence that weeds reduce yield, considering that we have data from only four plots for each treatment. The evidence does not, however, reach the levels usually considered convincing.

It is worth noting that the two-sample t test gives essentially the same result as the Wilcoxon test in Example 14.3 ($t = 1.554$, $P = 0.0937$). It is in fact somewhat unusual to find a strong disagreement between the conclusions reached by these two tests.

```
                Exact Wilcoxon rank-sum test
data: 0weeds and 3weeds
rank-sum statistic W = 23, n = 4, m = 4, p-value = 0.100
alternative hypothesis: true mu is greater than 0
```

FIGURE 14.3 Output from the S-Plus statistical software for the data in Example 14.1. This program uses the exact distribution for W when the samples are small and there are no tied observations.

The normal approximation

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:

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

continuity correction

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

EXAMPLE 14.4

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

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

The continuity correction (page 379) acts as if the whole number 23 occupies the entire interval from 22.5 to 23.5. We 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:

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

The continuity correction gives a result closer to the exact value $P = 0.10$.

We recommend always using either the exact distribution (from software or tables) or the continuity correction for the rank sum statistic W. The exact distribution is safer for small samples. As Example 14.4 illustrates, however, the normal approximation with the continuity correction is often adequate.

EXAMPLE 14.5

Mann-Whitney test

Figure 14.4 shows the output for our data from two more statistical programs. Minitab offers only the normal approximation, and it refers to the **Mann-Whitney test.** This is an alternative form of the Wilcoxon rank sum test. SAS carries out both the exact and approximate tests. SAS calls the rank sum S rather than W and gives the mean 18 and standard deviation 3.464 as well as the z statistic 1.299 (using the continuity correction). SAS gives the approximate two-sided P-value as 0.1939, so

```
Mann-Whitney Confidence Interval and Test

0 weeds    N =   4    Median =          169.45
3 weeds    N =   4    Median =          157.30
Point estimate for ETA1- ETA2 is          11.30
97.0 Percent C.I. for ETA1- ETA2 is (–11.40,23.80)
W = 23.0
Test of ETA1 = ETA2 vs. ETA1 > ETA2 is significant at 0.0970
```

(a)

```
                 Wilcoxon Scores (Rank Sums) for Variable YIELD
                          Classified by Variable WEEDS

                          Sum of        Expected         Std Dev          Mean
WEEDS       N             Scores        Under H0        Under H0         Score

0           4               23.0            18.0      3.46410162    5.75000000
3           4               13.0            18.0      3.46410162    3.25000000


Wilcoxon 2-Sample Test                 S = 23.0000


          Exact P-Values
          (One-sided)      Prob >= S          = 0.1000
          (Two-sided)      Prob >= |S - Mean| = 0.2000


          Normal Approximation (with Continuity Correction of .5)
          Z = 1.29904      Prob > |Z| = 0.1939
```

(b)

FIGURE 14.4 Output from the Minitab and SAS statistical software for the data in Example 14.1. (a) Minitab uses the normal approximation for the distribution of W. (b) SAS gives both the exact and approximate values.

the one-sided result is half this, $P = 0.0970$. This agrees with Minitab and (up to a small roundoff error) with our result in Example 14.4. This approximate P-value is close to the exact result $P = 0.1000$, given by SAS and in Figure 14.3.

What hypotheses does Wilcoxon test?

Our null hypothesis is that weeds do not affect yield. Our 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 use the two-sample t test for means. Our hypotheses then become

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

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

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

The Wilcoxon rank sum test provides a significance test for these hypotheses, but only if an additional assumption 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 no weeds except that it may slide to a different location on the scale of yields. The Minitab output in Figure 14.4(a) states the hypotheses in terms of population medians (which it calls "ETA") and also gives a confidence interval for the difference between the two population medians.

The same-shape assumption is too strict to be reasonable in practice. Recall that our preferred version of the two-sample t test does not require that the two populations have the same standard deviation—that is, it does not make a same-shape assumption. Fortunately, the Wilcoxon test also applies in a much more general and more useful setting. It tests hypotheses that we can state in words as

H_0: two distributions are the same
H_a: one has values that are systematically larger

Here is a more exact statement of the "systematically larger" alternative hypothesis. Take X_1 to be corn yield with no weeds and X_2 to be corn yield with 3 weeds per meter. These yields are random variables. That is, every time we plant a plot with no weeds, the yield is a value of the variable X_1. The probability that the yield is more than 160 bushels per acre when no weeds are present is $P(X_1 > 160)$. If weed-free yields are "systematically larger" than those with weeds, yields higher than 160 should be more likely with no weeds. That is, we should have

$$P(X_1 > 160) > P(X_2 > 160)$$

The alternative hypothesis says that this inequality holds not just for 160 but for *any* yield we care to specify. No weeds always puts more probability "to the right" of whatever yield we are interested in.[2]

This exact statement of the hypotheses we are testing is a bit awkward. The 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 reduce to comparing medians. Many texts and computer outputs state the hypotheses in terms of medians, sometimes ignoring the same-shape requirement. 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.

Ties

average ranks

The exact distribution for the Wilcoxon rank sum is obtained assuming that all observations in both samples take different values. This allows us to rank them all. 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:

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 if you want to use rank tests when the data contain tied values.

It is sometimes useful to use rank tests on data that have very many ties because the scale of measurement has only a few values. Here is an example.

EXAMPLE 14.6

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. Is there good evidence that men and women differ in their perceptions about food safety at fairs?[3]

We should first ask if the subjects in Example 14.6 are 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.)

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 are more concerned about food safety then the men:

	Response					
	1	2	3	4	5	Total
Female	6.6%	55.1%	25.5%	11.7%	1.0%	100%
Male	20.6%	53.3%	20.6%	4.7%	1.0%	100%

Is the difference between the genders statistically significant?

We might apply the chi-square test (Chapter 9). It is highly significant ($X^2 = 16.120$, df = 4, $P = 0.0029$). Although the chi-square test answers our general question, it ignores the ordering of the responses and so does not use all of the available information. We would really like to know whether men or women are more concerned about the safety of the food served. This question depends on the ordering of responses from least concerned to most concerned. We can use the Wilcoxon test for the hypotheses:

H_0: men and women do not differ in their responses
H_a: one of the two genders gives systematically larger responses than the other

The alternative hypothesis is two-sided. Because the responses can take only five values, there are very 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.

EXAMPLE 14.7

Figure 14.5 gives computer output for the Wilcoxon test. The rank sum for men (using average ranks for ties) is $W = 14{,}059.5$. The standardized value is $z = -3.33$, with two-sided P-value $P = 0.0009$. There is very strong evidence of a difference. Women are more concerned than men about the safety of food served at fairs.

With more than 100 observations in each group and no outliers, we might use the two-sample t even though responses take only five values. In fact, the results for Example 14.6 are $t = 3.3655$ with $P = 0.0009$. The P-value for two-sample t is the same as that for the Wilcoxon test. There is, however, another reason to prefer the rank test in this example. 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 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. Some statisticians avoid using t procedures when there is not a fully meaningful scale of measurement.

```
        Wilcoxon Scores (Rank Sums) for Variable SFAIR
                Classified by Variable GENDER

                         Sum of   Expected      Std Dev        Mean
GENDER         N         Scores   Under H0     Under H0       Score

Female       196    31996.5000     29792.0   661.161398  163.247449
Male         107    14059.5000     16264.0   661.161398  131.397196
                  Average Scores Were Used for Ties

      Wilcoxon 2-Sample Test (Normal Approximation)
      (with Continuity Correction of .5)

      S = 14059.5    Z = –3.33353    Prob > |Z| = 0.0009
```

FIGURE 14.5 Output from SAS for the food safety study of Example 14.6. The approximate two-sided P-value is 0.0009.

Limitations of nonparametric tests

The examples we have given illustrate the potential usefulness of nonparametric tests. Nonetheless, rank tests are of secondary importance relative to inference procedures based on the normal distribution.

- Nonparametric inference is largely restricted to simple settings. Normal inference extends to methods for use with complex experimental designs and multiple regression, but nonparametric tests do not. We stress normal inference in part because it leads on to more advanced statistics.
- Normal tests compare means and are accompanied by simple confidence intervals for means or differences between means. When we use nonparametric tests to compare medians, we can also give confidence intervals, though they are awkward to calculate without software. However, the usefulness of nonparametric tests is clearest in settings when they do not simply compare medians—see the discussion of "What hypotheses does Wilcoxon test?" In these settings, there is no measure of the *size* of the observed effect that is closely related to the rank test of the *statistical significance* of the effect.
- The robustness of normal tests for means implies that we rarely encounter data that require nonparametric procedures to obtain reasonably accurate P-values. The t and W tests give very similar results in our examples. Nonetheless, many statisticians would not use a t test in Example 14.6 because the response variable gives only the order of the responses.
- There are more modern and more effective ways to escape the assumption of normality, such as bootstrap methods (see page 427).

SUMMARY

Nonparametric tests do not require any specific form for the distribution of the population 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.

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.***

***P*-values** for the Wilcoxon 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).

SECTION 14.1 EXERCISES

Statistical software is very helpful in doing these exercises. If you do not have access to software, base your work on the normal approximation with continuity correction.

14.1 A study of early childhood education asked kindergarten students to retell two fairy tales that had been read to them earlier in the week. The 10 children in the study included 5 high-progress readers and 5 low-progress readers. Each child told two stories. Story 1 had been read to them; Story 2 had been read and also illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. Here are the data:[4]

Child	Progress	Story 1 score	Story 2 score
1	high	0.55	0.80
2	high	0.57	0.82
3	high	0.72	0.54
4	high	0.70	0.79
5	high	0.84	0.89
6	low	0.40	0.77
7	low	0.72	0.49
8	low	0.00	0.66
9	low	0.36	0.28
10	low	0.55	0.38

Is there evidence that the scores of high-progress readers are higher than those of low-progress readers when they retell a story they have heard without pictures (Story 1)?

(a) Make normal quantile plots for the 5 responses in each group. Are any major deviations from normality apparent?

(b) Carry out a two-sample t test. State hypotheses and give the two sample means, the t statistic and its P-value, and your conclusion.

(c) Carry out the Wilcoxon rank sum test. State hypotheses and give the rank sum W for high-progress readers, its P-value, and your conclusion. Do the t and Wilcoxon tests lead you to different conclusions?

14.2 Repeat the analysis of Exercise 14.1 for the scores when children retell a story they have heard and seen illustrated with pictures (Story 2).

14.3 Use the data in Exercise 14.1 for children telling Story 2 to carry out by hand the steps in the Wilcoxon rank sum test.

(a) Arrange the 10 observations in order and assign ranks. There are no ties.

(b) Find the rank sum W for the five high-progress readers. What are the mean and standard deviation of W under the null hypothesis that low-progress and high-progress readers do not differ?

(c) Standardize W to obtain a z statistic. Do a normal probability calculation with the continuity correction to obtain a one-sided P-value.

(d) The data for Story 1 contain tied observations. What ranks would you assign to the 10 scores for Story 1?

14.4 The corn yield study of Example 14.1 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 $\overline{x}$ and s are not resistant.

(a) Is there evidence that 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 14.1 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 for 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?

14.5 How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed. Part of the study involved burying 10 polyester strips in well-drained soil in the summer. Five of the strips, chosen at random, were dug up after 2 weeks; the other 5 were dug up after 16 weeks. Here are the breaking strengths in pounds:[5]

2 weeks	118	126	126	120	129
16 weeks	124	98	110	140	110

(a) Make a back-to-back stemplot. Does it appear reasonable to assume that the two distributions have the same shape?

(b) Is there evidence that breaking strengths are lower for strips buried longer?

14.6 A "subliminal" message is below our threshold of awareness but may nonetheless influence us. Can subliminal messages help students learn math? A group of students who had failed the mathematics part of the City University of New York Skills Assessment Test agreed to participate in a study to find out. All received a daily subliminal message, flashed on a screen too rapidly to be consciously read. The treatment group of 10 students 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." All students participated in a summer program designed to raise their math skills, and all took the assessment test again at the end of the program. Here are data on the subjects' scores before and after the program.[6]

Treatment group		Control group	
Pretest	Posttest	Pretest	Posttest
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		

(a) The study design was a randomized comparative experiment. Outline this design.

(b) Compare the gain in scores in the two groups, using a graph and numerical descriptions. Does it appear that the treatment group's scores rose more than the scores for the control group?

(c) Apply the Wilcoxon rank sum test to the posttest versus pretest differences. Note that there are some ties. What do you conclude?

14.7 "Conservationists have despaired over destruction of tropical rainforest by logging, clearing, and burning." These words begin a report on a statistical study of the effects of logging in Borneo.[7] Here are data on the number of tree species in 12 unlogged forest plots and 9 similar plots logged 8 years earlier:

Unlogged	22	18	22	20	15	21	13	13	19	13	19	15
Logged	17	4	18	14	18	15	15	10	12			

(a) Make a back-to-back stemplot of the data. Does there appear to be a difference in species counts for logged and unlogged plots?

(b) Does logging significantly reduce the number of species in a plot after 8 years? State hypotheses, do a Wilcoxon test, and state your conclusion.

14.8 Exercise 7.65 (page 546) studies the effect of piano lessons on the spatial-temporal reasoning of preschool children. The data there concern 34 children who took piano lessons and a control group of 44 children. The data take only small whole-number values. Use the Wilcoxon rank sum test (there are many ties) to decide whether piano lessons improve spatial-temporal reasoning.

14.9 Example 14.6 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. You can find the full data set online or on the CD as the file *eg14-006.dat*. It contains the responses of 303 people to several questions. The variables in this data set are (in order)

subject hfair sfair sfast srest gender

The variable "sfair" contains the responses described in the example concerning safety of food served at outdoor fairs and festivals. The variable "srest" contains responses to the same question asked about food served in restaurants. The variable "gender" contains 1 if the respondent is a woman, 2 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?

14.10 The data file used in Example 14.6 and Exercise 14.9 contains 303 rows, one for each of the 303 respondents. Each row contains the responses of one person to several questions. We wonder if people are more concerned about the safety of food served at fairs than they are about the safety of food served at restaurants. Explain carefully why we *cannot* answer this question by applying the Wilcoxon rank sum test to the variables "sfair" and "srest."

14.11 To study customers' attitudes toward secondhand stores, researchers interviewed samples of shoppers at two secondhand stores of the same chain in two cities. Here are data on the incomes of shoppers at the two stores, presented as a two-way table of counts:[8]

Income code	Income	City 1	City 2
1	Under $10,000	70	62
2	$10,000 to $19,999	52	63
3	$20,000 to $24,999	69	50
4	$25,000 to $34,999	22	19
5	$35,000 or more	28	24

(a) Is there a relationship between city and income? Use the chi-square test to answer this question.

(b) The chi-square test ignores the ordering of the income categories. The data file *ex14-11.dat* contains data on the 459 shoppers in this study. The first variable is the city (City1 or City2) and the second is the income code as it appears in the table here (1 to 5). Is there good evidence that shoppers in one city have systematically higher incomes than in the other?

14.2 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 meet a rank test for this setting.

EXAMPLE 14.8

A study of early childhood education asked kindergarten students to retell two fairy tales that had been read to them earlier in the week. Each child told two stories. The first had been read to them, and the second had been read but also illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. Here are the data for five "low-progress" readers in a pilot study:[9]

Child	1	2	3	4	5
Story 2	0.77	0.49	0.66	0.28	0.38
Story 1	0.40	0.72	0.00	0.36	0.55
Difference	0.37	−0.23	0.66	−0.08	−0.17

We wonder if illustrations improve how the children retell a story. We would like to test the hypotheses

H_0: scores have the same distribution for both stories
H_a: scores are systematically higher for Story 2

Because this is a matched pairs design, we base our inference on the differences. The matched pairs t test gives $t = 0.635$ with one-sided P-value $P = 0.280$. Displays of the data (Figure 14.6) suggest some lack of normality. We would therefore like to use a rank test.

Positive differences in Example 14.8 indicate that the child performed better telling Story 2. If scores are generally higher with illustrations, 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 value

absolute values of the differences, that is, their magnitudes without a sign. Here they are, with boldface indicating the positive values:

0.37 0.23 **0.66** 0.08 0.17

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.

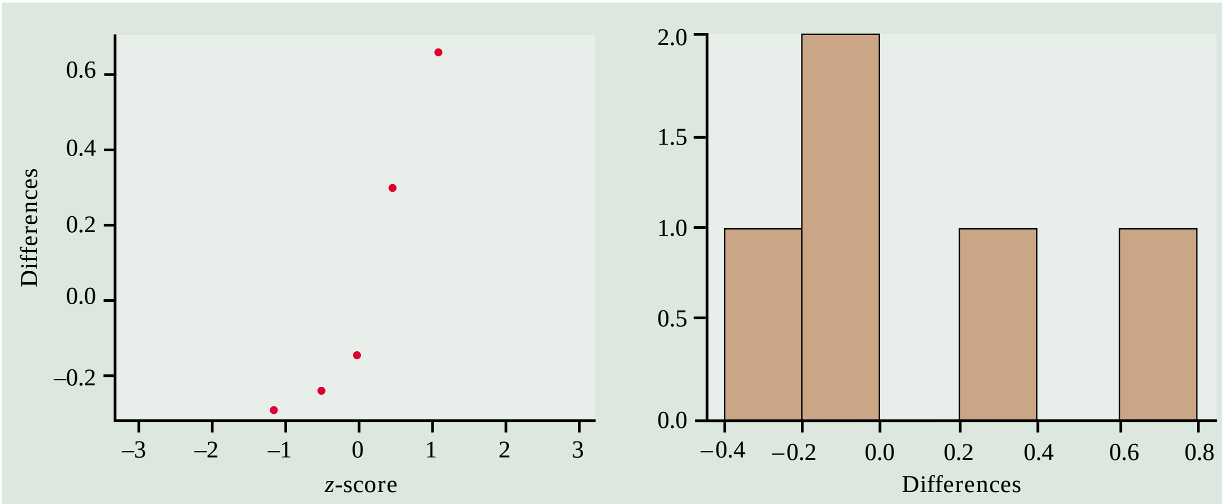

FIGURE 14.6 Normal quantile plot and histogram for the five differences in Example 14.8.

Absolute value	0.08	0.17	0.23	**0.37**	**0.66**
Rank	1	2	3	**4**	**5**

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 $W^+ = 9$.

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.

```
            Exact Wilcoxon signed-rank test

data: Story2-Story1

signed-rank statistic V = 9, n = 5, p-value = 0.4062

alternative hypothesis: true mu is greater than 0
```

(a)

```
Wilcoxon Signed Ranks Test

                          Ranks

                                  N      Mean     Sum of
STORY2-   Negative Ranks          3      Rank     Ranks
STORY1    Positive Ranks          2      2.00      6.00
          Ties                    0      4.50      9.00
          Total                   5

                   Test Statistics
                                      STORY2-
                                      STORY1
          Z                            -.405
          Asymp. Sig.(2-tailed)         .686
```

(b)

FIGURE 14.7 Output from (a) S-Plus and (b) SPSS for the storytelling study of Example 14.8. S-Plus reports the exact P-value $P = 0.4062$. SPSS uses the normal approximation without the continuity correction and so gives a less accurate P-value, $P = 0.343$ (one-sided).

EXAMPLE 14.9

In the storytelling study of Example 14.8, $n = 5$. If the null hypothesis (no systematic effect of illustrations) is true, the mean of the signed rank statistic is

$$\mu_{W^+} = \frac{n(n+1)}{4} = \frac{(5)(6)}{4} = 7.5$$

Our observed value $W^+ = 9$ is only slightly larger than this mean. The one-sided P-value is $P(W^+ \geq 9)$.

Figure 14.7 displays the output of two statistical programs. We see from Figure 14.7(a) that the one-sided P-value for the Wilcoxon signed rank test with $n = 5$ observations and $W^+ = 9$ is $P = 0.4062$. This result differs from the t test result $P = 0.280$, but both tell us that this very small sample gives no evidence that seeing illustrations improves the storytelling of low-progress readers.

The normal approximation

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 storytelling example, even though $n = 5$ is certainly not a large sample.

EXAMPLE 14.10

For $n = 5$ observations, we saw in Example 14.9 that $\mu_{W^+} = 7.5$. The standard deviation of W^+ under the null hypothesis is

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{(5)(6)(11)}{24}} = \sqrt{13.75} = 3.708$$

The continuity correction calculates the P-value $P(W^+ \geq 9)$ as $P(W^+ \geq 8.5)$, treating the value $W^+ = 9$ as occupying the interval from 8.5 to 9.5. We find the normal approximation for the P-value by standardizing and using the standard normal table:

$$P(W^+ \geq 8.5) = P\left(\frac{W^+ - 7.5}{3.708} \geq \frac{8.5 - 7.5}{3.708}\right) = P(Z \geq 0.27) = 0.394$$

Despite the small sample size, the normal approximation gives a result quite close to the exact value $P = 0.4062$. Figure 14.7(b) shows that the approximation is much less accurate without the continuity correction. This output reminds us not to trust software unless we know exactly what it does.

Ties

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. As in the case of the Wilcoxon rank sum, ties 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 14.11

Here are the golf scores of 12 members of a college women's golf team in two rounds of tournament play. (A golf score is the number of strokes required to complete the course, so that low scores are better.)

Player	1	2	3	4	5	6	7	8	9	10	11	12
Round 2	94	85	89	89	81	76	107	89	87	91	88	80
Round 1	89	90	87	95	86	81	102	105	83	88	91	79
Difference	5	−5	2	−6	−5	−5	5	−16	4	3	−3	1

Negative differences indicate better (lower) scores on the second round. We see that 6 of the 12 golfers improved their scores. We would like to test the hypotheses that in a large population of collegiate woman golfers

H_0: scores have the same distribution in rounds 1 and 2
H_a: scores are systematically lower or higher in round 2

A normal quantile plot of the differences (Figure 14.8) shows some irregularity and a low outlier. We will use the Wilcoxon signed rank test.

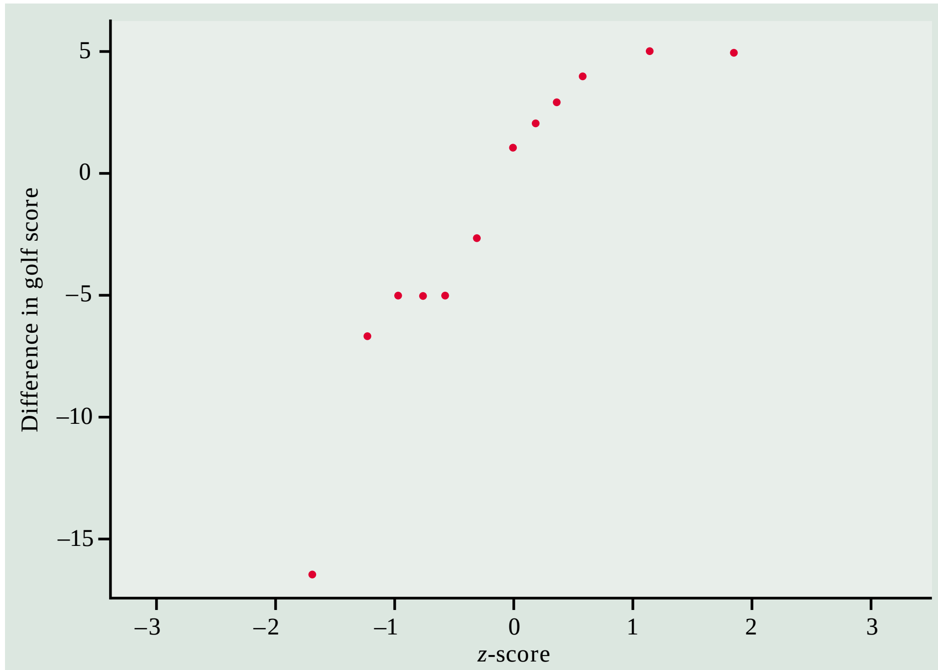

FIGURE 14.8 Normal quantile plot of the differences in scores for two rounds of a golf tournament, from Example 14.11.

The absolute values of the differences, with boldface indicating those that were negative, are

5 **5** 2 **6** **5** **5** 5 **16** 4 3 **3** 1

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

Absolute value	1	2	**3**	3	4	**5**	5	**5**	5	**5**	**6**	**16**
Rank	1	2	**3.5**	3.5	5	**8**	8	**8**	8	**8**	**11**	**12**

The Wilcoxon signed rank statistic is the sum of the ranks of the negative differences. (We could equally well use the sum of the ranks of the positive differences.) Its value is $W^+ = 50.5$.

EXAMPLE 14.12

Here are the two-sided *P*-values for the Wilcoxon signed rank test for the golf score data from several statistical programs:

Program	*P*-value
Minitab	$P = 0.388$
SAS	$P = 0.388$
S-PLUS	$P = 0.384$
SPSS	$P = 0.363$

All lead to the same practical conclusion: these data give no evidence for a systematic change in scores between rounds. However, the *P*-values reported differ a bit from program to program. The reason for the variations is that the programs use slightly different versions of the approximate calculations needed when ties are present. The exact result depends on which of these variations the programmer chooses to use.

For these data, the matched pairs t test gives $t = 0.9314$ with $P = 0.3716$. Once again, t and W^+ lead to the same conclusion.

SUMMARY

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** and the **sign test** are alternative tests 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).

SECTION 14.2 EXERCISES

Statistical software is very helpful in doing these exercises. If you do not have access to software, base your work on the normal approximation with continuity correction.

14.12 The concentration of carbon dioxide (CO_2) in the atmosphere is increasing rapidly due to our use of fossil fuels. Because plants use CO_2 to fuel photosynthesis, more CO_2 may cause trees and other plants to grow faster. An elaborate apparatus allows researchers to pipe extra CO_2 to a 30-meter circle of forest. They set up three pairs of circles in different parts of a forest in North Carolina. One of each pair received extra CO_2 for an entire growing season, and the other received ambient air. The response variable is the average growth in base area for trees in a circle, as a fraction of the starting area. Here are the data for one growing season:[10]

Pair	Control	Treatment
1	0.06528	0.08150
2	0.05232	0.06334
3	0.04329	0.05936

(a) Summarize the data. Does it appear that growth was faster in the treated plots?

(b) The researchers used a paired-sample *t* test to see if the data give good evidence of faster growth in the treated plots. State hypotheses, carry out the test, and state your conclusion.

(c) The sample is so small that we cannot assess normality. To be safe, we might use the Wilcoxon signed rank test. Carry out this test and report your result.

(d) The tests lead to very different conclusions. The primary reason is the lack of power of rank tests for very small samples. Explain to someone who knows no statistics what this means.

14.13 A student project asked subjects to step up and down for three minutes and measured their heart rates before and after the exercise. Here are data for five subjects and two treatments: stepping at a low rate (14 steps per minute) and at a medium rate (21 steps per minute). For each subject, we give the resting heart rate (beats per minutes) and the heart rate at the end of the exercise.[11]

	Low rate		Medium rate	
Subject	Resting	Final	Resting	Final
1	60	75	63	84
2	90	99	69	93
3	87	93	81	96
4	78	87	75	90
5	84	84	90	108

Does exercise at the low rate raise heart rate significantly? State hypotheses in terms of the median increase in heart rate and apply the Wilcoxon signed rank test. What do you conclude?

14.14 Do the data from the previous exercise give good reason to think that stepping at the medium rate increases heart rates more than stepping at the low rate?

(a) State hypotheses in terms of comparing the median increases for the two treatments. What is the proper rank test for these hypotheses?

(b) Carry out your test and state a conclusion.

14.15 Table 7.1 (page 498) presents the scores on a test of understanding of spoken French for a group of high school French teachers before and after a summer language institute. The improvements in scores between the pretest and the posttest for the 20 teachers were

2 0 6 6 3 3 2 3 −6 6 6 6 3 0 1 1 0 2 3 3

A normal quantile plot (Figure 7.7, page 503) shows granularity and a low outlier. We might wish to avoid the assumption of normality by using a rank test. Use the Wilcoxon signed rank procedure to reach a conclusion about the effect of the language institute. State hypotheses in words and report the statistic W^+, its P-value, and your conclusion. (Note that there are many ties in the data.)

14.16 Show the assignment of ranks and the calculation of the signed rank statistic W^+ for the data in Exercise 14.15. Remember that zeros are dropped from the data before ranking, so that n is the number of nonzero differences within pairs.

14.17 Example 14.6 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. The full data set is available online or on the CD as the file *eg14-006.dat.* It contains the responses of 303 people to several questions. The variables in this data set are (in order)

subject hfair sfair sfast srest gender

The variable "sfair" contains responses to the safety question described in Example 14.6. The variable "srest" contains responses to the same question asked about food served in restaurants. We suspect that restaurant food will appear safer than food served outdoors at a fair. Do the data give good evidence for this suspicion? (Give descriptive measures, a test statistic and its P-value, and your conclusion.)

14.18 The food safety survey data described in Example 14.6 and Exercise 14.17 also contain the responses of the 303 subjects to the same question asked about food served at fast-food restaurants. These responses are the values of the variable "sfast." Is there a systematic difference between the level of concern about food safety at outdoor fairs and at fast-food restaurants?

14.19 Exercise 7.37 (page 520) reports readings from 12 home radon detectors exposed to 105 picocuries per liter of radon:

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

We wonder if the median reading differs significantly from the true value 105.

(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 reading from home radon detectors:

$$H_0: \text{median} = 105$$
$$H_a: \text{median} \neq 105$$

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 105. (This is the one-sample version of the test.) What do you conclude?

14.20 Exercise 7.39 (page 520) gives data on the vitamin C content of 27 bags of wheat soy blend at the factory and five months later in Haiti. We want to know if vitamin C has been lost during transportation and storage. Describe what the data show about this question. Then use a rank test to see whether there has been a significant loss.

14.21 Exercise 7.40 (page 521) contains data from a student project that investigated whether right-handed people can turn a knob faster clockwise than they can counterclockwise. Describe what the data show, then state hypotheses and do a test that does not require normality. Report your conclusions carefully.

14.3 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 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 14.13

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:[12]

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

The summary statistics are

Weeds	n	Mean	Std. dev.
0	4	170.200	5.422
1	4	162.825	4.469
3	4	161.025	10.493
9	4	157.575	10.118

The sample standard deviations do not satisfy our rule of thumb that for safe use of ANOVA the largest should not exceed twice the smallest. Normal quantile plots (Figure 14.9) show that outliers are present in the yields for 3 and 9 weeds per meter. These are the correct yields for their plots, so we have no justification for removing them. We may want to use a nonparametric test.

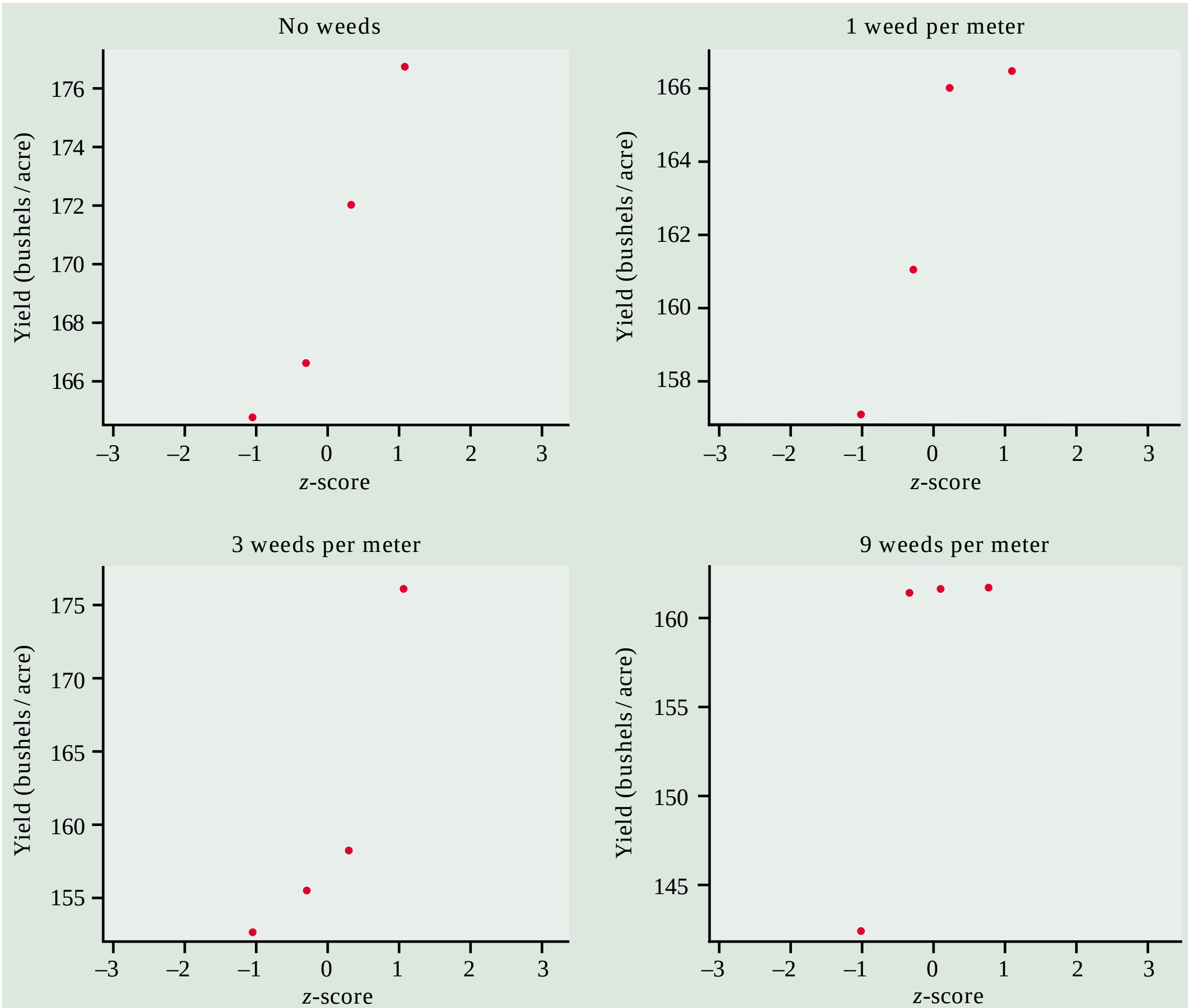

FIGURE 14.9 Normal quantile plots for the corn yields in the four treatment groups in Example 14.13.

Hypotheses and assumptions

The ANOVA F test concerns the means of the several populations represented by our samples. For Example 14.13, the ANOVA hypotheses are

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

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 assumption about data production (independent random samples from each population) remains important, but we can relax the normality assumption. 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 of 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 Kruskal-Wallis test

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 $n_1, n_2, \ldots, n_I$ from I 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 I populations have the same continuous distribution, H has approximately the chi-square distribution with $I - 1$ 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_I, 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 14.14

In Example 14.13, there are $I = 4$ 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 F) with df = 3, we find that the P-value lies in the interval $0.10 < P < 0.15$. This small experiment suggests that more weeds decrease yield but does not provide convincing evidence that weeds have an effect.

Figure 14.10 displays the output from the SAS statistical software, which gives the results $H = 5.5725$ and $P = 0.1344$. The software makes a small adjustment for the presence of ties that accounts for the slightly larger value

```
              Wilcoxon Scores (Rank Sums) for Variable YIELD
                      Classified by Variable WEEDS

                       Sum of      Expected         Std Dev            Mean
WEEDS      N           Scores      Under H0        Under H0           Score

0          4       52.5000000          34.0      8.24014563      13.1250000
1          4       33.5000000          34.0      8.24014563       8.3750000
3          4       25.0000000          34.0      8.24014563       6.2500000
9          4       25.0000000          34.0      8.24014563       6.2500000
                    Average Scores Were Used for Ties

              Kruskal-Wallis Test (Chi-Square Approximation)
                  CHISQ = 5.5725  DF = 3  Prob > CHISQ = 0.1344
```

FIGURE 14.10 Output from SAS for the Kruskal-Wallis test applied to the data in Example 14.13. SAS uses the chi-square approximation to obtain a *P*-value.

of H. The adjustment makes the chi-square approximation more accurate. It would be important if there were many ties.

As an option, SAS will calculate the exact *P*-value for the Kruskal-Wallis test. The result for Example 14.14 is $P = 0.1299$. This result required more than an hour of computing time. Fortunately, the chi-square approximation is quite accurate. The ordinary ANOVA F test gives $F = 1.73$ with $P = 0.2130$. 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.

SUMMARY

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, H for comparing **I** populations has approximately the chi-square distribution with $I - 1$ degrees of freedom. We use this approximate distribution to obtain *P*-values.

SECTION 14.3 EXERCISES

Statistical software is needed to do these exercises without unpleasant hand calculations. If you do not have access to software, find the Kruskal-Wallis statistic H by hand and use the chi-square table to get approximate P-values.

14.22 Exercise 12.11 presents the following data from a study of the loss of vitamin C in bread after baking:

Condition	Vitamin C (mg/100 g)	
Immediately after baking	47.62	49.79
One day after baking	40.45	43.46
Three days after baking	21.25	22.34
Five days after baking	13.18	11.65
Seven days after baking	8.51	8.13

The loss of vitamin C over time is clear, but with only 2 loaves of bread for each storage time we wonder if the differences among the groups are significant.

(a) Use the Kruskal-Wallis test to assess significance, then write a brief summary of what the data show.

(b) Because there are only 2 observations per group, we suspect that the common chi-square approximation to the distribution of the Kruskal-Wallis statistic may not be accurate. The exact P-value (from the SAS software) is $P = 0.0011$. Compare this with your P-value from (a). Is the difference large enough to affect your conclusion?

14.23 Exercise 12.23 discusses a study of the effect of exercise on bone density in growing rats. Ten rats were assigned to each of three treatments: a 60-centimeter "high jump," a 30-centimeter "low jump," and a control group with no jumping. Here are the bone densities (in milligrams per cubic centimeter) after 8 weeks of 10 jumps per day:

Group	Bone density (mg/cm^3)									
Control	611	621	614	593	593	653	600	554	603	569
Low jump	635	605	638	594	599	632	631	588	607	596
High jump	650	622	626	626	631	622	643	674	643	650

(a) The study was a randomized comparative experiment. Outline the design of this experiment.

(b) Make side-by-side stemplots for the three groups, with the stems lined up for easy comparison. The distributions are a bit irregular but not strongly nonnormal. We would usually use analysis of variance to assess the significance of the difference in group means.

(c) Do the Kruskal-Wallis test. Explain the distinction between the hypotheses tested by Kruskal-Wallis and ANOVA.

(d) Write a brief statement of your findings. Include a numerical comparison of the groups as well as your test result.

14.24 In Exercise 12.30 you used ANOVA to analyze the results of a study to see which of four colors best attracts cereal leaf beetles. Here are the data:

Color	Insects trapped					
Lemon yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

Because the samples are small, we will apply a nonparametric test.

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

(b) Find the median number of beetles trapped by boards of each color. Which colors appear more effective? Use the Kruskal-Wallis test to see if there are significant differences among the colors. What do you conclude?

14.25 Exercise 14.24 gives data on the counts of insects attracted by boards of four different colors. Carry out the Kruskal-Wallis test by hand, following these steps.

(a) What are I, the n_i, and N in this example?

(b) Arrange the counts in order and assign ranks. Be careful about ties. Find the sum of the ranks R_i for each color.

(c) 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.

14.26 Here are the breaking strengths (in pounds) of strips of polyester fabric buried in the ground for several lengths of time:[13]

Time	Breaking strength				
2 weeks	118	126	126	120	129
4 weeks	130	120	114	126	128
8 weeks	122	136	128	146	140
16 weeks	124	98	110	140	110

Breaking strength is a good measure of the extent to which the fabric has decayed.

(a) Find the standard deviations of the 4 samples. They do not meet our rule of thumb for applying ANOVA. In addition, the sample buried for 16 weeks contains an outlier. We will use the Kruskal-Wallis test.

(b) Find the medians of the four samples. What are the hypotheses for the Kruskal-Wallis test, expressed in terms of medians?

(c) Carry out the test and report your conclusion.

14.27 Example 14.6 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. The full data set is available online or on the CD as the file *eg14-006.dat*. It contains the responses of 303 people to several questions. The variables in this data set are (in order)

subject hfair sfair sfast srest gender

The variable "sfair" contains responses to the safety question described in Example 14.6. The variables "srest" and "sfast" contain responses to the same question asked about food served in restaurants and in fast-food chains. Explain carefully why we *cannot* use the Kruskal-Wallis test to see if there are systematic differences in perceptions of food safety in these three locations.

14.28 In Exercise 14.7 you compared the number of tree species in plots of land in a tropical rainforest that had never been logged with similar plots nearby that had been logged 8 years earlier. The researchers also counted species in plots that had been logged just 1 year earlier. Here are the counts of species:[14]

Plot type	Species count											
Unlogged	22	18	22	20	15	21	13	13	19	13	19	15
Logged 1 year ago	11	11	14	7	18	15	15	12	13	2	15	8
Logged 8 years ago	17	4	18	14	18	15	15	10	12			

(a) Use side-by-side stemplots to compare the distributions of number of trees per plot for the three groups of plots. Are there features that might prevent use of ANOVA? Also give the median number of trees per plot in the three groups.

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

14.29 In a study of heart disease in male federal employees, researchers classified 356 volunteer subjects according to their socioeconomic status (SES) and their smoking habits. There were three categories of SES: high, middle, and low. Individuals were asked whether they were current smokers, former smokers, or had never smoked. Here are the data, as a two-way table of counts:[15]

SES	Never (1)	Former (2)	Current (3)
High	68	92	51
Middle	9	21	22
Low	22	28	43

The data for all 356 subjects are stored in the file *ex14.29.dat* online and on the CD. Smoking behavior is stored numerically as 1, 2, or 3 using the codes given in the column headings above.

(a) Higher SES people in the United States smoke less as a group than lower SES people. Do these data show a relationship of this kind? Give percents that back your statements.

(b) Apply the chi-square test to see if there is a significant relationship between SES and smoking behavior.

(c) The chi-square test ignores the ordering of the responses. Use the Kruskal-Wallis test (with many ties) to test the hypothesis that some SES classes smoke systematically more than others.

CHAPTER 14 EXERCISES

14.30 Table 1.9 (page 59) presents data on the calorie and sodium content of selected brands of beef, meat, and poultry hot dogs. We will regard these brands as random samples from all brands available in food stores.

(a) Make stemplots of the calorie contents side by side, using the same stems for easy comparison. Give the five-number summaries for the three types of hot dog. What do the data suggest about the calorie content of different types of hot dog?

(b) Are any of the three distributions clearly not normal? Which ones, and why?

(c) Carry out a nonparametric test. Report your conclusions carefully.

14.31 Exercise 7.131 (page 569) reports data on the selling prices of 9 four-bedroom houses and 28 three-bedroom houses in West Lafayette, Indiana. We wonder if there is a difference between the average prices of three- and four-bedroom houses in this community.

(a) Make a normal quantile plot of the prices of three-bedroom houses. What kind of deviation from normality do you see?

(b) The t tests are quite robust. State the hypotheses for the proper t test, carry out the test, and present your results including appropriate data summaries.

(c) Carry out a nonparametric test. Once more state the hypotheses tested and present your results for both the test and the data summaries that should go with it.

14.32 Repeat the analysis of Exercise 14.30 for the sodium content of hot dogs.

Exercise 13.16 (page 822) reports data from a study of iron-deficiency anemia in Ethiopia. The issue is whether Ethiopian food loses more iron when cooked in some types of pots. Here are data on iron content (milligrams per 100 grams of food) for three types of food cooked in each of three types of pots:

	Iron content											
Type of pot	Meat				Legumes				Vegetables			
Aluminum	1.77	2.36	1.96	2.14	2.40	2.17	2.41	2.34	1.03	1.53	1.07	1.30
Clay	2.27	1.28	2.48	2.68	2.41	2.43	2.57	2.48	1.55	0.79	1.68	1.82
Iron	5.27	5.17	4.06	4.22	3.69	3.43	3.84	3.72	2.45	2.99	2.80	2.92

Exercises 14.33 to 14.35 use these data.

14.33 We want to know if the vegetable dish varies in iron content when cooked in aluminum, clay, and iron pots.

(a) Check the requirements for one-way ANOVA. Which requirements are a bit dubious in this setting?

(b) Instead of ANOVA, do a nonparametric test. Summarize your conclusions about the effect of pot material on iron content, including both descriptive measures and your test result.

14.34 There appears to be little difference between the iron content of food cooked in aluminum pots and food cooked in clay pots. Is there a significant difference between the iron content of meat cooked in aluminum and clay? Is the difference between aluminum and clay significant for legumes? Use nonparametric tests.

14.35 The data show that food cooked in iron pots has the highest iron content. They also suggest that the three types of food differ in iron content. Is there significant evidence that the three types of food differ in iron content when all are cooked in iron pots?

multiple comparisons

14.36 (**Optional**) As in ANOVA, we often want to carry out a **multiple comparisons** procedure following a Kruskal-Wallis test to tell us *which* groups differ significantly.[16] Here is a simple method: If we carry out k tests at fixed significance level $0.05/k$, the probability of *any* false rejection among the k tests is always no greater than 0.05. That is, to get overall significance level 0.05 for all of k comparisons, do each individual comparison at the $0.05/k$ level. In Exercise 14.30 you found a significant difference among the calorie contents of three types of hot dog. Now we will explore multiple comparisons.

(a) Write down all of the pairwise comparisons we can make, for example, beef versus meat. There are three possible pairwise comparisons.

(b) Carry out three Wilcoxon rank sum tests, one for each of the three pairs of hot dog types. What are the three two-sided *P*-values?

(c) For purposes of multiple comparisons, any of these three tests is significant if its *P*-value is no greater than 0.05/3 = 0.0167. Which pairs differ significantly at the overall 0.05 level?

14.37 (**Optional**) Exercise 14.36 outlines how to use the Wilcoxon rank sum test several times for multiple comparisons with overall significance level 0.05 for all comparisons together. In Exercise 14.24 you found that the numbers of beetles attracted by boards of four colors differ significantly. At the overall 0.05 level, which pairs of colors differ significantly? (*Hint:* There are 6 possible pairwise comparisons among 4 colors.)

NOTES

1. Data provided by Sam Phillips, Purdue University.

2. For purists, here is the precise definition: X_1 is *stochastically larger* than X_2 if

$$P(X_1 > a) \geq P(X_2 > a)$$

for all a, with strict inequality for at least one a. The Wilcoxon rank sum test is effective against this alternative in the sense that the power of the test approaches 1 (that is, the test becomes more certain to reject the null hypothesis) as the number of observations increases.

3. Data from Huey Chern Boo, "Consumers' perceptions and concerns about safety and healthfulness of food served at fairs and festivals," M.S. thesis, Purdue University, 1997.

4. Data provided by Susan Stadler, Purdue University.

5. From Sapna Aneja, "Biodeterioration of textile fibers in soil," M.S. thesis, Purdue University, 1994.

6. Data provided by Warren Page, New York City Technical College, from a study done by John Hudesman.

7. Data provided by Charles Cannon, Duke University. The study report is C. H. Cannon, D. R. Peart, and M. Leighton, "Tree species diversity in commercially logged Bornean rainforest," *Science*, 281 (1998), pp. 1366–1367.

8. From William D. Darley, "Store-choice behavior for pre-owned merchandise," *Journal of Business Research*, 27 (1993), pp. 17–31.

9. See Note 4.

10. Data for 1998 provided by Jason Hamilton, University of Illinois. The study report is Evan H. DeLucia et al., "Net primary production of a forest ecosystem with experimental CO_2 enhancement," *Science*, 284 (1999), pp. 1177–1179.

11. Simplified from the EESEE story "Stepping Up Your Heart Rate," on the CD.

12. See Note 1.

13. See Note 5.

14. See Note 7.

15. Ray H. Rosenman et al., "A 4-year prospective study of the relationship of different habitual vocational physical activity to risk and incidence of ischemic heart disease in volunteer male federal employees," in P. Milvey (ed.), *The Marathon: Physiological,* *Medical, Epidemiological and Psychological Studies*, New York Academy of Sciences, 301 (1977), pp. 627–641.

16. For more details on multiple comparisons (but not the simple procedure given here) see M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*, Wiley, 1973. This book is a useful reference on applied aspects of nonparametric inference in general.