CHAPTER FOURTEEN

Alternative Approaches to Inference

Introduction

In this final chapter we consider some inferential methods that are different in important ways from those considered earlier. Recall that many of the confidence intervals and test procedures developed in Chapters 9–12 were based on some sort of a normality assumption. As long as such an assumption is at least approximately satisfied, the actual confidence and significance levels will be at least approximately equal to the "nominal" levels, those prescribed by the experimenter through the choice of particular t or F critical values. However, if there is a substantial violation of the normality assumption, the actual levels may differ considerably from the nominal levels (e.g., the use of $t_{.025}$ in a confidence interval formula may actually result in a confidence level of only 88% rather than the nominal 95%). In the first three sections of this chapter, we develop *distribution-free* or *nonparametric* procedures that are valid for a wide variety of underlying distributions rather than being tied to normality. We have actually already introduced several such methods: the bootstrap intervals and permutation tests are valid without restrictive assumptions on the underlying distribution(s).

Section 14.4 introduces the Bayesian approach to inference. The standard *frequentist* view of inference is that the parameter of interest, θ, has a fixed but unknown value. Bayesians, however, regard θ as a random variable having a *prior* probability distribution that incorporates whatever is known about its value. Then to learn more about θ, a sample from the *conditional* distribution $f(x|\theta)$ is obtained, and Bayes' theorem is used to produce the *posterior* distribution of θ given the data $x_1, \ldots, x_n$. All Bayesian methods are based on this posterior distribution.

J.L. Devore and K.N. Berk, *Modern Mathematical Statistics with Applications*, Springer Texts in Statistics,
DOI 10.1007/978-1-4614-0391-3_14, © Springer Science+Business Media, LLC 2012

14.1 The Wilcoxon Signed-Rank Test

A research chemist replicated a particular experiment a total of 10 times and obtained the following values of reaction temperature, ordered from smallest to largest:

$$-.57 \quad -.19 \quad -.05 \quad .76 \quad 1.30 \quad 2.02 \quad 2.17 \quad 2.46 \quad 2.68 \quad 3.02$$

The distribution of reaction temperature is of course continuous. Suppose the investigator is willing to assume that this distribution is symmetric, so that the pdf satisfies $f(\tilde{\mu} + t) = f(\tilde{\mu} - t)$ for any $t > 0$, where $\tilde{\mu}$ is the median of the distribution (and also the mean μ provided that the mean exists). This condition on $f(x)$ simply says that the height of the density curve above a value any particular distance to the right of the median is the same as the height that same distance to the left of the median. The assumption of symmetry may at first thought seem quite bold, but remember that we have frequently assumed a normal distribution. Since a normal distribution is symmetric, the assumption of symmetry without any additional distributional specification is actually a weaker assumption than normality.

Let's now consider testing the null hypothesis that $\tilde{\mu} = 0$. This amounts to saying that a temperature of any particular magnitude, say 1.50, is no more likely to be positive (+1.50) than to be negative (−1.50). A glance at the data casts doubt on this hypothesis; for example, the sample median is 1.66, which is far larger in magnitude than any of the three negative observations.

Figure 14.1 shows graphs of two symmetric pdf's, one for which H_0 is true and the other for which the median of the distribution considerably exceeds 0. In the first case we expect the magnitudes of the negative observations in the sample to be comparable to those of the positive sample observations. However, in the second case observations of large absolute magnitude will tend to be positive rather than negative.

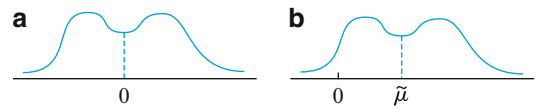


Figure 14.1 Distributions for which (a) $\tilde{\mu} = 0$; (b) $\tilde{\mu} \gg 0$

For the sample of ten reaction temperatures, let's for the moment disregard the signs of the observations and rank the absolute magnitudes from 1 to 10, with the smallest getting rank 1, the second smallest rank 2, and so on. Then apply the sign of each observation to the corresponding rank (so some signed ranks will be negative, e.g. −3, whereas others will be positive, e.g. 8). The test statistic will be $S_+ =$ the sum of the positively signed ranks.

Absolute Magnitude	.05	.19	.57	.76	1.30	2.02	2.17	2.46	2.68	3.02
Rank	1	2	3	4	5	6	7	8	9	10
Signed Rank	−1	−2	−3	4	5	6	7	8	9	10

$$s_+ = 4 + 5 + 6 + 7 + 8 + 9 + 10 = 49$$

When the median of the distribution is much greater than 0, most of the observations with large absolute magnitudes should be positive, resulting in positively signed ranks and a large value of s_+. On the other hand, if the median is 0, magnitudes of positively signed observations should be intermingled with those of negatively signed observations, in which case s_+ will not be very large. Thus we should reject $H_0: \tilde{\mu} = 0$ when s_+ is "quite large"— the rejection region should have the form $s_+ \geq c$.

The critical value c should be chosen so that the test has a desired significance level (type I error probability), such as .05 or .01. This necessitates finding the distribution of the test statistic S_+ when the null hypothesis is true. Let's consider $n = 5$, in which case there are $2^5 = 32$ ways of applying signs to the five ranks 1, 2, 3, 4, and 5 (each rank could have a − sign or a + sign). The key point is that when H_0 is true, *any collection of five signed ranks has the same chance as does any other collection*. That is, the smallest observation in absolute magnitude is equally likely to be positive or negative, the same is true of the second smallest observation in absolute magnitude, and so on. Thus the collection $-1, 2, 3, -4, 5$ of signed ranks is just as likely as the collection $1, 2, 3, 4, -5$, and just as likely as any one of the other 30 possibilities.

Table 14.1 lists the 32 possible signed-rank sequences when $n = 5$ along with the value s_+ for each sequence. This immediately gives the "null distribution" of S_+ displayed in Table 14.2. For example, Table 14.1 shows that three of the 32 possible sequences have $s_+ = 8$, so $P(S_+ = 8 \text{ when } H_0 \text{ is true}) = 1/32 + 1/32 + 1/32 = 3/32$. This null distribution appears in Table 14.2. Notice that it

Table 14.1 Possible signed-rank sequences for $n = 5$

Sequence					s_+	Sequence					s_+
−1	−2	−3	−4	−5	0	−1	−2	−3	+4	−5	4
+1	−2	−3	−4	−5	1	+1	−2	−3	+4	−5	5
−1	+2	−3	−4	−5	2	−1	+2	−3	+4	−5	6
−1	−2	+3	−4	−5	3	−1	−2	+3	+4	−5	7
+1	+2	−3	−4	−5	3	+1	+2	−3	+4	−5	7
+1	−2	+3	−4	−5	4	+1	−2	+3	+4	−5	8
−1	+2	+3	−4	−5	5	−1	+2	+3	+4	−5	9
+1	+2	+3	−4	−5	6	+1	+2	+3	+4	−5	10
−1	−2	−3	−4	+5	5	−1	−2	−3	+4	+5	9
+1	−2	−3	−4	+5	6	+1	−2	−3	+4	+5	10
−1	+2	−3	−4	+5	7	−1	+2	−3	+4	+5	11
−1	−2	+3	−4	+5	8	−1	−2	+3	+4	+5	12
+1	+2	−3	−4	+5	8	+1	+2	−3	+4	+5	12
+1	−2	+3	−4	+5	9	+1	−2	+3	+4	+5	13
−1	+2	+3	−4	+5	10	−1	+2	+3	+4	+5	14
+1	+2	+3	−4	+5	11	+1	+2	+3	+4	+5	15

Table 14.2 Null distribution of S_+ when $n = 5$

s_+	0	1	2	3	4	5	6	7
$p(s_+)$	1/32	1/32	1/32	2/32	2/32	3/32	3/32	3/32

s_+	8	9	10	11	12	13	14	15
$p(s_+)$	3/32	3/32	3/32	2/32	2/32	1/32	1/32	1/32

is symmetric about 7.5 [more generally, symmetrically distributed over the possible values $0,\ 1,\ 2,\ldots, n(n+1)/2$]. This symmetry is important in relating the rejection region of lower-tailed and two-tailed tests to that of an upper-tailed test.

For $n = 10$ there are $2^{10} = 1024$ possible signed rank sequences, so a listing would involve much effort. Each sequence, though, would have probability 1/1024 when H_0 is true, from which the distribution of S_+ when H_0 is true can be easily obtained.

We are now in a position to determine a rejection region for testing $H_0: \widetilde{\mu} = 0$ versus $H_a: \widetilde{\mu} > 0$ that has a suitably small significance level α. Consider the rejection region $R = \{s_+ : s_+ \geq 13\} = \{13, 14, 15\}$. Then

$$\begin{aligned}\alpha &= P(\text{reject } H_0 \text{ when } H_0 \text{ is true})\\ &= P(S_+ = 13, 14, \text{ or } 15 \text{ when } H_0 \text{ is true})\\ &= 1/32 + 1/32 + 1/32 = 3/32\\ &= .094\end{aligned}$$

so that $R = \{13, 14, 15\}$ specifies a test with approximate level .1. For the rejection region $\{14,\ 15\}$, $\alpha = 2/32 = .063$. For the sample $x_1 = .58,\ x_2 = 2.50,\ x_3 = -.21,\ x_4 = 1.23,\ x_5 = .97$, the signed rank sequence is -1, +2, +3, +4, +5, so $s_+ = 14$ and at level .063 H_0 would be rejected.

A General Description of the Wilcoxon Signed-Rank Test

Because the underlying distribution is assumed symmetric, $\mu = \widetilde{\mu}$, so we will state the hypotheses of interest in terms of μ rather than $\widetilde{\mu}$.[1]

ASSUMPTION

$X_1, X_2, \ldots, X_n$ is a random sample from a continuous and symmetric probability distribution with mean (and median) μ.

When the hypothesized value of μ is μ_0, the absolute differences $|x_1 - \mu_0|, \ldots |x_n - \mu_0|$, must be ranked from smallest to largest.

Null hypothesis: $H_0: \mu = \mu_0$

Test statistic value: $s_+ =$ the sum of the ranks associated with positive $(x_i - \mu_0)$'s

Alternative Hypothesis	**Rejection Region for Level α Test**
$H_a: \mu > \mu_0$	$s_+ \geq c_1$
$H_a: \mu < \mu_0$	$s_+ \leq c_2$ [where $c_2 = n(n+1)/2 - c_1$]
$H_a: \mu \neq \mu_0$	either $s_+ \geq c$ or $s_+ \leq n(n+1)/2 - c$

where the critical values c_1 and c obtained from Appendix Table A.12 satisfy $P(S_+ \geq c_1) \approx \alpha$ and $P(S_+ \geq c) \approx \alpha/2$ when H_0 is true.

---

[1] If the tails of the distribution are "too heavy," as was the case with the Cauchy distribution of Chapter 7, then μ will not exist. In such cases, the Wilcoxon test will still be valid for tests concerning $\widetilde{\mu}$.

Example 14.1 A producer of breakfast cereals wants to verify that a filler machine is operating correctly. The machine is supposed to fill one-pound boxes with 460 g, on the average. This is a little above the 453.6 g needed for one pound. When the contents are weighed, it is found that 15 boxes yield the following measurements:

454.4	470.8	447.5	453.2	462.6	445.0	455.9	458.2
461.6	457.3	452.0	464.3	459.2	453.5	465.8	

It is believed that deviations of any magnitude from 460 g are just as likely to be positive as negative (in accord with the symmetry assumption) but the distribution may not be normal. Therefore, the Wilcoxon signed-rank test will be used to see if the filler machine is calibrated correctly.

The hypotheses are H_0: $\mu = 460$ versus H_a: $\mu \neq 460$, where μ is the true average weight. Subtracting 460 from each measurement gives

−5.6	10.8	−12.5	−6.8	2.6	−15.0	−4.1	−1.8	1.6	−2.7
−8.0	4.3	−.8	−6.5	5.8					

The ranks are obtained by ordering these from smallest to largest without regard to sign.

Absolute Magnitude	.8	1.6	1.8	2.6	2.7	4.1	4.3	5.6	5.8	6.5	6.8	8.0	10.8	12.5	15.0
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sign	−	+	−	+	−	−	+	−	+	−	−	−	+	−	−

Thus $s_+ = 2 + 4 + 7 + 9 + 13 = 35$. From Appendix Table A.12, $P(S_+ \geq 95) = P(S_+ \leq 25) = .024$ when H_0 is true, so the two-tailed test with approximate level .05 rejects H_0 when either $s_+ \geq 95$ or ≤ 25 [the exact α is 2(.024) = .048]. Since $s_+ = 35$ is not in the rejection region, it cannot be concluded at level .05 that μ differs from 460. Even at level .094 (approximately .1), H_0 cannot be rejected, since $P(S_+ \leq 30) = P(S_+ \geq 90) = .047$ implies that s_+ values between 30 and 90 are not significant at that level. The P-value of the data is thus $>.1$. ■

Although a theoretical implication of the continuity of the underlying distribution is that ties will not occur, in practice they often do because of the discreteness of measuring instruments. If there are several data values with the same absolute magnitude, then they would be assigned the average of the ranks they would receive if they differed very slightly from one another. For example, if in Example 14.1 $x_8 = 458.2$ is changed to 458.4, then two different values of $(x_i - 460)$ would have absolute magnitude 1.6. The ranks to be averaged would be 2 and 3, so each would be assigned rank 2.5.

Paired Observations

When the data consisted of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the differences $D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n$ were normally distributed, in Chapter 10 we used a paired t test for hypotheses about the expected difference μ_D. If normality is not assumed, hypotheses about μ_D can be tested by using the Wilcoxon signed-rank test on the D_i's provided that the distribution of the differences is continuous and symmetric. *If X_i and Y_i both have continuous distributions that differ only with*

respect to their means (so the Y distribution is the X distribution shifted by $\mu_1 - \mu_2 = \mu_D$), *then* D_i *will have a continuous symmetric distribution* (it is *not* necessary for the X and Y distributions to be symmetric individually). The null hypothesis is $H_0: \mu_D = \Delta_0$, and the test statistic S_+ is the sum of the ranks associated with the positive $(D_i - \Delta_0)$'s.

Example 14.2 About 100 years ago an experiment was done to see if drugs could help people with severe insomnia ("The Action of Optical Isomers, II: Hyoscines," *J. Physiol.*, 1905: 501–510). There were 10 patients who had trouble sleeping, and each patient tried several medications. Here we compare just the control (no medication) and levo-hyoscine. Does the drug offer an improvement in average sleep time? The relevant hypotheses are H_0: $\mu_D = 0$ versus H_a: $\mu_D < 0$. Here are the sleep times, differences, and signed ranks.

Patient	1	2	3	4	5	6	7	8	9	10
Control	0.6	1.1	2.5	2.8	2.9	3.0	3.2	4.7	5.5	6.2
Drug	2.5	5.7	8.0	4.4	6.3	3.8	7.6	5.8	5.6	6.1
Difference	−1.9	−4.6	−5.5	−1.6	−3.4	−.8	−4.4	−1.1	−.1	.1
Signed rank	−6	−9	−10	−5	−7	−3	−8	−4	−1.5	1.5

Notice that there is a tie for the lowest rank, so the two lowest ranks are split between observations 9 and 10, and each receives rank 1.5. Appendix Table A.12 shows that for a test with significance level approximately .05, the null hypothesis should be rejected if $s_+ \le (10)(11)/2 - 44 = 11$. The test statistic value is 1.5, which falls in the rejection region. We therefore reject H_0 at significance level .05 in favor of the conclusion that the drug gives greater mean sleep time. The accompanying MINITAB output shows the test statistic value and also the corresponding P-value, which is $P(S_+ \le 1.5$ when H_0 is true).

```
Test of median = 0.000000 versus median < 0.000000
                N
                for     Wilcoxon              Estimated
        N       Test    Statistic       P     Median
diff    10      10      1.5          0.005    −2.250
```

■

Efficiency of the Wilcoxon Signed-Rank Test

When the underlying distribution being sampled is normal, either the t test or the signed-rank test can be used to test a hypothesis about μ. The t test is the best test in such a situation because among all level α tests it is the one having minimum β. It is generally agreed that there are many experimental situations in which normality can be reasonably assumed, as well as some in which it should not be. These two questions must be addressed in an attempt to compare the tests:

1. When the underlying distribution is normal (the "home ground" of the t test), how much is lost by using the signed-rank test?
2. When the underlying distribution is not normal, can a significant improvement be achieved by using the signed-rank test?

If the Wilcoxon test does not suffer much with respect to the t test on the "home ground" of the latter, and performs significantly better than the t test for a large number of other distributions, then there will be a strong case for using the Wilcoxon test.

Unfortunately, there are no simple answers to the two questions. Upon reflection, it is not surprising that the t test can perform poorly when the underlying distribution has "heavy tails" (i.e., when observed values lying far from μ are relatively more likely than they are when the distribution is normal). This is because the behavior of the t test depends on the sample mean and variance, which are both unstable in the presence of heavy tails. The difficulty in producing answers to the two questions is that β for the Wilcoxon test is very difficult to obtain and study for *any* underlying distribution, and the same can be said for the t test when the distribution is not normal. Even if β were easily obtained, any measure of efficiency would clearly depend on which underlying distribution was assumed. A number of different efficiency measures have been proposed by statisticians; one that many statisticians regard as credible is called **asymptotic relative efficiency** (ARE). The ARE of one test with respect to another is essentially the limiting ratio of sample sizes necessary to obtain identical error probabilities for the two tests. Thus if the ARE of one test with respect to a second equals .5, then when sample sizes are large, twice as large a sample size will be required of the first test to perform as well as the second test. Although the ARE does not characterize test performance for small sample sizes, the following results can be shown to hold:

1. When the underlying distribution is normal, the ARE of the Wilcoxon test with respect to the t test is approximately .95.
2. For any distribution, the ARE will be at least .86 and for many distributions will be much greater than 1.

We can summarize these results by saying that, in large-sample problems, the Wilcoxon test is never very much less efficient than the t test and may be much more efficient if the underlying distribution is far from normal. Although the issue is far from resolved in the case of sample sizes obtained in most practical problems, studies have shown that the Wilcoxon test performs reasonably and is thus a viable alternative to the t test.

Exercises Section 14.1 (1–8)

1. Reconsider the situation described in Exercise 32 of Section 9.2, and use the Wilcoxon test with $\alpha = .05$ to test the relevant hypotheses.
2. Use the Wilcoxon test to analyze the data given in Example 9.9.
3. The accompanying data is a subset of the data reported in the article "Synovial Fluid pH, Lactate, Oxygen and Carbon Dioxide Partial Pressure in Various Joint Diseases" (*Arthritis Rheum.*, 1971: 476–477). The observations are pH values of synovial fluid (which lubricates joints and tendons) taken from the knees of individuals suffering from arthritis. Assuming that true average pH for non-arthritic individuals is 7.39, test at level .05 to see whether the data indicates a difference between average pH values for arthritic and nonarthritic individuals.

   7.02 7.35 7.34 7.17 7.28 7.77 7.09
   7.22 7.45 6.95 7.40 7.10 7.32 7.14

4. A random sample of 15 automobile mechanics certified to work on a certain type of car was selected, and the time (in minutes) necessary for each one to diagnose a particular problem was determined, resulting in the following data:

   30.6 30.1 15.6 26.7 27.1 25.4 35.0 30.8
   31.9 53.2 12.5 23.2 8.8 24.9 30.2

   Use the Wilcoxon test at significance level .10 to decide whether the data suggests that true average diagnostic time is less than 30 minutes.
5. Both a gravimetric and a spectrophotometric method are under consideration for determining phosphate content of a particular material. Twelve samples of

the material are obtained, each is split in half, and a determination is made on each half using one of the two methods, resulting in the following data:

Sample	1	2	3	4
Gravimetric	54.7	58.5	66.8	46.1
Spectrophotometric	55.0	55.7	62.9	45.5

Sample	5	6	7	8
Gravimetric	52.3	74.3	92.5	40.2
Spectrophotometric	51.1	75.4	89.6	38.4

Sample	9	10	11	12
Gravimetric	87.3	74.8	63.2	68.5
Spectrophotometric	86.8	72.5	62.3	66.0

Use the Wilcoxon test to decide whether one technique gives on average a different value than the other technique for this type of material.

6. The signed-rank statistic can be represented as $S_+ = W_1 + W_2 + \cdots + W_n$, where $W_i = i$ if the sign of the $x_i - \mu_0$ with the ith largest absolute magnitude is positive (in which case i is included in S_+) and $W_i = 0$ if this value is negative ($i = 1, 2, 3, \ldots, n$). Furthermore, when H_0 is true, the W_i's are independent and $P(W = i) = P(W = 0) = .5$.

a. Use these facts to obtain the mean and variance of S_+ when H_0 is true. [*Hint*: The sum of the first n positive integers is $n(n+1)/2$, and the sum of the squares of the first n positive integers is $n(n+1)(2n+1)/6$.]

b. The W_i's are not identically distributed (e.g., possible values of W_2 are 2 and 0 whereas possible values of W_5 are 5 and 0), so our Central Limit Theorem for identically distributed and independent variables cannot be used here when n is large. However, a more general CLT can be used to assert that when H_0 is true and $n > 20$, S_+ has approximately a normal distribution with mean and variance obtained in (a). Use this to propose a large-sample standardized signed-rank test statistic and then an appropriate rejection region with level α for each of the three commonly encountered alternative hypotheses. [*Note*: When there are ties in the absolute magnitudes, it is still correct to standardize S_+ by subtracting the mean from (a), but there is a correction for the variance which can be found in books on nonparametric statistics.]

c. A particular type of steel beam has been designed to have a compressive strength (lb/in^2) of at least 50,000. An experimenter obtained a random sample of 25 beams and determined the strength of each one, resulting in the following data (expressed as deviations from 50,000):

−10	−27	36	−55	73	−77	−81
90	−95	−99	113	−127	−129	136
−150	−155	−159	165	−178	−183	−192
−199	−212	−217	−229			

Carry out a test using a significance level of approximately .01 to see if there is strong evidence that the design condition has been violated.

7. The accompanying 25 observations on fracture toughness of base plate of 18% nickel maraging steel were reported in the article "Fracture Testing of Weldments" (*ASTM Special Publ. No. 381*, 1965: 328–356). Suppose a company will agree to purchase this steel for a particular application only if it can be strongly demonstrated from experimental evidence that true average toughness exceeds 75. Assuming that the fracture toughness distribution is symmetric, state and test the appropriate hypotheses at level .05, and compute a P-value. [*Hint*: Use Exercise 6(b).]

69.5	71.9	72.6	73.1	73.3	73.5	74.1	74.2	75.3
75.5	75.7	75.8	76.1	76.2	76.2	76.9	77.0	77.9
78.1	79.6	79.7	80.1	82.2	83.7	93.7		

8. Suppose that observations $X_1, X_2, \ldots, X_n$ are made on a process at times $1, 2, \ldots, n$. On the basis of this data, we wish to test

H_0: the X_i's constitute an independent and identically distributed sequence

versus

H_a: X_{i+1} tends to be larger than X_i for $i = 1, \ldots, n$ (an increasing trend)

Suppose the X_i's are ranked from 1 to n. Then when H_a is true, larger ranks tend to occur later in the sequence, whereas if H_0 is true, large and small ranks tend to be mixed together. Let R_i be the rank of X_i and consider the test statistic $D = \sum_{i=1}^{n} (R_i - i)^2$.

Then small values of D give support to H_a (e.g., the smallest value is 0 for $R_1 = 1,\ R_2 = 2, \ldots, R_n = n$), so H_0 should be rejected in favor of H_a if $d \le c$. When H_0 is true, any sequence of ranks has probability $1/n!$. Use this to find c for which the test has a level as close to .10 as possible in the case $n = 4$. [*Hint*: List the 4! rank sequences, compute d for each one, and then obtain the null distribution of D. See the Lehmann book (in the chapter bibliography), for more information.]

14.2 The Wilcoxon Rank-Sum Test

When at least one of the sample sizes in a two-sample problem is small, the t test requires the assumption of normality (at least approximately). There are situations, though, in which an investigator would want to use a test that is valid even if the underlying distributions are quite nonnormal. We now describe such a test, called the **Wilcoxon rank-sum test**. An alternative name for the procedure is the Mann–Whitney test, although the Mann–Whitney test statistic is sometimes expressed in a slightly different form from that of the Wilcoxon test. The Wilcoxon test procedure is distribution-free because it will have the desired level of significance for a very large class of underlying distributions.

ASSUMPTIONS

$X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two independent random samples from continuous distributions with means μ_1 and μ_2, respectively. The X and Y distributions have the same shape and spread, the only possible difference between the two being in the values of μ_1 and μ_2.

When $H_0: \mu_1 - \mu_2 = \Delta_0$ is true, the X distribution is shifted by the amount Δ_0 to the right of the Y distribution; whereas when H_0 is false, the shift is by an amount other than Δ_0.

Development of the Test When $m = 3, n = 4$

Let's first test $H_0: \mu_1 - \mu_2 = 0$. If μ_1 is actually much larger than μ_2, then most of the observed x's will fall to the right of the observed y's. However, if H_0 is true, then the observed values from the two samples should be intermingled. The test statistic will provide a quantification of how much intermingling there is in the two samples.

Consider the case $m = 3$, $n = 4$. Then if all three observed x's were to the right of all four observed y's, this would provide strong evidence for rejecting H_0 in favor of $H_a: \mu_1 - \mu_2 \neq 0$, with a similar conclusion being appropriate if all three x's fall below all four of the y's. Suppose we pool the x's and y's into a combined sample of size $m + n = 7$ and rank these observations from smallest to largest, with the smallest receiving rank 1 and the largest, rank 7. If either most of the largest ranks or most of the smallest ranks were associated with X observations, we would begin to doubt H_0. This suggests the test statistic

$$W = \text{the sum of the ranks in the combined sample associated with } X \text{ observations} \tag{14.1}$$

For the values of m and n under consideration, the smallest possible value of W is $w = 1 + 2 + 3 = 6$ (if all three x's are smaller than all four y's), and the largest possible value is $w = 5 + 6 + 7 = 18$ (if all three x's are larger than all four y's).

As an example, suppose $x_1 = -3.10$, $x_2 = 1.67$, $x_3 = 2.01$, $y_1 = 5.27$, $y_2 = 1.89$, $y_3 = 3.86$, and $y_4 = .19$. Then the pooled ordered sample is -3.10, .19, 1.67, 1.89, 2.01, 3.86, and 5.27. The X ranks for this sample are 1 (for -3.10), 3 (for 1.67), and 5 (for 2.01), so the computed value of W is $w = 1 + 3 + 5 = 9$.

The test procedure based on the statistic (14.1) is to reject H_0 if the computed value w is "too extreme" — that is, $\geq c$ for an upper-tailed test, $\leq c$ for a lower-tailed test, and either $\geq c_1$ or $\leq c_2$ for a two-tailed test. The critical constant(s) c (c_1, c_2) should be chosen so that the test has the desired level of significance α. To see how this should be done, recall that when H_0 is true, all seven observations come from the same population. This means that under H_0, any possible triple of ranks associated with the three x's — such as (1, 4, 5), (3, 5, 6), or (5, 6, 7) — has the same probability as any other possible rank triple. Since there are $\binom{7}{3} = 35$ possible rank triples, under H_0 each rank triple has probability 1/35. From a list of all 35 rank triples and the w value associated with each, the probability distribution of W can immediately be determined. For example, there are four rank triples that have w value 11 — (1, 3, 7), (1, 4, 6), (2, 3, 6), and (2, 4, 5) — so $P(W = 11) = 4/35$. The summary of the listing and computations appears in Table 14.3.

Table 14.3 Probability distribution of W ($m = 3$, $n = 4$) when H_0 is true

w	6	7	8	9	10	11	12	13	14	15	16	17	18
$P(W = w)$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{4}{35}$	$\frac{4}{35}$	$\frac{5}{35}$	$\frac{4}{35}$	$\frac{4}{35}$	$\frac{3}{35}$	$\frac{2}{35}$	$\frac{1}{35}$	$\frac{1}{35}$

The distribution of Table 14.3 is symmetric about $w = (6 + 18)/2 = 12$, which is the middle value in the ordered list of possible W values. This is because the two rank triples (r, s, t) (with $r < s < t$) and $(8 - t, 8 - s, 8 - r)$ have values of w symmetric about 12, so for each triple with w value below 12, there is a triple with w value above 12 by the same amount.

If the alternative hypothesis is $H_a: \mu_1 - \mu_2 > 0$, then H_0 should be rejected in favor of H_a for large W values. Choosing as the rejection region the set of W values {17, 18}, $\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = P(W = 17 \text{ or } 18 \text{ when } H_0 \text{ is true}) = \frac{1}{35} + \frac{1}{35} = \frac{2}{35} = .057$; the region {17, 18} therefore specifies a test with level of significance approximately .05. Similarly, the region {6, 7}, which is appropriate for H_a: $\mu_1 - \mu_2 < 0$, has $\alpha = .057 \approx .05$. The region {6, 7, 17, 18}, which is appropriate for the two-sided alternative, has $\alpha = \frac{4}{35} = .114$. The W value for the data given several paragraphs previously was $w = 9$, which is rather close to the middle value 12, so H_0 would not be rejected at any reasonable level α for any one of the three H_a's.

General Description of the Rank-Sum Test

The null hypothesis $H_0: \mu_1 - \mu_2 = \Delta_0$ is handled by subtracting Δ_0 from each X_i and using the $(X_i - \Delta_0)$'s as the X_i's were previously used. Recalling that for any positive integer K, the sum of the first K integers is $K(K + 1)/2$, the smallest possible value of the statistic W is $m(m + 1)/2$, which occurs when the $(X_i - \Delta_0)$'s are all to the left of the Y sample. The largest possible value of W occurs when the $(X_i - \Delta_0)$'s lie entirely to the right of the Y's; in this case, $W = (n + 1) + \cdots + (m + n) = (\text{sum of first } m + n \text{ integers}) - (\text{sum of first } n \text{ integers})$, which gives

$m(m + 2n + 1)/2$. As with the special case $m = 3$, $n = 4$, the distribution of W is symmetric about the value that is halfway between the smallest and largest values; this middle value is $m(m + n + 1)/2$. Because of this symmetry, probabilities involving lower-tail critical values can be obtained from corresponding upper-tail values.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test statistic value : $w = \sum_{i=1}^{m} r_i$ where r_i = rank of $(x_i - \Delta_0)$ in the combined sample of $m + n$ $(x - \Delta_0)$'s and y's

Alternative Hypothesis	**Rejection Region**
$H_a : \mu_1 - \mu_2 > \Delta_0$	$w \geq c_1$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$w \leq m(m + n + 1) - c_1$
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	either $w \geq c$ or $w \leq m(m + n + 1) - c$

where $P(W \geq c_1$ when H_0 is true) $\approx \alpha$, $P(W \geq c$ when H_0 is true) $\approx \alpha/2$.

Because W has a discrete probability distribution, there will not always exist a critical value corresponding exactly to one of the usual levels of significance. Appendix Table A.13 gives upper-tail critical values for probabilities closest to .05, .025, .01, and .005, from which level .05 or .01 one- and two-tailed tests can be obtained. The table gives information only for $m = 3, 4, \ldots, 8$ and $n = m, m + 1, \ldots, 8$ (i.e., $3 \leq m \leq n \leq 8$). For values of m and n that exceed 8, a normal approximation can be used (Exercise 14). To use the table for small m and n, though, *the X and Y samples should be labeled so that* $m \leq n$.

Example 14.3 The urinary fluoride concentration (parts per million) was measured both for a sample of livestock grazing in an area previously exposed to fluoride pollution and for a similar sample grazing in an unpolluted region:

Polluted	21.3	18.7	23.0	17.1	16.8	20.9	19.7
Unpolluted	14.2	18.3	17.2	18.4	20.0		

Does the data indicate strongly that the true average fluoride concentration for livestock grazing in the polluted region is larger than for the unpolluted region? Use the Wilcoxon rank-sum test at level $\alpha = .01$.

The sample sizes here are 7 and 5. To obtain $m \leq n$, label the unpolluted observations as the x's ($x_1 = 14.2, \ldots, x_5 = 20.0$) and the polluted observations as the y's. Thus μ_1 is the true average fluoride concentration without pollution, and μ_2 is the true average concentration with pollution. The alternative hypothesis is $H_a : \mu_1 - \mu_2 < 0$ (pollution causes an increase in concentration), so a lower-tailed

test is appropriate. From Appendix Table A.13 with $m = 5$ and $n = 7$, $P(W \geq 47$ when H_0 is true) $\approx .01$. The critical value for the lower-tailed test is therefore $m(m + n + 1) - 47 = 5(13) - 47 = 18$; H_0 will now be rejected if $w \leq 18$. The pooled ordered sample follows; the computed W is $w = r_1 + r_2 + \cdots + r_5$ (where r_i is the rank of x_i) $= 1 + 5 + 4 + 6 + 9 = 25$. Since 25 is not ≤ 18, H_0 is not rejected at (approximately) level .01.

x	y	y	x	x	x	y	y	x	y	y	y
14.2	16.8	17.1	17.2	18.3	18.4	18.7	19.7	20.0	20.9	21.3	23.0
1	2	3	4	5	6	7	8	9	10	11	12

■

Ties are handled as suggested for the signed-rank test in the previous section.

Efficiency of the Wilcoxon Rank-Sum Test

When the distributions being sampled are both normal with $\sigma_1 = \sigma_2$, and therefore have the same shapes and spreads, either the pooled t test or the Wilcoxon test can be used (the two-sample t test assumes normality but not equal variances, so assumptions underlying its use are more restrictive in one sense and less in another than those for Wilcoxon's test). In this situation, the pooled t test is best among all possible tests in the sense of minimizing β for any fixed α. However, an investigator can never be absolutely certain that underlying assumptions are satisfied. It is therefore relevant to ask (1) how much is lost by using Wilcoxon's test rather than the pooled t test when the distributions are normal with equal variances and (2) how W compares to T in nonnormal situations.

The notion of test efficiency was discussed in the previous section in connection with the one-sample t test and Wilcoxon signed-rank test. The results for the two-sample tests are the same as those for the one-sample tests. When normality and equal variances both hold, the rank-sum test is approximately 95% as efficient as the pooled t test in large samples. That is, the t test will give the same error probabilities as the Wilcoxon test using slightly smaller sample sizes. On the other hand, the Wilcoxon test will always be at least 86% as efficient as the pooled t test and may be much more efficient if the underlying distributions are very nonnormal, especially with heavy tails. The comparison of the Wilcoxon test with the two-sample (unpooled) t test is less clear-cut. The t test is not known to be the best test in any sense, so it seems safe to conclude that as long as the population distributions have similar shapes and spreads, the behavior of the Wilcoxon test should compare quite favorably to the two-sample t test.

Lastly, we note that β calculations for the Wilcoxon test are quite difficult. This is because the distribution of W when H_0 is false depends not only on $\mu_1 - \mu_2$ but also on the shapes of the two distributions. For most underlying distributions, the nonnull distribution of W is virtually intractable. This is why statisticians have developed large-sample (asymptotic relative) efficiency as a means of comparing tests. With the capabilities of modern-day computer software, another approach to calculation of β is to carry out a simulation experiment.

Exercises Section 14.2 (9–16)

9. In an experiment to compare the bond strength of two different adhesives, each adhesive was used in five bondings of two surfaces, and the force necessary to separate the surfaces was determined for each bonding. For adhesive 1, the resulting values were 229, 286, 245, 299, and 250, whereas the adhesive 2 observations were 213, 179, 163, 247, and 225. Let μ_i denote the true average bond strength of adhesive type i. Use the Wilcoxon rank-sum test at level .05 to test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 > \mu_2$.

10. The article "A Study of Wood Stove Particulate Emissions" (*J. Air Pollut. Contr. Assoc.*, 1979: 724–728) reports the following data on burn time (hours) for samples of oak and pine. Test at level .05 to see whether there is any difference in true average burn time for the two types of wood.

Oak	1.72	.67	1.55	1.56	1.42	1.23	1.77	.48
Pine	.98	1.40	1.33	1.52	.73	1.20		

11. A modification has been made to the process for producing a certain type of "time-zero" film (film that begins to develop as soon as a picture is taken). Because the modification involves extra cost, it will be incorporated only if sample data strongly indicates that the modification has decreased true average developing time by more than 1 s. Assuming that the developing-time distributions differ only with respect to location if at all, use the Wilcoxon rank-sum test at level .05 on the accompanying data to test the appropriate hypotheses.

Original Process	8.6	5.1	4.5	5.4	6.3	6.6	5.7	8.5
Modified Process	5.5	4.0	3.8	6.0	5.8	4.9	7.0	5.7

12. The article "Measuring the Exposure of Infants to Tobacco Smoke" (*New Engl. J. Med.*, 1984: 1075–1078) reports on a study in which various measurements were taken both from a random sample of infants who had been exposed to household smoke and from a sample of unexposed infants. The accompanying data consists of observations on urinary concentration of cotinine, a major metabolite of nicotine (the values constitute a subset of the original data and were read from a plot that appeared in the article). Does the data suggest that true average cotinine level is higher in exposed infants than in unexposed infants by more than 25? Carry out a test at significance level .05.

Unexposed	8	11	12	14	20	43	111	
Exposed	35	56	83	92	128	150	176	208

13. Reconsider the situation described in Exercise 100 of Chapter 10 and the accompanying MINITAB output (the Greek letter eta is used to denote a median).

```
Mann-Whitney Confidence Interval and
Test
good      N = 8      Median = 0.540
poor      N = 8      Median = 2.400
Point estimate for ETA1 − ETA2 is
−1.155
95.9 % CI for ETA1 − ETA2 is(−3.160,
−0.409) W = 41.0
Test of ETA1 = ETA2 vs ETA1 < ETA2 is
significant at 0.0027
```

a. Verify that the value of MINITAB's test statistic is correct.

b. Carry out an appropriate test of hypotheses using a significance level of .01.

14. The Wilcoxon rank-sum statistic can be represented as $W = R_1 + R_2 + \cdots + R_m$, where R_i is the rank of $X_i - \Delta_0$ among all $m + n$ such differences. When H_0 is true, each R_i is equally likely to be one of the first $m + n$ positive integers; that is, R_i has a discrete uniform distribution on the values $1, 2, 3, \ldots, m + n$.

a. Determine the mean value of each R_i when H_0 is true and then show that the mean value of W is $m(m + n + 1)/2$. [*Hint*: Use the hint given in Exercise 6(a).]

b. The variance of each R_i is easily determined. However, the R_i's are not independent random variables because, for example, if $m = n = 10$ and we are told that $R_1 = 5$, then R_2 must be one of the *other* 19 integers between 1 and 20. However, if a and b are any two distinct positive integers between 1 and $m + n$ inclusive, it follows that $P(R_i = a \text{ and } R_j = b) = 1/[(m + n)(m + n - 1)]$ since two integers are being sampled without replacement from among $1, 2, \ldots, m + n$. Use this fact to show that $\text{Cov}(R_i, R_j) = -(m + n + 1)/12$ and then show that the variance of W is $mn(m + n + 1)/12$.

c. A central limit theorem for a sum of *non-independent* variables can be used to show that when $m > 8$ and $n > 8$, W has approximately a normal distribution with mean and variance given by the results of (a) and (b). Use this to

propose a large-sample standardized rank-sum test statistic and then describe the rejection region that has approximate significance level α for testing H_0 against each of the three commonly encountered alternative hypotheses. [*Note*: When there are ties in the observed values, a correction for the variance derived in (b) should be used in standardizing W; please consult a book on nonparametric statistics for the result.]

15. The accompanying data resulted from an experiment to compare the effects of vitamin C in orange juice and in synthetic ascorbic acid on the length of odontoblasts in guinea pigs over a 6-week period ("The Growth of the Odontoblasts of the Incisor Tooth as a Criterion of the Vitamin C Intake of the Guinea Pig," *J. Nutrit.*, 1947: 491–504). Use the Wilcoxon rank-sum test at level .01 to decide whether true average length differs for the two types of vitamin C intake. Compute also an approximate P-value. [*Hint*: See Exercise 14.]

Orange Juice	8.2	9.4	9.6	9.7	10.0	14.5
	15.2	16.1	17.6	21.5		
Ascorbic Acid	4.2	5.2	5.8	6.4	7.0	7.3
	10.1	11.2	11.3	11.5		

16. Test the hypotheses suggested in Exercise 15 using the following data:

Orange Juice	8.2	9.5	9.5	9.7	10.0	14.5
	15.2	16.1	17.6	21.5		
Ascorbic Acid	4.2	5.2	5.8	6.4	7.0	7.3
	9.5	10.0	11.5	11.5		

[*Hint*: See Exercise 14.]

14.3 Distribution-Free Confidence Intervals

The method we have used so far to construct a confidence interval (CI) can be described as follows: Start with a random variable (Z, T, χ^2, F, or the like) that depends on the parameter of interest and a probability statement involving the variable, manipulate the inequalities of the statement to isolate the parameter between random endpoints, and finally substitute computed values for random variables. Another general method for obtaining CIs takes advantage of a relationship between test procedures and CIs. A $100(1-\alpha)\%$ CI for a parameter θ can be obtained from a level α test for $H_0\colon \theta = \theta_0$ versus $H_a\colon \theta \neq \theta_0$. This method will be used to derive intervals associated with the Wilcoxon signed-rank test and the Wilcoxon rank-sum test.

Before using the method to derive new intervals, reconsider the t test and the t interval. Suppose a random sample of $n = 25$ observations from a normal population yields summary statistics $\bar{x} = 100$, $s = 20$. Then a 90% CI for μ is

$$\left(\bar{x} - t_{.05,24} \cdot \frac{s}{\sqrt{25}}, \bar{x} + t_{.05,24} \cdot \frac{s}{\sqrt{25}}\right) = (93.16, 106.84) \tag{14.2}$$

Suppose that instead of a CI, we had wished to test a hypothesis about μ. For $H_0\colon \mu = \mu_0$ versus $H_a\colon \mu \neq \mu_0$, the t test at level .10 specifies that H_0 should be rejected if t is either ≥ 1.711 or ≤ -1.711, where

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{25}} = \frac{100 - \mu_0}{20/\sqrt{25}} = \frac{100 - \mu_0}{4} \tag{14.3}$$

Consider now the null value $\mu_0 = 95$. Then $t = 1.25$, so H_0 is not rejected. Similarly, if $\mu_0 = 104$, then $t = -1$, so again H_0 is not rejected. However, if $\mu_0 = 90$, then $t = 2.5$, so H_0 is rejected, and if $\mu_0 = 108$, then $t = -2$, so H_0 is again rejected. By considering other values of μ_0 and the decision resulting from each one, the following general fact emerges: *Every number inside the*

interval (14.2) *specifies a value of* μ_0 *for which t of* (14.3) *leads to nonrejection of* H_0, *whereas every number outside interval* (14.2) *corresponds to a t for which* H_0 *is rejected*. That is, for the fixed values of n, $\bar{x}$, and s, the interval (14.2) is precisely the set of all μ_0 values for which testing $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ results in not rejecting H_0.

PROPOSITION

Suppose we have a level α test procedure for testing $H_0: \theta = \theta_0$ versus $H_a: \theta \neq \theta_0$. For fixed sample values, let A denote the set of all values θ_0 for which H_0 is not rejected. Then A is a $100(1 - \alpha)\%$ CI for θ.

There are actually pathological examples in which the set A defined in the proposition is not an interval of θ values, but instead the complement of an interval or something even stranger. To be more precise, we should really replace the notion of a CI with that of a confidence set. In the cases of interest here, the set A does turn out to be an interval.

The Wilcoxon Signed-Rank Interval

To test $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ using the Wilcoxon signed-rank test, where μ is the mean of a continuous symmetric distribution, the absolute values $|x_1 - \mu_0|, \ldots, |x_n - \mu_0|$ are ordered from smallest to largest, with the smallest receiving rank 1 and the largest, rank n. Each rank is then given the sign of its associated $x_i - \mu_0$, and the test statistic is the sum of the positively signed ranks. The two-tailed test rejects H_0 if s_+ is either $\geq c$ or $\leq n(n + 1)/2 - c$, where c is obtained from Appendix Table A.12 once the desired level of significance α is specified. For fixed $x_1, \ldots, x_n$, the $100(1 - \alpha)\%$ signed-rank interval will consist of all μ_0 for which $H_0: \mu = \mu_0$ is not rejected at level α. To identify this interval, it is convenient to express the test statistic S_+ in another form.

$$S_+ = \text{the number of pairwise averages } \left(X_i + X_j\right)/2 \text{ with } i \leq j \text{ that are } \geq \mu_0 \tag{14.4}$$

That is, if we average each x_j in the list with each x_i to its left, including $(x_j + x_j)/2$ (which is just x_j), and count the number of these averages that are $\geq \mu_0$, s_+ results. In moving from left to right in the list of sample values, we are simply averaging every pair of observations in the sample [again including $(x_j + x_j)/2$] exactly once, so the order in which the observations are listed before averaging is not important. The equivalence of the two methods for computing s_+ is not difficult to verify. The number of pairwise averages is $\binom{n}{2} + n$ (the first term due to averaging of different observations and the second due to averaging each x_i with itself), which equals $n(n + 1)/2$. If either too many or too few of these pairwise averages are $\geq \mu_0$, H_0 is rejected.

Example 14.4 The following observations are values of cerebral metabolic rate for rhesus monkeys: $x_1 = 4.51$, $x_2 = 4.59$, $x_3 = 4.90$, $x_4 = 4.93$, $x_5 = 6.80$, $x_6 = 5.08$, $x_7 = 5.67$. The 28 pairwise averages are, in increasing order,

4.51	4.55	4.59	4.705	4.72	4.745	4.76	4.795	4.835	4.90
4.915	4.93	4.99	5.005	5.08	5.09	5.13	5.285	5.30	5.375
5.655	5.67	5.695	5.85	5.865	5.94	6.235	6.80		

The first few and the last few of these are pictured on a measurement axis in Figure 14.2.

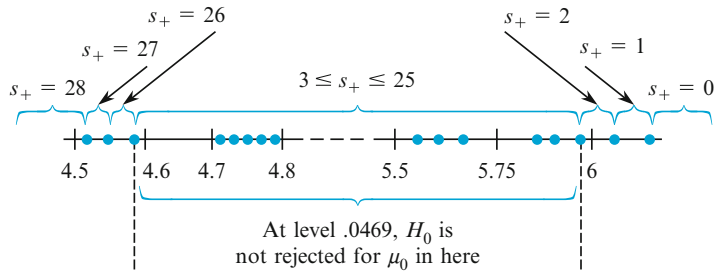


Figure 14.2 Plot of the data for Example 14.4

Because of the discreteness of the distribution of S_+, $\alpha = .05$ cannot be obtained exactly. The rejection region {0, 1, 2, 26, 27, 28} has $\alpha = .046$, which is as close as possible to .05, so the level is approximately .05. Thus if the number of pairwise averages $\geq \mu_0$ is between 3 and 25, inclusive, H_0 is not rejected. From Figure 14.2 the (approximate) 95% CI for μ is (4.59, 5.94). ■

In general, once the pairwise averages are ordered from smallest to largest, the endpoints of the Wilcoxon interval are two of the "extreme" averages. To express this precisely, let the smallest pairwise average be denoted by $\bar{x}_{(1)}$, the next smallest by $\bar{x}_{(2)}$, . . . , and the largest by $\bar{x}_{(n(n+1)/2)}$.

PROPOSITION

If the level α Wilcoxon signed-rank test for $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ is to reject H_0 if either $s_+ \geq c$ or $s_+ \leq n(n+1)/2 - c$, then a $100(1-\alpha)\%$ CI for μ is

$$(\bar{x}_{(n(n+1)/2-c+1)}, \bar{x}_{(c)}) \qquad (14.5)$$

In words, the interval extends from the dth smallest pairwise average to the dth largest average, where $d = n(n+1)/2 - c + 1$. Appendix Table A.14 gives the values of c that correspond to the usual confidence levels for $n = 5, 6, \ldots, 25$.

Example 14.5 (Example 14.4 continued) For $n = 7$, an 89.1% interval (approximately 90%) is obtained by using $c = 24$ (since the rejection region {0, 1, 2, 3, 4, 24, 25, 26, 27, 28} has $\alpha = .109$). The interval is $(\bar{x}_{(28-24+1)}, \bar{x}_{(24)}) = (\bar{x}_{(5)}, \bar{x}_{(24)}) = (4.72,\ 5.85)$, which extends from the fifth smallest to the fifth largest pairwise average. ■

The derivation of the interval depended on having a single sample from a continuous symmetric distribution with mean (median) μ. When the data is paired, the interval constructed from the differences $d_1, d_2, \ldots, d_n$ is a CI for the mean (median) difference μ_D. In this case, the symmetry of X and Y distributions need not be assumed; as long as the X and Y distributions have the same shape, the $X - Y$ distribution will be symmetric, so only continuity is required.

For $n > 20$, the large-sample approximation (Exercise 6) to the Wilcoxon test based on standardizing S_+ gives an approximation to c in (14.5). The result [for a $100(1 - \alpha)\%$ interval] is

$$c \approx \frac{n(n+1)}{4} + z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The efficiency of the Wilcoxon interval relative to the t interval is roughly the same as that for the Wilcoxon test relative to the t test. In particular, for large samples when the underlying population is normal, the Wilcoxon interval will tend to be slightly longer than the t interval, but if the population is quite nonnormal (symmetric but with heavy tails), then the Wilcoxon interval will tend to be much shorter than the t interval. And as we emphasized earlier in our discussion of bootstrapping, in the presence of nonnormality the actual confidence level of the t interval may differ considerably from the nominal (e.g., 95%) level.

The Wilcoxon Rank–Sum Interval

The Wilcoxon rank-sum test for testing $H_0 : \mu_1 - \mu_2 = \Delta_0$ is carried out by first combining the $(X_i - \Delta_0)$'s and Y_j's into one sample of size $m + n$ and ranking them from smallest (rank 1) to largest (rank $m + n$). The test statistic W is then the sum of the ranks of the $(X_i - \Delta_0)$'s. For the two-sided alternative, H_0 is rejected if w is either too small or too large.

To obtain the associated CI for fixed x_i's and y_j's, we must determine the set of all Δ_0 values for which H_0 is not rejected. This is easiest to do if we first express the test statistic in a slightly different form. The smallest possible value of W is $m(m + 1)/2$, corresponding to every $(X_i - \Delta_0)$ less than every Y_j, and there are mn differences of the form $(X_i - \Delta_0) - Y_j$. A bit of manipulation gives

$$\begin{aligned} W &= [\text{number of } (X_i - Y_j - \Delta_0)\text{'s} \geq 0] + \frac{m(m+1)}{2} \\ &= [\text{number of } (X_i - Y_j)\text{'s} \geq \Delta_0] + \frac{m(m+1)}{2} \end{aligned} \tag{14.6}$$

Thus rejecting H_0 if the number of $(x_i - y_j)$'s $\geq \Delta_0$ is either too small or too large is equivalent to rejecting H_0 for small or large w.

Expression (14.6) suggests that we compute $x_i - y_j$ for each i and j and order these mn differences from smallest to largest. Then if the null value Δ_0 is neither smaller than most of the differences nor larger than most, $H_0 : \mu_1 - \mu_2 = \Delta_0$ is not rejected. Varying Δ_0 now shows that a CI for $\mu_1 - \mu_2$ will have as its lower endpoint one of the ordered $(x_i - y_i)$'s, and similarly for the upper endpoint.

PROPOSITION

Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be the observed values in two independent samples from continuous distributions that differ only in location (and not in shape). With $d_{ij} = x_i - y_j$ and the ordered differences denoted by $d_{ij(1)}, d_{ij(2)}, \ldots, d_{ij(mn)}$, the general form of a $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$(d_{ij(mn-c+1)}, d_{ij(c)}) \tag{14.7}$$

where c is the critical constant for the two-tailed level α Wilcoxon rank-sum test.

Notice that the form of the Wilcoxon rank-sum interval (14.7) is very similar to the Wilcoxon signed-rank interval (14.5); (14.5) uses pairwise averages from a single sample, whereas (14.7) uses pairwise differences from two samples. Appendix Table A.15 gives values of c for selected values of m and n.

Example 14.6

The article "Some Mechanical Properties of Impregnated Bark Board" (*Forest Products J.*, 1977: 31–38) reports the following data on maximum crushing strength (psi) for a sample of epoxy-impregnated bark board and for a sample of bark board impregnated with another polymer:

Epoxy (x's)	10,860	11,120	11,340	12,130	14,380	13,070
Other (y's)	4,590	4,850	6,510	5,640	6,390	

Obtain a 95% CI for the true average difference in crushing strength between the epoxy-impregnated board and the other type of board.

From Appendix Table A.15, since the smaller sample size is 5 and the larger sample size is 6, $c = 26$ for a confidence level of approximately 95%. The d_{ij}'s appear in Table 14.4. The five smallest d_{ij}'s $[d_{ij(1)}, \ldots, d_{ij(5)}]$ are 4350, 4470, 4610, 4730, and 4830; and the five largest d_{ij}'s are (in descending order) 9790, 9530, 8740, 8480, and 8220. Thus the CI is $(d_{ij(5)}, d_{ij(26)}) = (4830, 8220)$.

Table 14.4 Differences (d_{ij}) for the rank-sum interval in Example 14.6

		y_j				
		4590	**4850**	**5640**	**6390**	**6510**
x_i	**10,860**	6270	6010	5220	4470	4350
	11,120	6530	6270	5480	4730	4610
	11,340	6750	6490	5700	4950	4830
	12,130	7540	7280	6490	5740	5620
	13,070	8480	8220	7430	6680	6560
	14,380	9790	9530	8740	7990	7870

■

When m and n are both large, the Wilcoxon test statistic has approximately a normal distribution (Exercise 14). This can be used to derive a large-sample approximation for the value c in interval (14.7). The result is

$$c \approx \frac{mn}{2} + z_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}} \tag{14.8}$$

As with the signed-rank interval, the rank-sum interval (14.7) is quite efficient with respect to the t interval; in large samples, (14.7) will tend to be only a bit longer than the t interval when the underlying populations are normal and may be considerably shorter than the t interval if the underlying populations have heavier tails than do normal populations. And once again, the actual confidence level for the t interval may be quite different from the nominal level in the presence of substantial nonnormality.

Exercises Section 14.3 (17–22)

17. The article "The Lead Content and Acidity of Christchurch Precipitation" (*New Zeal. J. Sci.*, 1980: 311–312) reports the accompanying data on lead concentration (μg/L) in samples gathered during eight different summer rainfalls: 17.0, 21.4, 30.6, 5.0, 12.2, 11.8, 17.3, and 18.8. Assuming that the lead-content distribution is symmetric, use the Wilcoxon signed-rank interval to obtain a 95% CI for μ.

18. Compute the 99% signed-rank interval for true average pH μ (assuming symmetry) using the data in Exercise 3. [*Hint*: Try to compute only those pairwise averages having relatively small or large values (rather than all 105 averages).]

19. Compute a CI for μ_D of Example 14.2 using the data given there; your confidence level should be roughly 95%.

20. The following observations are amounts of hydrocarbon emissions resulting from road wear of bias-belted tires under a 522-kg load inflated at 228 kPa and driven at 64 km/h for 6 h ("Characterization of Tire Emissions Using an Indoor Test Facility," *Rubber Chem. Tech.*, 1978: 7–25): .045, .117, .062, and .072. What confidence levels are achievable for this sample size using the signed-rank interval? Select an appropriate confidence level and compute the interval.

21. Compute the 90% rank-sum CI for $\mu_1 - \mu_2$ using the data in Exercise 9.

22. Compute a 99% CI for $\mu_1 - \mu_2$ using the data in Exercise 10.

14.4 Bayesian Methods

Consider making an inference about some parameter θ. The "frequentist" or "classical" approach, which we have followed until now in this book, is to regard the value of θ as fixed but unknown, observe data from a joint pmf or pdf $f(x_1, \ldots, x_n; \theta)$, and use the observations to draw appropriate conclusions. The Bayesian or "subjective" paradigm is different. Again the value of θ is unknown, but Bayesians say that all available information about it—intuition, data from past experiments, expert opinions, etc. —can be incorporated into a *prior distribution*, usually a prior pdf $g(\theta)$ since there will typically be a continuum of possible values of the parameter rather than just a discrete set. If there is substantial knowledge about θ, the prior will be quite peaked and highly concentrated about some central value, whereas a lack of information is shown by a relatively flat "uninformative" prior. These possibilities are illustrated in Figure 14.3.

In essence we are now thinking of the actual value of θ as the observed value of a random variable Θ, although unfortunately we ourselves don't get to observe the value. The (prior) distribution of this random variable is $g(\theta)$. Now, just as in

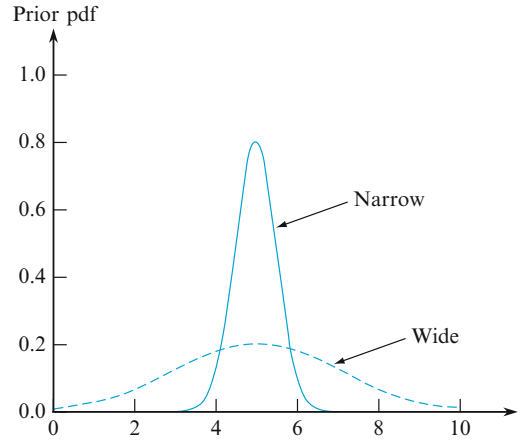


Figure 14.3 A narrow concentrated prior and a wider less informative prior

the frequentist scenario, an experiment is performed to obtain data. The joint pmf or pdf of the data given the value of θ is $p(x_1, \ldots, x_n \mid \theta)$ or $f(x_1, \ldots, x_n \mid \theta)$. We use a vertical line segment here rather than the earlier semicolon to emphasize that we are conditioning on the value of a random variable.

At this point, an appropriate version of Bayes' theorem is used to obtain $h(\theta|x_1,\ldots,x_n)$, the *posterior* distribution of the parameter. In the Bayesian world, this posterior distribution contains all current information about θ. In particular, the mean of this posterior distribution gives a point estimate of the parameter. An interval $[a, b]$ having posterior probability .95 gives a 95% *credibility* interval, the Bayesian analogue of a 95% confidence interval (but the interpretation is different). After presenting the necessary version of Bayes' Theorem, we illustrate the Bayesian approach with two examples.

Bayes' theorem here needs to be a bit more general than in Section 2.4 to allow for the possibility of continuous distributions. This version gives the posterior distribution $h(\theta \mid x_1, x_2, \ldots, x_n)$ as a product of the prior pdf times the conditional pdf, with a denominator to assure that the total posterior probability is 1:

$$h(\theta|x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n|\theta)g(\theta)}{\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n|\theta)g(\theta)d\theta}$$

Example 14.7 Suppose we want to make an inference about a population proportion p. Since the value of this parameter must be between 0 and 1, and the family of standard beta distributions is concentrated on the interval [0, 1], a particular beta distribution is a natural choice for a prior on p. In particular, consider data from a survey of 1574 American adults reported by the National Science Foundation in May 2002. Of those responding, 803 (51%) incorrectly said that antibiotics kill viruses. In accord with the discussion in Section 3.5, the data can be considered either a random sample of 1574 from the Bernoulli distribution (binomial with number of trials $= 1$) or a single observation from the binomial distribution with $n = 1574$. We use the latter approach here, but Exercise 23 involves showing that the Bernoulli approach is equivalent.

Assuming a beta prior for p on [0,1] with parameters a and b and the binomial distribution Bin($n = 1574$, p) for the data, we get for the posterior distribution,

$$h(p|x) = \frac{f(x|p)g(p)}{\int_{-\infty}^{\infty} f(x|p)g(p)dp} = \frac{\binom{n}{x}p^x(1-p)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}}{\int_0^1 \binom{n}{x}p^x(1-p)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}dp}.$$

The numerator can be written as

$$\binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(n+a+b)}\left[\frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)}p^{x+a-1}(1-p)^{n-x+b-1}\right].$$

Given that the part in square brackets is of the form of a beta pdf on [0, 1], its integral over this interval is 1. The part in front of the square brackets is shared by the numerator and denominator, and will therefore cancel. Thus

$$h(p|x) = \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)}p^{x+a-1}(1-p)^{n-x+b-1}$$

That is, the posterior distribution of p is itself a beta distribution with parameters $x + a$ and $n - x + b$.

If we were using the traditional non-Bayesian frequentist approach to statistics, and we wanted to give an estimate of p for this example, we would give the usual estimate from Section 8.2, $x/n = 803/1574 = .51$. The usual Bayesian estimate is the posterior mean, the expected value of p given the data. Recalling that the mean of the beta distribution on [0, 1] is $\alpha/(\alpha + \beta)$, we obtain

$$E(p|x) = (x + a)/(n + a + b) = (803 + a)/(1574 + a + b)$$

for the posterior mean.

Suppose that $a = b = 1$, so the beta prior distribution reduces to the uniform distribution on [0, 1]. Then $E(p|x) = (803 + 1)/(1574 + 2) = .51$, and in this case the Bayesian and frequentist results are essentially the same. It should be apparent that, if a and b are small compared to n, then the prior distribution will not matter much. Indeed, if a and b are close to 0 and positive, then $E(p|x) \approx x/n$. We should hesitate to set a and b equal to 0, because this would make the beta prior pdf not integrable, but it does nevertheless give a reasonable posterior distribution if x and $n - x$ are positive. When a prior distribution is not integrable it is said to be **improper**.

In Bayesian inference, is there an interval corresponding to the confidence interval for p given in Section 8.2? We have the posterior distribution for p, so we can take the central 95% of this distribution and call it a 95% credibility interval, as mentioned at the beginning of this section. In the case with a beta prior and $a = 1$, $b = 1$, we have a beta posterior with $\alpha = 804$, $\beta = 772$. Using the inverse cumulative beta distribution function from MINITAB (or almost any major statistical package) evaluated at .025 and .975, we obtain the interval [.4855, .5348]. For comparison the 95% confidence interval from Equation (8.10) of Section 8.2 is [.4855, .5348]. The intervals are not exactly the same, although they do agree to

four decimals. The simpler formula, Equation (8.11), gives the answer [.4855, .5349], which is very close because of the large sample size.

It is interesting that, although the frequentist and Bayesian intervals agree to four decimals, they have very different interpretations. For the Bayesian interval we can say that the probability is 95% that p is in the interval, given the data. However, this is not correct for the frequentist interval, because p is not random and the endpoints are not random after they have been specified, and therefore no probability statement is appropriate. Here the 95% applies to the aggregate of confidence intervals, of which in the long run 95% should include the true p.

The confidence intervals and credibility interval all include .5, so they allow the possibility that $p = .5$. Another way to view this possibility in Bayesian terms is to see whether the posterior distribution is consistent with $p = .5$. We actually consider the related hypothesis $p \leq .5$. Using $a = 1$ and $b = 1$ again, we find from MINITAB that the beta distribution with $\alpha = 804$ and $\beta = 772$ has probability .2100 of being less than or equal to .5. The corresponding one-tailed frequentist P-value is the probability, assuming $p = .5$, of at least 803 successes in 1574 trials, which is .2173. Both the Bayesian and frequentist values are much greater than .05, and there is no reason to reject .5 as a possible value for p.

To clarify the relationship between $E(p|x)$ and x/n, we can write $E(p|x)$ as a weighted average of the prior mean $a/(a + b)$ and x/n.

$$E(p|x) = \frac{a+b}{n+a+b} \cdot \frac{a}{a+b} + \frac{n}{n+a+b} \cdot \frac{x}{n}$$

The weights can be interpreted in terms of the sum of the two parameters of the beta distribution, which is often called the **concentration parameter**. The weights are proportional to the concentration parameter $a + b$ of the prior distribution and the number n of observations. The weight of the prior depends on the size of $a + b$ in relation to n, and the concentration parameter of the posterior distribution is the total $a + b + n$.

It is also useful to interpret the posterior pdf in terms of the concentration parameter. Because the first parameter is the sum $x + a$ and the second parameter is the sum $(n - x) + b$, the effect of a is to add to the number of successes and the effect of b is to add to the number of failures. In particular, setting a to 1 and b to 1 resulted in a posterior with the equivalent of 803 + 1 successes and (1574 – 803) + 1 failures, for a total of 1574 + 2 observations. From this viewpoint, the total observations are the $a + b$ provided by the prior plus the n provided by the data, and this addition also gives the concentration parameter of the posterior in terms of the concentration parameter of the prior.

How should we specify the prior distribution? The beta distribution is convenient, because it is easy with this specification to find the posterior distribution, but what about a and b? Suppose we have asked 10 adults about the effect of antibiotics on viruses, and it is reasonable to assume that the 10 are a random sample. If 6 of the 10 say that antibiotics kill viruses, then we set $a = 6$ and $b = 10 - 6 = 4$. That is, we have a beta distributed prior with parameters 6 and 4. Then the posterior distribution is beta with parameters 803 + 6 = 809 and (1574 – 803) + 4 = 775. The posterior is the same as if we had started with $a = 0$ and $b = 0$ and observed 809 who said that antibiotics kill viruses and 775 who

said no. In other words, observations can be incorporated into the prior and count just as if they were part of the NSF survey. ■

Life in the Bayesian world is sometimes more complicated. Perhaps the prior observations are not of a quality equivalent to that of the survey, but we would still like to use them to form a prior distribution. If we regard them as being only half as good, then we could use the same proportions but cut the a and b in half, using 3 and 2 instead of 6 and 4. There is certainly a subjective element to this, and it suggests why some statisticians are hesitant about using Bayesian methods. When everyone can agree about the prior distribution, there is little controversy about the Bayesian procedure, but when the prior is very much a matter of opinion people tend to disagree about its value.

Example 14.8 Assume a random sample $X_1, X_2, \ldots, X_n$ from the normal distribution with known variance, and assume a normal prior distribution for μ. In particular, consider the IQ scores of 18 first- grade boys,

113 108 140 113 115 146 136 107 108 119 132 127 118
108 103 103 122 111

from the private speech data introduced in Example 1.2. Because the IQ has a standard deviation of 15 nationwide, we can assume $\sigma = 15$ is valid here. For the prior distribution it is reasonable to use a mean of $\mu_0 = 110$, a ballpark figure for previous years in this school. It is harder to prescribe a standard deviation for the prior, but we will use $\sigma_0 = 7.5$. This is the standard deviation for the average of four independent observations if the individual standard deviation is 15. As a result, the effect on the posterior mean will turn out to be the same as if there were four additional observations with average 110.

To compute the posterior distribution of the mean μ, we use Bayes' theorem

$$h(\mu|x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n|\mu)g(\mu)}{\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n|\mu)g(\mu)d\mu}$$

The numerator is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\mu)g(\mu) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-.5(x_1-\mu)^2/\sigma^2} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-.5(x_n-\mu)^2/\sigma^2} \\ &\quad \times \frac{1}{\sqrt{2\pi}\sigma_0} e^{-.5(\mu-\mu_0)^2/\sigma_0^2} \\ &= \frac{1}{(2\pi)^{(n+1)/2}\sigma^n\sigma_0} e^{-.5[(x_1-\mu)^2/\sigma^2+\cdots+(x_n-\mu)^2/\sigma^2+(\mu-\mu_0)^2/\sigma_0^2]} \end{aligned}$$

The trick here is to complete the square in the exponent, which yields

$$(-.5/\sigma_1^2)(\mu - \mu_1)^2 + C$$

where C does not involve μ and

$$\sigma_1^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}, \quad \mu_1 = \frac{\frac{\sum x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

The posterior is then

$$h(\mu|x_1, x_2, \ldots, x_n) = \frac{\dfrac{\sigma_1}{(2\pi)^{n/2}\sigma^n\sigma_0} \cdot \dfrac{1}{(2\pi)^{.5}\sigma_1} e^{(-.5/\sigma_1^2)(\mu-\mu_1)^2} e^C}{\dfrac{\sigma_1}{(2\pi)^{n/2}\sigma^n\sigma_0} e^C \displaystyle\int_{-\infty}^{\infty} \dfrac{1}{(2\pi)^{.5}\sigma_1} e^{(-.5/\sigma_1^2)(\mu-\mu_1)^2} d\mu}$$

The integral is 1 because it is the area under a normal pdf, and the part in front of the integral cancels out, leaving a posterior distribution that is normal with mean μ_1 and standard deviation σ_1:

$$h(\mu|x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{.5}\sigma_1} e^{(-.5/\sigma_1^2)(\mu-\mu_1)^2}$$

Notice that the posterior mean μ_1 is a weighted average of the prior mean μ_0 and the data mean $\bar{x}$, with weights that are the reciprocals of the prior variance and the variance of $\bar{x}$. It makes sense to define the **precision** as the reciprocal of the variance because a lower variance implies a more precise measurement, and the weights then are the corresponding precisions. Furthermore, the posterior variance is the reciprocal of the sum of the reciprocals of the two variances, but this can be described much more simply by saying that the posterior precision is the sum of the prior precision plus the precision of $\bar{x}$.

Numerically, we have

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2} = \frac{1}{15^2/18} + \frac{1}{7.5^2} = .09778 = \frac{1}{10.227} = \frac{1}{3.198^2}$$

$$\mu_1 = \frac{\dfrac{n\bar{x}}{\sigma^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}} = \frac{\dfrac{18(118.28)}{15^2} + \dfrac{110}{7.5^2}}{\dfrac{18}{15^2} + \dfrac{1}{7.5^2}} = 116.77$$

The posterior distribution is normal with mean $\mu_1 = 116.77$ and standard deviation $\sigma_1 = 3.198$. The mean μ_1 is a weighted average of $\bar{x} = 118.28$ and $\mu_0 = 110$, so μ_1 is necessarily between them. As n becomes large the weight given to μ_0 declines, and μ_1 will be closer to $\bar{x}$.

Knowing the mean and standard deviation, we can use the normal distribution to find an interval with 95% probability for μ. This 95% credibility interval is [110.502, 123.038]. For comparison the 95% confidence interval using $\bar{x} = 118.28$ and $\sigma = 15$ is $\bar{x} \pm 1.96\sigma/\sqrt{n} = [111.35,\ 125.21]$. Notice that this interval must be wider. Because the precisions add to give the posterior precision, the posterior precision is greater than the prior precision and it is greater than the data precision. Therefore, it is guaranteed that the posterior standard deviation σ_1 will be less than σ_0 and less than the data standard deviation $\sigma/\sqrt{n}$.

Both the credibility interval and the confidence interval exclude 110, so we can be pretty sure that μ exceeds 110. Another way of looking at this is to calculate the posterior probability of μ being less than or equal to 110. Using $\mu_1 = 116.77$ and $\sigma_1 = 3.198$, we obtain the probability .0171, so this too supports the idea that μ exceeds 110.

How should we go about choosing μ_0 and σ_0 for the prior distribution? Suppose we have four prior observations for which the mean is 110. The standard

deviation of the mean is $15/\sqrt{4}$. We therefore choose $\mu_0 = 110$ and $\sigma_0 = 7.5$, the same values used for this example. If the four values are combined with the 18 values from the data set, then the mean of all 22 is $116.77 = \mu_1$ and the standard deviation is $15/\sqrt{22} = 3.198 = \sigma_1$. The 95% confidence interval for the mean, based on the average of all 22 observations, is the same as the Bayesian 95% credibility interval. This says that if you have some preliminary data values that are just as good as the regular data values that will be obtained, then base the prior distribution on the preliminary data. The posterior mean and its standard deviation will be the same as if the preliminary data were combined with the regular data, and the 95% credibility interval will be the same as the 95% confidence interval.

It should be emphasized that, even if the confidence interval is the same as the credibility interval, they have different interpretations. To interpret the Bayesian credibility interval, we can say that the probability is 95% that μ is in the interval [110.502, 123.038]. However, for the frequentist confidence interval such a probability statement does not make sense because μ and the endpoints of the interval are all constants after the interval has been calculated. Instead we have the more complicated interpretation that, in repeated realizations of the confidence interval, 95% of the intervals will include the true μ in the long run.

What should be done if there are no prior observations and there are no strong opinions about the prior mean μ_0? In this case the prior standard deviation σ_0 can be taken as some large number much bigger than σ, such as $\sigma_0 = 1000$ in our example. The result is that the prior will have essentially no effect, and the posterior distribution will be based on the data, $\mu_1 = \bar{x} = 118.28$ and $\sigma_1 = \sigma = 15$. The 95% credibility interval will be the same as the 95% confidence interval based on the 18 observations, [111.35, 125.21], but of course the interpretation is different. ■

In both examples it turned out that the posterior distribution has the same form as the prior distribution. When this happens we say that the prior distribution is **conjugate** to the data distribution. Exercises 31 and 32 offer additional examples of conjugate distributions.

Exercises Section 14.4 (23–32)

23. For the data of Example 14.7 assume a beta prior distribution and assume that the 1574 observations are a random sample from the Bernoulli distribution. Use Bayes' theorem to derive the posterior distribution, and compare your answer with the result of Example 14.7.

24. Here are the IQ scores for the 15 first-grade girls from the study mentioned in Example 14.8.

102	96	106	118	108	122	115	113
109	113	82	110	121	110	99	

Assume the same prior distribution used in Example 14.8, and assume that the data is a random sample from a normal distribution with mean μ and $\sigma = 15$.

a. Find the posterior distribution of μ.

b. Find a 95% credibility interval for μ.

c. Add four observations with average 110 to the data and find a 95% confidence interval for μ using the 19 observations. Compare with the result of (b).

d. Change the prior so the prior precision is very small but positive, and then recompute (a) and (b).

e. Find a 95% confidence interval for μ using the 15 observations and compare with the credibility interval of (d).

25. Laplace's rule of succession says that if there have been n Bernoulli trials and they have all been successes, then the probability of a success on the next trial is $(n+1)/(n+2)$. For the derivation Laplace used a beta prior with $a = 1$ and $b = 1$ for binomial data, as in Example 14.7.

a. Show that, if $a = 1$ and $b = 1$ and there are n successes in n trials, then the posterior mean of p is $(n+1)/(n+2)$.

b. Explain (a) in terms of total successes and failures; that is, explain the result in terms of two prior trials plus n later trials.

c. Laplace applied his rule of succession to compute the probability that the sun will rise tomorrow using 5000 years, or $n = 1{,}826{,}214$ days of history in which the sun rose every day. Is Laplace's method equivalent to including two prior days when the sun rose once and failed to rise once? Criticize the answer in terms of total successes and failures.

26. For the scenario of Example 14.8 assume the same normal prior distribution but assume that the data set is just one observation $\bar{x} = 118.28$ with standard deviation $\sigma/\sqrt{n} = 15/\sqrt{18} = 3.5355$. Use Bayes' theorem to derive the posterior distribution, and compare your answer with the result of Example 14.8.

27. Let X have the beta distribution on [0, 1] with parameters $\alpha = \nu_1/2$ and $\beta = \nu_2/2$, where $\nu_1/2$ and $\nu_2/2$ are positive integers. Define $Y = (X/\alpha)/[(1-X)/\beta]$. Show that Y has the F distribution with degrees of freedom ν_1, ν_2.

28. In a study by Erich Brandt of 70 restaurant bills, 40 of the 70 were paid using cash. We assume a random sample and estimate the posterior distribution of the binomial parameter p, the population proportion paying cash.

a. Use a beta prior distribution with $a = 2$ and $b = 2$.

b. Use a beta prior distribution with $a = 1$ and $b = 1$.

c. Use a beta prior distribution with a and b very small and positive.

d. Calculate a 95% credibility interval for p using (c). Is your interval compatible with $p = .5$?

e. Calculate a 95% confidence interval for p using Equation (8.10) of Section 8.2, and compare with the result of (d).

f. Calculate a 95% confidence interval for p using Equation (8.11) of Section 8.2, and compare with the results of (d) and (e).

g. Compare the interpretations of the credibility interval and the confidence intervals.

h. Based on the prior in (c), test the hypothesis $p \leq .5$ using the posterior distribution to find $P(p \leq .5)$.

29. Exercise 27 gives an alternative way of finding beta probabilities when software for the beta distribution is unavailable.

a. Use Exercise 27 together with the F table to obtain a 90% credibility interval for Exercise 28(c). [*Hint*: To find c such that .05 is the probability that F is to the left of c, reverse the degrees of freedom and take the reciprocal of the value for $\alpha = .05$.]

b. Repeat (a) using software for the beta distribution and compare with the result of (a).

30. If α and β are large, then the beta distribution can be approximated by the normal distribution using the beta mean and variance given in Section 4.5. This is useful in case beta distribution software is unavailable. Use the approximation to compute the credibility interval in Example 14.7.

31. Assume a random sample $X_1, X_2, \ldots, X_n$ from the Poisson distribution with mean λ. If the prior distribution for λ has a gamma distribution with parameters α and β, show that the posterior distribution is also gamma distributed. What are its parameters?

32. Consider a random sample $X_1, X_2, \ldots, X_n$ from the normal distribution with mean 0 and precision τ (use τ as a parameter instead of $\sigma^2 = 1/\tau$). Assume a gamma-distributed prior for τ and show that the posterior distribution of τ is also gamma. What are its parameters?

Supplementary Exercises (33–42)

33. The article "Effects of a Rice-Rich Versus Potato-Rich Diet on Glucose, Lipoprotein, and Cholesterol Metabolism in Noninsulin-Dependent Diabetics" (*Amer. J. Clin. Nutrit.*, 1984: 598–606) gives the accompanying data on cholesterol-synthesis rate for eight diabetic subjects. Subjects were fed a standardized diet with potato or rice as the major carbohydrate source. Participants received both diets for specified periods of time, with cholesterol-synthesis rate (mmol/day) measured at the end of each dietary period. The analysis presented in this article used a distribution-free test. Use such a test with significance level .05 to determine whether the true mean cholesterol-synthesis rate differs significantly for the two sources of carbohydrates.

Subject	1	2	3	4	5	6	7	8
Potato	1.88	2.60	1.38	4.41	1.87	2.89	3.96	2.31
Rice	1.70	3.84	1.13	4.97	.86	1.93	3.36	2.15

34. The study reported in "Gait Patterns During Free Choice Ladder Ascents" (*Hum. Movement Sci.*, 1983: 187–195) was motivated by publicity concerning the increased accident rate for individuals climbing ladders. A number of different gait patterns were used by subjects climbing a portable straight ladder according to specified instructions. The ascent times for seven subjects who used a lateral gait and six subjects who used a four-beat diagonal gait are given.

Lateral	.86	1.31	1.64	1.51	1.53	1.39	1.09
Diagonal	1.27	1.82	1.66	.85	1.45	1.24	

a. Carry out a test using $\alpha = .05$ to see whether the data suggests any difference in the true average ascent times for the two gaits.

b. Compute a 95% CI for the difference between the true average gait times.

35. The **sign test** is a very simple procedure for testing hypotheses about a population median assuming only that the underlying distribution is continuous. To illustrate, consider the following sample of 20 observations on component lifetime (hr):

1.7	3.3	5.1	6.9	12.6	14.4	16.4
24.6	26.0	26.5	32.1	37.4	40.1	40.5
41.5	72.4	80.1	86.4	87.5	100.2	

We wish to test the hypotheses $H_0: \tilde{\mu} = 25.0$ versus $H_a: \tilde{\mu} > 25.0$ The test statistic is $Y =$ the number of observations that exceed 25.

a. Consider rejecting H_0 if $Y \geq 15$. What is the value of α (the probability of a type I error) for this test? [*Hint*: Think of a "success" as a lifetime that exceeds 25.0. Then Y is the number of successes in the sample. What kind of a distribution does Y have when $\tilde{\mu} = 25.0$?]

b. What rejection region of the form $Y \geq c$ specifies a test with a significance level as close to .05 as possible? Use this region to carry out the test for the given data. [*Note*: The test statistic is the number of differences $X_i - 25.0$ that have positive signs, hence the name *sign test*.]

36. Refer to Exercise 35, and consider a confidence interval associated with the sign test, the **sign interval**. The relevant hypotheses are now $H_0: \tilde{\mu} = \tilde{\mu}_0$ versus $H_a: \tilde{\mu} \neq \tilde{\mu}_0$. Let's use the following rejection region: either $Y \geq 15$ or $Y \leq 5$.

a. What is the significance level for this test?

b. The confidence interval will consist of all values $\tilde{\mu}_0$ for which H_0 is not rejected. Determine the CI for the given data, and state the confidence level.

37. The single-factor ANOVA model considered in Chapter 11 assumed the observations in the ith sample were selected from a normal distribution with mean μ_i and variance σ^2, that is, $X_{ij} = \mu_i + \varepsilon_{ij}$ where the ε's are normal with mean 0 and variance σ^2. The normality assumption implies that the F test is not distribution-free. We now assume that the ε's all come from the same continuous, but not necessarily normal, distribution, and develop a distribution-free test of the null hypothesis that all I μ_i's are identical. Let $N = \sum J_i$, the total number of observations in the data set (there are J_i observations in the ith sample). Rank these N observations from 1 (the smallest) to N, and let $\overline{R}_i$ be the average of the ranks for the observations in the ith sample. When H_0 is true, we expect the rank of any particular observation and therefore also $\overline{R}_i$ to be $(N + 1)/2$. The data argues against H_0 when some of the $\overline{R}_i$'s differ considerably from $(N + 1)/2$. The *Kruskal–Wallis* test statistic is

$$K = \frac{12}{N(N+1)} \sum J_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2$$

When H_0 is true and either (1) $I = 3$, all $J_i \geq 6$ or (2) $I > 3$, all $J_i \geq 5$, the test statistic has approximately a chi-squared distribution with $I - 1$ df.

The accompanying observations on axial stiffness index resulted from a study of metal-plate connected trusses in which five different plate lengths—4 in., 6 in., 8 in., 10 in., and 12 in. — were used ("Modeling Joints Made with Light-Gauge Metal Connector Plates," *Forest Products J.*, 1979: 39–44).

$i = 1$ (4 in.):	309.2	309.7	311.0	316.8
	326.5	349.8	409.5	
$i = 2$ (6 in.):	331.0	347.2	348.9	361.0
	381.7	402.1	404.5	
$i = 3$ (8 in.):	351.0	357.1	366.2	367.3
	382.0	392.4	409.9	
$i = 4$ (10 in.):	346.7	362.6	384.2	410.6
	433.1	452.9	461.4	
$i = 5$ (12 in.):	407.4	410.7	419.9	441.2
	441.8	465.8	473.4	

Use the *K–W* test to decide at significance level .01 whether the true average axial stiffness index depends somehow on plate length.

38. The article "Production of Gaseous Nitrogen in Human Steady-State Conditions" (*J. Appl. Physiol.*, 1972: 155–159) reports the following observations on the amount of nitrogen expired (in liters) under four dietary regimens: (1) fasting, (2) 23% protein, (3) 32% protein, and (4) 67% protein. Use the Kruskal–Wallis test (Exercise 37) at level .05 to test equality of the corresponding μ_i's.

1.	4.079	4.859	3.540	5.047	3.298
	4.679	2.870	4.648	3.847	
2.	4.368	5.668	3.752	5.848	3.802
	4.844	3.578	5.393	4.374	
3.	4.169	5.709	4.416	5.666	4.123
	5.059	4.403	4.496	4.688	
4.	4.928	5.608	4.940	5.291	4.674
	5.038	4.905	5.208	4.806	

39. The model for the data from a randomized block experiment for comparing I treatments was $X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, where the α's are treatment effects, the β's are block effects, and the ε's were assumed normal with mean 0 and variance σ^2. We now replace normality by the assumption that the ε's have the same continuous distribution. A distribution-free test of the null hypothesis of no treatment effects, called *Friedman's test*, involves first ranking the observations in each block separately from 1 to I. The rank average $\overline{R}_i$ is then calculated for each of the I treatments. If H_0 is true, the expected value of each rank average is $(I + 1)/2$. The test statistic is

$$F_r = \frac{12J}{I(I+1)} \sum \left(\overline{R}_i - \frac{I+1}{2} \right)^2$$

For even moderate values of J, the test statistic has approximately a chi-squared distribution with $I - 1$ df when H_0 is true.

The article "Physiological Effects During Hypnotically Requested Emotions" (*Psychosomatic Med.*, 1963: 334–343) reports the following data (x_{ij}) on skin potential in millivolts when the emotions of fear, happiness, depression, and calmness were requested from each of eight subjects.

	Blocks (Subjects)			
	1	**2**	**3**	**4**
Fear	23.1	57.6	10.5	23.6
Happiness	22.7	53.2	9.7	19.6
Depression	22.5	53.7	10.8	21.1
Calmness	22.6	53.1	8.3	21.6
	5	**6**	**7**	**8**
Fear	11.9	54.6	21.0	20.3
Happiness	13.8	47.1	13.6	23.6
Depression	13.7	39.2	13.7	16.3
Calmness	13.3	37.0	14.8	14.8

Use Friedman's test to decide whether emotion has an effect on skin potential.

40. In an experiment to study the way in which different anesthetics affect plasma epinephrine concentration, ten dogs were selected and concentration was measured while they were under the influence of the anesthetics isoflurane, halothane, and cyclopropane ("Sympathoadrenal and Hemodynamic Effects of Isoflurane, Halothane, and Cyclopropane in Dogs," *Anesthesiology*, 1974: 465–470). Test at level .05 to see whether there is an anesthetic effect on concentration. [*Hint*: See Exercise 39.]

	Dog				
	1	**2**	**3**	**4**	**5**
Isoflurane	.28	.51	1.00	.39	.29
Halothane	.30	.39	.63	.38	.21
Cyclopropane	1.07	1.35	.69	.28	1.24
	6	**7**	**8**	**9**	**10**
Isoflurane	.36	.32	.69	.17	.33
Halothane	.88	.39	.51	.32	.42
Cyclopropane	1.53	.49	.56	1.02	.30

41. Suppose we wish to test

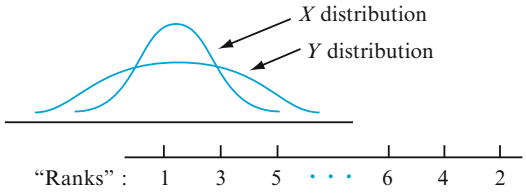
H_0: the X and Y distributions are identical

versus

H_a: the X distribution is less spread out than the Y distribution

The accompanying figure pictures X and Y distributions for which H_a is true. The Wilcoxon rank-sum test is not appropriate in this situation because when H_a is true as pictured, the Y's will tend to be at the extreme ends of the combined sample (resulting in small and large Y ranks), so

the sum of X ranks will result in a W value that is neither large nor small.



Consider modifying the procedure for assigning ranks as follows: After the combined sample of $m + n$ observations is ordered, the smallest observation is given rank 1, the largest observation is given rank 2, the second smallest is given rank 3, the second largest is given rank 4, and so on. Then if H_a is true as pictured, the X values will tend to be in the middle of the sample and thus receive large ranks. Let W' denote the sum of the X ranks and consider rejecting H_0 in favor of H_a when $w' \geq c$. When H_0 is true, every possible set of X ranks has the same probability, so W' *has the same distribution as does* W *when* H_0 *is true*. Thus c can be chosen from Appendix Table A.13 to yield a level α test. The accompanying data refers to medial muscle thickness for arterioles from the lungs of children who died from sudden infant death syndrome (x's) and a control group of children (y's). Carry out the test of H_0 versus H_a at level .05.

SIDS	4.0	4.4	4.8	4.9	
Control	3.7	4.1	4.3	5.1	5.6

Consult the Lehmann book (in the chapter bibliography) for more information on this test, called the *Siegel–Tukey test*.

42. The ranking procedure described in Exercise 41 is somewhat asymmetric, because the smallest observation receives rank 1 whereas the largest receives rank 2, and so on. Suppose both the smallest and the largest receive rank 1, the second smallest and second largest receive rank 2, and so on, and let W'' be the sum of the X ranks. The null distribution of W'' is not identical to the null distribution of W, so different tables are needed. Consider the case $m = 3$, $n = 4$. List all 35 possible orderings of the three X values among the seven observations (e.g., 1, 3, 7 or 4, 5, 6), assign ranks in the manner described, compute the value of W'' for each possibility, and then tabulate the null distribution of W''. For the test that rejects if $w'' \geq c$, what value of c prescribes approximately a level .10 test? This is the *Ansari–Bradley test*; for additional information, see the book by Hollander and Wolfe in the chapter bibliography.

Bibliography

Berry, Donald A., *Statistics: A Bayesian Perspective*, Brooks/Cole—Cengage Learning, Belmont, CA, 1996. An elementary introduction to Bayesian ideas and methodology.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin, *Bayesian Data Analysis* (2nd ed.), Chapman and Hall, London, 2003. An up-to-date survey of theoretical, practical, and computational issues in Bayesian inference.

Hollander, Myles, and Douglas Wolfe, *Nonparametric Statistical Methods* (2nd ed.), Wiley, New York, 1999. A very good reference on distribution-free methods with an excellent collection of tables.

Lehmann, Erich, *Nonparametrics: Statistical Methods Based on Ranks* (revised ed.), Springer, New York, 2006. An excellent discussion of the most important distribution-free methods, presented with a great deal of insightful commentary.