CHAPTER 11

Comparing Two Samples

11.1 Introduction

This chapter is concerned with methods for comparing samples from distributions that may be different and especially with methods for making inferences about how the distributions differ. In many applications, the samples are drawn under different conditions, and inferences must be made about possible effects of these conditions. We will be primarily concerned with effects that tend to increase or decrease the average level of response.

For example, in the end-of-chapter problems, we will consider some experiments performed to determine to what degree, if any, cloud seeding increases precipitation. In cloud-seeding experiments, some storms are selected for seeding, other storms are left unseeded, and the amount of precipitation from each storm is measured. This amount varies widely from storm to storm, and in the face of this natural variability, it is difficult to tell whether seeding has a systematic effect. The average precipitation from the seeded storms might be slightly higher than that from the unseeded storms, but a skeptic might not be convinced that the difference was due to anything but chance. We will develop statistical methods to deal with this type of problem based on a stochastic model that treats the amounts of precipitation as random variables. We will also see how a process of randomization allows us to make inferences about treatment effects even in the case where the observations are not modeled as samples from populations or probability laws.

This chapter will be concerned with analyzing measurements that are continuous in nature (such as temperature); Chapter 13 will take up the analysis of qualitative data. This chapter will conclude with some general discussion of the design and interpretation of experimental studies.

11.2 Comparing Two Independent Samples

In many experiments, the two samples may be regarded as being independent of each other. In a medical study, for example, a sample of subjects may be assigned to a particular treatment, and another independent sample may be assigned to a control (or placebo) treatment. This is often accomplished by randomly assigning individuals to the placebo and treatment groups. In later sections, we will discuss methods that are appropriate when there is some pairing, or dependence, between the samples, such as might occur if each person receiving the treatment were paired with an individual of similar weight in the control group.

Many experiments are such that if they were repeated, the measurements would not be exactly the same. To deal with this problem, a statistical model is often employed: The observations from the control group are modeled as independent random variables with a common distribution, F, and the observations from the treatment group are modeled as being independent of each other and of the controls and as having their own common distribution function, G. Analyzing the data thus entails making inferences about the comparison of F and G. In many experiments, the primary effect of the treatment is to change the overall level of the responses, so that analysis focuses on the difference of means or other location parameters of F and G. When only a small amount of data is available, it may not be practical to do much more than this.

11.2.1 Methods Based on the Normal Distribution

In this section, we will assume that a sample, $X_1, \ldots, X_n$, is drawn from a normal distribution that has mean μ_X and variance σ^2, and that an independent sample, $Y_1, \ldots, Y_m$, is drawn from another normal distribution that has mean μ_Y and the same variance, σ^2. If we think of the X's as having received a treatment and the Y's as being the control group, the effect of the treatment is characterized by the difference $\mu_X - \mu_Y$. A natural estimate of $\mu_X - \mu_Y$ is $\overline{X} - \overline{Y}$; in fact, this is the mle. Since $\overline{X} - \overline{Y}$ may be expressed as a linear combination of independent normally distributed random variables, it is normally distributed:

$$\overline{X} - \overline{Y} \sim N\left[\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right]$$

If σ^2 were known, a confidence interval for $\mu_X - \mu_Y$ could be based on

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

which follows a standard normal distribution. The confidence interval would be of the form

$$(\overline{X} - \overline{Y}) \pm z(\alpha/2)\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

This confidence interval is of the same form as those introduced in Chapters 7 and 8—a statistic ($\overline{X} - \overline{Y}$ in this case) plus or minus a multiple of its standard deviation.

Generally, σ^2 will not be known and must be estimated from the data by calculating the **pooled sample variance,**

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$

where $s_X^2 = (n-1)\sum_{i=1}^{n}(X_i - \overline{X})^2$ and similarly for s_Y^2. Note that s_p^2 is a weighted average of the sample variances of the X's and Y's, with the weights proportional to the degrees of freedom. This weighting is appropriate since if one sample is much larger than the other, the estimate of σ^2 from that sample is more reliable and should receive greater weight. The following theorem gives the distribution of a statistic that will be used for forming confidence intervals and performing hypothesis tests.

THEOREM **A**

Suppose that $X_1, \ldots, X_n$ are independent and normally distributed random variables with mean μ_X and variance σ^2, and that $Y_1, \ldots, Y_m$ are independent and normally distributed random variables with mean μ_Y and variance σ^2, and that the Y_i are independent of the X_i. The statistic

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

follows a t distribution with $m + n - 2$ degrees of freedom.

Proof

According to the definition of the t distribution in Section 6.2, we have to show that the statistic is the quotient of a standard normal random variable and the square root of an independent chi-square random variable divided by its $n + m - 2$ degrees of freedom. First, we note from Theorem B in Section 6.3 that $(n-1)s_X^2/\sigma^2$ and $(m-1)s_Y^2/\sigma^2$ are distributed as chi-square random variables with $n-1$ and $m-1$ degrees of freedom, respectively, and are independent since the X_i and Y_i are. Their sum is thus chi-square with $m + n - 2$ df. Now, we express the statistic as the ratio U/V, where

$$U = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

$$V = \sqrt{\left[\frac{(n-1)s_X^2}{\sigma^2} + \frac{(m-1)s_Y^2}{\sigma^2}\right]\frac{1}{m+n-2}}$$

U follows a standard normal distribution and from the preceding argument V has the distribution of the square root of a chi-square random variable divided by its degrees of freedom. The independence of U and V follows from Corollary A in Section 6.3. ■

It is convenient and suggestive to use the following notation for the estimated standard deviation (or standard error) of $\overline{X} - \overline{Y}$:

$$s_{\overline{X}-\overline{Y}} = s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

A confidence interval for $\mu_X - \mu_Y$ follows as a corollary to Theorem A.

COROLLARY **A**

Under the assumptions of Theorem A, a $100(1 - \alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is

$$(\overline{X} - \overline{Y}) \pm t_{m+n-2}(\alpha/2) s_{\overline{X}-\overline{Y}}$$ ■

EXAMPLE **A** Two methods, A and B, were used in a determination of the latent heat of fusion of ice (Natrella 1963). The investigators wanted to find out by how much the methods differed. The following table gives the change in total heat from ice at $-.72^\circ$C to water 0°C in calories per gram of mass:

Method A	Method B
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

It is fairly obvious from the table and from boxplots (Figure 11.1) that there is a difference between the two methods (we will test this more formally later). If we assume the conditions of Theorem A, we can form a 95% confidence interval to estimate the magnitude of the average difference between the two methods. From the table, we calculate

$$\overline{X}_A = 80.02 \qquad S_a = .024$$
$$\overline{X}_B = 79.98 \qquad S_b = .031$$
$$s_p^2 = \frac{12 \times S_a^2 + 7 \times S_b^2}{19} = .0007178$$
$$s_p = .027$$

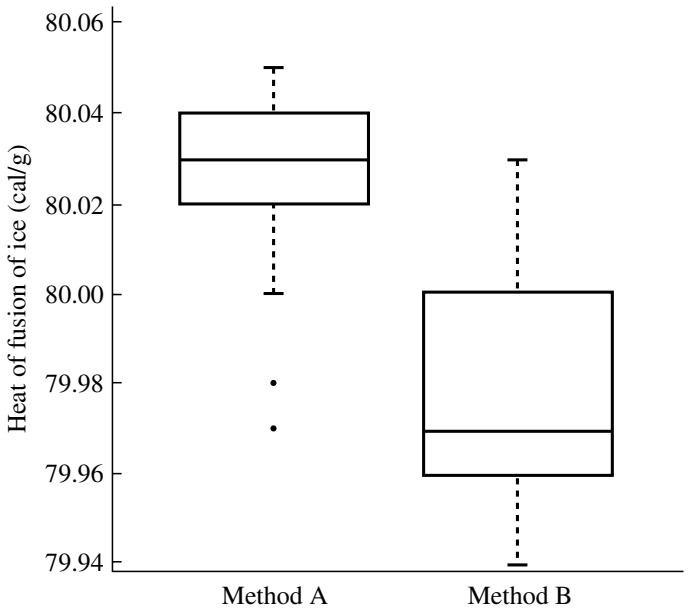

FIGURE **11.1** Boxplots of measurements of heat of fusion obtained by methods A and B.

Our estimate of the average difference of the two methods is $\overline{X}_A - \overline{X}_B = .04$ and its estimated standard error is

$$s_{\overline{X}_A - \overline{X}_B} = s_p \sqrt{\frac{1}{13} + \frac{1}{8}}$$
$$= .012$$

From Table 4 of Appendix B, the .975 quantile of the t distribution with 19 df is 2.093, so $t_{19}(.025) = 2.093$ and the 95% confidence interval is $(\overline{X}_A - \overline{X}_B) \pm t_{19}(.025)s_{\overline{X}_A - \overline{X}_B}$, or (.015, .065). The estimated standard error and the confidence interval quantify the uncertainty in the point estimate $\overline{X}_A - \overline{X}_B = .04$. ■

We will now discuss hypothesis testing for the two-sample problem. Although the hypotheses under consideration are different from those of Chapter 9, the general conceptual framework is the same (you should review that framework at this time). In the current case, the null hypothesis to be tested is

$$H_0: \mu_X = \mu_Y$$

This asserts that there is no difference between the distributions of the X's and Y's. If one group is a treatment group and the other a control, for example, this hypothesis asserts that there is no treatment effect. In order to conclude that there is a treatment effect, the null hypothesis must be rejected.

There are three common alternative hypotheses for the two-sample case:

$$H_1: \mu_X \neq \mu_Y$$
$$H_2: \mu_X > \mu_Y$$
$$H_3: \mu_X < \mu_Y$$

The first of these is a **two-sided alternative,** and the other two are **one-sided alternatives.** The first hypothesis is appropriate if deviations could in principle go in either direction, and one of the latter two is appropriate if it is believed that any deviation must be in one direction or the other. In practice, such a priori information is not usually available, and it is more prudent to conduct two-sided tests, as in Example A.

The test statistic that will be used to make a decision whether or not to reject the null hypothesis is

$$t = \frac{\overline{X} - \overline{Y}}{s_{\overline{X}-\overline{Y}}}$$

The t-statistic equals the multiple of its estimated standard deviation that $\overline{X} - \overline{Y}$ differs from zero. It plays the same role in the comparison of two samples as is played by the chi-square statistic in testing goodness of fit. Just as we rejected for large values of the chi-square statistic, we will reject in this case for extreme values of t. The distribution of t under H_0, its null distribution, is, from Theorem A, the t distribution with $m + n - 2$ degrees of freedom. Knowing this null distribution allows us to determine a rejection region for a test at level α, just as knowing that the null distribution of the chi-square statistic was chi-square with the appropriate degrees of freedom allowed the determination of a rejection region for testing goodness of fit. The rejection regions for the three alternatives just listed are

$$\text{For } H_1, |t| > t_{n+m-2}(\alpha/2)$$
$$\text{For } H_2, t > t_{n+m-2}(\alpha)$$
$$\text{For } H_3, t < -t_{n+m-2}(\alpha)$$

Note how the rejection regions are tailored to the particular alternatives and how knowing the null distribution of t allows us to determine the rejection region for any value of α.

EXAMPLE **B** Let us continue Example A. To test H_0: $\mu_A = \mu_B$ versus a two-sided alternative, we form and calculate the following test statistic:

$$t = \frac{\overline{X}_A - \overline{X}_B}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$
$$= 3.33$$

From Table 4 in Appendix B, $t_{19}(.005) = 2.861 < 3.33$. The two-sided test would thus reject at the level $\alpha = .01$. If there were no difference in the two conditions, differences as large or larger than that observed would occur only with probability less than .01—that is, the p-value is less than .01. There is little doubt that there is a difference between the two methods. ■

In Chapter 9, we developed a general duality between hypothesis tests and confidence intervals. In the case of the testing and confidence interval methods considered

in this section, the t test rejects if and only if the confidence interval does not include zero (see Problem 10 at the end of this chapter).

We will now demonstrate that the test of H_0 versus H_1 is equivalent to a likelihood ratio test. (The rather long argument is sketched here and should be read with paper and pencil in hand.) Ω is the set of all possible parameter values:

$$\Omega = \{-\infty < \mu_X < \infty, -\infty < \mu_Y < \infty, 0 < \sigma < \infty\}$$

The unknown parameters are $\theta = (\mu_X, \mu_Y, \sigma)$. Under H_0, $\theta \in \omega_0$, where $\omega_0 = \{\mu_X = \mu_Y, 0 < \sigma < \infty\}$. The likelihood of the two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ is

$$\text{lik}\left(\mu_X, \mu_Y, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2)[(X_i-\mu_X)^2/\sigma^2]} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2)[(Y_j-\mu_Y)^2/\sigma^2]}$$

and the log likelihood is

$$l\left(\mu_X, \mu_Y, \sigma^2\right) = -\frac{(m+n)}{2}\log 2\pi - \frac{(m+n)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i-\mu_X)^2 + \sum_{j=1}^{m}(Y_j-\mu_Y)^2\right]$$

We must maximize the likelihood under ω_0 and under Ω and then calculate the ratio of the two maximized likelihoods, or the difference of their logarithms.

Under ω_0, we have a sample of size $m + n$ from a normal distribution with unknown mean μ_0 and unknown variance σ_0^2. The mle's of μ_0 and σ_0^2 are thus

$$\hat{\mu}_0 = \frac{1}{m+n}\left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j\right)$$

$$\hat{\sigma}_0^2 = \frac{1}{m+n}\left[\sum_{i=1}^{n}(X_i-\hat{\mu}_0)^2 + \sum_{j=1}^{m}(Y_j-\hat{\mu}_0)^2\right]$$

The corresponding value of the maximized log likelihood is, after some cancellation,

$$l\left(\hat{\mu}_0, \hat{\sigma}_0^2\right) = -\frac{m+n}{2}\log 2\pi - \frac{m+n}{2}\log\hat{\sigma}_0^2 - \frac{m+n}{2}$$

To find the mle's $\hat{\mu}_X$, $\hat{\mu}_Y$, and $\hat{\sigma}_1^2$ under Ω, we first differentiate the log likelihood and obtain the equations

$$\sum_{i=1}^{n}(X_i-\hat{\mu}_X) = 0$$

$$\sum_{j=1}^{m}(Y_j-\hat{\mu}_Y) = 0$$

$$-\frac{m+n}{2\hat{\sigma}_1^2} + \frac{1}{2\hat{\sigma}_1^4}\left[\sum_{i=1}^{n}(X_i-\hat{\mu}_X)^2 + \sum_{j=1}^{m}(Y_j-\hat{\mu}_Y)^2\right] = 0$$

The mle's are, therefore,

$$\hat{\mu}_X = \overline{X}$$
$$\hat{\mu}_Y = \overline{Y}$$
$$\hat{\sigma}_1^2 = \frac{1}{m+n}\left[\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2 + \sum_{j=1}^{m}(Y_j - \hat{\mu}_Y)^2\right]$$

When these are substituted into the log likelihood, we obtain

$$l\left(\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_1^2\right) = -\frac{m+n}{2}\log 2\pi - \frac{m+n}{2}\log\hat{\sigma}_1^2 - \frac{m+n}{2}$$

The log of the likelihood ratio is thus

$$\frac{m+n}{2}\log\left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)$$

and the likelihood ratio test rejects for large values of

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{m}(Y_j - \hat{\mu}_0)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{j=1}^{m}(Y_j - \overline{Y})^2}$$

We now find an alternative expression for the numerator of this ratio, by using the identities

$$\sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \hat{\mu}_0)^2$$
$$\sum_{j=1}^{m}(Y_j - \hat{\mu}_0)^2 = \sum_{j=1}^{m}(Y_j - \overline{Y})^2 + m(\overline{Y} - \hat{\mu}_0)^2$$

We obtain

$$\hat{\mu}_0 = \frac{1}{m+n}(n\overline{X} + m\overline{Y})$$
$$= \frac{n}{m+n}\overline{X} + \frac{m}{m+n}\overline{Y}$$

Therefore,

$$\overline{X} - \hat{\mu}_0 = \frac{m(\overline{X} - \overline{Y})}{m+n}$$
$$\overline{Y} - \hat{\mu}_0 = \frac{n(\overline{Y} - \overline{X})}{m+n}$$

The alternative expression for the numerator of the ratio is thus

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{j=1}^{m}(Y_j - \overline{Y})^2 + \frac{mn}{m+n}(\overline{X} - \overline{Y})^2$$

and the test rejects for large values of

$$1 + \frac{mn}{m+n}\left(\frac{(\overline{X}-\overline{Y})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{j=1}^{m}(Y_j-\overline{Y})^2}\right)$$

or, equivalently, for large values of

$$\frac{|\overline{X}-\overline{Y}|}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{j=1}^{m}(Y_j-\overline{Y})^2}}$$

which is the t statistic apart from constants that do not depend on the data. Thus, the likelihood ratio test is equivalent to the t test, as claimed.

We have used the assumption that the two populations have the same variance. If the two variances are not assumed to be equal, a natural estimate of $\mathrm{Var}(\overline{X}-\overline{Y})$ is

$$\frac{s_X^2}{n}+\frac{s_Y^2}{m}$$

If this estimate is used in the denominator of the t statistic, the distribution of that statistic is no longer the t distribution. But it has been shown that its distribution can be closely approximated by the t distribution with degrees of freedom calculated in the following way and then rounded to the nearest integer:

$$\mathrm{df} = \frac{[(s_X^2/n)+(s_Y^2/m)]^2}{\frac{(s_X^2/n)^2}{n-1}+\frac{(s_Y^2/m)^2}{m-1}}$$

EXAMPLE **C** Let us rework Example B, but without the assumption that the variances are equal. Using the preceding formula, we find the degrees of freedom to be 12 rather than 19. The t statistic is 3.12. Since the .995 quantile of the t distribution with 12 df is 3.055 (Table 4 of Appendix B), the test still rejects at level $\alpha = .01$. ■

If the underlying distributions are not normal and the sample sizes are large, the use of the t distribution or the normal distribution is justified by the central limit theorem, and the probability levels of confidence intervals and hypothesis tests are approximately valid. In such a case, however, there is little difference between the t and normal distributions. If the sample sizes are small, however, and the distributions are not normal, conclusions based on the assumption of normality may not be valid. Unfortunately, if the sample sizes are small, the assumption of normality cannot be tested effectively unless the deviation is quite gross, as we saw in Chapter 9.

11.2.1.1 An Example—A Study of Iron Retention An experiment was performed to determine whether two forms of iron (Fe^{2+} and Fe^{3+}) are retained differently. (If one form of iron were retained especially well, it would be the better dietary supplement.) The investigators divided 108 mice randomly into 6 groups of 18 each; 3 groups were given Fe^{2+} in three different concentrations, 10.2, 1.2, and

.3 millimolar, and 3 groups were given Fe^{3+} at the same three concentrations. The mice were given the iron orally; the iron was radioactively labeled so that a counter could be used to measure the initial amount given. At a later time, another count was taken for each mouse, and the percentage of iron retained was calculated. The data for the two forms of iron are listed in the following table. We will look at the data for the concentration 1.2 millimolar. (In Chapter 12, we will discuss methods for analyzing all the groups simultaneously.)

Fe^{3+}			Fe^{2+}		
10.2	1.2	.3	10.2	1.2	.3
.71	2.20	2.25	2.20	4.04	2.71
1.66	2.93	3.93	2.69	4.16	5.43
2.01	3.08	5.08	3.54	4.42	6.38
2.16	3.49	5.82	3.75	4.93	6.38
2.42	4.11	5.84	3.83	5.49	8.32
2.42	4.95	6.89	4.08	5.77	9.04
2.56	5.16	8.50	4.27	5.86	9.56
2.60	5.54	8.56	4.53	6.28	10.01
3.31	5.68	9.44	5.32	6.97	10.08
3.64	6.25	10.52	6.18	7.06	10.62
3.74	7.25	13.46	6.22	7.78	13.80
3.74	7.90	13.57	6.33	9.23	15.99
4.39	8.85	14.76	6.97	9.34	17.90
4.50	11.96	16.41	6.97	9.91	18.25
5.07	15.54	16.96	7.52	13.46	19.32
5.26	15.89	17.56	8.36	18.4	19.87
8.15	18.3	22.82	11.65	23.89	21.60
8.24	18.59	29.13	12.45	26.39	22.25

As a summary of the data, boxplots (Figure 11.2) show that the data are quite skewed to the right. This is not uncommon with percentages or other variables that are bounded below by zero. Three observations from the Fe^{2+} group are flagged as possible outliers. The median of the Fe^{2+} group is slightly larger than the median of the Fe^{3+} groups, but the two distributions overlap substantially.

Another view of these data is provided by normal probability plots (Figure 11.3). These plots also indicate the skewness of the distributions. We should obviously doubt the validity of using normal distribution theory (for example, the t test) for this problem even though the combined sample size is fairly large (36).

The mean and standard deviation of the Fe^{2+} group are 9.63 and 6.69; for the Fe^{3+} group, the mean is 8.20 and the standard deviation is 5.45. To test the hypothesis that the two means are equal, we can use a t test without assuming that the population standard deviations are equal. The approximate degrees of freedom, calculated as described at the end of Section 11.2.1, are 32. The t statistic is .702, which corresponds to a p-value of .49 for a two-sided test; if the two populations had the same mean, values of the t statistic this large or larger would occur 49% of the time. There is thus insufficient evidence to reject the null hypothesis. A 95% confidence interval for the

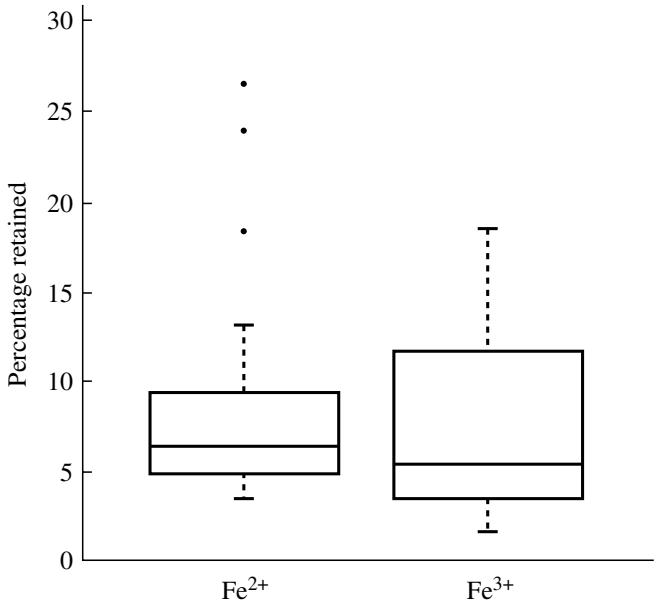

FIGURE **11.2** Boxplots of the percentages of iron retained for the two forms.

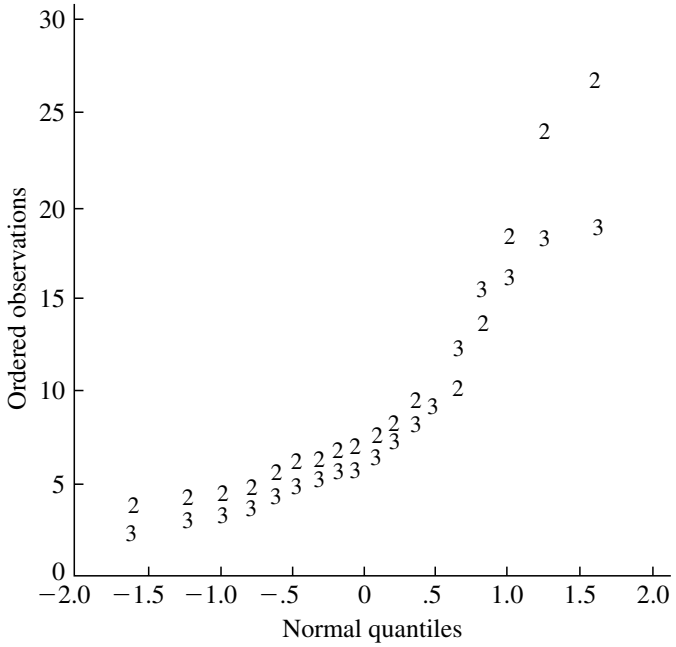

FIGURE **11.3** Normal probability plots of iron retention data.

difference of the two population means is $(-2.7, 5.6)$. But the t test assumes that the underlying populations are normally distributed, and we have seen there is reason to doubt this assumption.

It is sometimes advocated that skewed data be transformed to a more symmetric shape before normal theory is applied. Transformations such as taking the log or

the square root can be effective in symmetrizing skewed distributions because they spread out small values and compress large ones. Figures 11.4 and 11.5 show boxplots and normal probability plots for the natural logs of the iron retention data we have been considering. The transformation was fairly successful in symmetrizing these distributions, and the probability plots are more linear than those in Figure 11.3, although some curvature is still evident.

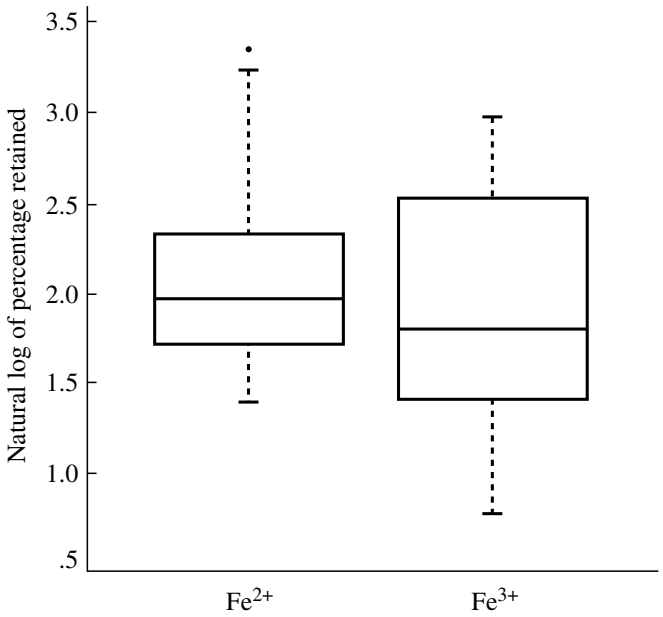

FIGURE **11.4** Boxplots of natural logs of percentages of iron retained.

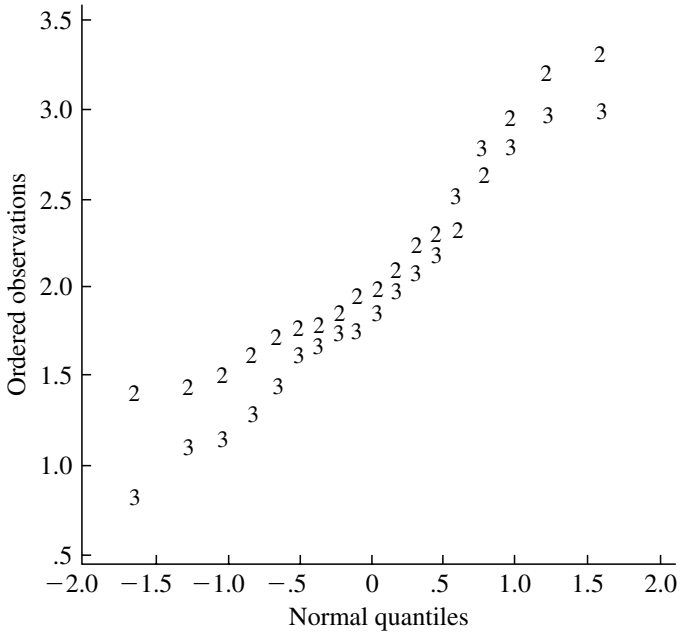

FIGURE **11.5** Normal probability plots of natural logs of iron retention data.

The following model is natural for the log transformation:

$$\begin{aligned} X_i &= \mu_X(1+\varepsilon_i), \qquad i = 1, \ldots, n \\ Y_j &= \mu_Y(1+\delta_j), \qquad j = 1, \ldots, m \\ \log X_i &= \log \mu_X + \log(1+\varepsilon_i) \\ \log Y_j &= \log \mu_Y + \log(1+\delta_j) \end{aligned}$$

Here the ε_i and δ_j are independent random variables with mean zero. This model implies that if the variances of the errors are σ^2, then

$$\begin{aligned} E(X_i) &= \mu_X \\ E(Y_j) &= \mu_Y \\ \sigma_X &= \mu_X \sigma \\ \sigma_Y &= \mu_Y \sigma \end{aligned}$$

or that

$$\frac{\sigma_X}{\mu_X} = \frac{\sigma_Y}{\mu_Y}$$

If the ε_i and δ_j have the same distribution, $\text{Var}(\log X) = \text{Var}(\log Y)$. The ratio of the standard deviation of a distribution to the mean is called the **coefficient of variation (CV);** it expresses the standard deviation as a fraction of the mean. Coefficients of variation are sometimes expressed as percentages. For the iron retention data we have been considering, the CV's are .69 and .67 for the Fe^{2+} and Fe^{3+} groups; these values are quite close. These data are quite "noisy"—the standard deviation is nearly 70% of the mean for both groups.

For the transformed iron retention data, the means and standard deviations are given in the following table:

	Fe^{2+}	Fe^{3+}
Mean	2.09	1.90
Standard Deviation	.659	.574

For the transformed data, the t statistic is .917, which gives a p-value of .37. Again, there is no reason to reject the null hypothesis. A 95% confidence interval is $(-.61, .23)$. Using the preceding model, this is a confidence interval for

$$\log \mu_X - \log \mu_Y = \log\left(\frac{\mu_X}{\mu_Y}\right)$$

The interval is

$$-.61 \le \log\left(\frac{\mu_X}{\mu_Y}\right) \le .23$$

or

$$.54 \le \frac{\mu_X}{\mu_Y} \le 1.26$$

Other transformations, such as raising all values to some power, are sometimes used. Attitudes toward the use of transformations vary: Some view them as a very

useful tool in statistics and data analysis, and others regard them as questionable manipulation of the data.

11.2.2 Power

Calculations of power are an important part of planning experiments in order to determine how large sample sizes should be. The power of a test is the probability of rejecting the null hypothesis when it is false. The power of the two-sample t test depends on four factors:

1. The real difference, $\Delta = |\mu_X - \mu_Y|$. The larger this difference, the greater the power.
2. The significance level α at which the test is done. The larger the significance level, the more powerful the test.
3. The population standard deviation σ, which is the amplitude of the "noise" that hides the "signal." The smaller the standard deviation, the larger the power.
4. The sample sizes n and m. The larger the sample sizes, the greater the power.

Before continuing, you should try to understand intuitively why these statements are true. We will express them quantitatively below.

The necessary sample sizes can be determined from the significance level of the test, the standard deviation, and the desired power against an alternative hypothesis,

$$H_1: \mu_X - \mu_Y = \Delta$$

To calculate the power of a t test exactly, special tables of the noncentral t distribution are required. But if the sample sizes are reasonably large, one can perform approximate power calculations based on the normal distribution, as we will now demonstrate.

Suppose that σ, α, and Δ are given and that the samples are both of size n. Then

$$\begin{aligned} \text{Var}(\overline{X} - \overline{Y}) &= \sigma^2 \left(\frac{1}{n} + \frac{1}{n} \right) \\ &= \frac{2\sigma^2}{n} \end{aligned}$$

The test at level α of $H_0: \mu_X = \mu_Y$ against the alternative $H_1: \mu_X \neq \mu_Y$ is based on the test statistic

$$Z = \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{2/n}}$$

The rejection region for this test is $|Z| > z(\alpha/2)$, or

$$|\overline{X} - \overline{Y}| > z(\alpha/2)\sigma\sqrt{\frac{2}{n}}$$

The power of the test if $\mu_X - \mu_Y = \Delta$ is the probability that the test statistic falls in the rejection region, or

$$P\left[|\overline{X} - \overline{Y}| > z(\alpha/2)\sigma\sqrt{\frac{2}{n}}\right]$$
$$= P\left[\overline{X} - \overline{Y} > z(\alpha/2)\sigma\sqrt{\frac{2}{n}}\right] + P\left[\overline{X} - \overline{Y} < -z(\alpha/2)\sigma\sqrt{\frac{2}{n}}\right]$$

since the two events are mutually exclusive. Both probabilities on the right-hand side are calculated by standardizing. For the first one, we have

$$P\left[\overline{X} - \overline{Y} > z(\alpha/2)\sigma\sqrt{\frac{2}{n}}\right] = P\left[\frac{(\overline{X} - \overline{Y}) - \Delta}{\sigma\sqrt{2/n}} > \frac{z(\alpha/2)\sigma\sqrt{2/n} - \Delta}{\sigma\sqrt{2/n}}\right]$$
$$= 1 - \Phi\left[z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]$$

where Φ is the standard normal cdf. Similarly, the second probability is

$$\Phi\left[-z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]$$

Thus, the probability that the test statistic falls in the rejection region is equal to

$$1 - \Phi\left[z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right] + \Phi\left[-z(\alpha/2) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]$$

Typically, as Δ moves away from zero, one of these terms will be negligible with respect to the other. For example, if Δ is greater than zero, the first term will be dominant. For fixed n, this expression can be evaluated as a function of Δ; or for fixed Δ, it can be evaluated as a function of n.

EXAMPLE **A** As an example, let us consider a situation similar to an idealized form of the iron retention experiment. Assume that we have samples of size 18 from two normal distributions whose standard deviations are both 5, and we calculate the power for various values of Δ when the null hypothesis is tested at a significance level of .05. The results of the calculations are displayed in Figure 11.6. We see from the plot that if the mean difference in retention is only 1%, the probability of rejecting the null hypothesis is quite small, only 9%. A mean difference of 5% in retention rate gives a more satisfactory power of 85%.

Suppose that we wanted to be able to detect a difference of $\Delta = 1$ with probability .9. What sample size would be necessary? Using only the dominant term in the expression for the power, the sample size should be such that

$$\Phi\left(1.96 - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right) = .1$$

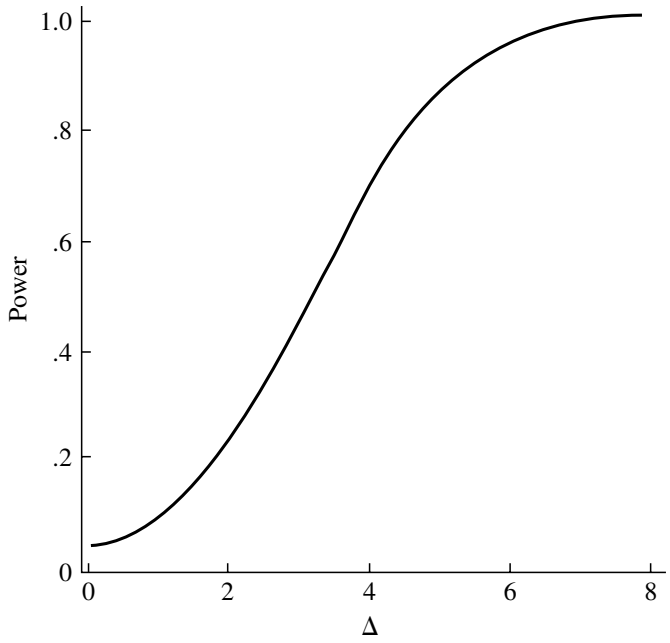

FIGURE **11.6** Plot of power versus Δ.

From the tables for the normal distribution, $.1 = \Phi(-1.28)$, so

$$1.96 - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}} = -1.28$$

Solving for n, we find that the necessary sample size would be 525! This is clearly unfeasible; if in fact the experimenters wanted to detect such a difference, some modification of the experimental technique to reduce σ would be necessary. ■

11.2.3 A Nonparametric Method—The Mann-Whitney Test

Nonparametric methods do not assume that the data follow any particular distributional form. Many of them are based on replacement of the data by ranks. With this replacement, the results are invariant under any monotonic transformation; in comparison, we saw that the p-value of a t test may change if the log of the measurements is analyzed rather than the measurements on the original scale. Replacing the data by ranks also has the effect of moderating the influence of outliers.

For purposes of discussion, we will develop the **Mann-Whitney test** (also sometimes called the Wilcoxon rank sum test) in a specific context. Suppose that we have $m + n$ experimental units to assign to a treatment group and a control group. The assignment is made at random: n units are randomly chosen and assigned to the control, and the remaining m units are assigned to the treatment. We are interested in testing the null hypothesis that the treatment has no effect. If the null hypothesis is true, then any difference in the outcomes under the two conditions is due to the randomization.

A test statistic is calculated in the following way. First, we group all $m + n$ observations together and rank them in order of increasing size (we will assume for simplicity that there are no ties, although the argument holds even in the presence of ties). We next calculate the sum of the ranks of those observations that came from the control group. If this sum is too small or too large, we will reject the null hypothesis.

It is easiest to see how the procedure works by considering a very small example. Suppose that a treatment and a control are to be compared: Of four subjects, two are randomly assigned to the treatment and the other two to the control, and the following responses are observed (the ranks of the observations are shown in parentheses):

Treatment	Control
1 (1)	6 (4)
3 (2)	4 (3)

The sum of the ranks of the control group is $R = 7$, and the sum of the ranks of the treatment group is 3. Does this discrepancy provide convincing evidence of a systematic difference between treatment and control, or could it be just due to chance? To answer this question, we calculate the probability of such a discrepancy if the treatment had no effect at all, so that the difference was entirely due to the particular randomization—this is the null hypothesis. The key idea of the Mann-Whitney test is that we can explicitly calculate the distribution of R under the null hypothesis, since under this hypothesis every assignment of ranks to observations is equally likely and we can enumerate all $4! = 24$ such assignments. In particular, each of the $\binom{7}{2} = 6$ assignments of ranks to the control group shown in the following table is equally likely:

Ranks	R
$\{1, 2\}$	3
$\{1, 3\}$	4
$\{1, 4\}$	5
$\{2, 3\}$	5
$\{2, 4\}$	6
$\{3, 4\}$	7

From this table, we see that under the null hypothesis, the distribution of R (its null distribution) is:

r	3	4	5	6	7
$P(R = r)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

In particular, $P(R = 7) = \frac{1}{6}$, so this discrepancy would occur one time out of six purely on the basis of chance.

The small example of the previous paragraph has been laid out for pedagogical reasons, the point being that we could in principle go through similar calculations for any sample sizes m and n. Suppose that there are n observations in the treatment group and m in the control group. If the null hypothesis holds, every assignment of ranks to the $m + n$ observations is equally likely, and hence each of the $\binom{m+n}{m}$ possible assignments of ranks to the control group is equally likely. For each of these assignments, we can calculate the sum of the ranks and thus determine the null distribution of the test statistic—the sum of the ranks of the control group.

It is important to note that we have not made any assumption that the observations from the control and treatment groups are samples from a probability distribution. Probability has entered in only as a result of the random assignment of experimental units to treatment and control groups (this is similar to the way that probability enters into survey sampling). We should also note that, although we chose the sum of control ranks as the test statistic, any other test statistic could have been used and its null distribution computed in the same fashion. The rank sum is easy to compute and is sensitive to a treatment effect that tends to make responses larger or smaller. Also, its null distribution has to be computed only once and tabled; if we worked with the actual numerical values, the null distribution would depend on those particular values.

Tables of the null distribution of the rank sum are widely available and vary in format. Note that because the sum of the two rank sums is the sum of the integers from 1 to $m + n$, which is $[(m + n)(m + n + 1)/2]$, knowing one rank sum tells us the other. Some tables are given in terms of the rank sum of the smaller of the two groups, and some are in terms of the smaller of the two rank sums (the advantage of the latter scheme is that only one tail of the distribution has to be tabled). Table 8 of Appendix B makes use of additional symmetries. Let n_1 be the smaller sample size and let R be the sum of the ranks from that sample. Let $R' = n_1(m + n + 1) - R$ and $R^* = \min(R, R')$. The table gives critical values for R^*. (Fortunately, such fussy tables are largely obsolete with the increasing use of computers.)

When it is more appropriate to model the control values, $X_1, \ldots, X_n$, as a sample from some probability distribution F and the experimental values, $Y_1, \ldots, Y_m$, as a sample from some distribution G, the Mann-Whitney test is a test of the null hypothesis H_0: $F = G$. The reasoning is exactly the same: Under H_0, any assignment of ranks to the pooled $m + n$ observations is equally likely, etc.

We have assumed here that there are no ties among the observations. If there are only a small number of ties, tied observations are assigned average ranks (the average of the ranks for which they are tied); the significance levels are not greatly affected.

EXAMPLE **A** Let us illustrate the Mann-Whitney test by referring to the data on latent heats of fusion of ice considered earlier (Example A in Section 11.2.1). The sample sizes are fairly small (13 and 8), so in the absence of any prior knowledge concerning the adequacy of the assumption of a normal distribution, it would seem safer to use a nonparametric

method. The following table exhibits the ranks given to the measurements for each method (refer to Example A in Section 11.2.1 for the original data):

Method A	Method B
7.5	11.5
19.0	1.0
11.5	7.5
19.0	4.5
15.5	4.5
15.5	15.5
19.0	2.0
4.5	4.5
21.0	
15.5	
11.5	
9.0	
11.5	

Note how the ties were handled. For example, the four observations with the value 79.97 tied for ranks 3, 4, 5, and 6 were each assigned the rank of $4.5 = (3 + 4 + 5 + 6)/4$.

Table 8 of Appendix B is used as follows. The sum of the ranks of the smaller sample is $R = 51$.

$$\begin{aligned} R' &= 8(8 + 13 + 1) - R \\ &= 125 \end{aligned}$$

Thus, $R^* = 51$. From the table, 53 is the critical value for a two-tailed test with $\alpha = .01$, and 60 is the critical value for $\alpha = .05$. The Mann-Whitney test thus rejects at the .01 significance level. ■

Let T_Y denote the sum of the ranks of $Y_1, Y_2, \ldots, Y_m$. Using results from Chapter 7, we can easily find $E(T_Y)$ and $\text{Var}(T_Y)$ under the null hypothesis $F = G$.

THEOREM A

If $F = G$,

$$E(T_Y) = \frac{m(m + n + 1)}{2}$$

$$\text{Var}(T_Y) = \frac{mn(m + n + 1)}{12}$$

Proof

Under the null hypothesis, T_Y is the sum of a random sample of size m drawn without replacement from a population consisting of the integers $\{1, 2, \ldots, m+n\}$. T_Y thus equals m times the average of such a sample. From Theorems A and B of Section 7.3.1,

$$E(T_Y) = m\mu$$
$$\text{Var}(T_Y) = m\sigma^2\left(\frac{N-m}{N-1}\right)$$

where $N = m + n$ is the size of the population, and μ and σ^2 are the population mean and variance. Now, using the identities

$$\sum_{k=1}^{N} k = \frac{N(N+1)}{2}$$
$$\sum_{k=1}^{N} k^2 = \frac{N(N+1)(2N+1)}{6}$$

we find that for the population $\{1, 2, \ldots, m+n\}$

$$\mu = \frac{N+1}{2}$$
$$\sigma^2 = \frac{N^2-1}{12}$$

The result then follows after algebraic simplification. ■

Unlike the t test, the Mann-Whitney test does not depend on an assumption of normality. Since the actual numerical values are replaced by their ranks, the test is insensitive to outliers, whereas the t test is sensitive. It has been shown that even when the assumption of normality holds, the Mann-Whitney test is nearly as powerful as the t test and it is thus generally preferable, especially for small sample sizes.

The Mann-Whitney test can also be derived starting from a different point of view. Suppose that the X's are a sample from F and the Y's a sample from G, and consider estimating, as a measure of the effect of the treatment,

$$\pi = P(X < Y)$$

where X and Y are independently distributed with distribution functions F and G, respectively. The value π is the probability that an observation from the distribution F is smaller than an independent observation from the distribution G.

If, for example, F and G represent lifetimes of components that have been manufactured according to two different conditions, π is the probability that a component of one type will last longer than a component of the other type. An estimate of π can be obtained by comparing all n values of X to all m values of Y and calculating the

proportion of the comparisons for which X was less than Y:

$$\hat{\pi} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}$$

where

$$Z_{ij} = \begin{cases} 1, & \text{if } X_i < Y_j \\ 0, & \text{otherwise} \end{cases}$$

To see the relationship of $\hat{\pi}$ to the rank sum introduced earlier, we will find it convenient to work with

$$V_{ij} = \begin{cases} 1, & \text{if } X_{(i)} < Y_{(j)} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$\sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij}$$

since the V_{ij} are just a reordering of the Z_{ij}. Also,

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} = \ & (\text{number of } X\text{'s that are less than } Y_{(1)}) \\ & + (\text{number of } X\text{'s that are less than } Y_{(2)}) \\ & + \cdots + \ (\text{number of } X\text{'s that are less than } Y_{(m)}) \end{aligned}$$

If the rank of $Y_{(k)}$ in the combined sample is denoted by R_{yk}, then the number of X's less than $Y_{(1)}$ is $R_{y1} - 1$, the number of X's less than $Y_{(2)}$ is $R_{y2} - 2$, etc. Therefore,

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} &= (R_{y1} - 1) + (R_{y2} - 2) + \cdots + (R_{ym} - m) \\ &= \sum_{i=1}^{m} R_{yi} - \sum_{i=1}^{m} i \\ &= \sum_{i=1}^{m} R_{yi} - \frac{m(m+1)}{2} \\ &= T_y - \frac{m(m+1)}{2} \end{aligned}$$

Thus, $\hat{\pi}$ may be expressed in terms of the rank sum of the Y's (or in terms of the rank sum of the X's, since the two rank sums add up to a constant).

From Theorem A, we have

COROLLARY **A**

Under the null hypothesis H_0: $F = G$,

$$E(U_Y) = \frac{mn}{2}$$

$$\text{Var}(U_Y) = \frac{mn(m+n+1)}{12}$$

■

For m and n both greater than 10, the null distribution of U_Y is quite well approximated by a normal distribution,

$$\frac{U_Y - E(U_Y)}{\sqrt{\text{Var}(U_Y)}} \sim N(0, 1)$$

(Note that this does not follow immediately from the ordinary central limit theorem; although U_Y is a sum of random variables, they are not independent.) Similarly, the distribution of the rank sum of the X's or Y's may be approximated by a normal distribution, since these rank sums differ from U_Y only by constants.

EXAMPLE **B** Referring to Example A, let us use a normal approximation to the distribution of the rank sum from method B. For $n = 13$ and $m = 8$, we have from Corollary A that under the null hypothesis,

$$E(T) = \frac{8(8+13+1)}{2} = 88$$

$$\sigma_T = \sqrt{\frac{8 \times 13(8+13+1)}{12}} = 13.8$$

T is the sum of the ranks from method B, or 51, and the normalized test statistic is

$$\frac{T - E(T)}{\sigma_T} = -2.68$$

From the tables of the normal distribution, this corresponds to a p-value of .007 for a two-sided test, so the null hypothesis is rejected at level $\alpha = .01$, just as it was when we used the exact distribution. For this set of data, we have seen that the t test with the assumption of equal variances, the t test without that assumption, the exact Mann-Whitney test, and the approximate Mann-Whitney test all reject at level $\alpha = .01$. ■

The Mann-Whitney test can be inverted to form confidence intervals. Let us consider a "shift" model: $G(x) = F(x - \Delta)$. This model says that the effect of the treatment (the Y's) is to add a constant Δ to what the response would have been with no treatment (the X's). (This is a very simple model, and we have already seen cases for which it is not appropriate.) We now derive a confidence interval for Δ. To test H_0: $F = G$, we used the statistic U_Y equal to the number of the $X_i - Y_j$ that are less than zero. To test the hypothesis that the shift parameter is Δ, we can similarly use

$$U_Y(\Delta) = \#[X_i - (Y_j - \Delta) < 0] = \#(Y_j - X_i > \Delta)$$

It can be shown that the null distribution of $U_Y(\Delta)$ is symmetric about $mn/2$:

$$P\left(U_Y(\Delta) = \frac{mn}{2} + k\right) = P\left(U_Y(\Delta) = \frac{mn}{2} - k\right)$$

for all integers k. Suppose that $k = k(\alpha)$ is such that $P(k \le U_Y(\Delta) \le mn - k) = 1 - \alpha$; the level α test then accepts for such $U_Y(\Delta)$. By the duality of confidence intervals and hypothesis tests, a $100(1 - \alpha)\%$ confidence interval for Δ is thus

$$C = \{\Delta \mid k \le U_Y(\Delta) \le mn - k\}$$

C consists of the set of values Δ for which the null hypothesis would not be rejected.

We can find an explicit form for this confidence interval. Let $D_{(1)}, D_{(2)}, \ldots, D_{(mn)}$ denote the ordered mn differences $Y_j - X_i$. We will show that

$$C = [D_{(k)}, D_{(mn-k+1)})$$

To see this, first suppose that $\Delta = D_{(k)}$. Then

$$\begin{aligned} U_Y(\Delta) &= \#(X_i - Y_j + \Delta < 0) \\ &= \#(Y_j - X_i > \Delta) \\ &= mn - k \end{aligned}$$

Similarly, if $\Delta = D_{(mn-k+1)}$,

$$\begin{aligned} U_Y(\Delta) &= \#(Y_j - X_i > \Delta) \\ &= k \end{aligned}$$

(You might find it helpful to consider the case $m = 3$, $n = 2$, $k = 2$.)

EXAMPLE **C** We return to the data on iron retention (Section 11.2.1.1). The earlier analysis using the t test rested on the assumption that the populations were normally distributed, which, in fact, seemed rather dubious. The Mann-Whitney test does not make this assumption. The sum of the ranks of the Fe^{2+} group is used as a test statistic (we could have as easily used the U statistic). The rank sum is 362. Using the normal approximation to the null distribution of the rank sum, we get a p-value of .36. Again, there is insufficient evidence to reject the null hypothesis that there is no differential retention. The 95% confidence interval for the shift between the two distributions is $(-1.6, 3.7)$, which overlaps zero substantially. Note that this interval is shorter than

the interval based on the t distribution; the latter was inflated by the contributions of the large observations to the sample variance. ■

We close this section with an illustration of the use of the bootstrap in a two-sample problem. As before, suppose that $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are two independent samples from distributions F and G, respectively, and that $\pi = P(X < Y)$ is estimated by $\hat{\pi}$. How can the standard error of $\hat{\pi}$ be estimated and how can an approximate confidence interval for π be constructed? (Note that the calculations of Theorem A are not directly relevant, since they are done under the assumption that $F = G$.)

The problem can be approached in the following way: First suppose for the moment that F and G were known. Then the sampling distribution of $\hat{\pi}$ and its standard error could be estimated by simulation. A sample of size n would be generated from F, an independent sample of size m would be generated from G, and the resulting value of $\hat{\pi}$ would be computed. This procedure would be repeated many times, say B times, producing $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_B$. A histogram of these values would be an indication of the sampling distribution of $\hat{\pi}$ and their standard deviation would be an estimate of the standard error of $\hat{\pi}$.

Of course, this procedure cannot be implemented, because F and G are not known. But as in the previous chapter, an approximation can be obtained by using the empirical distributions F_n and G_n in their places. This means that a bootstrap value of $\hat{\pi}$ is generated by randomly selecting n values from $X_1, X_2, \ldots, X_n$ with replacement, m values from $Y_1, Y_2, \ldots, Y_m$ with replacement and calculating the resulting value of $\hat{\pi}$. In this way, a bootstrap sample $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_B$ is generated.

11.2.4 Bayesian Approach

We consider a Bayesian approach to the model, which stipulates that the X_i are i.i.d. normal with mean μ_X and precision ξ; and the Y_j are i.i.d. normal with mean μ_Y, precision ξ, and independent of the X_i. In general, a prior joint distribution assigned to (μ_X, μ_Y, ξ) would be multiplied by the likelihood and normalized to integrate to 1 to produce a three-dimensional joint posterior distribution for (μ_X, μ_Y, ξ). The marginal joint distribution of (μ_X, μ_Y) could be obtained by integrating out ξ. The marginal distribution of $\mu_X - \mu_Y$ could then be obtained by another integration as in Section 3.6.1. Several integrations would thus have to be done, either analytically or numerically. Special Monte Carlo methods have been devised for high dimensional Bayesian problems, but we will not consider them here.

An approximate result can be obtained using improper priors. We take (μ_X, μ_Y, ξ) to be independent. The means μ_X and μ_Y are given improper priors that are constant on $(-\infty, \infty)$, and ξ is given the improper prior $f_\Xi(\xi) = \xi^{-1}$. The posterior is thus proportional to the likelihood multiplied by ξ^{-1}:

$$f_{\text{post}}(\mu_X, \mu_Y, \xi) \propto \xi^{\frac{n+m}{2}-1} \exp\left(-\frac{\xi^{m+n}}{2}\left[\sum_{i=1}^{n}(x_i - \mu_X)^2 + \sum_{j=1}^{m}(y_j - \mu_Y)^2\right]\right)$$

Next, using $\sum_{i=1}^{n}(x_i-\mu_X)^2=(n-1)s_x^2+n(\mu_X-\bar{x})^2$ and the analogous expression for the y_j, we have

$$f_{\text{post}}(\mu_X,\mu_Y,\xi)\propto \xi^{\frac{n+m}{2}-1}\exp\left(-\frac{\xi}{2}\left[(n-1)s_x^2+(m-1)s_y^2\right]\right)$$
$$\times\exp\left(-\frac{n\xi}{2}(\mu_X-\bar{x})^2\right)\exp\left(-\frac{m\xi}{2}(\mu_Y-\bar{y})^2\right)$$

From the form of this expression as a function of μ_X and μ_Y, we see that for fixed ξ, μ_X and μ_Y are independent normally distributed with means $\bar{x}$ and $\bar{y}$ and precisions $n\xi$ and $m\xi$. Their difference, $\mu_X-\mu_Y$, is thus normally distributed with mean $\bar{x}-\bar{y}$ and variance $\xi^{-1}(n^{-1}+m^{-1})$.

With further analysis similar to that of Section 8.6, it can be shown that the marginal posterior distribution of $\Delta=\mu_X-\mu_Y$ can be related to the t distribution:

$$\frac{\Delta-(\bar{x}-\bar{y})}{s_p\sqrt{n^{-1}+m^{-1}}}\sim t_{n+m-2}$$

Although formally similar to Theorem A of Section 11.2.1, the interpretation is different: $\bar{x}-\bar{y}$ and s_p are random in Theorem A but are fixed here, and $\Delta=\mu_X-\mu_Y$ is random here but fixed in Theorem A. The Bayesian formalism makes probability statements about Δ given the observed data.

The posterior probability that $\Delta>0$ can thus be found using the t distribution. Let T denote a random variable with a t_{m+n-2} distribution. Then, denoting the observations by X and Y

$$P(\Delta>0\mid X,Y)=P\left(\frac{\Delta-(\bar{x}-\bar{y})}{s_p\sqrt{n^{-1}+m^{-1}}}\geq\frac{-(\bar{x}-\bar{y})}{s_p\sqrt{n^{-1}+m^{-1}}}\mid X,Y\right)$$
$$=P\left(T\geq\frac{\bar{y}-\bar{x}}{s_p\sqrt{n^{-1}+m^{-1}}}\right)$$

Letting X denote the measurements of method A, and Y denote the measurements of method B in Example A of Section 11.2.1, we find that for that example,

$$P(\Delta>0|X,Y)=t_{19}(-3.33)=.998$$

This posterior probability is very close to 1.0, and there is thus little doubt that the mean of method A is larger than the mean of method B.

The confidence interval calculated in Section 11.2.1 is formally similar but has a different interpretation under the Bayesian model, which concludes that

$$P(.015\leq\Delta\leq.065|X,Y)=.95$$

by integration of the posterior t distribution over a region containing 95% of the probability.

11.3 Comparing Paired Samples

In Section 11.2, we considered the problem of analyzing two independent samples. In many experiments, the samples are paired. In a medical experiment, for example,

subjects might be matched by age or weight or severity of condition, and then one member of each pair randomly assigned to the treatment group and the other to the control group. In a biological experiment, the paired subjects might be littermates. In some applications, the pair consists of a "before" and an "after" measurement on the same object. Since pairing causes the samples to be dependent, the analysis of Section 11.2 does not apply.

Pairing can be an effective experimental technique, as we will now demonstrate by comparing a paired design and an unpaired design. First, we consider the paired design. Let us denote the pairs as (X_i, Y_i), where $i = 1, \ldots, n$, and assume the X's and Y's have means μ_X and μ_Y and variances σ_X^2 and σ_Y^2. We will assume that different pairs are independently distributed and that $\text{Cov}(X_i, Y_i) = \sigma_{XY}$. We will work with the differences $D_i = X_i - Y_i$, which are independent with

$$\begin{aligned} E(D_i) &= \mu_X - \mu_Y \\ \text{Var}(D_i) &= \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY} \\ &= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y \end{aligned}$$

when ρ is the correlation of members of a pair. A natural estimate of $\mu_X - \mu_Y$ is $\overline{D} = \overline{X} - \overline{Y}$, the average difference. From the properties of D_i, it follows that

$$\begin{aligned} E(\overline{D}) &= \mu_X - \mu_Y \\ \text{Var}(\overline{D}) &= \frac{1}{n}\left(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y\right) \end{aligned}$$

Suppose, on the other hand, that an experiment had been done by taking a sample of n X's and an independent sample of n Y's. Then $\mu_X - \mu_Y$ would be estimated by $\overline{X} - \overline{Y}$ and

$$\begin{aligned} E(\overline{X} - \overline{Y}) &= \mu_X - \mu_Y \\ \text{Var}(\overline{X} - \overline{Y}) &= \frac{1}{n}\left(\sigma_X^2 + \sigma_Y^2\right) \end{aligned}$$

Comparing the variances of the two estimates, we see that the variance of $\overline{D}$ is smaller if the correlation is positive—that is, if the X's and Y's are positively correlated. In this circumstance, pairing is the more effective experimental design. In the simple case in which $\sigma_X = \sigma_Y = \sigma$, the two variances may be more simply expressed as

$$\text{Var}(\overline{D}) = \frac{2\sigma^2(1-\rho)}{n}$$

in the paired case and as

$$\text{Var}(\overline{X} - \overline{Y}) = \frac{2\sigma^2}{n}$$

in the unpaired case, and the relative efficiency is

$$\frac{\text{Var}(\overline{D})}{\text{Var}(\overline{X} - \overline{Y})} = 1 - \rho$$

If the correlation coefficient is .5, for example, a paired design with n pairs of subjects yields the same precision as an unpaired design with $2n$ subjects per treatment. This additional precision results in shorter confidence intervals and more powerful tests if the degrees of freedom for estimating σ^2 are sufficiently large.

We next present methods based on the normal distribution for analyzing data from paired designs and then a nonparametric, rank-based method.

11.3.1 Methods Based on the Normal Distribution

In this section, we assume that the differences are a sample from a normal distribution with

$$E(D_i) = \mu_X - \mu_Y = \mu_D$$
$$\text{Var}(D_i) = \sigma_D^2$$

Generally, σ_D will be unknown, and inferences will be based on

$$t = \frac{\overline{D} - \mu_D}{s_{\overline{D}}}$$

which follows a t distribution with $n - 1$ degrees of freedom. Following familiar reasoning, a $100(1 - \alpha)\%$ confidence interval for μ_D is

$$\overline{D} \pm t_{n-1}(\alpha/2)s_{\overline{D}}$$

A two-sided test of the null hypothesis H_0: $\mu_D = 0$ (the natural null hypothesis for testing no treatment effect) at level α has the rejection region

$$|\overline{D}| > t_{n-1}(\alpha/2)s_{\overline{D}}$$

If the sample size n is large, the approximate validity of the confidence interval and hypothesis test follows from the central limit theorem. If the sample size is small and the true distribution of the differences is far from normal, the stated probability levels may be considerably in error.

EXAMPLE A To study the effect of cigarette smoking on platelet aggregation, Levine (1973) drew blood samples from 11 individuals before and after they smoked a cigarette and measured the extent to which the blood platelets aggregated. Platelets are involved in the formation of blood clots, and it is known that smokers suffer more often from disorders involving blood clots than do nonsmokers. The data are shown in the following table, which gives the maximum percentage of all the platelets that aggregated after being exposed to a stimulus.

Before	After	Difference
25	27	2
25	29	4
27	37	10
44	56	12
30	46	16
67	82	15
53	57	4
53	80	27
52	61	9
60	59	−1
28	43	15

From the column of differences, $\overline{D} = 10.27$ and $s_{\overline{D}} = 2.40$. The uncertainty in $\overline{D}$ is quantified in $s_{\overline{D}}$ or in a confidence interval. Since $t_{10}(.05) = 1.812$, a 90% confidence interval is $\overline{D} \pm 1.812 s_{\overline{D}}$, or (5.9, 14.6). We can also formally test the null hypothesis that means before and after are the same. The t statistic is $10.27/2.40 = 4.28$, and since $t_{10}(.005) = 3.169$, the p-value of a two-sided test is less than .01. There is little doubt that smoking increases platelet aggregation.

The experiment was actually more complex than we have indicated. Some subjects also smoked cigarettes made of lettuce leaves and "smoked" unlit cigarettes. (You should reflect on why these additional experiments were done.)

Figure 11.7 is a plot of the after values versus the before values. They are correlated, with a correlation coefficient of .90. Pairing was a natural and effective experimental design in this case. ■

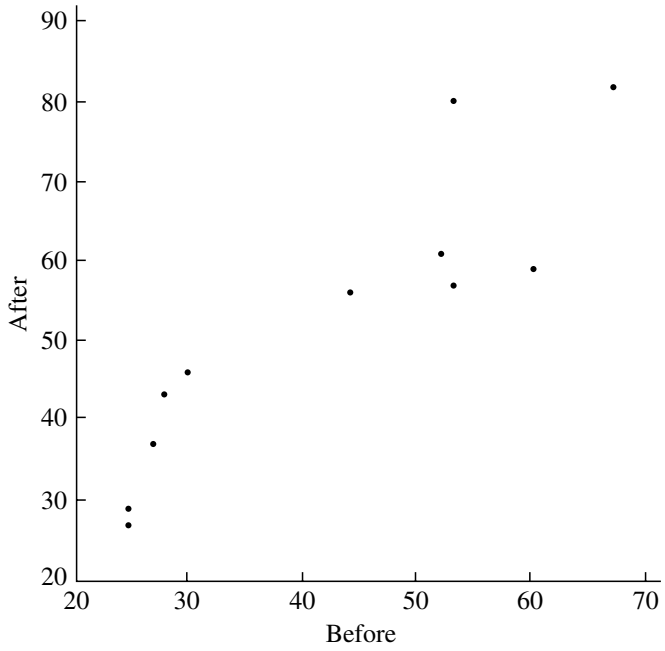

FIGURE **11.7** Plot of platelet aggregation after smoking versus aggregation before smoking.

11.3.2 A Nonparametric Method—The Signed Rank Test

A nonparametric test based on ranks can be constructed for paired samples. We illustrate the calculation with a very small example. Suppose there are four pairs, corresponding to "before" and "after" measurements listed in the following table:

Before	After	Difference	|Difference|	Rank	Signed Rank
25	27	2	2	2	2
29	25	−4	4	3	−3
60	59	−1	1	1	−1
27	37	10	10	4	4

The test statistic is calculated by the following steps:

1. Calculate the differences, D_i, and the absolute values of the differences and rank the latter.
2. Restore the signs of the differences to the ranks, obtaining signed ranks.
3. Calculate W_+, the sum of those ranks that have positive signs. For the table, this sum is $W_+ = 2 + 4 = 6$.

The idea behind the **signed rank test** (sometimes called the Wilcoxon signed rank test) is intuitively simple. If there is no difference between the two paired conditions, we expect about half the D_i to be positive and half negative, and W_+ will not be too small or too large. If one condition tends to produce larger values than the other, W_+ will tend to be more extreme. We therefore can use W_+ as a test statistic and reject for extreme values.

Before continuing, we need to specify more precisely the null hypothesis we are testing with the signed rank test: H_0 states that the distribution of the D_i is symmetric about zero. This will be true if the members of pairs of experimental units are assigned randomly to treatment and control conditions, and the treatment has no effect at all.

As usual, in order to define a rejection region for a test at level α, we need to know the sampling distribution of W_+ if the null hypothesis is true. The rejection region will be located in the tails of this null distribution in such a way that the test has level α. The null distribution may be calculated in the following way. If H_0 is true, it makes no difference which member of the pair corresponds to treatment and which to control. The difference $X_i - Y_i = D_i$ has the same distribution as the difference $Y_i - X_i = -D_i$, so the distribution of D_i is symmetric about zero. The kth largest value of D is thus equally likely to be positive or negative, and any particular assignment of signs to the integers $1, \ldots, n$ (the ranks) is equally likely. There are 2^n such assignments, and for each we can calculate W_+. We obtain a list of 2^n values (not all distinct) of W_+, each of which occurs with probability $1/2^n$. The probability of each distinct value of W_+ may thus be calculated, giving the desired null distribution.

The preceding argument has assumed that the D_i are a sample from some continuous probability distribution. If we do not wish to regard the X_i and Y_i as random variables and if the assignments to treatment and control have been made at random, the hypothesis that there is no treatment effect may be tested in exactly the same

manner, except that inferences are based on the distribution induced by the randomization, as was done for the Mann-Whitney test.

The null distribution of W_+ is calculated by many computer packages, and tables are also available.

The signed rank test is a nonparametric version of the paired sample t test. Unlike the t test, it does not depend on an assumption of normality. Since differences are replaced by ranks, it is insensitive to outliers, whereas the t test is sensitive. It has been shown that even when the assumption of normality holds, the signed rank test is nearly as powerful as the t test. The nonparametric method is thus generally preferable, especially for small sample sizes.

EXAMPLE **A** The signed rank test can be applied to the data on platelet aggregation considered previously (Example A in Section 11.3.1). In this case, it is easier to work with W_- rather than W_+, since W_- is clearly 1. From Table 9 of Appendix B, the two-sided test is significant at $\alpha = .01$. ■

If the sample size is greater than 20, a normal approximation to the null distribution can be used. To find this, we calculate the mean and variance of W_+.

THEOREM **A**

Under the null hypothesis that the D_i are independent and symmetrically distributed about zero,

$$E(W_+) = \frac{n(n+1)}{4}$$

$$\text{Var}(W_+) = \frac{n(n+1)(2n+1)}{24}$$

Proof

To facilitate the calculation, we represent W_+ in the following way:

$$W_+ = \sum_{k=1}^{n} k I_k$$

where

$$I_k = \begin{cases} 1, & \text{if the } k\text{th largest } |D_i| \text{ has } D_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

Under H_0, the I_k are independent Bernoulli random variables with $p = \frac{1}{2}$, so

$$E(I_k) = \frac{1}{2}$$

$$\text{Var}(I_k) = \frac{1}{4}$$

We thus have

$$E(W_+) = \frac{1}{2}\sum_{k=1}^{n} k = \frac{n(n+1)}{4}$$

$$\text{Var}(W_+) = \frac{1}{4}\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{24}$$

as was to be shown. ■

If some of the differences are equal to zero, the most common technique is to discard those observations. If there are ties, each $|D_i|$ is assigned the average value of the ranks for which it is tied. If there are not too many ties, the significance level of the test is not greatly affected. If there are a large number of ties, modifications must be made. For further information on these matters, see Hollander and Wolfe (1973) or Lehmann (1975).

11.3.3 An Example—Measuring Mercury Levels in Fish

Kacprzak and Chvojka (1976) compared two methods of measuring mercury levels in fish. A new method, which they called "selective reduction," was compared to an established method, referred to as "the permanganate method." One advantage of selective reduction is that it allows simultaneous measurement of both inorganic mercury and methyl mercury. The mercury in each of 25 juvenile black marlin was measured by both techniques. The 25 measurements for each method (in ppm of mercury) and the differences are given in the following table.

Fish	Selective Reduction	Permanganate	Difference	Signed Rank
1	.32	.39	.07	+15.5
2	.40	.47	.07	+15.5
3	.11	.11	.00	
4	.47	.43	−.04	−11
5	.32	.42	.10	+19
6	.35	.30	−.05	−13.5
7	.32	.43	.11	+20
8	.63	.98	.35	+23
9	.50	.86	.36	+24
10	.60	.79	.19	+22
11	.38	.33	−.05	−13.5
12	.46	.45	−.01	−2.5

(*Continued*)

Fish	Selective Reduction	Permanganate	Difference	Signed Rank
13	.20	.22	.02	+6.5
14	.31	.30	−.01	−2.5
15	.62	.60	−.02	−6.5
16	.52	.53	.01	+2.5
17	.77	.85	.08	+17.5
18	.23	.21	−.02	−6.5
19	.30	.33	.03	+9.0
20	.70	.57	−.13	−21
21	.41	.43	.02	+6.5
22	.53	.49	−.04	−11
23	.19	.20	.01	+2.5
24	.31	.35	.04	+11
25	.48	.40	−.08	−17.5

In analyzing such data, it is often informative to check whether the differences depend in some way on the level or size of the quantity being measured. The differences versus the permanganate values are plotted in Figure 11.8. This plot is quite interesting. It appears that the differences are small for low permanganate values and larger for higher permanganate values. It is striking that the differences are all positive and large for the highest four values. The investigators do not comment on these phenomena. It is not uncommon for the size of fluctuations to increase as the value being measured increases; the percent error may remain nearly constant but the actual error does not. For this reason, data of this nature are often analyzed on a log scale.

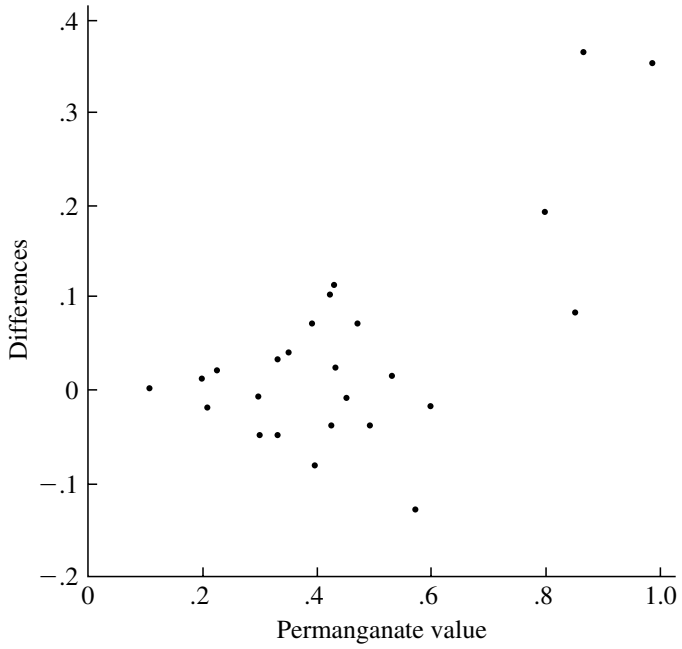

FIGURE **11.8** Plot of differences versus permanganate values.

Because the observations are paired (two measurements on each fish), we will use the paired t test for a parametric test. The sample size is large enough that the test should be robust against nonnormality. The mean difference is .04, and the standard deviation of the differences is .116. The t statistic is 1.724; with 24 degrees of freedom, this corresponds to a p-value of .094 for a two-sided test. Although this p-value is fairly small, the evidence against H_0: $\mu_D = 0$ is not overwhelming. The test does not reject at the significance level .05.

The signed ranks are shown in the last column of the table above. Note that the single zero difference was set aside, and also note how the tied ranks were handled. The test statistic W_+ is 194.5. Under H_0, its mean and variance are

$$E(W_+) = \frac{24 \times 25}{4} = 150$$

$$\text{Var}(W_+) = \frac{24 \times 25 \times 49}{24} = 1225$$

Since n is greater than 20, we use the normalized test statistic, or

$$Z = \frac{W_+ - E(W_+)}{\sqrt{\text{Var}(W_+)}} = 1.27$$

The p-value for a two-sided test from the normal approximation is .20, which is not strong evidence against the null hypothesis. It is possible to correct for the presence of ties, but in this case the correction only amounts to changing the standard deviation of W_+ from 35 to 34.95.

Neither the parametric nor the nonparametric test gives conclusive evidence that there is any systematic difference between the two methods of measurement. The informal graphical analysis does suggest, however, that there may be a difference for high concentrations of mercury.

11.4 Experimental Design

This section covers some basic principles of the interpretation and design of experimental studies and illustrates them with case studies.

11.4.1 Mammary Artery Ligation

A person with coronary artery disease suffers from chest pain during exercise because the constricted arteries cannot deliver enough oxygen to the heart. The treatment of ligating the mammary arteries enjoyed a brief vogue; the basic idea was that ligating these arteries forced more blood to flow into the heart. This procedure had the advantage of being quite simple surgically, and it was widely publicized in an article in *Reader's Digest* (Ratcliffe 1957). Two years later, the results of a more careful study (Cobb et al. 1959) were published. In this study, a control group and an experimental group were established in the following way. When a prospective patient entered surgery, the surgeon made the necessary preliminary incisions prior to tying off the mammary artery. At that point, the surgeon opened a sealed envelope that contained instructions about whether to complete the operation by tying off the artery. Neither

the patient nor his attending physician knew whether the operation had actually been carried out. The study showed essentially no difference after the operation between the control group (no ligation) and the experimental group (ligation), although there was some suggestion that the control group had done better.

The Ratcliffe and Cobb studies differ in that in the earlier one there was no control group and thus no benchmark by which to gauge improvement. The reported improvement of the patients in this earlier study could have been due to the placebo effect, which we discuss next. The design of the later study protected against possible unconscious biases by randomly assigning the control and experimental groups and by concealing from the patients and their physicians the actual nature of the treatment. Such a design is called a double-blind, randomized controlled experiment.

11.4.2 The Placebo Effect

The **placebo effect** refers to the effect produced by any treatment, including dummy pills (placebos), when the subject believes that he or she has been given an effective treatment. The possibility of a placebo effect makes the use of a blind design necessary in many experimental investigations.

The placebo effect may not be due entirely to psychological factors, as was shown in an interesting experiment by Levine, Gordon, and Fields (1978). A group of subjects had teeth extracted. During the extraction, they were given nitrous oxide and local anesthesia. In the recovery room, they rated the amount of pain they were experiencing on a numerical scale. Two hours after surgery, the subjects were given a placebo and were again asked to rate their pain. An hour later, some of the subjects were given a placebo and some were given naloxone, a morphine antagonist. It is known that there are specific receptors to morphine in the brain and that the body can also release endorphins that bind to these sites. Naloxone blocks the morphine receptors. In the study, it was found that when those subjects who responded positively to the placebo received naloxone, they experienced an increase in pain that made their pain levels comparable to those of the patients who did not respond to the placebo. The implication is that those who responded to the placebo had produced endorphins, the actions of which were subsequently blocked by the naloxone.

An instance of the placebo effect was demonstrated by a psychologist, Claude Steele (2002), who gave a math exam to a group of male and female undergraduates at Stanford University. One group (treatment) was told that the exam was gender-neutral, and the other group (controls) was not so informed. The men outperformed the women in the control group. In the treatment group, men and women performed equally well. Men in the treatment group did worse than men in the control group. (*Economist* Feb 21, 2002).

11.4.3 The Lanarkshire Milk Experiment

The importance of the randomized assignment of individuals (or other experimental units) to treatment and control groups is illustrated by a famous study known as the Lanarkshire milk experiment. In the spring of 1930, an experiment was carried out in Lanarkshire, Scotland, to determine the effect of providing free milk to schoolchildren. In each participating school, some children (treatment group) were given free milk

and others (controls) were not. The assignment of children to control or treatment was initially done at random; however, teachers were allowed to use their judgment in switching children between treatment and control to obtain a better balance of undernourished and well-nourished individuals in the groups.

A paper by Gosset (1931), who published under the name Student (as in Student's *t* test), is a very interesting critique of the experiment. An examination of the data revealed that at the start of the experiment the controls were heavier and taller. Student conjectured that the teachers, perhaps unconsciously, had adjusted the initial randomization in a manner that placed more of the undernourished children in the treatment group. A further complication was caused by weighing the children with their clothes on. The experimental data were weight gains measured in late spring relative to early spring or late winter. The more well-to-do children probably tended to be better nourished and may have had heavier winter clothing than the poor children. Thus, the well-to-do children's weight gains were vitiated as a result of differences in clothing, which may have influenced comparisons between the treatment and control groups.

11.4.4 The Portacaval Shunt

Cirrhosis of the liver, to which alcoholics are prone, is a condition in which resistance to blood flow causes blood pressure in the liver to build up to dangerously high levels. Vessels may rupture, which may cause death. Surgeons have attempted to relieve this condition by connecting the portal artery, which feeds the liver, to the vena cava, one of the main veins returning to the heart, thus reducing blood flow through the liver. This procedure, called the Portacaval shunt, had been used for more than 20 years when Grace, Muench, and Chalmers (1966) published an examination of 51 studies of the method. They examined the design of each study (presence or absence of a control group and presence or absence of randomization) and the investigators' conclusions (categorized as markedly enthusiastic, moderately enthusiastic, or not enthusiastic). The results are summarized in the following table, which speaks for itself:

	Enthusiasm		
Design	Marked	Moderate	None
No controls	24	7	1
Nonrandomized controls	10	3	2
Randomized controls	0	1	3

The differences between the experiments that used controls and those that did not is not entirely surprising, because the placebo effect was probably operating. The importance of randomized assignment to treatment and control groups is illustrated by comparing the conclusions for the randomized and nonrandomized controlled experiments. Randomization can help to ensure against subtle unconscious biases that may creep into an experiment. For example, a physician might tend to recommend surgery for patients who are somewhat more robust than the average. Articulate

patients might be more likely to have an influence on the decision as to which group they are assigned to.

11.4.5 FD&C Red No. 40

This discussion follows Lagakos and Mosteller (1981). During the middle and late 1970s, experiments were conducted to determine possible carcinogenic effects of a widely used food coloring, FD&C Red No. 40. One of the experiments involved 500 male and 500 female mice. Both genders were divided into five groups: two control groups, a low-dose group, a medium-dose group, and a high-dose group. The mice were bred in the following way: Males and females were paired and before and during mating were given their prescribed dose of Red No. 40. The regime was continued during gestation and weaning of the young. From litters that had at least three pups of each sex, three of each sex were selected randomly and continued on their parents' dosage throughout their lives. After 109–111 weeks, all the mice still living were killed. The presence or absence of reticuloendothelial tumors was of particular interest. Although there were significant differences between some of the treatment groups, the results were rather confusing. For example, there was a significant difference between the incidence rates for the two male control groups, and among the males the medium-dose group had the lowest incidence.

Several experts were asked to examine the results of this and other experiments. Among them were Lagakos and Mosteller, who requested information on how the cages that housed the mice were arranged. There were three racks of cages, each containing five rows of seven cages in the front and five rows of seven cages in the back. Five mice were housed in each cage. The mice were assigned to the cages in a systematic way: The first male control group was in the top of the front of rack 1; the first female control group was in the bottom of the front of rack 1; and so on, ending with the high-dose females in the bottom of the back of rack 3 (Figure 11.9). Lagakos and Mosteller showed that there were effects due to cage position that could not be explained by gender or by dosage group. A random assignment of cage positions would have eliminated this confounding. Lagakos and Mosteller also suggested some experimental designs to systematically control for cage position.

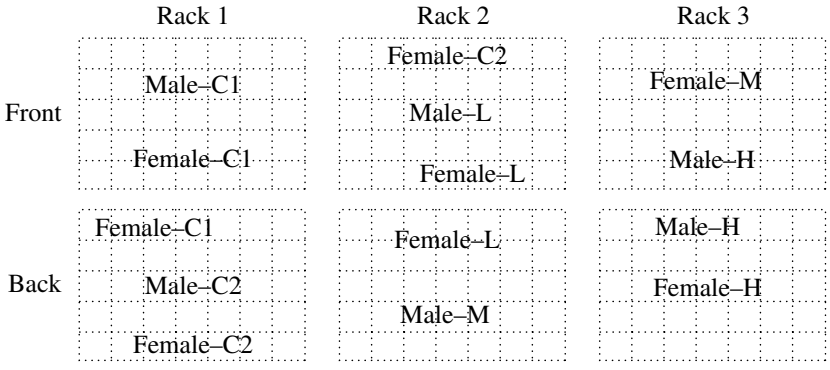

FIGURE **11.9** Location of mice cages in racks.

It was also possible that a litter effect might be complicating the analysis, since littermates received the same treatment and littermates of the same sex were housed in the same or contiguous cages. In the presence of a litter effect, mice from the same litter might show less variability than that present among mice from different litters. This reduces the effective sample size—in the extreme case in which littermates react identically, the effective sample size is the number of litters, not the total number of mice. One way around this problem would have been to use only one mouse from each litter.

The presence of a possible selection bias is another problem. Because mice were included in the experiment only if they came from a litter with at least three males and three females, offspring of possibly less healthy parents were excluded. This could be a serious problem since exposure to Red No. 40 might affect the parents' health and the birth process. If, for example, among the high-dose mice, only the most hardy produced large enough litters, their offspring might be hardier than the controls' offspring.

11.4.6 Further Remarks on Randomization

As well as guarding against possible biases on the part of the experimenter, the process of randomization tends to balance any factors that may be influential but are not explicitly controlled in the experiment. Time is often such a factor; background variables such as temperature, equipment calibration, line voltage, and chemical composition can change slowly with time. In experiments that are run over some period of time, therefore, it is important to randomize the assignments to treatment and control over time. Time is not the only factor that should be randomized, however. In agricultural experiments, the positions of test plots in a field are often randomly assigned. In biological experiments with test animals, the locations of the animals' cages may have an effect, as illustrated in the preceding section.

Although rarer than in other areas, randomized experiments have been carried out in the social sciences as well (*Economist* Feb 28, 2002). Randomized trials have been used to evaluate such programs as driver training, as well as the criminal justice system and reduced classroom size. In evaluations of "whole-language" approaches to reading (in which children are taught to read by evaluating contextual clues rather than breaking down words), 52 randomized studies carried out by the National Reading Panel in 2000 showed that effective reading instruction requires phonics. Randomized studies of "scared straight" programs, in which juvenile delinquents are introduced to prison inmates, suggested that the likelihood of subsequent arrests is actually increased by such programs.

Generally, if it is anticipated that a variable will have a significant effect, that variable should be included as one of the controlled factors in the experimental design. The matched-pairs design of this chapter can be used to control for a single factor. To control for more than one factor, factorial designs, which are briefly introduced in the next chapter, may be used.

11.4.7 Observational Studies, Confounding, and Bias in Graduate Admissions

It is not always possible to conduct controlled experiments or use randomization. In evaluating some medical therapies, for example, a randomized, controlled experiment would be unethical if one therapy was strongly believed to be superior. For many problems of psychological interest (effects of parental modes of discipline, for example), it is impossible to conduct controlled experiments. In such situations, recourse is often made to observational studies. Hospital records may be examined to compare the outcomes of different therapies, or psychological records of children raised in different ways may be analyzed. Although such studies may be valuable, the results are seldom unequivocal. Because there is no randomization, it is always possible that the groups under comparison differ in respects other than their "treatments."

As an example, let us consider a study of gender bias in admissions to graduate school at the University of California at Berkeley (Bickel and O'Connell 1975). In the fall of 1973, 8442 men applied for admission to graduate studies at Berkeley, and 44% were admitted; 4321 women applied, and 35% were admitted. If the men and women were similar in every respect other than sex, this would be strong evidence of sex bias. This was not a controlled, randomized experiment, however; sex was not randomly assigned to the applicants. As will be seen, the male and female applicants differed in other respects, which influenced admission.

The following table shows admission rates for the six most popular majors on the Berkeley campus.

	Men		Women	
Major	Number of Applicants	Percentage Admitted	Number of Applicants	Percentage Admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	34
F	373	6	341	7

If the percentages admitted are compared, women do not seem to be unfavorably treated. But when the combined admission rates for all six majors are calculated, it is found that 44% of the men and only 30% of the women were admitted, which seems paradoxical. The resolution of the paradox lies in the observation that the women tended to apply to majors that had low admission rates (C through F) and the men to majors that had relatively high admission rates (A and B). This factor was not controlled for, because the study was observational in nature; it was also "confounded" with the factor of interest, sex; randomization, had it been possible, would have tended to balance out the confounded factor.

Confounding also plays an important role in studies of the effect of coffee drinking. Several studies have claimed to show a significant association of coffee

consumption with coronary disease. Clearly, randomized, controlled trials are not possible here—a randomly selected individual cannot be told that he or she is in the treatment group and must drink 10 cups of coffee a day for the next five years. Also, it is known that heavy coffee drinkers also tend to smoke more than average, so smoking is confounded with coffee drinking. Hennekens et al. (1976) review several studies in this area.

11.4.8 Fishing Expeditions

Another problem that sometimes flaws observational studies, and controlled experiments as well, is that they engage in "fishing expeditions." For example, consider a hypothetical study of the effects of birth control pills. In such a case, it would be impossible to assign women to a treatment or a placebo at random, but a nonrandomized study might be conducted by carefully matching controls to treatments on such factors as age and medical history. The two groups might be followed up on for some time, with many variables being recorded for each subject such as blood pressure, psychological measures, and incidences of various medical problems. After termination of the study, the two groups might be compared on each of these variables, and it might be found, say, that there was a "significant difference" in the incidence of melanoma. The problem with this "significant finding" is the following. Suppose that 100 independent two-sample t tests are conducted at the .05 level and that, in fact, all the null hypotheses are true. We would expect that five of the tests would produce a "significant" result. Although each of the tests has probability .05 of type I error, as a collection they do not simultaneously have $\alpha = .05$. The combined significance level is the probability that at least one of the null hypotheses is rejected:

$$\begin{aligned}\alpha &= P\{\text{at least one } H_0 \text{ rejected}\}\\ &= 1 - P\{\text{no } H_0 \text{ rejected}\}\\ &= 1 - .95^{100} = .994\end{aligned}$$

Thus, with very high probability, at least one "significant" result will be found, even if all the null hypotheses are true.

There are no simple cures for this problem. One possibility is to regard the results of a fishing expedition as merely providing suggestions for further experiments. Alternatively, and in the same spirit, the data could be split randomly into two halves, one half for fishing in and the other half to be locked safely away, unexamined. "Significant" results from the first half could then be tested on the second half. A third alternative is to conduct each individual hypothesis test at a small significance level. To see how this works, suppose that all null hypotheses are true and that each of n null hypotheses is tested at level α. Let R_i denote the event that the ith null hypothesis is rejected, and let α^* denote the overall probability of a type I error. Then

$$\begin{aligned}\alpha^* &= P\{R_1 \text{ or } R_2 \text{ or } \cdots \text{ or } R_n\}\\ &\le P\{R_1\} + P\{R_2\} + \cdots + P\{R_n\}\\ &= n\alpha\end{aligned}$$

Thus, if each of the n null hypotheses is tested at level α/n, the overall significance level is less than or equal to α. This is often called the **Bonferroni method.**

11.5 Concluding Remarks

This chapter was concerned with the problem of comparing two samples. Within this context, the fundamental statistical concepts of estimation and hypothesis testing, which were introduced in earlier chapters, were extended and utilized. The chapter also showed how informal descriptive and data analytic techniques are used in supplementing more formal analysis of data. Chapter 12 will extend the techniques of this chapter to deal with multisample problems. Chapter 13 is concerned with similar problems that arise in the analysis of qualitative data.

We considered two types of experiments, those with two independent samples and those with matched pairs. For the case of independent samples, we developed the t test, based on an assumption of normality, as well as a modification of the t test that takes into account possibly unequal variances. The Mann-Whitney test, based on ranks, was presented as a nonparametric method, that is, a method that is not based on an assumption of a particular distribution. Similarly, for the matched-pairs design, we developed a parametric t test and a nonparametric test, the signed rank test.

We discussed methods based on an assumption of normality and rank methods, which do not make this assumption. It turns out, rather surprisingly, that even if the normality assumption holds, the rank methods are quite powerful relative to the t test. Lehmann (1975) shows that the efficiency of the rank tests relative to that of the t test—that is, the ratio of sample sizes required to attain the same power—is typically around .95 if the distributions are normal. Thus, a rank test using a sample of size 100 is as powerful as a t test based on 95 observations. Collecting the extra 5 pieces of data is a small price to pay for a safeguard against nonnormality.

The bootstrap appeared again in this chapter. Indeed, uses of this recently developed technique are finding applications in a great variety of statistical problems. In contrast with earlier chapters, where bootstrap samples were generated from one distribution, here we have bootstrapped from two empirical distributions.

The chapter concluded with a discussion of experimental design, which emphasized the importance of incorporating controls and randomization in investigations. Possible problems associated with observational studies were discussed. Finally, the difficulties encountered in making many comparisons from a single data set were pointed out; such problems of multiplicity will come up again in Chapter 12.

11.6 Problems

1. A computer was used to generate four random numbers from a normal distribution with a set mean and variance: 1.1650, .6268, .0751, .3516. Five more random normal numbers with the same variance but perhaps a different mean were then generated (the mean may or may not actually be different): .3035, 2.6961, 1.0591, 2.7971, 1.2641.

a. What do you think the means of the random normal number generators were? What do you think the difference of the means was?

b. What do you think the variance of the random number generator was?

c. What is the estimated standard error of your estimate of the difference of the means?

d. Form a 90% confidence interval for the difference of the means of the random number generators.
e. In this situation, is it more appropriate to use a one-sided test or a two-sided test of the equality of the means?
f. What is the p-value of a two-sided test of the null hypothesis of equal means?
g. Would the hypothesis that the means were the same versus a two-sided alternative be rejected at the significance level $\alpha = .1$?
h. Suppose you know that the variance of the normal distribution was $\sigma^2 = 1$. How would your answers to the preceding questions change?

2. The difference of the means of two normal distributions with equal variance is to be estimated by sampling an equal number of observations from each distribution. If it were possible, would it be better to halve the standard deviations of the populations or double the sample sizes?

3. In Section 11.2.1, we considered two methods of estimating $\text{Var}(\overline{X} - \overline{Y})$. Under the assumption that the two population variances were equal, we estimated this quantity by

$$s_p^2 \left(\frac{1}{n} + \frac{1}{m} \right)$$

and without this assumption by

$$\frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

Show that these two estimates are identical if $m = n$.

4. Respond to the following:

> Using the t distribution is absolutely ridiculous—another example of deliberate mystification! It's valid when the populations are normal and have equal variance. If the sample sizes were so small that the t distribution were practically different from the normal distribution, you would be unable to check these assumptions.

5. Respond to the following:

> Here is another example of deliberate mystification—the idea of formulating and testing a null hypothesis. Let's take Example A of Section 11.2.1. It seems to me that it is inconceivable that the expected values of *any* two methods of measurement could be *exactly* equal. It is certain that there will be subtle differences at the very least. What is the sense, then, in testing H_0: $\mu_X = \mu_Y$?

6. Respond to the following:

> I have two batches of numbers and I have a corresponding $\bar{x}$ and $\bar{y}$. Why should I test whether they are equal when I can just see whether they are or not?

7. In the development of Section 11.2.1, where are the following assumptions used? (1) $X_1, X_2, \ldots, X_n$ are independent random variables; (2) $Y_1, Y_2, \ldots, Y_n$ are independent random variables; (3) the X's and Y's are independent.

8. An experiment to determine the efficacy of a drug for reducing high blood pressure is performed using four subjects in the following way: two of the subjects are chosen at random for the control group and two for the treatment group. During the course of treatment with the drug, the blood pressure of each of the subjects in the treatment group is measured for ten consecutive days as is the blood pressure of each of the subjects in the control group.

a. In order to test whether the treatment has an effect, do you think it is appropriate to use the two-sample t test with $n = m = 20$?

b. Do you think it is appropriate to use the Mann-Whitney test with $n = m = 20$?

9. Referring to the data in Section 11.2.1.1, compare iron retention at concentrations of 10.2 and .3 millimolar using graphical procedures and parametric and nonparametric tests. Write a brief summary of your conclusions.

10. Verify that the two-sample t test at level α of H_0: $\mu_X = \mu_Y$ versus H_A: $\mu_X \neq \mu_Y$ rejects if and only if the confidence interval for $\mu_X - \mu_Y$ does not contain zero.

11. Explain how to modify the t test of Section 11.2.1 to test H_0: $\mu_X = \mu_Y + \Delta$ versus H_A: $\mu_X \neq \mu_Y + \Delta$ where Δ is specified.

12. An equivalence between hypothesis tests and confidence intervals was demonstrated in Chapter 9. In Chapter 10, a nonparametric confidence interval for the median, η, was derived. Explain how to use this confidence interval to test the hypothesis H_0: $\eta = \eta_0$. In the case where $\eta_0 = 0$, show that using this approach on a sample of differences from a paired experiment is equivalent to the **sign test.** The sign test counts the number of positive differences and uses the fact that in the case that the null hypothesis is true, the distribution of the number of positive differences is binomial with $(n, .5)$. Apply the sign test to the data from the measurement of mercury levels, listed in Section 11.3.3.

13. Let $X_1, \ldots, X_{25}$ be i.i.d. $N(.3, 1)$. Consider testing the null hypothesis H_0: $\mu = 0$ versus H_A: $\mu > 0$ at significance level $\alpha = .05$. Compare the power of the sign test and the power of the test based on normal theory assuming that σ is known.

14. Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. To test the null hypothesis H_0: $\mu = \mu_0$, the t test is often used:

$$t = \frac{\overline{X} - \mu_0}{s_{\overline{X}}}$$

Under H_0, t follows a t distribution with $n - 1$ df. Show that the likelihood ratio test of this H_0 is equivalent to the t test.

15. Suppose that n measurements are to be taken under a treatment condition and another n measurements are to be taken independently under a control condition. It is thought that the standard deviation of a single observation is about 10 under both conditions. How large should n be so that a 95% confidence interval for $\mu_X - \mu_Y$ has a width of 2? Use the normal distribution rather than the t distribution, since n will turn out to be rather large.

16. Referring to Problem 15, how large should n be so that the test of H_0: $\mu_X = \mu_Y$ against the one-sided alternative H_A: $\mu_X > \mu_Y$ has a power of .5 if $\mu_X - \mu_Y = 2$ and $\alpha = .10$?

17. Consider conducting a two-sided test of the null hypothesis H_0: $\mu_X = \mu_Y$ as described in Problem 16. Sketch power curves for (a) $\alpha = .05$, $n = 20$; (b) $\alpha = .10$, $n = 20$; (c) $\alpha = .05$, $n = 40$; (d) $\alpha = .10$, $n = 40$. Compare the curves.

18. Two independent samples are to be compared to see if there is a difference in the population means. If a total of m subjects are available for the experiment, how should this total be allocated between the two samples in order to (a) provide the shortest confidence interval for $\mu_X - \mu_Y$ and (b) make the test of H_0: $\mu_X = \mu_Y$ as powerful as possible? Assume that the observations in the two samples are normally distributed with the same variance.

19. An experiment is planned to compare the mean of a control group to the mean of an independent sample of a group given a treatment. Suppose that there are to be 25 samples in each group. Suppose that the observations are approximately normally distributed and that the standard deviation of a single measurement in either group is $\sigma = 5$.

a. What will the standard error of $\overline{Y} - \overline{X}$ be?
b. With a significance level $\alpha = .05$, what is the rejection region of the test of the null hypothesis H_0: $\mu_Y = \mu_X$ versus the alternative H_A: $\mu_Y > \mu_X$?
c. What is the power of the test if $\mu_Y = \mu_X + 1$?
d. Suppose that the p-value of the test turns out to be 0.07. Would the test reject at significance level $\alpha = .10$?
e. What is the rejection region if the alternative is H_A: $\mu_Y \neq \mu_X$? What is the power if $\mu_Y = \mu_X + 1$?

20. Consider Example A of Section 11.3.1 using a Bayesian model. As in the example, use a normal model for the differences and also use an improper prior for the expected difference and the precision (as in the case of unknown mean and variance in Section 8.6). Find the posterior probability that the expected difference is positive. Find a 90% posterior credibility interval for the expected difference.

21. A study was done to compare the performances of engine bearings made of different compounds (McCool 1979). Ten bearings of each type were tested. The following table gives the times until failure (in units of millions of cycles):

Type I	Type II
3.03	3.19
5.53	4.26
5.60	4.47
9.30	4.53
9.92	4.67
12.51	4.69
12.95	12.78
15.21	6.79
16.04	9.37
16.84	12.75

a. Use normal theory to test the hypothesis that there is no difference between the two types of bearings.
b. Test the same hypothesis using a nonparametric method.
c. Which of the methods—that of part (a) or that of part (b)—do you think is better in this case?
d. Estimate π, the probability that a type I bearing will outlast a type II bearing.
e. Use the bootstrap to estimate the sampling distribution of $\hat{\pi}$ and its standard error.
f. Use the bootstrap to find an approximate 90% confidence interval for π.

22. An experiment was done to compare two methods of measuring the calcium content of animal feeds. The standard method uses calcium oxalate precipitation followed by titration and is quite time-consuming. A new method using flame photometry is faster. Measurements of the percent calcium content made by each method of 118 routine feed samples (Heckman 1960) are contained in the file `calcium`. Analyze the data to see if there is any systematic difference between the two methods. Use both parametric and nonparametric tests and graphical methods.

23. Let $X_1, \ldots, X_n$ be i.i.d. with cdf F, and let $Y_1, \ldots, Y_m$ be i.i.d. with cdf G. The hypothesis to be tested is that $F = G$. Suppose for simplicity that $m + n$ is even so that in the combined sample of X's and Y's, $(m + n)/2$ observations are less than the median and $(m + n)/2$ are greater.

a. As a test statistic, consider T, the number of X's less than the median of the combined sample. Show that T follows a hypergeometric distribution under the null hypothesis:

$$P(T = t) = \frac{\binom{(m+n)/2}{t}\binom{(m+n)/2}{n-t}}{\binom{m+n}{n}}$$

Explain how to form a rejection region for this test.
b. Show how to find a confidence interval for the difference between the median of F and the median of G under the shift model, $G(x) = F(x - \Delta)$. (*Hint:* Use the order statistics.)
c. Apply the results (a) and (b) to the data of Problem 21.

24. Find the exact null distribution of the Mann-Whitney statistic, U_Y, in the case where $m = 3$ and $n = 2$.

25. Referring to Example A in Section 11.2.1, (a) if the smallest observation for method B (79.94) is made arbitrarily small, will the t test still reject? (b) If the largest observation for method B (80.03) is made arbitrarily large, will the t test still reject? (c) Answer the same questions for the Mann-Whitney test.

26. Let $X_1, \ldots, X_n$ be a sample from an $N(0, 1)$ distribution and let $Y_1, \ldots, Y_n$ be an independent sample from an $N(1, 1)$ distribution.

a. Determine the expected rank sum of the X's.
b. Determine the variance of the rank sum of the X's.

27. Find the exact null distribution of W_+ in the case where $n = 4$.

28. For $n = 10, 20$, and 30, find the .05 and .01 critical values for a two-sided signed rank test from the tables and then by using the normal approximation. Compare the values.

29. (Permutation Test for Means) Here is another view on hypothesis testing that we will illustrate with Example A of Section 11.2.1. We ask whether the measurements produced by methods A and B are identical or exchangeable in the following sense. There are $13 + 8 = 21$ measurements in all and there are $\binom{21}{8}$, or about 2×10^5, ways that 8 of these could be assigned to method B. Is the particular assignment we have observed unusual among these in the sense that the means of the two samples are unusually different?

a. It's not inconceivable, but it may be asking too much for you to generate all $\binom{21}{8}$ partitions. So just choose a random sample of these partitions, say of size 1000, and make a histogram of the resulting values of $\overline{X}_A - \overline{X}_B$. Where on this distribution does the value of $\overline{X}_A - \overline{X}_B$ that was actually observed fall? Compare to the result of Example B of Section 11.2.1.

b. In what way is this procedure similar to the Mann-Whitney test?

30. Use the bootstrap to estimate the standard error of and a confidence interval for $\overline{X}_A - \overline{X}_B$ and compare to the result of Example A of Section 11.2.1.

31. In Section 11.2.3, if $F = G$, what are $E(\hat{\pi})$ and $\text{Var}(\hat{\pi})$? Would there be any advantage in using equal sample sizes $m = n$ in estimating π or does it make no difference?

32. If $X \sim N(\mu_X, \sigma_X^2)$ and Y is independent $N(\mu_Y, \sigma_Y^2)$, what is $\pi = P(X < Y)$ in terms of μ_X, μ_Y, σ_X, and σ_Y?

33. To compare two variances in the normal case, let $X_1, \ldots, X_n$ be i.i.d. $N(\mu_X, \sigma_X^2)$, and let $Y_1, \ldots, Y_m$ be i.i.d. $N(\mu_Y, \sigma_Y^2)$, where the X's and Y's are independent samples. Argue that under H_0: $\sigma_X = \sigma_Y$,

$$\frac{s_X^2}{s_Y^2} \sim F_{n-1,\,m-1}$$

a. Construct rejection regions for one- and two-sided tests of H_0.

b. Construct a confidence interval for the ratio σ_X^2/σ_Y^2.

c. Apply the results of parts (a) and (b) to Example A in Section 11.2.1. (*Caution:* This test and confidence interval are not robust against violations of the assumption of normality.)

34. This problem contrasts the power functions of paired and unpaired designs. Graph and compare the power curves for testing H_0: $\mu_X = \mu_Y$ for the following two designs.

a. Paired: $\text{Cov}(X_i, Y_i) = 50$, $\sigma_X = \sigma_Y = 10$, $i = 1, \ldots, 25$.

b. Unpaired: $X_1, \ldots, X_{25}$ and $Y_1, \ldots, Y_{25}$ are independent with variance as in part (a).

35. An experiment was done to measure the effects of ozone, a component of smog. A group of 22 seventy-day-old rats were kept in an environment containing ozone for 7 days, and their weight gains were recorded. Another group of 23 rats of a similar age were kept in an ozone-free environment for a similar time, and their weight gains were recorded. The data (in grams) are given below. Analyze the data to determine the effect of ozone. Write a summary of your conclusions. [This problem is from Doksum and Sievers (1976) who provide an interesting analysis.]

Controls			Ozone		
41.0	38.4	24.9	10.1	6.1	20.4
25.9	21.9	18.3	7.3	14.3	15.5
13.1	27.3	28.5	−9.9	6.8	28.2
−16.9	17.4	21.8	17.9	−12.9	14.0
15.4	27.4	19.2	6.6	12.1	15.7
22.4	17.7	26.0	39.9	−15.9	54.6
29.4	21.4	22.7	−14.7	44.1	−9.0
26.0	26.6		−9.0		

36. Lin, Sutton, and Qurashi (1979) compared microbiological and hydroxylamine methods for the analysis of ampicillin dosages. In one series of experiments, pairs of tablets were analyzed by the two methods. The data in the following table give the percentages of claimed amount of ampicillin found by the two methods in several pairs of tablets. What are $\overline{X} - \overline{Y}$ and $s_{\overline{X}-\overline{Y}}$? If the pairing had been erroneously ignored and it had been assumed that the two samples were independent, what would have been the estimate of the standard deviation of $\overline{X} - \overline{Y}$? Analyze the data to determine if there is a systematic difference between the two methods.

Microbiological Method	Hydroxylamine Method
97.2	97.2
105.8	97.8
99.5	96.2
100.0	101.8
93.8	88.0
79.2	74.0
72.0	75.0
72.0	67.5
69.5	65.8
20.5	21.2
95.2	94.8
90.8	95.8
96.2	98.0
96.2	99.0
91.0	100.2

37. Stanley and Walton (1961) ran a controlled clinical trial to investigate the effect of the drug stelazine on chronic schizophrenics. The trials were conducted on chronic schizophrenics in two closed wards. In each of the wards, the patients were divided into two groups matched for age, length of time in the hospital, and score on a behavior rating sheet. One member of each pair was given stelazine, and the other a placebo. Only the hospital pharmacist knew which member of each pair received the actual drug. The following table gives the behavioral rating scores for the patients at the beginning of the trial and after 3 mo. High scores are good.

Ward A			
Stelazine		Placebo	
Before	After	Before	After
2.3	3.1	2.4	2.0
2.0	2.1	2.2	2.6
1.9	2.45	2.1	2.0
3.1	3.7	2.9	2.0
2.2	2.54	2.2	2.4
2.3	3.72	2.4	3.18
2.8	4.54	2.7	3.0
1.9	1.61	1.9	2.54
1.1	1.63	1.3	1.72

Ward B			
Stelazine		Placebo	
Before	After	Before	After
1.9	1.45	1.9	1.91
2.3	2.45	2.4	2.54
2.0	1.81	2.0	1.45
1.6	1.72	1.5	1.45
1.6	1.63	1.5	1.54
2.6	2.45	2.7	1.54
1.7	2.18	1.7	1.54

a. For each of the wards, test whether stelazine is associated with improvement in the patients' scores.

b. Test if there is any difference in improvement between the wards. [These data are also presented in Lehmann (1975), who discusses methods of combining the data from the wards.]

38. Bailey, Cox, and Springer (1978) used high-pressure liquid chromatography to measure the amounts of various intermediates and by-products in food dyes. The following table gives the percentages added and found for two substances in the dye FD&C Yellow No. 5. Is there any evidence that the amounts found differ systematically from the amounts added?

Sulfanilic Acid		Pyrazolone-T	
Percentage Added	Percentage Found	Percentage Added	Percentage Found
.048	.060	.035	.031
.096	.091	.087	.084
.20	.16	.19	.16
.19	.16	.19	.17
.096	.091	.16	.15
.18	.19	.032	.040
.080	.070	.060	.076
.24	.23	.13	.11
0	0	.080	.082
.040	.042	0	0
.060	.056		

39. An experiment was done to test a method for reducing faults on telephone lines (Welch 1987). Fourteen matched pairs of areas were used. The following table shows the fault rates for the control areas and for the test areas:

Test	Control
676	88
206	570
230	605
256	617
280	653
433	2913
337	924
466	286
497	1098
512	982
794	2346
428	321
452	615
512	519

a. Plot the differences versus the control rate and summarize what you see.
b. Calculate the mean difference, its standard deviation, and a confidence interval.
c. Calculate the median difference and a confidence interval and compare to the previous result.
d. Do you think it is more appropriate to use a t test or a nonparametric method to test whether the apparent difference between test and control could be due to chance? Why? Carry out both tests and compare.

40. Biological effects of magnetic fields are a matter of current concern and research. In an early study of the effects of a strong magnetic field on the development of mice (Barnothy 1964), 10 cages, each containing three 30-day-old albino female

mice, were subjected for a period of 12 days to a field with an average strength of 80 Oe/cm. Thirty other mice housed in 10 similar cages were not placed in a magnetic field and served as controls. The following table shows the weight gains, in grams, for each of the cages.

a. Display the data graphically with parallel dotplots. (Draw two parallel number lines and put dots on one corresponding to the weight gains of the controls and on the other at points corresponding to the gains of the treatment group.)
b. Find a 95% confidence interval for the difference of the mean weight gains.
c. Use a t test to assess the statistical significance of the observed difference. What is the p-value of the test?
d. Repeat using a nonparametric test.
e. What is the difference of the median weight gains?
f. Use the bootstrap to estimate the standard error of the difference of median weight gains.
g. Form a confidence interval for the difference of median weight gains based on the bootstrap approximation to the sampling distribution.

Field Present	Field Absent
22.8	23.5
10.2	31.0
20.8	19.5
27.0	26.2
19.2	26.5
9.0	25.2
14.2	24.5
19.8	23.8
14.5	27.8
14.8	22.0

41. The *Hodges-Lehmann shift estimate* is defined to be $\hat{\Delta} = \text{median}(X_i - Y_j)$, where $X_1, X_2, \ldots, X_n$ are independent observations from a distribution F and $Y_1, Y_2, \ldots, Y_m$ are independent observations from a distribution G and are independent of the X_i.

a. Show that if F and G are normal distributions, then $E(\hat{\Delta}) = \mu_X - \mu_Y$.
b. Why is $\hat{\Delta}$ robust to outliers?
c. What is $\hat{\Delta}$ for the previous problem and how does it compare to the differences of the means and of the medians?
d. Use the bootstrap to approximate the sampling distribution and the standard error of $\hat{\Delta}$.
e. From the bootstrap approximation to the sampling distribution, form an approximate 90% confidence interval for $\hat{\Delta}$.

42. Use the data of Problem 40 of Chapter 10.

a. Estimate π, the probability that more rain will fall from a randomly selected seeded cloud than from a randomly selected unseeded cloud.

b. Use the bootstrap to estimate the standard error of $\hat{\pi}$.

c. Use the bootstrap to form an approximate confidence interval for π.

43. Suppose that $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are two independent samples. As a measure of the difference in location of the two samples, the difference of the 20% trimmed means is used. Explain how the bootstrap could be used to estimate the standard error of this difference.

44. Interest in the role of vitamin C in mental illness in general and schizophrenia in particular was spurred by a paper of Linus Pauling in 1968. This exercise takes its data from a study of plasma levels and urinary vitamin C excretion in schizophrenic patients (Subotičanec et al. 1986). Twenty schizophrenic patients and 15 controls with a diagnosis of neurosis of different origin who had been patients at the same hospital for a minimum of 2 months were selected for the study. Before the experiment, all the subjects were on the same basic hospital diet. A sample of 2 ml of venous blood for vitamin C determination was drawn from each subject before breakfast and after the subjects had emptied their bladders. Each subject was then given 1 g ascorbic acid dissolved in water. No foods containing ascorbic acid were available during the test. For the next 6 h all urine was collected from the subjects for assay of vitamin C. A second blood sample was also drawn 2 h after the dose of vitamin C.

The following two tables show the plasma concentrations (mg/dl).

Schizophrenics		Nonschizophrenics	
0 h	2 h	0 h	2 h
.55	1.22	1.27	2.00
.60	1.54	.09	.41
.21	.97	1.64	2.37
.09	.45	.23	.41
1.01	1.54	.18	.79
.24	.75	.12	.94
.37	1.12	.85	1.72
1.01	1.31	.69	1.75
.26	.92	.78	1.60
.30	1.27	.63	1.80
.26	1.08	.50	2.08
.10	1.19	.62	1.58
.42	.64	.19	.86
.11	.30	.66	1.92
.14	.24	.91	1.54
.20	.89		
.09	.24		
.32	1.68		
.24	.99		
.25	.67		

a. Graphically compare the two groups at the two times and for the difference in concentration at the two times.
b. Use the t test to assess the strength of the evidence for differences between the two groups at 0 h, at 2 h, and the difference 2 h − 0 h.
c. Use the Mann-Whitney test to test the hypotheses of (b).

The following tables show the amounts of urinary vitamin C, both total and milligrams per kilogram of body weight, for the two groups:

Schizophrenics		Nonschizophrenics	
Total	mg/kg	Total	mg/kg
16.6	.19	289.4	3.96
33.3	.44	0.0	0.00
34.1	.39	620.4	7.95
0.0	.00	0.0	0.00
119.8	1.75	8.5	.10
.1	.01	5.5	.09
25.3	.27	43.2	.91
359.3	5.99	91.7	1.00
6.6	.10	200.9	3.46
.4	.01	113.8	2.01
62.8	.68	102.2	1.50
.2	.01	108.2	1.98
13.0	.15	36.9	.49
0.0	0.00	122.0	1.72
0.0	0.00	101.9	1.52
5.9	.10		
.1	.01		
6.0	.07		
32.1	.42		
0.0	0.00		

d. Use descriptive statistics and graphical presentations to compare the two groups with respect to total excretion and mg/kg body weight. Do the data look normally distributed?
e. Use a t test to compare the two groups on both variables. Is the normality assumption reasonable?
f. Use the Mann-Whitney test to compare the two groups. How do the results compare with those obtained in part (e)?

The lower levels of plasma vitamin C in the schizophrenics before administration of ascorbic acid could be attributed to several factors. Interindividual differences in the intake of meals cannot be excluded, despite the fact that all patients were offered the same food. A more interesting possibility is that the differences are the result of poorer resorption or of higher ascorbic acid utilization in schizophrenics. In order to answer this question, another

experiment was run on 15 schizophrenics and 15 controls. All subjects were given 70 mg of ascorbic acid daily for 4 weeks before the ascorbic acid loading test. The following table shows the concentration of plasma vitamin C (mg/dl) and the 6-h urinary excretion (mg) after administration of 1 g ascorbic acid.

Schizophrenics		Controls	
Plasma	Urine	Plasma	Urine
.72	86.20	1.02	190.14
1.11	21.55	.86	149.76
.96	182.07	.78	285.27
1.23	88.28	1.38	244.93
.76	76.58	.95	184.45
.75	18.81	1.00	135.34
1.26	50.02	.47	157.74
.64	107.74	.60	125.65
.67	.09	1.15	164.98
1.05	113.23	.86	99.65
1.28	34.38	.61	86.29
.54	8.44	1.01	142.23
.77	109.03	.77	144.60
1.11	144.44	.77	265.40
.51	172.09	.94	28.26

g. Use graphical methods and descriptive statistics to compare the two groups with respect to plasma concentrations and urinary excretion.

h. Use the t test to compare the two groups on the two variables. Does the normality assumption look reasonable?

i. Compare the two groups using the Mann-Whitney test.

45. This and the next two problems are based on discussions and data in Le Cam and Neyman (1967), which is devoted to the analysis of weather modification experiments. The examples illustrate some ways in which principles of experimental design have been used in this field. During the summers of 1957 through 1960, a series of randomized cloud-seeding experiments were carried out in the mountains of Arizona. Of each pair of successive days, one day was randomly selected for seeding to be done. The seeding was done during a two-hour to four-hour period starting at midday, and rainfall during the afternoon was measured by a network of 29 gauges. The data for the four years are given in the following table (in inches). Observations in this table are listed in chronological order.

a. Analyze the data for each year and for the years pooled together to see if there appears to be any effect due to seeding. You should use graphical descriptive methods to get a qualitative impression of the results and hypothesis tests to assess the significance of the results.

b. Why should the day on which seeding is to be done be chosen at random rather than just alternating seeded and unseeded days? Why should the days be paired at all, rather than just deciding randomly which days to seed?

1957		1958		1959		1960	
Seeded	Unseeded	Seeded	Unseeded	Seeded	Unseeded	Seeded	Unseeded
0	.154	.152	.013	.015	0	0	.010
.154	0	0	0	0	0	0	0
.003	.008	0	.445	0	.086	.042	.057
.084	.033	.002	0	.021	.006	0	0
.002	.035	.007	.079	0	.115	0	.093
.157	.007	.013	.006	.004	.090	0	.183
.010	.140	.161	.008	.010	0	.152	0
0	.022	0	.001	0	0	0	0
.002	0	.274	.001	.055	0	0	0
.078	.074	.001	.025	.004	.076	0	0
.101	.002	.122	.046	.053	.090	0	0
.169	.318	.101	.007	0	0	0	0
.139	.096	.012	.019	0	.078	.008	0
.172	0	.002	0	.090	.121	.040	.060
0	0	.066	0	.028	1.027	.003	.102
0	.050	.040	.012	0	.104	.011	.041
				.032	.023		
				.133	.172		
				.083	.002		
					0	0	

46. The National Weather Bureau's ACN cloud-seeding project was carried out in the states of Oregon and Washington. Cloud seeding was accomplished by dispersing dry ice from an aircraft; only clouds that were deemed "ripe" for seeding were candidates for seeding. On each occasion, a decision was made at random whether to seed, the probability of seeding being $\frac{2}{3}$. This resulted in 22 seeded and 13 control cases. Three types of targets were considered, two of which are dealt with in this problem. Type I targets were large geographical areas downwind from the seeding; type II targets were sections of type I targets located so as to have, theoretically, the greatest sensitivity to cloud seeding. The following table gives the average target rainfalls (in inches) for the seeded and control cases, listed in chronological order. Is there evidence that seeding has an effect on either type of target? In what ways is the design of this experiment different from that of the one in Problem 45?

Control Cases		Seeded Cases	
Type I	Type II	Type I	Type II
.0080	.0000	.1218	.0200
.0046	.0000	.0403	.0163
.0549	.0053	.1166	.1560
.1313	.0920	.2375	.2885
.0587	.0220	.1256	.1483
.1723	.1133	.1400	.1019
.3812	.2880	.2439	.1867
.1720	.0000	.0072	.0233
.1182	.1058	.0707	.1067
.1383	.2050	.1036	.1011
.0106	.0100	.1632	.2407
.2126	.2450	.0788	.0666
.1435	.1529	.0365	.0133
		.2409	.2897
		.0408	.0425
		.2204	.2191
		.1847	.0789
		.3332	.3570
		.0676	.0760
		.1097	.0913
		.0952	.0400
		.2095	.1467

47. During 1963 and 1964, an experiment was carried out in France; its design differed somewhat from those of the previous two problems. A 1500-km target area was selected, and an adjacent area of about the same size was designated as the control area; 33 ground generators were used to produce silver iodide to seed the target area. Precipitation was measured by a network of gauges for each suitable "rainy period," which was defined as a sequence of periods of continuous precipitation between dry spells of a specified length. When a forecaster determined that the situation was favorable for seeding, he telephoned an order to a service agent, who then opened a sealed envelope that contained an order to actually seed or not. The envelopes had been prepared in advance, using a table of random numbers. The following table gives precipitation (in inches) in the target and control areas for the seeded and unseeded periods.

a. Analyze the data, which are listed in chronological order, to see if there is an effect of seeding.

b. The analysis done by the French investigators used the square root transformation in order to make normal theory more applicable. Do you think that taking the square root was an effective transformation for this purpose?

c. Reflect on the nature of this design. In particular, what advantage is there to using the control area? Why not just compare seeded and unseeded periods on the target area?

Seeded		Unseeded	
Target	Control	Target	Control
1.6	1.0	1.1	2.2
28.1	27.0	3.5	5.2
7.8	.3	2.6	0.0
4.0	6.0	2.6	2.0
9.6	12.6	9.8	4.9
0.2	0.5	5.6	8.5
18.7	8.7	.1	3.5
16.5	21.5	0.0	1.1
4.6	13.9	17.7	11.0
9.3	6.7	19.4	19.8
3.5	4.5	8.9	5.3
0.1	0.7	10.6	8.9
11.5	8.7	10.2	4.5
0.0	0.0	16.0	13.0
9.3	10.7	9.7	21.1
5.5	4.7	21.4	15.9
70.2	29.1	6.1	19.5
0.7	1.9	24.3	16.3
38.6	34.7	20.9	6.3
11.3	10.2	60.2	47.0
3.3	2.7	15.2	10.8
8.9	2.8	2.7	4.8
11.1	4.3	0.3	0.0
64.3	38.7	12.2	5.7
16.6	11.1	2.2	5.1
7.3	6.5	23.3	30.6
3.2	3.0	9.9	3.7
23.9	13.6		
0.6	0.1		

48. Proteinuria, the presence of excess protein in urine, is a symptom of renal (kidney) distress among diabetics. Taguma et al. (1985) studied the effects of captopril for treating proteinuria in diabetics. Urinary protein was measured for 12 patients before and after eight weeks of captopril therapy. The amounts of urinary protein (in g/24 hrs) before and after therapy are shown in the following table. What can you conclude about the effect of captopril? Consider using parametric or nonparametric methods and analyzing the data on the original scale or on a log scale.

Before	After
24.6	10.1
17.0	5.7
16.0	5.6
10.4	3.4
8.2	6.5
7.9	0.7
8.2	6.5
7.9	0.7
5.8	6.1
5.4	4.7
5.1	2.0
4.7	2.9

49. Egyptian researchers, Kamal et al. (1991), took a sample of 126 police officers subject to inhalation of vehicle exhaust in downtown Cairo and found an average blood level concentration of lead equal to 29.2 μg/dl with a standard deviation of 7.5 μg/dl. A sample of 50 policemen from a suburb, Abbasia, had an average concentration of 18.2 μg/dl and a standard deviation of 5.8 μg/dl. Form a confidence interval for the population difference and test the null hypothesis that there is no difference in the populations.

50. The file `bodytemp` contains normal body temperature readings (degrees Fahrenheit) and heart rates (beats per minute) of 65 males (coded by 1) and 65 females (coded by 2) from Shoemaker (1996).

a. Using normal theory, form a 95% confidence interval for the difference of mean body temperatures between males and females. Is the use of the normal approximation reasonable?

b. Using normal theory, form a 95% confidence interval for the difference of mean heart rates between males and females. Is the use of the normal approximation reasonable?

c. Use both parametric and nonparametric tests to compare the body temperatures and heart rates. What do you conclude?

51. A common symptom of otitis-media (inflamation of the middle ear) in young children is the prolonged presence of fluid in the middle ear, called *middle-ear effusion.* It is hypothesized that breast-fed babies tend to have less prolonged effusions than do bottle-fed babies. Rosner (2006) presents the results of a study of 24 pairs of infants who were matched according to sex, socioeconomic status, and type of medication taken. One member of each pair was bottle-fed and the other was breast-fed. The file `ears` gives the durations (in days) of middle-ear effusions after the first episode of otitis-media.

a. Examine the data using graphical methods and summarize your conclusions.

b. In order to test the hypothesis of no difference, do you think it is more appropriate to use a parametric or a nonparametric test? Carry out a test. What do you conclude?

52. The media often present short reports of the results of experiments. To the critical reader or listener, such reports often raise more questions than they answer. Comment on possible pitfalls in the interpretation of each of the following.

a. It is reported that patients whose hospital rooms have a window recover faster than those whose rooms do not.

b. Nonsmoking wives whose husbands smoke have a cancer rate twice that of wives whose husbands do not smoke.

c. A 2-year study in North Carolina found that 75% of all industrial accidents in the state happened to workers who had skipped breakfast.

d. A school integration program involved busing children from minority schools to majority (primarily white) schools. Participation in the program was voluntary. It was found that the students who were bused scored lower on standardized tests than did their peers who chose not to be bused.

e. When a group of students were asked to match pictures of newborns with pictures of their mothers, they were correct 36% of the time.

f. A survey found that those who drank a moderate amount of beer were healthier than those who totally abstained from alcohol.

g. A 15-year study of more than 45,000 Swedish soldiers revealed that heavy users of marijuana were six times more likely than nonusers to develop schizophrenia.

h. A University of Wisconsin study showed that within 10 years of the wedding, 38% of those who had lived together before marriage had split up, compared to 27% of those who had married without a "trial period."

i. A study of nearly 4,000 elderly North Carolinians has found that those who attended religious services every week were 46% less likely to die over a six-year period than people who attended less often or not at all, according to researchers at Duke University Medical Center.

53. Explain why in Levine's experiment (Example A in Section 11.3.1) subjects also smoked cigarettes made of lettuce leaves and unlit cigarettes.

54. This example is taken from an interesting article by Joiner (1981) and from data in Ryan, Joiner, and Ryan (1976). The National Institute of Standards and Technology supplies standard materials of many varieties to manufacturers and other parties, who use these materials to calibrate their own testing equipment. Great pains are taken to make these reference materials as homogeneous as possible. In an experiment, a long homogeneous steel rod was cut into 4-inch lengths, 20 of which were randomly selected and tested for oxygen content. Two measurements were made on each piece. The 40 measurements were made over a period of 5 days, with eight measurements per day. In order to avoid possible bias from time-related trends, the sequence of measurements was randomized. The file `steelrods` contains the measurements. There is an unexpected systematic source of variability in these data. Can you find it by making an appropriate plot? Would this effect have been detectable if the measurements had not been randomized over time?