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:

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

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

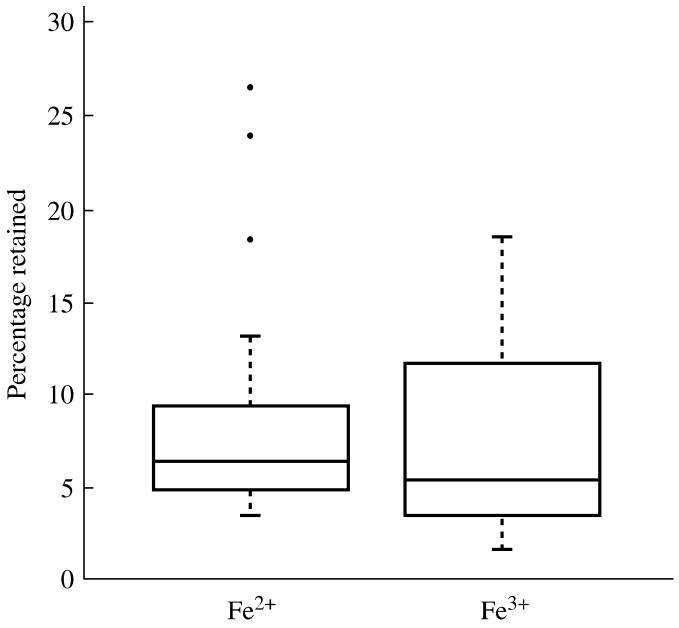

FIGURE **11.2** Boxplots of the percentages of iron retained for the two forms.

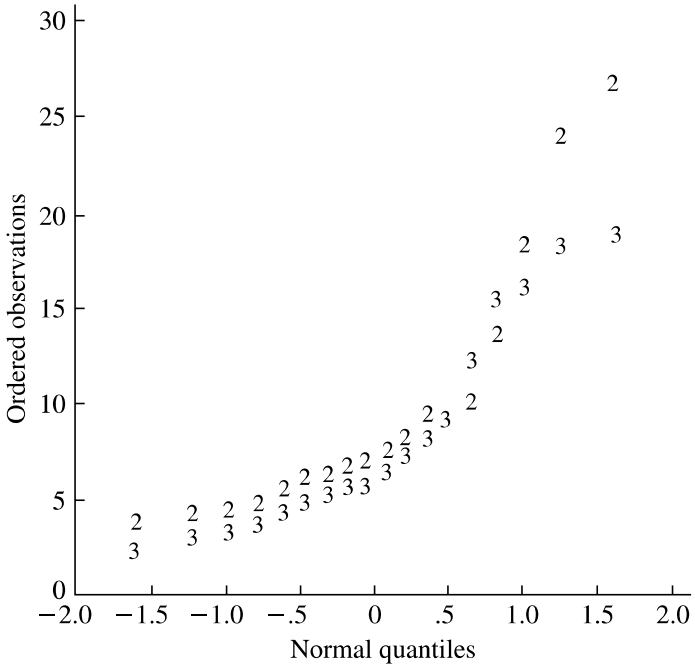

FIGURE **11.3** Normal probability plots of iron retention data.

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.

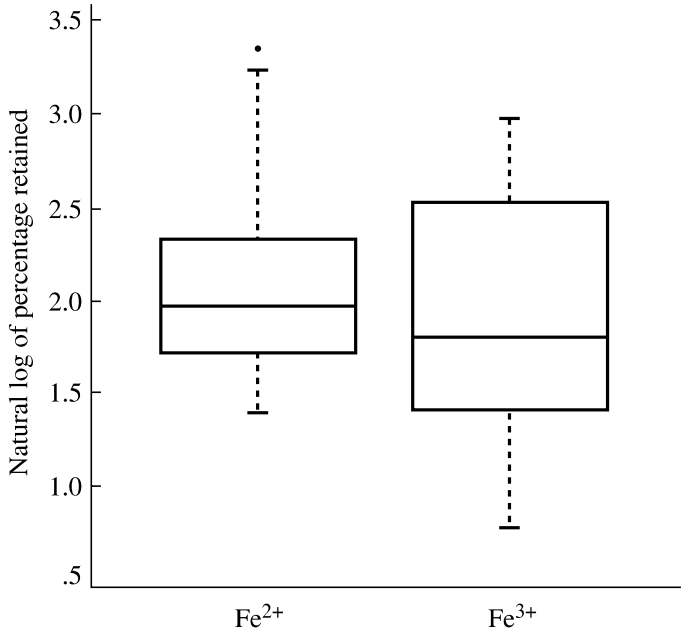

FIGURE **11.4** Boxplots of natural logs of percentages of iron retained.

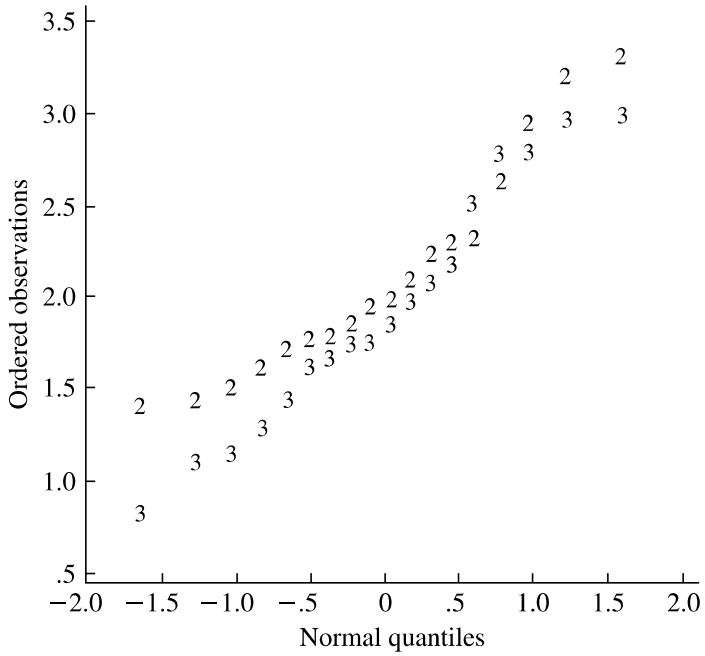

FIGURE **11.5** Normal probability plots of natural logs of iron retention data.

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

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

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

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

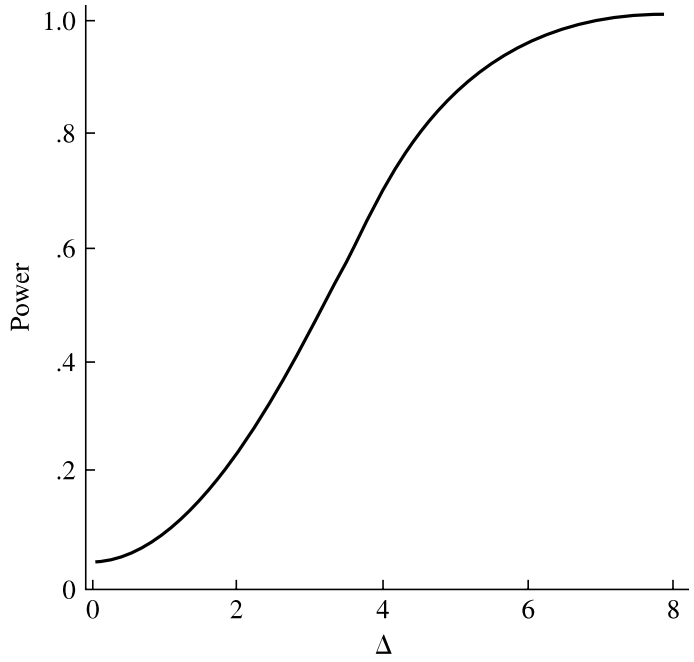

FIGURE **11.6** Plot of power versus Δ.

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.