Chapter 8

Tests of Statistical Hypotheses

8.1 TESTS ABOUT ONE MEAN

We begin this chapter on tests of statistical hypotheses with an application in which we define many of the terms associated with testing.

Example 8.1-1 Let X equal the breaking strength of a steel bar. If the bar is manufactured by process I, X is $N(50, 36)$, i.e., X is normally distributed with $\mu = 50$ and $\sigma^2 = 36$. It is hoped that if process II (a new process) is used, X will be $N(55, 36)$. Given a large number of steel bars manufactured by process II, how could we test whether the five-unit increase in the mean breaking strength was realized?

In this problem, we are assuming that X is $N(\mu, 36)$ and μ is equal to 50 or 55. We want to test the **simple null hypothesis** H_0: $\mu = 50$ against the **simple alternative hypothesis** H_1: $\mu = 55$. Note that each of these hypotheses completely specifies the distribution of X. That is, H_0 states that X is $N(50, 36)$ and H_1 states that X is $N(55, 36)$. (If the alternative hypothesis had been H_1: $\mu > 50$, it would be a **composite hypothesis**, because it is composed of all normal distributions with = 36 and means greater than 50.) In order to test which of the two hypotheses, H_0 or H_1, is true, we shall set up a rule based on the breaking strengths $x_1, x_2, \ldots, x_n$ of n bars (the observed values of a random sample of size n from this new normal distribution). The rule leads to a decision to accept or reject H_0; hence, it is necessary to partition the sample space into two parts—say, C and C'—so that if $(x_1, x_2, \ldots, x_n) \in C$, H_0 is rejected, and if $(x_1, x_2, \ldots, x_n) \in C'$, H_0 is accepted (not rejected). The rejection region C for H_0 is called the **critical region** for the test. Often, the partitioning of the sample space is specified in terms of the values of a statistic called the **test statistic**. In this example, we could let $\overline{X}$ be the test statistic and, say, take $C = \{(x_1, x_2, \ldots, x_n) : \bar{x} \geq 53\}$; that is, we will reject H_0 if $\bar{x} \geq 53$. If $(x_1, x_2, \ldots, x_n) \in C$ when H_0 is true, H_0 would be rejected when it is true, a **Type I error**. If $(x_1, x_2, \ldots, x_n) \in C'$ when H_1 is true, H_0 would be accepted (i.e., not rejected) when in fact H_1 is true, a **Type II error**. The probability of a Type I error is called the **significance level** of the test and is denoted by α. That is, $\alpha = P[(X_1, X_2, \ldots, X_n) \in C; H_0]$ is the probability that $(X_1, X_2, \ldots, X_n)$

falls into C when H_0 is true. The probability of a Type II error is denoted by β; that is, $\beta = P[(X_1, X_2, \ldots, X_n) \in C'; H_1]$ is the probability of accepting (failing to reject) H_0 when it is false.

As an illustration, suppose $n = 16$ bars were tested and $C = \{\overline{x} : \overline{x} \geq 53\}$. Then $\overline{X}$ is $N(50, 36/16)$ when H_0 is true and is $N(55, 36/16)$ when H_1 is true. Thus,

$$\begin{aligned}\alpha = P(\overline{X} \geq 53; H_0) &= P\left(\frac{\overline{X} - 50}{6/4} \geq \frac{53 - 50}{6/4}; H_0\right) \\ &= 1 - \Phi(2) = 0.0228\end{aligned}$$

and

$$\begin{aligned}\beta = P(\overline{X} < 53; H_1) &= P\left(\frac{\overline{X} - 55}{6/4} < \frac{53 - 55}{6/4}; H_1\right) \\ &= \Phi\left(-\frac{4}{3}\right) = 1 - 0.9087 = 0.0913.\end{aligned}$$

Figure 8.1-1 shows the graphs of the probability density functions of $\overline{X}$ when H_0 and H_1, respectively, are true. Note that by changing the critical region, C, it is possible to decrease (increase) the size of α but this leads to an increase (decrease) in the size of β. Both α and β can be decreased if the sample size n is increased. ■

Through another example, we define a p-value obtained in testing a hypothesis about a mean.

Example 8.1-2 Assume that the underlying distribution is normal with unknown mean μ but known variance $\sigma^2 = 100$. Say we are testing the simple null hypothesis H_0: $\mu = 60$ against the composite alternative hypothesis H_1: $\mu > 60$ with a sample mean $\overline{X}$ based on $n = 52$ observations. Suppose that we obtain the observed sample mean of $\overline{x} = 62.75$. If we compute the probability of obtaining an $\overline{X}$ of that value of 62.75 or greater when $\mu = 60$, then we obtain the **p-value** associated with $\overline{x} = 62.75$. That is,

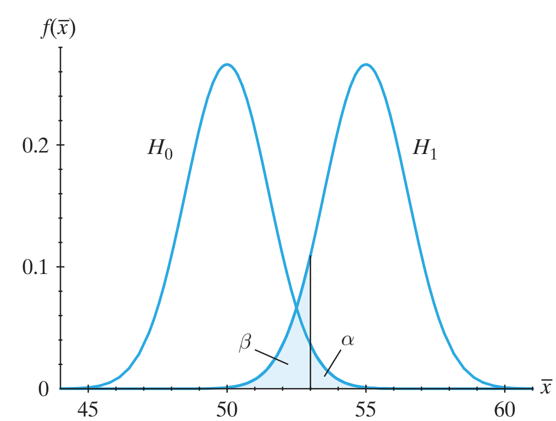

Figure 8.1-1 pdf of $\overline{X}$ under H_0 and H_1

$$\begin{aligned} p\text{-value} &= P(\overline{X} \ge 62.75;\ \mu = 60) \\ &= P\left(\frac{\overline{X} - 60}{10/\sqrt{52}} \ge \frac{62.75 - 60}{10/\sqrt{52}};\ \mu = 60\right) \\ &= 1 - \Phi\left(\frac{62.75 - 60}{10/\sqrt{52}}\right) = 1 - \Phi(1.983) = 0.0237. \end{aligned}$$

If this p-value is small, we tend to reject the hypothesis H_0: $\mu = 60$. For example, rejecting H_0: $\mu = 60$ if the p-value is less than or equal to $\alpha = 0.05$ is exactly the same as rejecting H_0 if

$$\overline{x} \ge 60 + (1.645)\left(\frac{10}{\sqrt{52}}\right) = 62.281.$$

Here

$$p\text{-value} = 0.0237 < \alpha = 0.05 \qquad \text{and} \qquad \overline{x} = 62.75 > 62.281.$$

To help the reader keep the definition of p-value in mind, we note that it can be thought of as that **tail-end probability**, under H_0, of the distribution of the statistic (here $\overline{X}$) beyond the observed value of the statistic. (See Figure 8.1-2 for the p-value associated with $\overline{x} = 62.75$.)

If the alternative were the two-sided H_1: $\mu \ne 60$, then the p-value would have been double 0.0237; that is, then the p-value $= 2(0.0237) = 0.0474$ because we include both tails.

When we sample from a normal distribution, the null hypothesis is generally of the form H_0: $\mu = \mu_0$. There are three possibilities of interest for a composite alternative hypothesis: (i) that μ has increased, or H_1: $\mu > \mu_0$; (ii) that μ has decreased, or H_1: $\mu < \mu_0$; and (iii) that μ has changed, but it is not known whether it has increased or decreased, which leads to the two-sided alternative hypothesis, or H_1: $\mu \ne \mu_0$.

To test H_0: $\mu = \mu_0$ against one of these three alternative hypotheses, a random sample is taken from the distribution and an observed sample mean, $\overline{x}$, that is close to μ_0 supports H_0. The closeness of $\overline{x}$ to μ_0 is measured in terms of standard deviations of $\overline{X}$, $\sigma/\sqrt{n}$, when σ is known, a measure that is sometimes called the **standard error of the mean**. Thus, the test statistic could be defined by

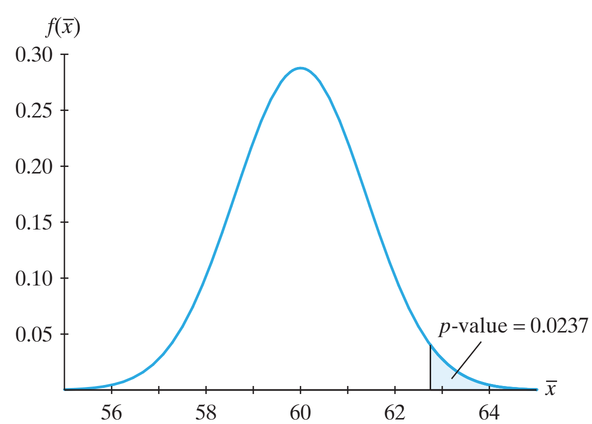

Figure 8.1-2 Illustration of p-value

Table 8.1-1 Tests of hypotheses about one mean, variance known

H_0	H_1	Critical Region
$\mu = \mu_0$	$\mu > \mu_0$	$z \ge z_\alpha$ or $\overline{x} \ge \mu_0 + z_\alpha \sigma/\sqrt{n}$
$\mu = \mu_0$	$\mu < \mu_0$	$z \le -z_\alpha$ or $\overline{x} \le \mu_0 - z_\alpha \sigma/\sqrt{n}$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\lvert z\rvert \ge z_{\alpha/2}$ or $\lvert\overline{x} - \mu_0\rvert \ge z_{\alpha/2}\sigma/\sqrt{n}$

$$Z = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}, \tag{8.1-1}$$

and the critical regions, at a significance level α, for the three respective alternative hypotheses would be (i) $z \ge z_\alpha$, (ii) $z \le -z_\alpha$, and (iii) $|z| \ge z_{\alpha/2}$. In terms of $\overline{x}$, these three critical regions become (i) $\overline{x} \ge \mu_0 + z_\alpha(\sigma/\sqrt{n}\,)$, (ii) $\overline{x} \le \mu_0 - z_\alpha(\sigma/\sqrt{n}\,)$, and (iii) $|\overline{x} - \mu_0| \ge z_{\alpha/2}(\sigma/\sqrt{n}\,)$.

The three tests and critical regions are summarized in Table 8.1-1. The underlying assumption is that the distribution is $N(\mu, \sigma^2)$ and σ^2 is known.

It is usually the case that the variance σ^2 is not known. Accordingly, we now take a more realistic position and assume that the variance is unknown. Suppose our null hypothesis is H_0: $\mu = \mu_0$ and the two-sided alternative hypothesis is H_1: $\mu \ne \mu_0$. Recall from Section 7.1, for a random sample $X_1, X_2, \ldots, X_n$ taken from a normal distribution $N(\mu, \sigma^2)$, a confidence interval for μ is based on

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

This suggests that T might be a good statistic to use for the test of H_0: $\mu = \mu_0$ with μ replaced by μ_0. In addition, it is the natural statistic to use if we replace σ^2/n by its unbiased estimator S^2/n in $(\overline{X} - \mu_0)/\sqrt{\sigma^2/n}$ in Equation 8.1-1. If $\mu = \mu_0$, we know that T has a t distribution with $n-1$ degrees of freedom. Thus, with $\mu = \mu_0$,

$$P[\,|T| \ge t_{\alpha/2}(n-1)] = P\left[\frac{|\overline{X} - \mu_0|}{S/\sqrt{n}} \ge t_{\alpha/2}(n-1)\right] = \alpha.$$

Accordingly, if $\overline{x}$ and s are, respectively, the sample mean and sample standard deviation, then the rule that rejects H_0: $\mu = \mu_0$ and accepts H_1: $\mu \ne \mu_0$ if and only if

$$|t| = \frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \ge t_{\alpha/2}(n-1)$$

provides a test of this hypothesis with significance level α. Note that this rule is equivalent to rejecting H_0: $\mu = \mu_0$ if μ_0 is not in the open $100(1-\alpha)\%$ confidence interval

$$\left(\overline{x} - t_{\alpha/2}(n-1)\left[s/\sqrt{n}\,\right], \overline{x} + t_{\alpha/2}(n-1)\left[s/\sqrt{n}\,\right]\right).$$

Table 8.1-2 summarizes tests of hypotheses for a single mean, along with the three possible alternative hypotheses, when the underlying distribution is $N(\mu, \sigma^2)$, σ^2 is unknown, $t = (\overline{x} - \mu_0)/(s/\sqrt{n}\,)$, and $n \le 30$. If $n > 30$, we use Table 8.1-1 for approximate tests, with σ replaced by s.

Table 8.1-2 Tests of hypotheses for one mean, variance unknown

H_0	H_1	Critical Region
$\mu = \mu_0$	$\mu > \mu_0$	$t \ge t_\alpha(n-1)$ or $\bar{x} \ge \mu_0 + t_\alpha(n-1)s/\sqrt{n}$
$\mu = \mu_0$	$\mu < \mu_0$	$t \le -t_\alpha(n-1)$ or $\bar{x} \le \mu_0 - t_\alpha(n-1)s/\sqrt{n}$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\lvert t\rvert \ge t_{\alpha/2}(n-1)$ or $\lvert\bar{x} - \mu_0\rvert \ge t_{\alpha/2}(n-1)s/\sqrt{n}$

Example 8.1-3

Let X (in millimeters) equal the growth in 15 days of a tumor induced in a mouse. Assume that the distribution of X is $N(\mu, \sigma^2)$. We shall test the null hypothesis H_0: $\mu = \mu_0 = 4.0$ mm against the two-sided alternative hypothesis H_1: $\mu \ne 4.0$. If we use $n = 9$ observations and a significance level of $\alpha = 0.10$, the critical region is

$$|t| = \frac{|\bar{x} - 4.0|}{s/\sqrt{9}} \ge t_{\alpha/2}(8) = 1.860.$$

If we are given that $n = 9, \bar{x} = 4.3$, and $s = 1.2$, we see that

$$t = \frac{4.3 - 4.0}{1.2/\sqrt{9}} = \frac{0.3}{0.4} = 0.75.$$

Thus,

$$|t| = |0.75| < 1.860,$$

and we accept (do not reject) H_0: $\mu = 4.0$ at the $\alpha = 10\%$ significance level. (See Figure 8.1-3.) The p-value is the two-sided probability of $|T| \ge 0.75$, namely,

$$p\text{-value} = P(|T| \ge 0.75) = 2P(T \ge 0.75).$$

With our t tables with eight degrees of freedom, we cannot find this p-value exactly. It is about 0.50, because

$$P(|T| \ge 0.706) = 2P(T \ge 0.706) = 0.50.$$

However, Minitab gives a p-value of 0.4747. (See Figure 8.1-3.) ■

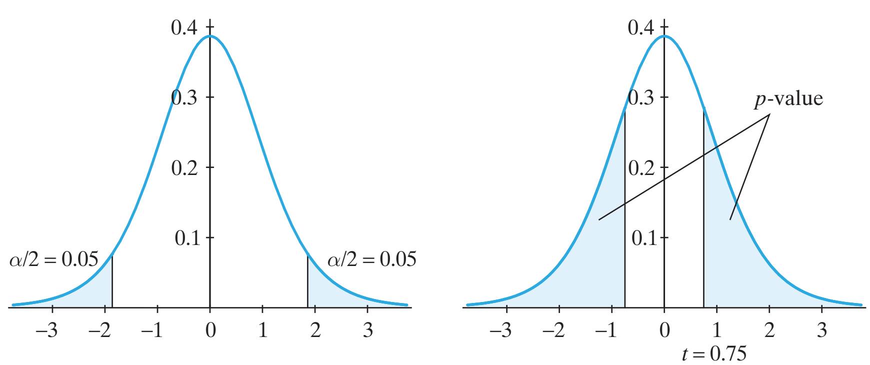

Figure 8.1-3 Test about mean of tumor growths

REMARK In discussing the test of a statistical hypothesis, the word *accept* H_0 might better be replaced by *do not reject* H_0. That is, if, in Example 8.1-3, $\bar{x}$ is close enough to 4.0 so that we accept $\mu = 4.0$, we do not want that acceptance to imply that μ is actually equal to 4.0. We want to say that the data do not deviate enough from $\mu = 4.0$ for us to reject that hypothesis; that is, we do not reject $\mu = 4.0$ with these observed data. With this understanding, we sometimes use *accept*, and sometimes *fail to reject* or *do not reject*, the null hypothesis. ■

The next example illustrates the use of the t statistic with a one-sided alternative hypothesis.

Example 8.1-4 In attempting to control the strength of the wastes discharged into a nearby river, a paper firm has taken a number of measures. Members of the firm believe that they have reduced the oxygen-consuming power of their wastes from a previous mean μ of 500 (measured in parts per million of permanganate). They plan to test H_0: $\mu = 500$ against H_1: $\mu < 500$, using readings taken on $n = 25$ consecutive days. If these 25 values can be treated as a random sample, then the critical region, for a significance level of $\alpha = 0.01$, is

$$t = \frac{\bar{x} - 500}{s/\sqrt{25}} \le -t_{0.01}(24) = -2.492.$$

The observed values of the sample mean and sample standard deviation were $\bar{x} = 308.8$ and $s = 115.15$. Since

$$t = \frac{308.8 - 500}{115.15/\sqrt{25}} = -8.30 < -2.492,$$

we clearly reject the null hypothesis and accept H_1: $\mu < 500$. Note, however, that although an improvement has been made, there still might exist the question of whether the improvement is adequate. The one-sided 99% confidence interval for μ, namely,

$$[0, 308.8 + 2.492(115.25/\sqrt{25}\,)] = [0,\ 366.191],$$

provides an upper bound for μ and may help the company answer this question. ■

Oftentimes, there is interest in comparing the means of two different distributions or populations. We must consider two situations: that in which X and Y are dependent and that in which X and Y are independent. We consider the independent case in the next section.

If X and Y are dependent, let $W = X - Y$, and the hypothesis that $\mu_X = \mu_Y$ would be replaced with the hypothesis H_0: $\mu_W = 0$. For example, suppose that X and Y equal the resting pulse rate for a person before and after taking an eight-week program in aerobic dance. We would be interested in testing H_0: $\mu_W = 0$ (no change) against H_1: $\mu_W > 0$ (the aerobic dance program decreased the mean resting pulse rate). Because X and Y are measurements on the same person, X and Y are clearly dependent. If we can assume that the distribution of W is (approximately) $N(\mu_W, \sigma^2)$, then we can choose to use the appropriate t test for a single mean from Table 8.1-2. This is often called a **paired** t **test**.

Example 8.1-5

Twenty-four girls in the 9th and 10th grades were put on an ultraheavy rope-jumping program. Someone thought that such a program would increase their speed in the 40-yard dash. Let W equal the difference in time to run the 40-yard dash—the "before-program time" minus the "after-program time." Assume that the distribution of W is approximately $N(\mu_W, \sigma_W^2)$. We shall test the null hypothesis H_0: $\mu_W = 0$ against the alternative hypothesis H_1: $\mu_W > 0$. The test statistic and the critical region that has an $\alpha = 0.05$ significance level are given by

$$t = \frac{\overline{w} - 0}{s_w/\sqrt{24}} \geq t_{0.05}(23) = 1.714.$$

The following data give the difference in time that it took each girl to run the 40-yard dash, with positive numbers indicating a faster time after the program:

0.28	0.01	0.13	0.33	−0.03	0.07	−0.18	−0.14
−0.33	0.01	0.22	0.29	−0.08	0.23	0.08	0.04
−0.30	−0.08	0.09	0.70	0.33	−0.34	0.50	0.06

For these data, $\overline{w} = 0.0788$ and $s_w = 0.2549$. Thus, the observed value of the test statistic is

$$t = \frac{0.0788 - 0}{0.2549/\sqrt{24}} = 1.514.$$

Since $1.514 < 1.714$, the null hypothesis is not rejected. Note, however, that $t_{0.10}(23) = 1.319$ and $t = 1.514 > 1.319$. Hence, the null hypothesis would be rejected at an $\alpha = 0.10$ significance level. Another way of saying this is that

$$0.05 < p\text{-value} < 0.10.$$

It would be instructive to draw a figure illustrating this double inequality. ■

There are two ways of viewing a statistical test. One of these is through the p-value of the test; this approach is becoming more popular and is included in most computer printouts, so we mention it again. After observing the test statistic, we can say that the p-value is the probability, under the hypothesis H_0, of the test statistic being at least as extreme (in the direction of rejection of H_0) as the observed one. That is, the p-value is the tail-end probability. As an illustration, say a golfer averages about 90 for an 18-hole round, with a standard deviation of 3, and she takes some lessons to improve. To test her possible improvement, namely, H_0: $\mu = 90$, against H_1: $\mu < 90$, she plays $n = 16$ rounds of golf. Assume a normal distribution with $\sigma = 3$. If the golfer averaged $\overline{x} = 87.9375$, then

$$p\text{-value} = P(\overline{X} \leq 87.9375) = P\left(\frac{\overline{X} - 90}{3/4} \leq \frac{87.9375 - 90}{3/4}\right) = 0.0030.$$

The fact that the p-value is less than 0.05 is equivalent to the fact that $\overline{x} < 88.77$, because $P(\overline{X} \leq 88.77;\ \mu = 90) = 0.05$. Since $\overline{x} = 87.9375$ is an observed value of a random variable, namely, $\overline{X}$, it follows that the p-value, a function of $\overline{x}$, is also an observed value of a random variable. That is, before the random experiment is performed, the probability that the p-value is less than or equal to α is approximately equal to α when the null hypothesis is true. Many statisticians believe that the observed p-value provides an understandable measure of the truth of H_0: The smaller the p-value, the less they believe in H_0.

Two additional examples of the p-value may be based on Examples 8.1-3 and 8.1-4. In two-sided tests for means and proportions, the p-value is the probability of the extreme values in both directions. With the mouse data (Example 8.1-3), the p-value is

$$p\text{-value} = P(|T| \geq 0.75).$$

In Table VI in Appendix B, we see that if T has a t distribution with eight degrees of freedom, then $P(T \geq 0.706) = 0.25$. Thus, $P(|T| \geq 0.706) = 0.50$ and the p-value will be a little smaller than 0.50. In fact, $P(|T| \geq 0.75) = 0.4747$ (a probability that was found with Minitab), which is not less than $\alpha = 0.10$; hence, we do not reject H_0 at that significance level. In the example concerned with waste (Example 8.1-4), the p-value is essentially zero, since $P(T \leq -8.30) \approx 0$, where T has a t distribution with 24 degrees of freedom. Consequently, we reject H_0.

The other way of looking at tests of hypotheses is through the consideration of confidence intervals, particularly for two-sided alternatives and the corresponding tests. For example, with the mouse data (Example 8.1-3), a 90% confidence interval for the unknown mean is

$$4.3 \pm (1.86)(1.2)/\sqrt{9}, \qquad \text{or} \qquad [3.56, 5.04],$$

since $t_{0.05}(8) = 1.86$. Note that this confidence interval covers the hypothesized value $\mu = 4.0$ and we do not reject H_0: $\mu = 4.0$. If the confidence interval did not cover $\mu = 4.0$, then we would have rejected H_0: $\mu = 4.0$. Many statisticians believe that estimation is much more important than tests of hypotheses and accordingly approach statistical tests through confidence intervals. For one-sided tests, we use one-sided confidence intervals.

Exercises

8.1-1. Assume that IQ scores for a certain population are approximately $N(\mu, 100)$. To test H_0: $\mu = 110$ against the one-sided alternative hypothesis H_1: $\mu > 110$, we take a random sample of size $n = 16$ from this population and observe $\overline{x} = 113.5$.

(a) Do we accept or reject H_0 at the 5% significance level?

(b) Do we accept or reject H_0 at the 10% significance level?

(c) What is the p-value of this test?

8.1-2. Assume that the weight of cereal in a "12.6-ounce box" is $N(\mu, 0.2^2)$. The Food and Drug Association (FDA) allows only a small percentage of boxes to contain less than 12.6 ounces. We shall test the null hypothesis H_0: $\mu = 13$ against the alternative hypothesis H_1: $\mu < 13$.

(a) Use a random sample of $n = 25$ to define the test statistic and the critical region that has a significance level of $\alpha = 0.025$.

(b) If $\overline{x} = 12.9$, what is your conclusion?

(c) What is the p-value of this test?

8.1-3. Let X equal the Brinell hardness measurement of ductile iron subcritically annealed. Assume that the distribution of X is $N(\mu, 100)$. We shall test the null hypothesis H_0: $\mu = 170$ against the alternative hypothesis H_1: $\mu > 170$, using $n = 25$ observations of X.

(a) Define the test statistic and a critical region that has a significance level of $\alpha = 0.05$. Sketch a figure showing this critical region.

(b) A random sample of $n = 25$ observations of X yielded the following measurements:

170	167	174	179	179	156	163	156	187
156	183	179	174	179	170	156	187	
179	183	174	187	167	159	170	179	

Calculate the value of the test statistic and state your conclusion clearly.

(c) Give the approximate p-value of this test.

8.1-4. Let X equal the thickness of spearmint gum manufactured for vending machines. Assume that the distribution of X is $N(\mu, \sigma^2)$. The target thickness is 7.5

hundredths of an inch. We shall test the null hypothesis H_0: $\mu = 7.5$ against a two-sided alternative hypothesis, using 10 observations.

(a) Define the test statistic and critical region for an $\alpha = 0.05$ significance level. Sketch a figure illustrating this critical region.

(b) Calculate the value of the test statistic and state your decision clearly, using the following $n = 10$ thicknesses in hundredths of an inch for pieces of gum that were selected randomly from the production line:

7.65	7.60	7.65	7.70	7.55
7.55	7.40	7.40	7.50	7.50

(c) Is $\mu = 7.50$ contained in a 95% confidence interval for μ?

8.1-5. The mean birth weight of infants in the United States is $\mu = 3315$ grams. Let X be the birth weight (in grams) of a randomly selected infant in Jerusalem. Assume that the distribution of X is $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. We shall test the null hypothesis H_0: $\mu = 3315$ against the alternative hypothesis H_1: $\mu < 3315$, using $n = 30$ randomly selected Jerusalem infants.

(a) Define a critical region that has a significance level of $\alpha = 0.05$.

(b) If the random sample of $n = 30$ yielded $\bar{x} = 3189$ and $s = 488$, what would be your conclusion?

(c) What is the approximate p-value of your test?

8.1-6. Let X equal the forced vital capacity (FVC) in liters for a female college student. (The FVC is the amount of air that a student can force out of her lungs.) Assume that the distribution of X is approximately $N(\mu, \sigma^2)$. Suppose it is known that $\mu = 3.4$ liters. A volleyball coach claims that the FVC of volleyball players is greater than 3.4. She plans to test her claim with a random sample of size $n = 9$.

(a) Define the null hypothesis.

(b) Define the alternative (coach's) hypothesis.

(c) Define the test statistic.

(d) Define a critical region for which $\alpha = 0.05$. Draw a figure illustrating your critical region.

(e) Calculate the value of the test statistic given that the random sample yielded the following FVCs:

3.4	3.6	3.8	3.3	3.4	3.5	3.7	3.6	3.7

(f) What is your conclusion?

(g) What is the approximate p-value of this test?

8.1-7. Vitamin B_6 is one of the vitamins in a multiple vitamin pill manufactured by a pharmaceutical company. The pills are produced with a mean of 50 mg of vitamin B_6 per pill. The company believes that there is a deterioration of 1 mg/month, so that after 3 months it expects that $\mu = 47$. A consumer group suspects that $\mu < 47$ after 3 months.

(a) Define a critical region to test H_0: $\mu = 47$ against H_1: $\mu < 47$ at an $\alpha = 0.05$ significance level based on a random sample of size $n = 20$.

(b) If the 20 pills yielded a mean of $\bar{x} = 46.94$ with a standard deviation of $s = 0.15$, what is your conclusion?

(c) What is the approximate p-value of this test?

8.1-8. A company that manufactures brackets for an automaker regularly selects brackets from the production line and performs a torque test. The goal is for mean torque to equal 125. Let X equal the torque and assume that X is $N(\mu, \sigma^2)$. We shall use a sample of size $n = 15$ to test H_0: $\mu = 125$ against a two-sided alternative hypothesis.

(a) Give the test statistic and a critical region with significance level $\alpha = 0.05$. Sketch a figure illustrating the critical region.

(b) Use the following observations to calculate the value of the test statistic and state your conclusion:

128	149	136	114	126	142	124	136
122	118	122	129	118	122	129	

8.1-9. The ornamental ground cover *Vinca minor* is spreading rapidly through the Hope College Biology Field Station because it can outcompete the small, native woody vegetation. In an attempt to discover whether *Vinca minor* utilized natural chemical weapons to inhibit the growth of the native vegetation, Hope biology students conducted an experiment in which they treated 33 sunflower seedlings with extracts taken from *Vinca minor* roots for several weeks and then measured the heights of the seedlings. Let X equal the height of one of these seedlings and assume that the distribution of X is $N(\mu, \sigma^2)$. The observed growths (in cm) were

11.5	11.8	15.7	16.1	14.1	10.5	15.2	19.0	12.8	12.4	19.2
13.5	16.5	13.5	14.4	16.7	10.9	13.0	15.1	17.1	13.3	12.4
8.5	14.3	12.9	11.1	15.0	13.3	15.8	13.5	9.3	12.2	10.3

The students also planted some control sunflower seedlings that had a mean height of 15.7 cm. We shall test the null hypothesis H_0: $\mu = 15.7$ against the alternative hypothesis H_1: $\mu < 15.7$.

(a) Calculate the value of the test statistic and give limits for the p-value of this test.

(b) What is your conclusion?

(c) Find an approximate 98% one-sided confidence interval that gives an upper bound for μ.

8.1-10. In a mechanical testing lab, Plexiglass® strips are stretched to failure. Let X equal the change in length in mm before breaking. Assume that the distribution of X is $N(\mu, \sigma^2)$. We shall test the null hypothesis H_0: $\mu = 5.70$ against the alternative hypothesis H_1: $\mu > 5.70$, using $n = 8$ observations of X.

(a) Define the test statistic and a critical region that has a significance level of $\alpha = 0.05$. Sketch a figure showing this critical region.

(b) A random sample of eight observations of X yielded the following data:

5.71 5.80 6.03 5.87 6.22 5.92 5.57 5.83

Calculate the value of the test statistic and state your conclusion clearly.

(c) Give the approximate value of or bounds for the p-value of this test.

8.1-11. A vendor of milk products produces and sells low-fat dry milk to a company that uses it to produce baby formula. In order to determine the fat content of the milk, both the company and the vendor take an observation from each lot and test it for fat content in percent. Ten sets of paired test results are as follows:

Lot Number	Company Test Results (x)	Vendor Test Results (y)
1	0.50	0.79
2	0.58	0.71
3	0.90	0.82
4	1.17	0.82
5	1.14	0.73
6	1.25	0.77
7	0.75	0.72
8	1.22	0.79
9	0.74	0.72
10	0.80	0.91

Let μ_D denote the mean of the difference $x - y$. Test H_0: $\mu_D = 0$ against H_1: $\mu_D > 0$, using a paired t test with the differences. Let $\alpha = 0.05$.

8.1-12. To test whether a golf ball of brand A can be hit a greater distance off the tee than a golf ball of brand B, each of 17 golfers hit a ball of each brand, 8 hitting ball A before ball B and 9 hitting ball B before ball A. The results in yards are as follows:

Golfer	Distance for Ball A	Distance for Ball B	Golfer	Distance for Ball A	Distance for Ball B
1	265	252	10	274	260
2	272	276	11	274	267
3	246	243	12	269	267
4	260	246	13	244	251
5	274	275	14	212	222
6	263	246	15	235	235
7	255	244	16	254	255
8	258	245	17	224	231
9	276	259			

Assume that the differences of the paired A distance and B distance are approximately normally distributed, and test the null hypothesis H_0: $\mu_D = 0$ against the alternative hypothesis H_1: $\mu_D > 0$, using a paired t test with the 17 differences. Let $\alpha = 0.05$.

8.1-13. A company that manufactures motors receives reels of 10,000 terminals per reel. Before using a reel of terminals, 20 terminals are randomly selected to be tested. The test is the amount of pressure needed to pull the terminal apart from its mate. This amount of pressure should continue to increase from test to test as the terminal is "roughed up." (Since this kind of testing is destructive testing, a terminal that is tested cannot be used in a motor.) Let W equal the difference of the pressures: "test No. 1 pressure" minus "test No. 2 pressure." Assume that the distribution of W is $N(\mu_W, \sigma_W^2)$. We shall test the null hypothesis H_0: $\mu_W = 0$ against the alternative hypothesis H_1: $\mu_W < 0$, using 20 pairs of observations.

(a) Give the test statistic and a critical region that has a significance level of $\alpha = 0.05$. Sketch a figure illustrating this critical region.

(b) Use the following data to calculate the value of the test statistic, and state your conclusion clearly:

Terminal	Test 1	Test 2	Terminal	Test 1	Test 2
1	2.5	3.8	11	7.3	8.2
2	4.0	3.9	12	7.2	6.6
3	5.2	4.7	13	5.9	6.8
4	4.9	6.0	14	7.5	6.6
5	5.2	5.7	15	7.1	7.5
6	6.0	5.7	16	7.2	7.5
7	5.2	5.0	17	6.1	7.3
8	6.6	6.2	18	6.3	7.1
9	6.7	7.3	19	6.5	7.2
10	6.6	6.5	20	6.5	6.7

(c) What would the conclusion be if $\alpha = 0.01$?

(d) What is the approximate p-value of this test?

8.1-14. A researcher claims that she can reduce the variance of $N(\mu, 100)$ by a new manufacturing process. If S^2 is the variance of a random sample of size n from this new distribution, she tests $H_0\colon \sigma^2 = 100$ against $H_1\colon \sigma^2 < 100$ by rejecting H_0 if $(n-1)S^2/100 \le \chi^2_{1-\alpha}(n-1)$ since $(n-1)S^2/100$ is $\chi^2(n-1)$ when H_0 is true.

(a) If $n = 23$, $s^2 = 32.52$, and $\alpha = 0.025$, would she reject H_0?

(b) Based on the same distributional result, what would be a reasonable test of $H_0\colon \sigma^2 = 100$ against a two-sided alternative hypothesis $H_1\colon \sigma^2 \ne 100$ when $\alpha = 0.05$?

8.1-15. Let $X_1, X_2, \ldots, X_{19}$ be a random sample of size $n = 19$ from the normal distribution $N(\mu, \sigma^2)$.

(a) Find a critical region, C, of size $\alpha = 0.05$ for testing $H_0\colon \sigma^2 = 30$ against $H_1\colon \sigma^2 = 80$.

(b) Find the approximate value of β, the probability of a Type II error, for the critical region C of part (a).

8.2 TESTS OF THE EQUALITY OF TWO MEANS

Let independent random variables X and Y have normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively. There are times when we are interested in testing whether the distributions of X and Y are the same. So if the assumption of normality is valid, we would be interested in testing whether the two means are equal. (A test for the equality of the two variances is given in the next section.)

When X and Y are independent and normally distributed, we can test hypotheses about their means with the same t statistic that we used to construct a confidence interval for $\mu_X - \mu_Y$ in Section 7.2. Recall that the t statistic used to construct the confidence interval assumed that the variances of X and Y were equal. (That is why we shall consider a test for the equality of two variances in the next section.)

We begin with an example and then give a table that lists some hypotheses and critical regions. A botanist is interested in comparing the growth response of dwarf pea stems against two different levels of the hormone indoleacetic acid (IAA). Using 16-day-old pea plants, the botanist obtains 5-mm sections and floats these sections on solutions with different hormone concentrations to observe the effect of the hormone on the growth of the pea stem. Let X and Y denote, respectively, the independent growths that can be attributed to the hormone during the first 26 hours after sectioning for $(0.5)(10)^{-4}$ and 10^{-4} levels of concentration of IAA. The botanist would like to test the null hypothesis $H_0\colon \mu_X - \mu_Y = 0$ against the alternative hypothesis $H_1\colon \mu_X - \mu_Y < 0$. If we can assume that X and Y are independent and normally distributed with a common variance, and if we assume respective random samples of sizes n and m, then we can find a test based on the statistic

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\{[(n-1)S_X^2 + (m-1)S_Y^2]/(n+m-2)\}(1/n + 1/m)}} \tag{8.2-1}$$

$$= \frac{\overline{X} - \overline{Y}}{S_P\sqrt{1/n + 1/m}},$$

where

$$S_P = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}\,. \tag{8.2-2}$$

Now, T has a t distribution with $r = n + m - 2$ degrees of freedom when H_0 is true and the variances are equal. Thus, the hypothesis H_0 will be rejected in favor of H_1 if the observed value of T is less than $-t_\alpha(n+m-2)$.

Example 8.2-1 In the preceding discussion, the botanist measured the growths of pea stem segments, in millimeters, for $n = 11$ observations of X:

0.8	1.8	1.0	0.1	0.9	1.7	1.0	1.4	0.9	1.2	0.5

She did the same with $m = 13$ observations of Y:

1.0	0.8	1.6	2.6	1.3	1.1	2.4
1.8	2.5	1.4	1.9	2.0	1.2	

For these data, $\bar{x} = 1.03$, $s_x^2 = 0.24$, $\bar{y} = 1.66$, and $s_y^2 = 0.35$. The critical region for testing H_0: $\mu_X - \mu_Y = 0$ against H_1: $\mu_X - \mu_Y < 0$ is $t \le -t_{0.05}(22) = -1.717$, where t is the two-sample t found in Equation 8.2-1. Since

$$t = \frac{1.03 - 1.66}{\sqrt{\{[10(0.24) + 12(0.35)]/(11 + 13 - 2)\}(1/11 + 1/13)}}$$

$$= -2.81 \ < \ -1.717,$$

H_0 is clearly rejected at an $\alpha = 0.05$ significance level. Notice that the approximate p-value of this test is 0.005, because $-t_{0.005}(22) = -2.819$. (See Figure 8.2-1.) Notice also that the sample variances do not differ too much; thus, most statisticians would use this two-sample t test.

It is instructive to construct box-and-whisker diagrams to gain a visual comparison of the two samples. For these two sets of data, the five-number summaries (minimum, three quartiles, maximum) are

0.1	0.8	1.0	1.4	1.8

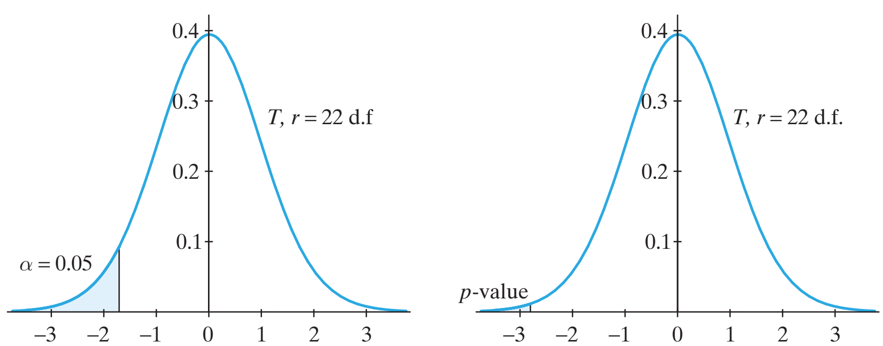

Figure 8.2-1 Critical region and p-value for pea stem growths

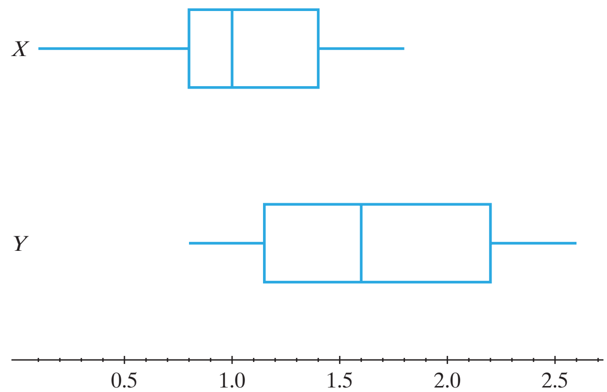

Figure 8.2-2 Box plots for pea stem growths

for the X sample and

$$0.8 \qquad 1.15 \qquad 1.6 \qquad 2.2 \qquad 2.6$$

for the Y sample. The two box plots are shown in Figure 8.2-2. ■

Assuming independent random samples of sizes n and m, let $\bar{x}$, $\bar{y}$, and s_p^2 represent the observed unbiased estimates of the respective parameters μ_X, μ_Y, and $\sigma_X^2 = \sigma_Y^2$ of two normal distributions with a common variance. Then α-level tests of certain hypotheses are given in Table 8.2-1 when $\sigma_X^2 = \sigma_Y^2$. If the common-variance assumption is violated, but not too badly, the test is satisfactory, but the significance levels are only approximate. The t statistic and s_p are given in Equations 8.2-1 and 8.2-2, respectively.

REMARK Again, to emphasize the relationship between confidence intervals and tests of hypotheses, we note that each of the tests in Table 8.2-1 has a corresponding confidence interval. For example, the first one-sided test is equivalent to saying that we reject H_0: $\mu_X - \mu_Y = 0$ if zero is not in the one-sided confidence interval with lower bound

$$\bar{x} - \bar{y} - t_\alpha(n+m-2)s_p\sqrt{1/n + 1/m}.$$ ■

Table 8.2-1 Tests of hypotheses for equality of two means

H_0	H_1	Critical Region
$\mu_X = \mu_Y$	$\mu_X > \mu_Y$	$t \ge t_\alpha(n+m-2)$ or $\bar{x} - \bar{y} \ge t_\alpha(n+m-2)s_p\sqrt{1/n + 1/m}$
$\mu_X = \mu_Y$	$\mu_X < \mu_Y$	$t \le -t_\alpha(n+m-2)$ or $\bar{x} - \bar{y} \le -t_\alpha(n+m-2)s_p\sqrt{1/n + 1/m}$
$\mu_X = \mu_Y$	$\mu_X \ne \mu_Y$	$\lvert t\rvert \ge t_{\alpha/2}(n+m-2)$ or $\lvert\bar{x} - \bar{y}\rvert \ge t_{\alpha/2}(n+m-2)s_p\sqrt{1/n + 1/m}$

Example 8.2-2 A product is packaged by a machine with 24 filler heads numbered 1 to 24, with the odd-numbered heads on one side of the machine and the even on the other side. Let X and Y equal the fill weights in grams when a package is filled by an odd-numbered head and an even-numbered head, respectively. Assume that the distributions of X and Y are $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively, and that X and Y are independent. We would like to test the null hypothesis H_0: $\mu_X - \mu_Y = 0$ against the alternative hypothesis H_1: $\mu_X - \mu_Y \neq 0$. To perform the test, after the machine has been set up and is running, we shall select one package at random from each filler head and weigh it. The test statistic is that given by Equation 8.2-1 with $n = m = 12$. At an $\alpha = 0.10$ significance level, the critical region is $|t| \geq t_{0.05}(22) = 1.717$.

For the $n = 12$ observations of X, namely,

1071	1076	1070	1083	1082	1067
1078	1080	1075	1084	1075	1080

$\bar{x} = 1076.75$ and $s_x^2 = 29.30$. For the $m = 12$ observations of Y, namely,

1074	1069	1075	1067	1068	1079
1082	1064	1070	1073	1072	1075

$\bar{y} = 1072.33$ and $s_y^2 = 26.24$. The calculated value of the test statistic is

$$t = \frac{1076.75 - 1072.33}{\sqrt{\dfrac{11(29.30) + 11(26.24)}{22}\left(\dfrac{1}{12} + \dfrac{1}{12}\right)}} = 2.05.$$

Since

$$|t| = |2.05| = 2.05 > 1.717,$$

the null hypothesis is rejected at an $\alpha = 0.10$ significance level. Note, however, that

$$|t| = 2.05 < 2.074 = t_{0.025}(22),$$

so that the null hypothesis would not be rejected at an $\alpha = 0.05$ significance level. That is, the p-value is between 0.05 and 0.10.

Again, it is instructive to construct box plots on the same graph for these two sets of data. The box plots in Figure 8.2-3 were constructed with the use of the five-number summary for the observations of X (1067, 1072, 1077, 1081.5, and 1084) and the five-number summary for the observations of Y (1064, 1068.25, 1072.5, 1075, and 1082). It looks like additional sampling would be advisable to test that the filler heads on the two sides of the machine are filling in a similar manner. If not, some corrective action needs to be taken. ■

We would like to give two modifications of tests about two means. First, if we are able to assume that we know the variances of X and Y, then the appropriate test statistic to use for testing H_0: $\mu_X = \mu_Y$ is

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}}, \tag{8.2-3}$$

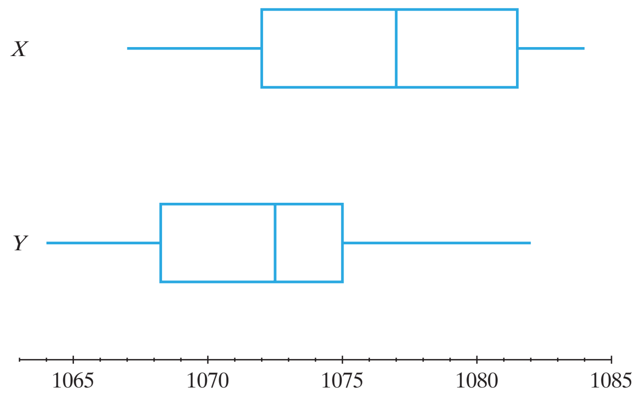

Figure 8.2-3 Box plots for fill weights

which has a standard normal distribution when the null hypothesis is true and, of course, when the populations are normally distributed. Second, if the variances are unknown and the sample sizes are large, replace σ_X^2 with S_X^2 and σ_Y^2 with S_Y^2 in Equation 8.2-3. The resulting statistic will have an approximate $N(0,1)$ distribution.

Example 8.2-3

The target thickness for Fruit Flavored Gum and for Fruit Flavored Bubble Gum is 6.7 hundredths of an inch. Let the independent random variables X and Y equal the respective thicknesses of these gums in hundredths of an inch, and assume that their distributions are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively. Because bubble gum has more elasticity than regular gum, it seems as if it would be harder to roll it out to the correct thickness. Thus, we shall test the null hypothesis H_0: $\mu_X = \mu_Y$ against the alternative hypothesis H_1: $\mu_X < \mu_Y$, using samples of sizes $n = 50$ and $m = 40$.

Because the variances are unknown and the sample sizes are large, the test statistic that is used is

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_X^2}{50} + \dfrac{S_Y^2}{40}}}.$$

At an approximate significance level of $\alpha = 0.01$, the critical region is

$$z \le -z_{0.01} = -2.326.$$

The observed values of X were

6.85	6.60	6.70	6.75	6.75	6.90	6.85	6.90	6.70	6.85
6.60	6.70	6.75	6.70	6.70	6.70	6.55	6.60	6.95	6.95
6.80	6.80	6.70	6.75	6.60	6.70	6.65	6.55	6.55	6.60
6.60	6.70	6.80	6.75	6.60	6.75	6.50	6.75	6.70	6.65
6.70	6.70	6.55	6.65	6.60	6.65	6.60	6.65	6.80	6.60

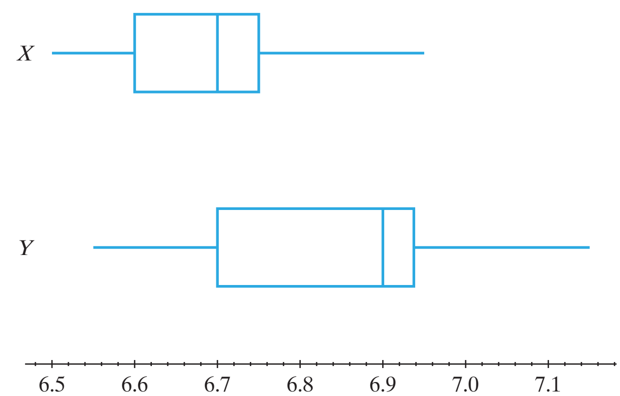

Figure 8.2-4 Box plots for gum thicknesses

for which $\bar{x} = 6.701$ and $s_x = 0.108$. The observed values of Y were

7.10	7.05	6.70	6.75	6.90	6.90	6.65	6.60	6.55	6.55
6.85	6.90	6.60	6.85	6.95	7.10	6.95	6.90	7.15	7.05
6.70	6.90	6.85	6.95	7.05	6.75	6.90	6.80	6.70	6.75
6.90	6.90	6.70	6.70	6.90	6.90	6.70	6.70	6.90	6.95

for which $\bar{y} = 6.841$ and $s_y = 0.155$. Since the calculated value of the test statistic is

$$z = \frac{6.701 - 6.841}{\sqrt{0.108^2/50 + 0.155^2/40}} = -4.848 < -2.326,$$

the null hypothesis is clearly rejected.

The box-and-whisker diagrams in Figure 8.2-4 were constructed with the use of the five-number summary of the observations of X (6.50, 6.60, 6.70, 6.75, and 6.95) and the five-number summary of the observations of Y (6.55, 6.70, 6.90, 6.94, and 7.15). This graphical display also confirms our conclusion. ■

REMARK To have satisfactory tests, our assumptions must be satisfied reasonably well. As long as the underlying distributions have finite means and variances and are not highly skewed, the normal assumptions are not too critical, as $\overline{X}$ and $\overline{Y}$ have approximate normal distributions by the central limit theorem. As distributions become nonnormal and highly skewed, the sample mean and sample variance become more dependent, and that causes problems in using the Student's t as an approximating distribution for T. In these cases, some of the nonparametric methods described later could be used. (See Section 8.4.)

When the distributions are close to normal, but the variances seem to differ by a great deal, the t statistic should again be avoided, particularly if the sample sizes are also different. In that case, use Z or the modification produced by substituting the sample variances for the distribution variances. In the latter situation, if n and m are large enough, there is no problem. With small n and m, most statisticians would use Welch's suggestion (or other modifications of it); that is, they would use an approximating Student's t distribution with r degrees of freedom, where r is given by Equation 7.2-1. We actually give a test for the equality of variances that could be employed to decide whether to use T or a modification of Z. However, most statisticians do not place much confidence in this test of $\sigma_X^2 = \sigma_Y^2$ and would use a modification of Z (possibly Welch's) if they suspected that the variances differed greatly. Alternatively, nonparametric methods described in Section 8.4 could be used. ■

Exercises

(In some of the exercises that follow, we must make assumptions such as the existence of normal distributions with equal variances.)

8.2-1. The botanist in Example 8.2-1 is really interested in testing for synergistic interaction. That is, given the two hormones gibberellin (GA_3) and indoleacetic acid (IAA), let X_1 and X_2 equal the growth responses (in mm) of dwarf pea stem segments to GA_3 and IAA, respectively and separately. Also, let $X = X_1 + X_2$ and let Y equal the growth response when both hormones are present. Assuming that X is $N(\mu_X, \sigma^2)$ and Y is $N(\mu_Y, \sigma^2)$, the botanist is interested in testing the hypothesis H_0: $\mu_X = \mu_Y$ against the alternative hypothesis of synergistic interaction H_1: $\mu_X < \mu_Y$.

(a) Using $n = m = 10$ observations of X and Y, define the test statistic and the critical region. Sketch a figure of the t pdf and show the critical region on your figure. Let $\alpha = 0.05$.

(b) Given $n = 10$ observations of X, namely,

2.1 2.6 2.6 3.4 2.1 1.7 2.6 2.6 2.2 1.2

and $m = 10$ observations of Y, namely,

3.5 3.9 3.0 2.3 2.1 3.1 3.6 1.8 2.9 3.3

calculate the value of the test statistic and state your conclusion. Locate the test statistic on your figure.

(c) Construct two box plots on the same figure. Does this confirm your conclusion?

8.2-2. Let X and Y denote the weights in grams of male and female common gallinules, respectively. Assume that X is $N(\mu_X, \sigma_X^2)$ and Y is $N(\mu_Y, \sigma_Y^2)$.

(a) Given $n = 16$ observations of X and $m = 13$ observations of Y, define a test statistic and a critical region for testing the null hypothesis H_0: $\mu_X = \mu_Y$ against the one-sided alternative hypothesis H_1: $\mu_X > \mu_Y$. Let $\alpha = 0.01$. (Assume that the variances are equal.)

(b) Given that $\bar{x} = 415.16$, $s_x^2 = 1356.75$, $\bar{y} = 347.40$, and $s_y^2 = 692.21$, calculate the value of the test statistic and state your conclusion.

(c) Although we assumed that $\sigma_X^2 = \sigma_Y^2$, let us say we suspect that that equality is not valid. Thus, use the test proposed by Welch.

8.2-3. Let X equal the weight in grams of a Low-Fat Strawberry Kudo and Y the weight of a Low-Fat Blueberry Kudo. Assume that the distributions of X and Y are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively. Let

21.7 21.0 21.2 20.7 20.4 21.9 20.2 21.6 20.6

be $n = 9$ observations of X, and let

21.5 20.5 20.3 21.6 21.7 21.3 23.0
21.3 18.9 20.0 20.4 20.8 20.3

be $m = 13$ observations of Y. Use these observations to answer the following questions:

(a) Test the null hypothesis H_0: $\mu_X = \mu_Y$ against a two-sided alternative hypothesis. You may select the significance level. Assume that the variances are equal.

(b) Construct and interpret box-and-whisker diagrams to support your conclusions.

8.2-4. Among the data collected for the World Health Organization air quality monitoring project is a measure of suspended particles, in μg/m^3. Let X and Y equal the concentration of suspended particles in μg/m^3 in the city centers (commercial districts), of Melbourne and Houston, respectively. Using $n = 13$ observations of X and $m = 16$ observations of Y, we shall test H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X < \mu_Y$.

(a) Define the test statistic and critical region, assuming that the variances are equal. Let $\alpha = 0.05$.

(b) If $\bar{x} = 72.9$, $s_x = 25.6$, $\bar{y} = 81.7$, and $s_y = 28.3$, calculate the value of the test statistic and state your conclusion.

(c) Give bounds for the p-value of this test.

8.2-5. Some nurses in county public health conducted a survey of women who had received inadequate prenatal care. They used information from birth certificates to select mothers for the survey. The mothers selected were divided into two groups: 14 mothers who said they had five or fewer prenatal visits and 14 mothers who said they had six or more prenatal visits. Let X and Y equal the respective birth weights of the babies from these two sets of mothers, and assume that the distribution of X is $N(\mu_X, \sigma^2)$ and the distribution of Y is $N(\mu_Y, \sigma^2)$.

(a) Define the test statistic and critical region for testing H_0: $\mu_X - \mu_Y = 0$ against H_1: $\mu_X - \mu_Y < 0$. Let $\alpha = 0.05$.

(b) Given that the observations of X were

49	108	110	82	93	114	134
114	96	52	101	114	120	116

and the observations of Y were

133	108	93	119	119	98	106
131	87	153	116	129	97	110

calculate the value of the test statistic and state your conclusion.

(c) Approximate the p-value.

(d) Construct box plots on the same figure for these two sets of data. Do the box plots support your conclusion?

8.2-6. Let X and Y equal the forces required to pull stud No. 3 and stud No. 4 out of a window that has been manufactured for an automobile. Assume that the distributions of X and Y are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively.

(a) If $m = n = 10$ observations are selected randomly, define a test statistic and a critical region for testing H_0: $\mu_X - \mu_Y = 0$ against a two-sided alternative hypothesis. Let $\alpha = 0.05$. Assume that the variances are equal.

(b) Given $n = 10$ observations of X, namely,

111 120 139 136 138 149 143 145 111 123

and $m = 10$ observations of Y, namely,

152 155 133 134 119 155 142 146 157 149

calculate the value of the test statistic and state your conclusion clearly.

(c) What is the approximate p-value of this test?

(d) Construct box plots on the same figure for these two sets of data. Do the box plots confirm your decision in part (b)?

8.2-7. Let X and Y equal the number of milligrams of tar in filtered and nonfiltered cigarettes, respectively. Assume that the distributions of X and Y are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively. We shall test the null hypothesis H_0: $\mu_X - \mu_Y = 0$ against the alternative hypothesis H_1: $\mu_X - \mu_Y < 0$, using random samples of sizes $n = 9$ and $m = 11$ observations of X and Y, respectively.

(a) Define the test statistic and a critical region that has an $\alpha = 0.01$ significance level. Sketch a figure illustrating this critical region.

(b) Given $n = 9$ observations of X, namely,

0.9 1.1 0.1 0.7 0.4 0.9 0.8 1.0 0.4

and $m = 11$ observations of Y, namely,

1.5 0.9 1.6 0.5 1.4 1.9 1.0 1.2 1.3 1.6 2.1

calculate the value of the test statistic and state your conclusion clearly. Locate the value of the test statistic on your figure.

8.2-8. Let X and Y denote the tarsus lengths of male and female grackles, respectively. Assume that X is $N(\mu_X, \sigma_X^2)$ and Y is $N(\mu_Y, \sigma_Y^2)$. Given that $n = 25$, $\bar{x} = 33.80$, $s_x^2 = 4.88$, $m = 29$, $\bar{y} = 31.66$, and $s_y^2 = 5.81$, test the null hypothesis H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X > \mu_Y$ with $\alpha = 0.01$.

8.2-9. When a stream is turbid, it is not completely clear due to suspended solids in the water. The higher the turbidity, the less clear is the water. A stream was studied on 26 days, half during dry weather (say, observations of X) and the other half immediately after a significant rainfall (say, observations of Y). Assume that the distributions of X and Y are $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively. The following turbidities were recorded in units of NTUs (nephelometric turbidity units):

x:	2.9	14.9	1.0	12.6	9.4	7.6	3.6
	3.1	2.7	4.8	3.4	7.1	7.2	
y:	7.8	4.2	2.4	12.9	17.3	10.4	5.9
	4.9	5.1	8.4	10.8	23.4	9.7	

(a) Test the null hypothesis H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X < \mu_Y$. Give bounds for the p-value and state your conclusion.

(b) Draw box-and-whisker diagrams on the same graph. Does this figure confirm your answer?

8.2-10. Plants convert carbon dioxide (CO_2) in the atmosphere, along with water and energy from sunlight, into the energy they need for growth and reproduction. Experiments were performed under normal atmospheric air conditions and in air with enriched CO_2 concentrations to determine the effect on plant growth. The plants were given the same amount of water and light for a four-week period. The following table gives the plant growths in grams:

Normal Air	4.67	4.21	2.18	3.91	4.09	5.24	2.94	4.71
	4.04	5.79	3.80	4.38				
Enriched Air	5.04	4.52	6.18	7.01	4.36	1.81	6.22	5.70

On the basis of these data, determine whether CO_2-enriched atmosphere increases plant growth.

8.2-11. Let X equal the fill weight in April and Y the fill weight in June for an 8-pound box of bleach. We shall test the null hypothesis H_0: $\mu_X - \mu_Y = 0$ against the alternative hypothesis H_1: $\mu_X - \mu_Y > 0$ given that $n = 90$ observations of X yielded $\overline{x} = 8.10$ and $s_x = 0.117$ and $m = 110$ observations of Y yielded $\overline{y} = 8.07$ and $s_y = 0.054$.

(a) What is your conclusion if $\alpha = 0.05$?
HINT: Do the variances seem to be equal?

(b) What is the approximate p-value of this test?

8.2-12. Let X and Y denote the respective lengths of male and female green lynx spiders. Assume that the distributions of X and Y are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively, and that $\sigma_Y^2 > \sigma_X^2$. Thus, use the modification of Z to test the hypothesis H_0: $\mu_X - \mu_Y = 0$ against the alternative hypothesis H_1: $\mu_X - \mu_Y < 0$.

(a) Define the test statistic and a critical region that has a significance level of $\alpha = 0.025$.

(b) Using the data given in Exercise 7.2-5, calculate the value of the test statistic and state your conclusion.

(c) Draw two box-and-whisker diagrams on the same figure. Does your figure confirm the conclusion of this exercise?

8.2-13. Students looked at the effect of a certain fertilizer on plant growth. The students tested this fertilizer on one group of plants (Group A) and did not give fertilizer to a second group (Group B). The growths of the plants, in mm, over six weeks were as follows:

Group A:	55	61	33	57	17	46	50	42	71	51	63
Group B:	31	27	12	44	9	25	34	53	33	21	32

(a) Test the null hypothesis that the mean growths are equal against the alternative that the fertilizer enhanced growth. Assume that the variances are equal.

(b) Construct box plots of the two sets of growths on the same graph. Does this confirm your answer to part (a)?

8.2-14. An ecology laboratory studied tree dispersion patterns for the sugar maple, whose seeds are dispersed by the wind, and the American beech, whose seeds are dispersed by mammals. In a plot of area 50 m by 50 m, they measured distances between like trees, yielding the following distances in meters for 19 American beech trees and 19 sugar maple trees:

American beech:	5.00	5.00	6.50	4.25	4.25	8.80	6.50
	7.15	6.15	2.70	2.70	11.40	9.70	
	6.10	9.35	2.85	4.50	4.50	6.50	
sugar maple:	6.00	4.00	6.00	6.45	5.00	5.00	5.50
	2.35	2.35	3.90	3.90	5.35	3.15	
	2.10	4.80	3.10	5.15	3.10	6.25	

(a) Test the null hypothesis that the means are equal against the one-sided alternative that the mean for the distances between beech trees is greater than that between maple trees.

(b) Construct two box plots to confirm your answer.

8.2-15. Say X and Y are independent random variables with distributions that are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. We wish to test H_0: $\sigma_X^2 = \sigma_Y^2$ against H_1: $\sigma_X^2 > \sigma_Y^2$.

(a) Argue that, if H_0 is true, the ratio of the two variances of the samples of sizes n and m, S_X^2/S_Y^2, has an $F(n-1, m-1)$ distribution.

(b) If $n = m = 31$, $\overline{x} = 8.153$, $s_x^2 = 1.410$, $\overline{y} = 5.917$, $s_y^2 = 0.4399$, $s_x^2/s_y^2 = 3.2053$, and $\alpha = 0.01$, show that H_0 is rejected and H_1 is accepted since $3.2053 > 2.39$.

(c) Where did the 2.39 come from?

8.2-16. To measure air pollution in a home, let X and Y equal the amount of suspended particulate matter (in μg/m^3) measured during a 24-hour period in a home in which there is no smoker and a home in which there is a smoker, respectively. We shall test the null hypothesis H_0: $\sigma_X^2/\sigma_Y^2 = 1$ against the one-sided alternative hypothesis H_1: $\sigma_X^2/\sigma_Y^2 > 1$.

(a) If a random sample of size $n = 9$ yielded $\overline{x} = 93$ and $s_x = 12.9$ while a random sample of size $m = 11$ yielded $\overline{y} = 132$ and $s_y = 7.1$, define a critical region and give your conclusion if $\alpha = 0.05$.

(b) Now test H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X < \mu_Y$ if $\alpha = 0.05$.

8.2-17. Consider the distributions $N(\mu_X, 400)$ and $N(\mu_Y, 225)$. Let $\theta = \mu_X - \mu_Y$. Say $\overline{x}$ and $\overline{y}$ denote the observed means of two independent random samples, each of size n, from the respective distributions. Say we reject H_0: $\theta = 0$ and accept H_1: $\theta > 0$ if $\overline{x} - \overline{y} \geq c$. Let $K(\theta)$ be the power function of the test. Find n and c so that $K(0) = 0.05$ and $K(10) = 0.90$, approximately.

8.3 TESTS ABOUT PROPORTIONS

Suppose a manufacturer of a certain printed circuit observes that approximately a proportion $p = 0.06$ of the circuits fail. An engineer and statistician working together suggest some changes that might improve the design of the product. To test this new procedure, it was agreed that $n = 200$ circuits would be produced by the proposed method and then checked. Let Y equal the number of these 200 circuits that fail. Clearly, if the number of failures, Y, is such that $Y/200$ is about equal to 0.06, then it seems that the new procedure has not resulted in an improvement. Also, on the one hand, if Y is small, so that $Y/200$ is about 0.02 or 0.03, we might believe that the new method is better than the old. On the other hand, if $Y/200$ is 0.09 or 0.10, the proposed method has perhaps caused a greater proportion of failures.

What we need to establish is a formal rule that tells us when to accept the new procedure as an improvement. In addition, we must know the consequences of this rule. As an example of such a rule, we could accept the new procedure as an improvement if $Y \le 7$ or $Y/n \le 0.035$. We do note, however, that the probability of failure could still be about $p = 0.06$ even with the new procedure, and yet we could observe 7 or fewer failures in $n = 200$ trials. That is, we could erroneously accept the new method as being an improvement when, in fact, it was not. This decision is a mistake we call a Type I error. By contrast, the new procedure might actually improve the product so that p is much smaller, say, $p = 0.03$, and yet we could observe $y = 9$ failures, so that $y/200 = 0.045$. Thus, we could, again erroneously, not accept the new method as resulting in an improvement when, in fact, it had. This decision is a mistake we call a Type II error. We must study the probabilities of these two types of errors to understand fully the consequences of our rule.

Let us begin by modeling the situation. If we believe that these trials, conducted under the new procedure, are independent, and that each trial has about the same probability of failure, then Y is binomial $b(200, p)$. We wish to make a statistical inference about p using the unbiased estimator $\widehat{p} = Y/200$. Of course, we could construct a one-sided confidence interval—say, one that has 95% confidence of providing an upper bound for p—and obtain

$$\left[0, \widehat{p} + 1.645\sqrt{\frac{\widehat{p}\,(1-\widehat{p}\,)}{200}}\,\right].$$

This inference is appropriate and many statisticians simply make it. If the limits of this confidence interval contain 0.06, they would not say that the new procedure is necessarily better, at least until more data are taken. If, however, the upper limit of the confidence interval is less than 0.06, then those same statisticians would feel 95% confident that the true p is now less than 0.06. Hence, they would support the conclusion that the new procedure has improved the manufacturing of the printed circuits in question.

While this use of confidence intervals is highly appropriate, and later we indicate the relationship of confidence intervals to tests of hypotheses, every student of statistics should also have some understanding of the basic concepts in the latter area. Here, in our illustration, we are testing whether the probability of failure has or has not decreased from 0.06 when the new manufacturing procedure is used. The null hypothesis is H_0: $p = 0.06$ and the alternative hypothesis is H_1: $p < 0.06$. Since, in our illustration, we make a Type I error if $Y \le 7$ when, in fact, $p = 0.06$, we can calculate the probability of this error. We denote that probability by α and call it the significance level of the test. Under our assumptions, it is

$$\alpha = P(Y \le 7;\, p = 0.06) = \sum_{y=0}^{7} \binom{200}{y} (0.06)^y (0.94)^{200-y}.$$

Since n is rather large and p is small, these binomial probabilities can be approximated very well by Poisson probabilities with $\lambda = 200(0.06) = 12$. That is, from the Poisson table, the probability of a Type I error is

$$\alpha \approx \sum_{y=0}^{7} \frac{12^y e^{-12}}{y!} = 0.090.$$

Thus, the approximate significance level of this test is $\alpha = 0.090$. (Using the binomial distribution, we find that the exact value of α is 0.0829, which you can easily verify with Minitab.)

This value of α is reasonably small. However, what about the probability of a Type II error in case p has been improved to, say, 0.03? This error occurs if $Y > 7$ when, in fact, $p = 0.03$; hence, its probability, denoted by β, is

$$\beta = P(Y > 7;\, p = 0.03) = \sum_{y=8}^{200} \binom{200}{y} (0.03)^y (0.97)^{200-y}.$$

Again, we use the Poisson approximation, here with $\lambda = 200(0.03) = 6$, to obtain

$$\beta \approx 1 - \sum_{y=0}^{7} \frac{6^y e^{-6}}{y!} = 1 - 0.744 = 0.256.$$

(The binomial distribution tells us that the exact probability is 0.2539, so the approximation is very good.) The engineer and the statistician who created the new procedure probably are not too pleased with this answer. That is, they might note that if their new procedure of manufacturing circuits has actually decreased the probability of failure to 0.03 from 0.06 (*a big improvement*), there is still a good chance, 0.256, that H_0: $p = 0.06$ is accepted and their improvement rejected. In Section 8.5, more will be said about modifying tests so that satisfactory values of the probabilities of the two types of errors, namely, α and β, can be obtained; however, to decrease both of them, we need larger sample sizes.

Without worrying more about the probability of the Type II error here, we present a frequently used procedure for testing H_0: $p = p_0$, where p_0 is some specified probability of success. This test is based upon the fact that the number of successes Y in n independent Bernoulli trials is such that Y/n has an approximate normal distribution $N[p_0, p_0(1-p_0)/n]$, provided that H_0: $p = p_0$ is true and n is large. Suppose the alternative hypothesis is H_1: $p > p_0$; that is, it has been hypothesized by a research worker that something has been done to increase the probability of success. Consider the test of H_0: $p = p_0$ against H_1: $p > p_0$ that rejects H_0 and accepts H_1 if and only if

$$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \ge z_\alpha.$$

That is, if Y/n exceeds p_0 by z_α standard deviations of Y/n, we reject H_0 and accept the hypothesis H_1: $p > p_0$. Since, under H_0, Z is approximately $N(0,1)$, the approximate probability of this occurring when H_0: $p = p_0$ is true is α. So the significance level of this test is approximately α.

If the alternative is H_1: $p < p_0$ instead of H_1: $p > p_0$, then the appropriate α-level test is given by $Z \le -z_\alpha$. Hence, if Y/n is smaller than p_0 by z_α standard deviations of Y/n, we accept H_1: $p < p_0$.

Example 8.3-1 It was claimed that many commercially manufactured dice are not fair because the "spots" are really indentations, so that, for example, the 6-side is lighter than the 1-side. Let p equal the probability of rolling a 6 with one of these dice. To test H_0: $p = 1/6$ against the alternative hypothesis H_1: $p > 1/6$, several such dice will be rolled to yield a total of $n = 8000$ observations. Let Y equal the number of times that 6 resulted in the 8000 trials. The test statistic is

$$Z = \frac{Y/n - 1/6}{\sqrt{(1/6)(5/6)/n}} = \frac{Y/8000 - 1/6}{\sqrt{(1/6)(5/6)/8000}}.$$

If we use a significance level of $\alpha = 0.05$, the critical region is

$$z \geq z_{0.05} = 1.645.$$

The results of the experiment yielded $y = 1389$, so the calculated value of the test statistic is

$$z = \frac{1389/8000 - 1/6}{\sqrt{(1/6)(5/6)/8000}} = 1.67.$$

Since

$$z = 1.67 > 1.645,$$

the null hypothesis is rejected, and the experimental results indicate that these dice favor a 6 more than a fair die would. (You could perform your own experiment to check out other dice.) ■

There are times when a two-sided alternative is appropriate; that is, here we test $H_0\colon p = p_0$ against $H_1\colon p \neq p_0$. For example, suppose that the pass rate in the usual beginning statistics course is p_0. There has been an intervention (say, some new teaching method) and it is not known whether the pass rate will increase, decrease, or stay about the same. Thus, we test the null (no-change) hypothesis $H_0\colon p = p_0$ against the two-sided alternative $H_1\colon p \neq p_0$. A test with the approximate significance level α for doing this is to reject $H_0\colon p = p_0$ if

$$|Z| = \frac{|Y/n - p_0|}{\sqrt{p_0(1-p_0)/n}} \geq z_{\alpha/2},$$

since, under H_0, $P(|Z| \geq z_{\alpha/2}) \approx \alpha$. These tests of approximate significance level α are summarized in Table 8.3-1. The rejection region for H_0 is often called the critical region of the test, and we use that terminology in the table.

The p-value associated with a test is the probability, under the null hypothesis H_0, that the test statistic (a random variable) is equal to or exceeds the observed value (a constant) of the test statistic in the direction of the alternative hypothesis.

Table 8.3-1 Tests of hypotheses for one proportion

H_0	H_1	Critical Region
$p = p_0$	$p > p_0$	$z = \dfrac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_\alpha$
$p = p_0$	$p < p_0$	$z = \dfrac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \leq -z_\alpha$
$p = p_0$	$p \neq p_0$	$\|z\| = \dfrac{\|y/n - p_0\|}{\sqrt{p_0(1-p_0)/n}} \geq z_{\alpha/2}$

Rather than select the critical region ahead of time, the p-value of a test can be reported and the reader then makes a decision. In Example 8.3-1, the value of the test statistic was $z = 1.67$. Because the alternative hypothesis was H_1: $p > 1/6$, the p-value is

$$P(Z \geq 1.67) = 0.0475.$$

Note that this p-value is less than $\alpha = 0.05$, which would lead to the rejection of H_0 at an $\alpha = 0.05$ significance level. If the alternative hypothesis were two sided, H_1: $p \neq 1/6$, then the p-value would be $P(|Z| \geq 1.67) = 0.095$ and would not lead to the rejection of H_0 at $\alpha = 0.05$.

Often there is interest in tests about p_1 and p_2, the probabilities of success for two different distributions or the proportions of two different populations having a certain characteristic. For example, if p_1 and p_2 denote the respective proportions of homeowners and renters who vote in favor of a proposal to reduce property taxes, a politician might be interested in testing H_0: $p_1 = p_2$ against the one-sided alternative hypothesis H_1: $p_1 > p_2$.

Let Y_1 and Y_2 represent, respectively, the numbers of observed successes in n_1 and n_2 independent trials with probabilities of success p_1 and p_2. Recall that the distribution of $\widehat{p}_1 = Y_1/n_1$ is approximately $N[p_1, p_1(1-p_1)/n_1]$ and the distribution of $\widehat{p}_2 = Y_2/n_2$ is approximately $N[p_2, p_2(1-p_2)/n_2]$. Thus, the distribution of $\widehat{p}_1 - \widehat{p}_2 = Y_1/n_1 - Y_2/n_2$ is approximately $N[p_1-p_2, p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2]$. It follows that the distribution of

$$Z = \frac{Y_1/n_1 - Y_2/n_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \tag{8.3-1}$$

is approximately $N(0,1)$. To test H_0: $p_1 - p_2 = 0$ or, equivalently, H_0: $p_1 = p_2$, let $p = p_1 = p_2$ be the common value under H_0. We shall estimate p with $\widehat{p} = (Y_1+Y_2)/(n_1+n_2)$. Replacing p_1 and p_2 in the denominator of Equation 8.3-1 with this estimate, we obtain the test statistic

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - 0}{\sqrt{\widehat{p}\,(1-\widehat{p}\,)(1/n_1 + 1/n_2)}},$$

which has an approximate $N(0,1)$ distribution for large sample sizes when the null hypothesis is true.

The three possible alternative hypotheses and their critical regions are summarized in Table 8.3-2.

Table 8.3-2 Tests of Hypotheses for two proportions

H_0	H_1	Critical Region
$p_1 = p_2$	$p_1 > p_2$	$z = \dfrac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\,(1-\widehat{p}\,)(1/n_1+1/n_2)}} \geq z_\alpha$
$p_1 = p_2$	$p_1 < p_2$	$z = \dfrac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\,(1-\widehat{p}\,)(1/n_1+1/n_2)}} \leq -z_\alpha$
$p_1 = p_2$	$p_1 \neq p_2$	$\lvert z\rvert = \dfrac{\lvert\widehat{p}_1 - \widehat{p}_2\rvert}{\sqrt{\widehat{p}\,(1-\widehat{p}\,)(1/n_1+1/n_2)}} \geq z_{\alpha/2}$

REMARK In testing both H_0: $p = p_0$ and H_0: $p_1 = p_2$, statisticians sometimes use different denominators for z. For tests of single proportions, $\sqrt{p_0(1-p_0)/n}$ can be replaced by $\sqrt{(y/n)(1-y/n)/n}$, and for tests of the equality of two proportions, the following denominator can be used:

$$\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

We do not have a strong preference one way or the other since the two methods provide about the same numerical result. The substitutions do provide better estimates of the standard deviations of the numerators when the null hypotheses are clearly false. There is some advantage to this result if the null hypothesis is likely to be false. In addition, the substitutions tie together the use of confidence intervals and tests of hypotheses. For example, if the null hypothesis is H_0: $p = p_0$, then the alternative hypothesis H_1: $p < p_0$ is accepted if

$$z = \frac{\widehat{p} - p_0}{\sqrt{\dfrac{\widehat{p}\,(1-\widehat{p}\,)}{n}}} \le -z_\alpha.$$

This formula is equivalent to the statement that

$$p_0 \notin \left[0, \widehat{p} + z_\alpha\sqrt{\frac{\widehat{p}\,(1-\widehat{p}\,)}{n}}\right),$$

where the latter is a one-sided confidence interval providing an upper bound for p. Or if the alternative hypothesis is H_1: $p \ne p_0$, then H_0 is rejected if

$$\frac{|\widehat{p} - p_0|}{\sqrt{\dfrac{\widehat{p}\,(1-\widehat{p}\,)}{n}}} \ge z_{\alpha/2}.$$

This inequality is equivalent to

$$p_0 \notin \left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p}\,)}{n}}, \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p}\,)}{n}}\right),$$

where the latter is a confidence interval for p. However, using the forms given in Tables 8.3-1 and 8.3-2, we do get better approximations to α-level significance tests. Thus, there are trade-offs, and it is difficult to say that one is better than the other. Fortunately, the numerical answers are about the same.

In the second situation in which the estimates of p_1 and p_2 are the observed $\widehat{p_1} = y_1/n_1$ and $\widehat{p_2} = y_2/n_2$, we have, with large values of n_1 and n_2, an approximate 95% confidence interval for $p_1 - p_2$ given by

$$\frac{y_1}{n_1} - \frac{y_2}{n_2} \pm 1.96\sqrt{\frac{(y_1/n_1)(1-y_1/n_1)}{n_1} + \frac{(y_2/n_2)(1-y_2/n_2)}{n_2}}.$$

If $p_1 - p_2 = 0$ is not in this interval, we reject H_0: $p_1 - p_2 = 0$ at the $\alpha = 0.05$ significance level. This is equivalent to saying that we reject H_0: $p_1 - p_2 = 0$ if

$$\frac{\left|\dfrac{y_1}{n_1} - \dfrac{y_2}{n_2}\right|}{\sqrt{\dfrac{(y_1/n_1)(1-y_1/n_1)}{n_1} + \dfrac{(y_2/n_2)(1-y_2/n_2)}{n_2}}} \geq 1.96.$$

In general, if the estimator $\widehat{\theta}$ (often, the maximum likelihood estimator) of θ has an approximate (sometimes exact) normal distribution $N(\theta, \sigma^2_{\widehat{\theta}})$, then H_0: $\theta = \theta_0$ is rejected in favor of H_1: $\theta \neq \theta_0$ at the approximate (sometimes exact) α significance level if

$$\theta_0 \notin (\,\widehat{\theta} - z_{\alpha/2}\,\sigma_{\widehat{\theta}}\,,\,\widehat{\theta} + z_{\alpha/2}\,\sigma_{\widehat{\theta}}\,)$$

or, equivalently,

$$\frac{|\widehat{\theta} - \theta_0|}{\sigma_{\widehat{\theta}}} \geq z_{\alpha/2}.$$

Note that $\sigma_{\widehat{\theta}}$ often depends upon some unknown parameter that must be estimated and substituted in $\sigma_{\widehat{\theta}}$ to obtain $\widehat{\sigma_{\widehat{\theta}}}$. Sometimes $\sigma_{\widehat{\theta}}$ or its estimate is called the **standard error** of $\widehat{\theta}$. This was the case in our last illustration when, with $\theta = p_1 - p_2$ and $\widehat{\theta} = \widehat{p_1} - \widehat{p_2}$, we substituted y_1/n_1 for p_1 and y_2/n_2 for p_2 in

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

to obtain the standard error of $\widehat{p_1} - \widehat{p_2} = \widehat{\theta}$. ■

Exercises

8.3-1. Let Y be $b(100, p)$. To test H_0: $p = 0.08$ against H_1: $p < 0.08$, we reject H_0 and accept H_1 if and only if $Y \leq 6$.

(a) Determine the significance level α of the test.

(b) Find the probability of the Type II error if, in fact, $p = 0.04$.

8.3-2. A bowl contains two red balls, two white balls, and a fifth ball that is either red or white. Let p denote the probability of drawing a red ball from the bowl. We shall test the simple null hypothesis H_0: $p = 3/5$ against the simple alternative hypothesis H_1: $p = 2/5$. Draw four balls at random from the bowl, one at a time and with replacement. Let X equal the number of red balls drawn.

(a) Define a critical region C for this test in terms of X.

(b) For the critical region C defined in part (a), find the values of α and β.

8.3-3. Let Y be $b(192, p)$. We reject H_0: $p = 0.75$ and accept H_1: $p > 0.75$ if and only if $Y \geq 152$. Use the normal approximation to determine

(a) $\alpha = P(Y \geq 152; p = 0.75)$.

(b) $\beta = P(Y < 152)$ when $p = 0.80$.

8.3-4. Let p denote the probability that, for a particular tennis player, the first serve is good. Since $p = 0.40$, this player decided to take lessons in order to increase p. When the lessons are completed, the hypothesis H_0: $p = 0.40$ will be tested against H_1: $p > 0.40$ on the basis of $n = 25$ trials. Let y equal the number of first serves that are good, and let the critical region be defined by $C = \{y : y \geq 13\}$.

(a) Determine $\alpha = P(Y \geq 13; p = 0.40)$. Use Table II in the appendix.

(b) Find $\beta = P(Y < 13)$ when $p = 0.60$; that is, $\beta = P(Y \leq 12; p = 0.60)$. Use Table II.

8.3-5. If a newborn baby has a birth weight that is less than 2500 grams (5.5 pounds), we say that the baby has a low birth weight. The proportion of babies with a low birth weight is an indicator of lack of nutrition for the mothers. For the United States, approximately 7% of babies have a low birth weight. Let p equal the proportion of babies born in the Sudan who weigh less than 2500 grams. We shall test the null hypothesis H_0: $p = 0.07$ against the alternative hypothesis H_1: $p > 0.07$. In a random sample of $n = 209$ babies, $y = 23$ weighed less than 2500 grams.

(a) What is your conclusion at a significance level of $\alpha = 0.05$?

(b) What is your conclusion at a significance level of $\alpha = 0.01$?

(c) Find the p-value for this test.

8.3-6. It was claimed that 75% of all dentists recommend a certain brand of gum for their gum-chewing patients. A consumer group doubted this claim and decided to test H_0: $p = 0.75$ against the alternative hypothesis H_1: $p < 0.75$, where p is the proportion of dentists who recommend that brand of gum. A survey of 390 dentists found that 273 recommended the given brand of gum.

(a) Which hypothesis would you accept if the significance level is $\alpha = 0.05$?

(b) Which hypothesis would you accept if the significance level is $\alpha = 0.01$?

(c) Find the p-value for this test.

8.3-7. The management of the Tigers baseball team decided to sell only low-alcohol beer in their ballpark to help combat rowdy fan conduct. They claimed that more than 40% of the fans would approve of this decision. Let p equal the proportion of Tiger fans on opening day who approved of the decision. We shall test the null hypothesis H_0: $p = 0.40$ against the alternative hypothesis H_1: $p > 0.40$.

(a) Define a critical region that has an $\alpha = 0.05$ significance level.

(b) If, out of a random sample of $n = 1278$ fans, $y = 550$ said that they approved of the new policy, what is your conclusion?

8.3-8. Let p equal the proportion of drivers who use a seat belt in a state that does not have a mandatory seat belt law. It was claimed that $p = 0.14$. An advertising campaign was conducted to increase this proportion. Two months after the campaign, $y = 104$ out of a random sample of $n = 590$ drivers were wearing their seat belts. Was the campaign successful?

(a) Define the null and alternative hypotheses.

(b) Define a critical region with an $\alpha = 0.01$ significance level.

(c) What is your conclusion?

8.3-9. According to a population census in 1986, the percentage of males who are 18 or 19 years old and are married was 3.7%. We shall test whether this percentage increased from 1986 to 1988.

(a) Define the null and alternative hypotheses.

(b) Define a critical region that has an approximate significance level of $\alpha = 0.01$. Sketch a standard normal pdf to illustrate this critical region.

(c) If $y = 20$ out of a random sample of $n = 300$ males, each 18 or 19 years old, were married (*U.S. Bureau of the Census, Statistical Abstract of the United States: 1988*), what is your conclusion? Show the calculated value of the test statistic on your figure in part (b).

8.3-10. Because of tourism in the state, it was proposed that public schools in Michigan begin after Labor Day. To determine whether support for this change was greater than 65%, a public poll was taken. Let p equal the proportion of Michigan adults who favor a post–Labor Day start. We shall test H_0: $p = 0.65$ against H_1: $p > 0.65$.

(a) Define a test statistic and an $\alpha = 0.025$ critical region.

(b) Given that 414 out of a sample of 600 favor a post–Labor Day start, calculate the value of the test statistic.

(c) Find the p-value and state your conclusion.

(d) Find a 95% one-sided confidence interval that gives a lower bound for p.

8.3-11. A machine shop that manufactures toggle levers has both a day and a night shift. A toggle lever is defective if a standard nut cannot be screwed onto the threads. Let p_1 and p_2 be the proportion of defective levers among those manufactured by the day and night shifts, respectively. We shall test the null hypothesis, H_0: $p_1 = p_2$, against a two-sided alternative hypothesis based on two random samples, each of 1000 levers taken from the production of the respective shifts.

(a) Define the test statistic and a critical region that has an $\alpha = 0.05$ significance level. Sketch a standard normal pdf illustrating this critical region.

(b) If $y_1 = 37$ and $y_2 = 53$ defectives were observed for the day and night shifts, respectively, calculate the value of the test statistic. Locate the calculated test statistic on your figure in part (a) and state your conclusion.

8.3-12. Let p equal the proportion of yellow candies in a package of mixed colors. It is claimed that $p = 0.20$.

(a) Define a test statistic and critical region with a significance level of $\alpha = 0.05$ for testing H_0: $p = 0.20$ against a two-sided alternative hypothesis.

(b) To perform the test, each of 20 students counted the number of yellow candies, y, and the total number of candies, n, in a 48.1-gram package, yielding the following ratios, y/n: 8/56, 13/55, 12/58, 13/56, 14/57, 5/54, 14/56, 15/57, 11/54, 13/55, 10/57, 8/59, 10/54, 11/55, 12/56, 11/57, 6/54, 7/58, 12/58, 14/58. If each individual tests H_0: $p = 0.20$, what proportion of the students rejected the null hypothesis?

(c) If we may assume that the null hypothesis is true, what proportion of the students would you have expected to reject the null hypothesis?

(d) For each of the 20 ratios in part (b), a 95% confidence interval for p can be calculated. What proportion of these 95% confidence intervals contain $p = 0.20$?

(e) If the 20 results are pooled so that $\sum_{i=1}^{20} y_i$ equals the number of yellow candies and $\sum_{i=1}^{20} n_i$ equals the total sample size, do we reject H_0: $p = 0.20$?

8.3-13. Let p_m and p_f be the respective proportions of male and female white-crowned sparrows that return to their hatching site. Give the endpoints for a 95% confidence interval for $p_m - p_f$ if 124 out of 894 males and 70 out of 700 females returned (*The Condor*, 1992, pp. 117–133). Does your result agree with the conclusion of a test of H_0: $p_1 = p_2$ against H_1: $p_1 \neq p_2$ with $\alpha = 0.05$?

8.3-14. For developing countries in Africa and the Americas, let p_1 and p_2 be the respective proportions of babies with a low birth weight (below 2500 grams). We shall test H_0: $p_1 = p_2$ against the alternative hypothesis H_1: $p_1 > p_2$.

(a) Define a critical region that has an $\alpha = 0.05$ significance level.

(b) If respective random samples of sizes $n_1 = 900$ and $n_2 = 700$ yielded $y_1 = 135$ and $y_2 = 77$ babies with a low birth weight, what is your conclusion?

(c) What would your decision be with a significance level of $\alpha = 0.01$?

(d) What is the p-value of your test?

8.3-15. Each of six students has a deck of cards and selects a card randomly from his or her deck.

(a) Show that the probability of at least one match is equal to 0.259.

(b) Now let each of the students randomly select an integer from 1–52, inclusive. Let p equal the probability of at least one match. Test the null hypothesis H_0: $p = 0.259$ against an appropriate alternative hypothesis. Give a reason for your alternative.

(c) Perform this experiment a large number of times. What is your conclusion?

8.3-16. Let p be the fraction of engineers who do not understand certain basic statistical concepts. Unfortunately, in the past, this number has been high, about $p = 0.73$. A new program to improve the knowledge of statistical methods has been implemented, and it is expected that under this program p would decrease from the aforesaid 0.73 value. To test H_0: $p = 0.73$ against H_1: $p < 0.73$, 300 engineers in the new program were tested and 204 (i.e., 68%) did not comprehend certain basic statistical concepts. Compute the p-value to determine whether this result indicates progress. That is, can we reject H_0 is favor of H_1? Use $\alpha = 0.05$.

8.4 THE WILCOXON TESTS

As mentioned earlier in the text, at times it is clear that the normality assumptions are not met and that other procedures, sometimes referred to as **nonparametric** or **distribution-free** methods, should be considered. For example, suppose some hypothesis, say, H_0: $m = m_0$, against H_1: $m \neq m_0$, is made about the unknown median, m, of a continuous-type distribution. From the data, we could construct a $100(1 - \alpha)\%$ confidence interval for m, and if m_0 is not in that interval, we would reject H_0 at the α significance level.

Now let X be a continuous-type random variable and let m denote the median of X. To test the hypothesis H_0: $m = m_0$ against an appropriate alternative hypothesis, we could also use a **sign test**. That is, if $X_1, X_2, \ldots, X_n$ denote the observations of a random sample from this distribution, and if we let Y equal the number of negative differences among $X_1 - m_0, X_2 - m_0, \ldots, X_n - m_0$, then Y has the binomial distribution $b(n, 1/2)$ under H_0 and is the test statistic for the sign test. If Y is too large or too small, we reject H_0: $m = m_0$.

Example 8.4-1

Let X denote the length of time in seconds between two calls entering a call center. Let m be the unique median of this continuous-type distribution. We test the null hypothesis H_0: $m = 6.2$ against the alternative hypothesis H_1: $m < 6.2$. Table II in Appendix B tells us that if Y is the number of lengths of time between calls in a random sample of size 20 that are less than 6.2, then the critical region $C = \{y : y \geq 14\}$ has a significance level of $\alpha = 0.0577$. A random sample of size 20 yielded the following data:

6.8	5.7	6.9	5.3	4.1	9.8	1.7	7.0
2.1	19.0	18.9	16.9	10.4	44.1	2.9	2.4
4.8	18.9	4.8	7.9				

Since $y = 9$, the null hypothesis is not rejected.

The sign test can also be used to test the hypothesis that two possibly dependent continuous-type random variables X and Y are such that $p = P(X > Y) = 1/2$. To test the hypothesis H_0: $p = 1/2$ against an appropriate alternative hypothesis, consider the independent pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Let W denote the number of pairs for which $X_k - Y_k > 0$. When H_0 is true, W is $b(n, 1/2)$, and the test can be based upon the statistic W. For example, say X is the length of the right foot of a person and Y the length of the corresponding left foot. Thus, there is a natural pairing, and here H_0: $p = P(X > Y) = 1/2$ suggests that either foot of a particular individual is equally likely to be longer.

One major objection to the sign test is that it does not take into account the magnitude of the differences $X_1 - m_0, \ldots, X_n - m_0$. We now discuss a **test of Wilcoxon** that does take into account the magnitude of the differences $|X_k - m_0|$, $k = 1, 2, \ldots, n$. However, in addition to assuming that the random variable X is of the continuous type, we must also assume that the pdf of X is symmetric about the median in order to find the distribution of this new statistic. Because of the continuity assumption, we assume, in the discussion which follows, that no two observations are equal and that no observation is equal to the median.

We are interested in testing the hypothesis H_0: $m = m_0$, where m_0 is some given constant. With our random sample $X_1, X_2, \ldots, X_n$, we rank the absolute values $|X_1 - m_0|, |X_2 - m_0|, \ldots, |X_n - m_0|$ in ascending order according to magnitude. That is, for $k = 1, 2, \ldots, n$, we let R_k denote the rank of $|X_k - m_0|$ among $|X_1 - m_0|, |X_2 - m_0|, \ldots, |X_n - m_0|$. Note that $R_1, R_2, \ldots, R_n$ is a permutation of the first n positive integers, $1, 2, \ldots, n$. Now, with each R_k, we associate the sign of the difference $X_k - m_0$; that is, if $X_k - m_0 > 0$, we use R_k, but if $X_k - m_0 < 0$, we use $-R_k$. The Wilcoxon statistic W is the sum of these n signed ranks, and therefore is often called the **Wilcoxon signed rank statistic**.

Example 8.4-2

Suppose the lengths of $n = 10$ sunfish are

$$x_i\colon\ 5.0\ \ 3.9\ \ 5.2\ \ 5.5\ \ 2.8\ \ 6.1\ \ 6.4\ \ 2.6\ \ 1.7\ \ 4.3$$

We shall test H_0: $m = 3.7$ against the alternative hypothesis H_1: $m > 3.7$. Thus, we have

$x_k - m_0$:	1.3,	0.2,	1.5,	1.8,	−0.9,	2.4,	2.7,	−1.1,	−2.0,	0.6
$\lvert x_k - m_0\rvert$:	1.3,	0.2,	1.5,	1.8,	0.9,	2.4,	2.7,	1.1,	2.0,	0.6
Ranks:	5,	1,	6,	7,	3,	9,	10,	4,	8,	2
Signed Ranks:	5,	1,	6,	7,	−3,	9,	10,	−4,	−8,	2

Therefore, the Wilcoxon statistic is equal to

$$W = 5 + 1 + 6 + 7 - 3 + 9 + 10 - 4 - 8 + 2 = 25.$$

Incidentally, the positive answer seems reasonable because the number of the 10 lengths that are less than 3.7 is 3, which is the statistic used in the sign test. ■

If the hypothesis H_0: $m = m_0$ is true, about one half of the differences would be negative and thus about one half of the signs would be negative. Hence, it seems that the hypothesis H_0: $m = m_0$ is supported if the observed value of W is close to zero. If the alternative hypothesis is H_1: $m > m_0$, we would reject H_0 if the observed $W = w$ is too large, since, in this case, the larger deviations $|X_k - m_0|$ would usually be associated with observations for which $x_k - m_0 > 0$. That is, the critical region would be of the form $\{w\colon\ w \ge c_1\}$. If the alternative hypothesis is H_1: $m < m_0$, the critical region would be of the form $\{w\colon\ w \le c_2\}$. Of course, the critical region would be of the form $\{w\colon\ w \le c_3 \text{ or } w \ge c_4\}$ for a two-sided alternative hypothesis H_1: $m \ne m_0$. In order to find the values of c_1, c_2, c_3, and c_4 that yield desired significance levels, it is necessary to determine the distribution of W under H_0. Accordingly, we consider certain characteristics of this distribution.

When H_0: $m = m_0$ is true,

$$P(X_k < m_0) = P(X_k > m_0) = \frac{1}{2}, \qquad k = 1, 2, \ldots, n.$$

Hence, the probability is $1/2$ that a negative sign is associated with the rank R_k of $|X_k - m_0|$. Moreover, the assignments of these n signs are independent because $X_1, X_2, \ldots, X_n$ are mutually independent. In addition, W is a sum that contains the integers $1, 2, \ldots, n$, each with a positive or negative sign. Since the underlying distribution is symmetric, it seems intuitively obvious that W has the same distribution as the random variable

$$V = \sum_{k=1}^{n} V_k,$$

where $V_1, V_2, \ldots, V_n$ are independent and

$$P(V_k = k) = P(V_k = -k) = \frac{1}{2}, \qquad k = 1, 2, \ldots, n.$$

That is, V is a sum that contains the integers $1, 2, \ldots, n$, and these integers receive their algebraic signs by independent assignments.

Since W and V have the same distribution, their means and variances are equal, and we can easily find those of V. Now, the mean of V_k is

$$E(V_k) = -k\left(\frac{1}{2}\right) + k\left(\frac{1}{2}\right) = 0;$$

thus,

$$E(W) = E(V) = \sum_{k=1}^{n} E(V_k) = 0.$$

The variance of V_k is

$$\text{Var}(V_k) = E(V_k^2) = (-k)^2\left(\frac{1}{2}\right) + (k)^2\left(\frac{1}{2}\right) = k^2.$$

Hence,

$$\operatorname{Var}(W) = \operatorname{Var}(V) = \sum_{k=1}^{n} \operatorname{Var}(V_k) = \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

We shall not try to find the distribution of W in general, since that pmf does not have a convenient expression. However, we demonstrate how we could find the distribution of W (or V) with enough patience and computer support. Recall that the moment-generating function of V_i is

$$M_k(t) = e^{t(-k)}\left(\frac{1}{2}\right) + e^{t(+k)}\left(\frac{1}{2}\right) = \frac{e^{-kt} + e^{kt}}{2}, \qquad k = 1, 2, \ldots, n.$$

Let $n = 2$; then the moment-generating function of $V_1 + V_2$ is

$$M(t) = E[e^{t(V_1+V_2)}].$$

From the independence of V_1 and V_2, we obtain

$$\begin{aligned} M(t) &= E(e^{tV_1})E(e^{tV_2}) \\ &= \left(\frac{e^{-t} + e^{t}}{2}\right)\left(\frac{e^{-2t} + e^{2t}}{2}\right) \\ &= \frac{e^{-3t} + e^{-t} + e^{t} + e^{3t}}{4}. \end{aligned}$$

This means that each of the points $-3, -1, 1, 3$ in the support of $V_1 + V_2$ has probability $1/4$.

Next let $n = 3$; then the moment-generating function of $V_1 + V_2 + V_3$ is

$$\begin{aligned} M(t) &= E[e^{t(V_1+V_2+V_3)}] \\ &= E[e^{t(V_1+V_2)}]E(e^{tV_3}) \\ &= \left(\frac{e^{-3t} + e^{-t} + e^{t} + e^{3t}}{4}\right)\left(\frac{e^{-3t} + e^{3t}}{2}\right) \\ &= \frac{e^{-6t} + e^{-4t} + e^{-2t} + 2e^{0} + e^{2t} + e^{4t} + e^{6t}}{8}. \end{aligned}$$

Thus, the points $-6, -4, -2, 0, 2, 4$, and 6 in the support of $V_1 + V_2 + V_3$ have the respective probabilities $1/8, 1/8, 1/8, 2/8, 1/8, 1/8$, and $1/8$. Obviously, this procedure can be continued for $n = 4, 5, 6, \ldots$, but it is rather tedious. Fortunately, however, even though $V_1, V_2, \ldots, V_n$ are not identically distributed random variables, the sum V of them still has an approximate normal distribution for large samples. To obtain this normal approximation for V (or W), a more general form of the central limit theorem, due to Liapounov, can be used which allows us to say that the standardized random variable

$$Z = \frac{W - 0}{\sqrt{n(n+1)(2n+1)/6}}$$

is approximately $N(0, 1)$ when H_0 is true. We accept this theorem without proof, so that we can use this normal distribution to approximate probabilities such as

$P(W \geq c; H_0) \approx P(Z \geq z_\alpha; H_0)$ when the sample size n is sufficiently large. The next example illustrates this approximation.

Example 8.4-3 The moment-generating function of W or of V is given by

$$M(t) = \prod_{i=1}^{n} \frac{e^{-kt} + e^{kt}}{2}.$$

Using a computer algebra system such as *Maple*, we can expand $M(t)$ and find the coefficients of e^{kt}, which is equal to $P(W = k)$. In Figure 8.4-1, we have drawn a probability histogram for the distribution of W along with the approximating $N[0, n(n+1)(2n+1)/6]$ pdf for $n = 4$ (a poor approximation) and for $n = 10$. It is important to note that the widths of the rectangles in the probability histogram are equal to 2, so the "half-unit correction for continuity" mentioned in Section 5.7 now is equal to 1. ■

Example 8.4-4 Let m be the median of a symmetric distribution of the continuous type. To test the hypothesis H_0: $m = 160$ against the alternative hypothesis H_1: $m > 160$, we take a random sample of size $n = 16$. For an approximate significance level of $\alpha = 0.05$, H_0 is rejected if the computed $W = w$ is such that

$$z = \frac{w}{\sqrt{16(17)(33)/6}} \geq 1.645,$$

or

$$w \geq 1.645\sqrt{\frac{16(17)(33)}{6}} = 63.626.$$

Say the observed values of a random sample are 176.9, 158.3, 152.1, 158.8, 172.4, 169.8, 159.7, 162.7, 156.6, 174.5, 184.4, 165.2, 147.8, 177.8, 160.1, and 160.5. In Table 8.4-1, the magnitudes of the differences $|x_k - 160|$ have been ordered and ranked. Those differences $x_k - 160$ which were negative have been underlined, and the ranks are under the ordered values. For this set of data,

$$w = 1 - 2 + 3 - 4 - 5 + 6 + \cdots + 16 = 60.$$

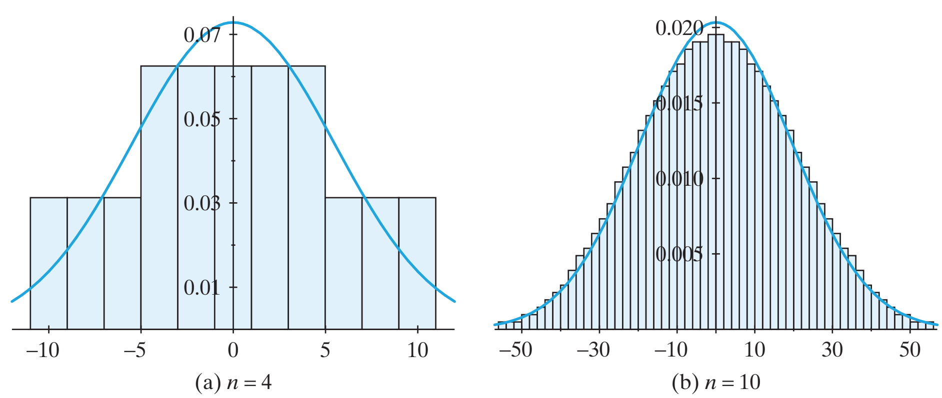

Figure 8.4-1 The Wilcoxon distribution

Table 8.4-1 Ordered absolute differences from 160

0.1	<u>0.3</u>	0.5	<u>1.2</u>	<u>1.7</u>	2.7	<u>3.4</u>	5.2
1	2	3	4	5	6	7	8
<u>7.9</u>	9.8	<u>12.2</u>	12.4	14.5	16.9	17.8	24.4
9	10	11	12	13	14	15	16

Since $60 < 63.626$, H_0 is not rejected at the 0.05 significance level. It is interesting to note that H_0 would have been rejected at $\alpha = 0.10$, since, with a unit correction made for continuity, the approximate p-value is

$$\begin{aligned} p\text{-value} &= P(W \ge 60) \\ &= P\left(\frac{W-0}{\sqrt{(16)(17)(33)/6}} \ge \frac{59-0}{\sqrt{(16)(17)(33)/6}}\right) \\ &\approx P(Z \ge 1.525) \;=\; 0.0636. \end{aligned}$$

(*Maple* produces a p-value equal to $4{,}251/65{,}536 = 0.0649$.) Such a p-value would indicate that the data are too few to reject H_0, but if the pattern continues, we shall most certainly reject H_0 with a larger sample size. ■

Although theoretically we could ignore the possibilities that $x_k = m_0$ for some k and that $|x_k - m_0| = |x_j - m_0|$ for some $k \ne j$, these situations do occur in applications. Usually, in practice, if $x_k = m_0$ for some k, that observation is deleted and the test is performed with a reduced sample size. If the absolute values of the differences from m_0 of two or more observations are equal, each observation is assigned the average of the corresponding ranks. The change this causes in the distribution of W is not very great, provided that the number of ties is relatively small; thus, we continue using the same normal approximation.

We now give an example that has some tied observations.

Example 8.4-5 We consider some paired data for percentage of body fat measured at the beginning and the end of a semester. Let m equal the median of the differences, $x - y$. We shall use the Wilcoxon statistic to test the null hypothesis H_0: $m = 0$ against the alternative hypothesis H_1: $m > 0$ with the differences given below. Since there are $n = 25$ nonzero differences, we reject H_0 if

$$z = \frac{w - 0}{\sqrt{(25)(26)(51)/6}} \ge 1.645$$

or, equivalently, if

$$w \ge 1.645\sqrt{\frac{(25)(26)(51)}{6}} = 122.27$$

at an approximate $\alpha = 0.05$ significance level. The 26 differences are

1.8	−3.1	0.1	1.1	0.6	−5.1	9.2	0.2	0.4
0.0	1.9	−0.4	−1.5	1.4	−1.0	2.2	0.8	−0.4
2.0	−5.8	−3.4	−2.3	3.0	2.7	0.2	3.2	

Table 8.4-2 Ordered absolute values, changes in percentage of body fat

0.1	0.2	0.2	0.4	$\underline{0.4}$	$\underline{0.4}$	0.6	0.8	$\underline{1.0}$	1.1	1.4	$\underline{1.5}$	1.8
1	2.5	2.5	5	5	5	7	8	9	10	11	12	13
1.9	2.0	2.2	$\underline{2.3}$	2.7	3.0	$\underline{3.1}$	3.2	$\underline{3.4}$	$\underline{5.1}$	$\underline{5.8}$	9.2	
14	15	16	17	18	19	20	21	22	23	24	25	

Table 8.4-2 lists the ordered nonzero absolute values, with those that were originally negative underlined. The rank is under each observation. Note that in the case of ties, the average of the ranks of the tied measurements is given.

The value of the Wilcoxon statistic is

$$w = 1 + 2.5 + 2.5 + 5 - 5 - 5 + \cdots + 25 = 51.$$

Since $51 < 122.27$, we fail to reject the null hypothesis. The approximate p-value of this test, using the continuity correction, is

$$\begin{aligned} p\text{-value} &= P(W \geq 51) \\ &\approx P\left(Z \geq \frac{50 - 0}{\sqrt{(25)(26)(51)/6}}\right) = P(Z \geq 0.673) = 0.2505. \end{aligned}$$

Another method due to Wilcoxon for testing the equality of two distributions of the continuous type uses the magnitudes of the observations. For this test, it is assumed that the respective cdfs F and G have the same shape and spread but possibly different locations; that is, there exists a constant c such that $F(x) = G(x + c)$ for all x. To proceed with the test, place the combined sample of $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ in increasing order of magnitude. Assign the ranks $1, 2, 3, \ldots, n_1 + n_2$ to the ordered values. In the case of ties, assign the average of the ranks associated with the tied values. Let w equal the sum of the ranks of $y_1, y_2, \ldots, y_{n_2}$. If the distribution of Y is shifted to the right of that of X, the values of Y would tend to be larger than the values of X and w would usually be larger than expected when $F(z) = G(z)$. If m_X and m_Y are the respective medians, the critical region for testing H_0: $m_X = m_Y$ against H_1: $m_X < m_Y$ would be of the form $w \geq c$. Similarly, if the alternative hypothesis is $m_X > m_Y$, the critical region would be of the form $w \leq c$.

We shall not derive the distribution of W. However, if n_1 and n_2 are both greater than 7, and there are no ties, a normal approximation can be used. With $F(z) = G(z)$, the mean and variance of W are

$$\mu_W = \frac{n_2(n_1 + n_2 + 1)}{2}$$

and

$$\text{Var}(W) = \frac{n_1 n_2(n_1 + n_2 + 1)}{12},$$

and the statistic

$$Z = \frac{W - n_2(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2(n_1 + n_2 + 1)/12}}$$

is approximately $N(0, 1)$.

Example 8.4-6 The weights of of the contents of $n_1 = 8$ and $n_2 = 8$ tins of cinnamon packaged by companies A and B, respectively, selected at random, yielded the following observations of X and Y:

x:	117.1	121.3	127.8	121.9	117.4	124.5	119.5	115.1
y:	123.5	125.3	126.5	127.9	122.1	125.6	129.8	117.2

The critical region for testing H_0: $m_X = m_Y$ against H_1: $m_X < m_Y$ is of the form $w \geq c$. Since $n_1 = n_2 = 8$, at an approximate $\alpha = 0.05$ significance level H_0 is rejected if

$$z = \frac{w - 8(8+8+1)/2}{\sqrt{[(8)(8)(8+8+1)]/12}} \geq 1.645,$$

or

$$w \geq 1.645\sqrt{\frac{(8)(8)(17)}{12}} + 4(17) = 83.66.$$

To calculate the value of W, it is sometimes helpful to construct a **back-to-back stem-and-leaf display**. In such a display, the stems are put in the center and the leaves go to the left and the right. (See Table 8.4-3.)

Reading from this two-sided stem-and-leaf display, we show the combined sample in Table 8.4-4, with the Company B (y) weights underlined. The ranks are given beneath the values.

From Table 8.4-4, the computed W is

$$w = 3 + 8 + 9 + 11 + 12 + 13 + 15 + 16 = 87 > 83.66.$$

Table 8.4-3 Back-to-back stem-and-leaf diagram of weights of cinnamon

x Leaves		Stems	y Leaves	
	51	$11f$		
74	71	$11s$	72	
	95	$11\bullet$		
19	13	$12*$		
		$12t$	21	35
	45	$12f$	53	56
	78	$12s$	65	79
		$12\bullet$	98	

Multiply numbers by 10^{-1}.

Table 8.4-4 Combined ordered samples

115.1	117.1	117.2	117.4	119.5	121.3	121.9	122.1
1	2	3	4	5	6	7	8
123.5	124.5	125.3	125.6	126.5	127.8	127.9	129.8
9	10	11	12	13	14	15	16

Thus, H_0 is rejected. Finally, making a half-unit correction for continuity, we see that the p-value of this test is

$$\begin{aligned} p\text{-value} &= P(W \geq 87) \\ &= P\left(\frac{W-68}{\sqrt{90.667}} \geq \frac{86.5-68}{\sqrt{90.667}}\right) \\ &\approx P(Z \geq 1.943) \;=\; 0.0260. \end{aligned}$$

Exercises

8.4-1. It is claimed that the median weight m of certain loads of candy is 40,000 pounds.

(a) Use the following 13 observations and the Wilcoxon statistic to test the null hypothesis H_0: $m = 40{,}000$ against the one-sided alternative hypothesis H_1: $m < 40{,}000$ at an approximate significance level of $\alpha = 0.05$:

41,195 39,485 41,229 36,840 38,050 40,890 38,345

34,930 39,245 31,031 40,780 38,050 30,906

(b) What is the approximate p-value of this test?

(c) Use the sign test to test the same hypothesis.

(d) Calculate the p-value from the sign test and compare it with the p-value obtained from the Wilcoxon test.

8.4-2. A course in economics was taught to two groups of students, one in a classroom situation and the other online. There were 24 students in each group. The students were first paired according to cumulative grade point averages and background in economics, and then assigned to the courses by a flip of a coin. (The procedure was repeated 24 times.) At the end of the course each class was given the same final examination. Use the Wilcoxon test to test the hypothesis that the two methods of teaching are equally effective against a two-sided alternative. The differences in the final scores for each pair of students were as follows (the online student's score was subtracted from the corresponding classroom student's score):

14	−4	−6	−2	−1	18
6	12	8	−4	13	7
2	6	21	7	−2	11
−3	−14	−2	17	−4	−5

8.4-3. Let X equal the weight (in grams) of a Hershey's grape-flavored Jolly Rancher. Denote the median of X by m. We shall test H_0: $m = 5.900$ against H_1: $m > 5.900$. A random sample of size $n = 25$ yielded the following ordered data:

5.625 5.665 5.697 5.837 5.863 5.870 5.878 5.884 5.908

5.967 6.019 6.020 6.029 6.032 6.037 6.045 6.049

6.050 6.079 6.116 6.159 6.186 6.199 6.307 6.387

(a) Use the sign test to test the hypothesis.

(b) Use the Wilcoxon test statistic to test the hypothesis.

(c) Use a t test to test the hypothesis.

(d) Write a short comparison of the three tests.

8.4-4. The outcomes on $n = 10$ simulations of a Cauchy random variable were −1.9415, 0.5901, −5.9848, −0.0790, −0.7757, −1.0962, 9.3820, −74.0216, −3.0678, and 3.8545. For the Cauchy distribution, the mean does not exist, but for this one, the median is believed to equal zero. Use the Wilcoxon test and these data to test H_0: $m = 0$ against the alternative hypothesis H_1: $m \neq 0$. Let $\alpha \approx 0.05$.

8.4-5. Let x equal a student's GPA in the fall semester and y the same student's GPA in the spring semester. Let m equal the median of the differences, $x - y$. We shall test the null hypothesis H_0: $m = 0$ against an appropriate alternative hypothesis that you select on the basis of your past experience. Use a Wilcoxon test and the following 15 observations of paired data to test H_0:

x	y	x	y
2.88	3.22	3.98	3.76
3.67	3.49	4.00	3.96
2.76	2.54	3.39	3.52
2.34	2.17	2.59	2.36
2.46	2.53	2.78	2.62
3.20	2.98	2.85	3.06
3.17	2.98	3.25	3.16
2.90	2.84		

8.4-6. Let m equal the median of the posttest grip strengths in the right arms of male freshmen in a study of health dynamics. We shall use observations on $n = 15$ such students to test the null hypothesis H_0: $m = 50$ against the alternative hypothesis H_1: $m > 50$.

(a) Using the Wilcoxon statistic, define a critical region that has an approximate significance level of $\alpha = 0.05$.

(b) Given the observed values

58.0	52.5	46.0	57.5	52.0	45.5	65.5	71.0
57.0	54.0	48.0	58.0	35.5	44.0	53.0	

what is your conclusion?

(c) What is the p-value of this test?

8.4-7. Let X equal the weight in pounds of a "1-pound" bag of carrots. Let m equal the median weight of a population of these bags. Test the null hypothesis H_0: $m = 1.14$ against the alternative hypothesis H_1: $m > 1.14$.

(a) With a sample of size $n = 14$, use the Wilcoxon statistic to define a critical region. Use $\alpha \approx 0.10$.

(b) What would be your conclusion if the observed weights were

1.12	1.13	1.19	1.25	1.06	1.31	1.12
1.23	1.29	1.17	1.20	1.11	1.18	1.23

(c) What is the p-value of your test?

8.4-8. A pharmaceutical company is interested in testing the effect of humidity on the weight of pills that are sold in aluminum packaging. Let X and Y denote the respective weights of pills and their packaging (in grams), when the packaging is good and when it is defective, after the pill has spent 1 week in a chamber containing 100% humidity and heated to 30 °C.

(a) Use the Wilcoxon test to test H_0: $m_X = m_Y$ against H_0: $m_X - m_Y < 0$ on the following random samples of $n_1 = 12$ observations of X and $n_2 = 12$ observations of Y:

x:	0.7565	0.7720	0.7776	0.7750	0.7494	0.7615
	0.7741	0.7701	0.7712	0.7719	0.7546	0.7719
y:	0.7870	0.7750	0.7720	0.7876	0.7795	0.7972
	0.7815	0.7811	0.7731	0.7613	0.7816	0.7851

What is the p-value?

(b) Construct and interpret a q–q plot of these data. HINT: This is a q–q plot of the empirical distribution of X against that of Y.

8.4-9. Let us compare the failure times of a certain type of light bulb produced by two different manufacturers, X and Y, by testing 10 bulbs selected at random from each of the outputs. The data, in hundreds of hours used before failure, are

x:	5.6	4.6	6.8	4.9	6.1	5.3	4.5	5.8	5.4	4.7
y:	7.2	8.1	5.1	7.3	6.9	7.8	5.9	6.7	6.5	7.1

(a) Use the Wilcoxon test to test the equality of medians of the failure times at the approximate 5% significance level. What is the p-value?

(b) Construct and interpret a q–q plot of these data. HINT: This is a q–q plot of the empirical distribution of X against that of Y.

8.4-10. Let X and Y denote the heights of blue spruce trees, measured in centimeters, growing in two large fields. We shall compare these heights by measuring 12 trees selected at random from each of the fields. Take $\alpha \approx 0.05$, and use the statistic W—the sum of the ranks of the observations of Y in the combined sample—to test the hypothesis H_0: $m_X = m_Y$ against the alternative hypothesis H_1: $m_X < m_Y$ on the basis of the following $n_1 = 12$ observations of X and $n_2 = 12$ observations of Y.

x:	90.4	77.2	75.9	83.2	84.0	90.2
	87.6	67.4	77.6	69.3	83.3	72.7
y:	92.7	78.9	82.5	88.6	95.0	94.4
	73.1	88.3	90.4	86.5	84.7	87.5

8.4-11. Let X and Y equal the sizes of grocery orders from, respectively, a south-side and a north-side food store of the same chain. We shall test the null hypothesis

H_0: $m_X = m_Y$ against a two-sided alternative, using the following ordered observations:

x:	5.13	8.22	11.81	13.77	15.36
	23.71	31.39	34.65	40.17	75.58
y:	4.42	6.47	7.12	10.50	12.12
	12.57	21.29	33.14	62.84	72.05

(a) Use the Wilcoxon test when $\alpha = 0.05$. What is the p-value of this two-sided test?

(b) Construct a q–q plot and interpret it. HINT: This is a q–q plot of the empirical distribution of X against that of Y.

8.4-12. A charter bus line has 48-passenger and 38-passenger buses. Let m_{48} and m_{38} denote the median number of miles traveled per day by the respective buses. With $\alpha = 0.05$, use the Wilcoxon statistic to test H_0: $m_{48} = m_{38}$ against the one-sided alternative H_1: $m_{48} > m_{38}$. Use the following data, which give the numbers of miles traveled per day for respective random samples of sizes 9 and 11:

48-passenger buses:	331	308	300	414	253	
	323	452	396	104		
38-passenger buses:	248	393	260	355	279	184
	386	450	432	196	197	

8.4-13. A company manufactures and packages soap powder in 6-pound boxes. The quality assurance department was interested in comparing the fill weights of packages from the east and west lines. Taking random samples from the two lines, the department obtained the following weights:

East line (x):	6.06	6.04	6.11	6.06	6.06
	6.07	6.06	6.08	6.05	6.09
West line (y):	6.08	6.03	6.04	6.07	6.11
	6.08	6.08	6.10	6.06	6.04

(a) Let m_X and m_Y denote the median weights for the east and west lines, respectively. Test H_0: $m_X = m_Y$ against a two-sided alternative hypothesis, using the Wilcoxon test with $\alpha \approx 0.05$. Find the p-value of this two-sided test.

(b) Construct and interpret a q–q plot of these data.

8.4-14. In Exercise 8.2-13, data are given that show the effect of a certain fertilizer on plant growth. The growths of the plants in mm over six weeks are repeated here, where Group A received fertilizer and Group B did not:

Group A:	55	61	33	57	17	46	50	42	71	51	63
Group B:	31	27	12	44	9	25	34	53	33	21	32

We shall test the hypothesis that fertilizer enhanced the growth of the plants.

(a) Construct a back-to-back stem-and-leaf display in which the stems are put down the center of the diagram and the Group A leaves go to the left while the Group B leaves go to the right.

(b) Calculate the value of the Wilcoxon statistic and give your conclusion.

(c) How does this result compare with that using the t test in Exercise 8.2-13?

8.4-15. With $\alpha = 0.05$, use the Wilcoxon statistic to test H_0: $m_X = m_Y$ against a two-sided alternative. Use the following observations of X and Y, which have been ordered for your convenience:

x:	−2.3864	−2.2171	−1.9148	−1.9097	−1.4883
	−1.2007	−1.1077	−0.3601	0.4325	1.0598
	1.3035	1.5241	1.7133	1.7656	2.4912
y:	−1.7613	−0.9391	−0.7437	−0.5530	−0.2469
	0.0647	0.2031	0.3219	0.3579	0.6431
	0.6557	0.6724	0.6762	0.9041	1.3571

8.4-16. Data were collected during a step-direction experiment in the biomechanics laboratory at Hope College. The goal of the study is to establish differences in stepping responses between healthy young and healthy older adults. In one part of the experiment, the subjects are told in what direction they should take a step. Then, when given a signal, the subject takes a step in that direction, and the time it takes for them to lift their foot to take the step is measured. The direction is repeated a few times throughout the testing, and for each subject, a mean of all the "liftoff" times in a certain direction is calculated. The mean liftoff times (in thousandths of a second) for the anterior direction, ordered for your convenience, are as follows:

Young Subjects	397	433	450	468	485	488	498	504	561
	565	569	576	577	579	581	586	696	
Older Subjects	463	538	549	573	588	590	594	626	627
	653	674	728	818	835	863	888	936	

(a) Construct a back-to-back stem-and-leaf display. Use stems $3\bullet, 4*, \ldots, 9*$.

(b) Use the Wilcoxon statistic to test the null hypothesis that the response times are equal against the alternative that the times for the young subjects are less than that for the older subjects.

(c) What outcome does a t test give?

8.4-17. Some measurements in mm were made on a species of spiders, named *Sosippus floridanus*, that are native to Florida. There are 10 female spiders and 10 male spiders. The body lengths and the lengths of their front and back legs are repeated here:

Female body lengths	Female front legs	Female back legs	Male body lengths	Male front legs	Male back legs
11.06	15.03	19.29	12.26	21.22	25.54
13.87	17.96	22.74	11.66	18.62	23.94
12.93	17.56	21.28	12.53	18.62	23.94
15.08	21.22	25.54	13.00	19.95	25.80
17.82	22.61	28.86	11.79	19.15	25.40
14.14	20.08	25.14	12.46	19.02	25.27
12.26	16.49	20.22	10.65	17.29	22.21
17.82	18.75	24.61	10.39	17.02	21.81
20.17	23.01	28.46	12.26	18.49	23.41
16.88	22.48	28.59	14.07	22.61	28.86

In this exercise, we shall use the Wilcoxon statistic to compare the sizes of the female and male spiders. For each of the following instructions, construct back-to-back stem-and-leaf displays:

(a) Test the null hypothesis that the body lengths of female and male spiders are equal against the alternative hypothesis that female spiders are longer.

(b) Test the null hypothesis that the lengths of the front legs of the female and male spiders are equal against a two-sided alternative.

(c) Test the null hypothesis that the lengths of the back legs of the female and male spiders are equal against a two-sided alternative.

8.4-18. In Exercise 8.2-10, growth data are given for plants in normal air and for plants in CO_2-enriched air. Those data are repeated here:

Normal Air (x)	4.67 4.21 2.18 3.91 4.09 5.24 2.94 4.71 4.04 5.79 3.80 4.38
Enriched Air (y)	5.04 4.52 6.18 7.01 4.36 1.81 6.22 5.70

In this exercise, we shall test the null hypothesis that the medians are equal, namely, H_0: $m_X = m_Y$, against the alternative hypothesis H_1: $m_X < m_Y$. You may select the significance level. However, give the approximate p-value or state clearly why you arrived at a particular conclusion, for each of the tests. Show your work.

(a) What is your conclusion from the Wilcoxon test?

(b) What was your conclusion from the t test in Exercise 8.2-10?

(c) Write a comparison of these two tests.

8.5 POWER OF A STATISTICAL TEST

In Chapter 8, we gave several tests of fairly common statistical hypotheses in such a way that we described the significance level α and the p-values of each. Of course, those tests were based on good (sufficient) statistics of the parameters, when the latter exist. In this section, we consider the probability of making the other type of error: accepting the null hypothesis H_0 when the alternative hypothesis H_1 is true. This consideration leads to ways to find most powerful tests of the null hypothesis H_0 against the alternative hypothesis H_1.

The first example introduces a new concept using a test about p, the probability of success. The sample size is kept small so that Table II in Appendix B can be used to find probabilities. The application is one that you can actually perform.

Example 8.5-1 Assume that when given a name tag, a person puts it on either the right or left side. Let p equal the probability that the name tag is placed on the right side. We shall test the null hypothesis H_0: $p = 1/2$ against the composite alternative hypothesis H_1: $p < 1/2$. (Included with the null hypothesis are those values of p which are greater than 1/2; that is, we could think of H_0 as H_0: $p \geq 1/2$.) We shall give name tags to a random sample of $n = 20$ people, denoting the placements of their name tags with Bernoulli random variables, $X_1, X_2, \ldots, X_{20}$, where $X_{i=1}$ if a person places the name tag on the right and $X_i = 0$ if a person places the name tag on the left. For our test statistic, we can then use $Y = \sum_{i=1}^{20} X_i$, which has the binomial distribution $b(20, p)$. Say the critical region is defined by $C = \{y : y \le 6\}$ or, equivalently, by $\{(x_1, x_2, \ldots, x_{20}) : \sum_{i=1}^{20} x_i \le 6\}$. Since Y is $b(20, 1/2)$ if $p = 1/2$, the significance level of the corresponding test is

$$\alpha = P\left(Y \le 6;\, p = \frac{1}{2}\right) = \sum_{y=0}^{6} \binom{20}{y} \left(\frac{1}{2}\right)^{20} = 0.0577,$$

from Table II in Appendix B. Of course, the probability β of a Type II error has different values, with different values of p selected from the composite alternative hypothesis $H_1\colon p < 1/2$. For example, with $p = 1/4$,

$$\beta = P\left(7 \le Y \le 20;\, p = \frac{1}{4}\right) = \sum_{y=7}^{20} \binom{20}{y} \left(\frac{1}{4}\right)^{y} \left(\frac{3}{4}\right)^{20-y} = 0.2142,$$

whereas with $p = 1/10$,

$$\beta = P\left(7 \le Y \le 20;\, p = \frac{1}{10}\right) = \sum_{y=7}^{20} \binom{20}{y} \left(\frac{1}{10}\right)^{y} \left(\frac{9}{10}\right)^{20-y} = 0.0024.$$

Instead of considering the probability β of accepting H_0 when H_1 is true, we could compute the probability K of rejecting H_0 when H_1 is true. After all, β and $K = 1-\beta$ provide the same information. Since K is a function of p, we denote this explicitly by writing $K(p)$. The probability

$$K(p) = \sum_{y=0}^{6} \binom{20}{y} p^y (1-p)^{20-y}, \qquad 0 < p \le \frac{1}{2},$$

is called the **power function of the test**. Of course, $\alpha = K(1/2) = 0.0577$, $1-K(1/4) = 0.2142$, and $1 - K(1/10) = 0.0024$. The value of the power function at a specified p is called the **power** of the test at that point. For instance, $K(1/4) = 0.7858$ and $K(1/10) = 0.9976$ are the powers at $p = 1/4$ and $p = 1/10$, respectively. An acceptable power function assumes small values when H_0 is true and larger values when p differs much from $p = 1/2$. (See Figure 8.5-1 for a graph of this power function.) ■

In Example 8.5-1, we introduced the new concept of the power function of a test. We now show how the sample size can be selected so as to create a test with appropriate power.

Example 8.5-2 Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the normal distribution $N(\mu, 100)$, which we can suppose is a possible distribution of scores of students in a statistics course that uses a new method of teaching (e.g., computer-related materials). We wish to decide between $H_0\colon \mu = 60$ (the *no-change* hypothesis because, let us say, this was the mean by the previous method of teaching) and the research worker's hypothesis $H_1\colon \mu > 60$. Let us consider a sample of size $n = 25$. Of course, the sample mean $\overline{X}$ is the maximum likelihood estimator of μ; thus, it seems reasonable to base our decision on this statistic. Initially, we use the rule to reject H_0 and accept H_1 if and only if $\bar{x} \ge 62$. What are the consequences of this test? These are summarized in the power function of the test.

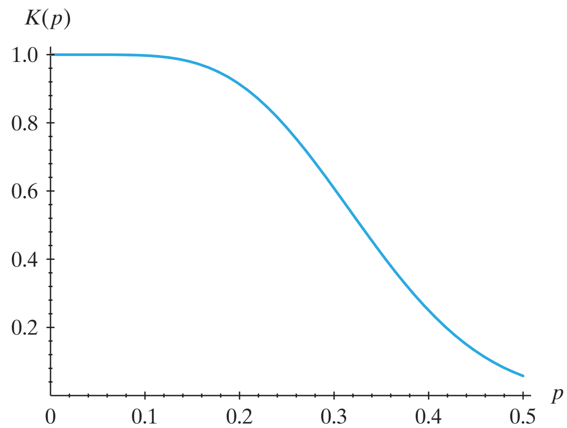

Figure 8.5-1 Power function: $K(p) = P(Y \le 6; p)$, where Y is $b(20, p)$

We first find the probability of rejecting H_0: $\mu = 60$ for various values of $\mu \ge 60$. The probability of rejecting H_0 is given by

$$K(\mu) = P(\overline{X} \ge 62; \mu),$$

because this test calls for the rejection of H_0: $\mu = 60$ when $\overline{x} \ge 62$. When the new process has the general mean μ, $\overline{X}$ has the normal distribution $N(\mu, 100/25 = 4)$. Accordingly,

$$\begin{aligned} K(\mu) &= P\left(\frac{\overline{X} - \mu}{2} \ge \frac{62 - \mu}{2}; \mu\right) \\ &= 1 - \Phi\left(\frac{62 - \mu}{2}\right), \qquad 60 \le \mu, \end{aligned}$$

is the probability of rejecting H_0: $\mu = 60$ by using this particular test. Several values of $K(\mu)$ are given in Table 8.5-1. Figure 8.5-2 depicts the graph of the function $K(\mu)$.

Table 8.5-1 Values of the power function

μ	$K(\mu)$
60	0.1587
61	0.3085
62	0.5000
63	0.6915
64	0.8413
65	0.9332
66	0.9772

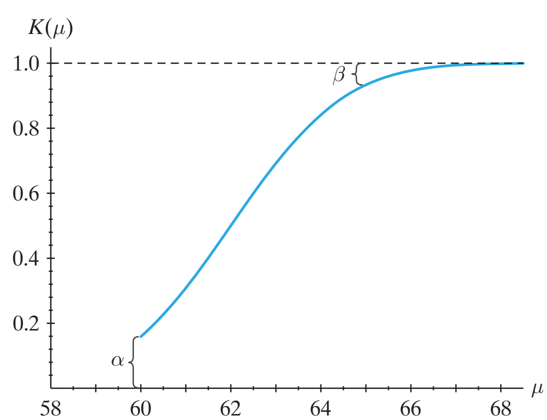

Figure 8.5-2 Power function $K(\mu) = 1 - \Phi([62 - \mu]/2)$

The probability $K(\mu)$ of rejecting H_0: $\mu = 60$ is called the *power function* of the test. At the value μ_1 of the parameter, $K(\mu_1)$ is the power at μ_1. The power at $\mu = 60$ is $K(60) = 0.1587$, and this is the probability of rejecting H_0: $\mu = 60$ when H_0 is true. That is, $K(60) = 0.1587 = \alpha$ is the probability of a Type I error and is called the *significance level* of the test.

The power at $\mu = 65$ is $K(65) = 0.9332$, and this is the probability of making the correct decision (namely, rejecting H_0: $\mu = 60$ when $\mu = 65$). Hence, we are pleased that here it is large. When $\mu = 65$, $1 - K(65) = 0.0668$ is the probability of not rejecting H_0: $\mu = 60$ when $\mu = 65$; that is, it is the probability of a Type II error and is denoted by $\beta = 0.0668$. These α- and β-values are displayed in Figure 8.5-2. Clearly, the probability $\beta = 1 - K(\mu_1)$ of a Type II error depends on which value—say, μ_1—is taken in the alternative hypothesis H_1: $\mu > 60$. Thus, while $\beta = 0.0668$ when $\mu = 65$, β is equal to $1 - K(63) = 0.3085$ when $\mu = 63$.

Frequently, statisticians like to have the significance level α smaller than 0.1587—say, around 0.05 or less—because it is a probability of an error, namely, a Type I error. Thus, if we would like $\alpha = 0.05$, then, with $n = 25$, we can no longer use the critical region $\bar{x} \ge 62$; rather, we use $\bar{x} \ge c$, where c is selected such that

$$K(60) = P(\overline{X} \ge c;\ \mu = 60) = 0.05.$$

However, when $\mu = 60$, $\overline{X}$ is $N(60, 4)$, and it follows that

$$\begin{aligned} K(60) &= P\left(\frac{\overline{X} - 60}{2} \ge \frac{c - 60}{2};\ \mu = 60\right) \\ &= 1 - \Phi\left(\frac{c - 60}{2}\right) = 0.05. \end{aligned}$$

From Table Va in Appendix B, we have

$$\frac{c - 60}{2} = 1.645 = z_{0.05} \qquad \text{and} \qquad c = 60 + 3.29 = 63.29.$$

Although this change reduces α from 0.1587 to 0.05, it increases β at $\mu = 65$ from 0.0668 to

$$\begin{aligned} \beta &= 1 - P(\overline{X} \ge 63.29;\ \mu = 65) \\ &= 1 - P\left(\frac{\overline{X} - 65}{2} \ge \frac{63.29 - 65}{2};\ \mu = 65\right) \\ &= \Phi(-0.855) = 0.1963. \end{aligned}$$

In general, without changing the sample size or the type of test of the hypothesis, a decrease in α causes an increase in β, and a decrease in β causes an increase in α. *Both* probabilities α and β of the two types of errors can be decreased only by increasing the sample size or, in some way, constructing a better test of the hypothesis.

For example, if $n = 100$ and we desire a test with significance level $\alpha = 0.05$, then, since $\overline{X}$ is $N(\mu, 100/100 = 1)$,

$$\alpha = P(\overline{X} \ge c;\ \mu = 60) = 0.05$$

means that

$$P\left(\frac{\overline{X} - 60}{1} \ge \frac{c - 60}{1};\ \mu = 60\right) = 0.05$$

and $c - 60 = 1.645$. Thus, $c = 61.645$. The power function is

$$\begin{aligned} K(\mu) &= P(\overline{X} \geq 61.645;\ \mu) \\ &= P\left(\frac{\overline{X}-\mu}{1} \geq \frac{61.645-\mu}{1};\ \mu\right) \ = \ 1 - \Phi(61.645-\mu). \end{aligned}$$

In particular, this means that at $\mu = 65$,

$$\beta = 1 - K(\mu) = \Phi(61.645 - 65) = \Phi(-3.355) = 0.0004;$$

so, with $n = 100$, both α and β have decreased from their respective original values of 0.1587 and 0.0668 when $n = 25$.

Rather than guess at the value of n, an ideal power function determines the sample size. Let us use a critical region of the form $\overline{x} \geq c$. Further, suppose that we want $\alpha = 0.025$ and, when $\mu = 65$, $\beta = 0.05$. Thus, since $\overline{X}$ is $N(\mu, 100/n)$, it follows that

$$0.025 = P(\overline{X} \geq c;\ \mu = 60) = 1 - \Phi\left(\frac{c-60}{10/\sqrt{n}}\right)$$

and

$$0.05 = 1 - P(\overline{X} \geq c;\ \mu = 65) = \Phi\left(\frac{c-65}{10/\sqrt{n}}\right).$$

That is,

$$\frac{c-60}{10/\sqrt{n}} = 1.96 \qquad \text{and} \qquad \frac{c-65}{10/\sqrt{n}} = -1.645.$$

Solving these equations simultaneously for c and $10/\sqrt{n}$, we obtain

$$c = 60 + 1.96\,\frac{5}{3.605} = 62.718;$$

$$\frac{10}{\sqrt{n}} = \frac{5}{3.605}.$$

Hence,

$$\sqrt{n} = 7.21 \qquad \text{and} \qquad n = 51.98.$$

Since n must be an integer, we would use $n = 52$ and thus obtain $\alpha \approx 0.025$ and $\beta \approx 0.05$.

The next example is an extension of Example 8.5-1.

Example 8.5-3 To test H_0: $p = 1/2$ against H_1: $p < 1/2$, we take a random sample of Bernoulli trials, $X_1, X_2, \ldots, X_n$, and use for our test statistic $Y = \sum_{i=1}^{n} X_i$, which has a binomial distribution $b(n, p)$. Let the critical region be defined by $C = \{y : y \leq c\}$. The power function for this test is defined by $K(p) = P(Y \leq c;\ p)$. We shall find the values of n and c so that $K(1/2) \approx 0.05$ and $K(1/4) \approx 0.90$. That is, we would like the significance level to be $\alpha = K(1/2) = 0.05$ and the power at $p = 1/4$ to equal 0.90. We proceed as follows: Since

$$0.05 = P\left(Y \leq c;\ p = \frac{1}{2}\right) = P\left(\frac{Y - n/2}{\sqrt{n(1/2)(1/2)}} \leq \frac{c - n/2}{\sqrt{n(1/2)(1/2)}}\right),$$

it follows that

$$(c - n/2)/\sqrt{n/4} \approx -1.645;$$

and since

$$0.90 = P\left(Y \le c;\, p = \frac{1}{4}\right) = P\left(\frac{Y - n/4}{\sqrt{n(1/4)(3/4)}} \le \frac{c - n/4}{\sqrt{n(1/4)(3/4)}}\right),$$

it follows that

$$(c - n/4)/\sqrt{3n/16} \approx 1.282.$$

Therefore,

$$\frac{n}{4} \approx 1.645\sqrt{\frac{n}{4}} + 1.282\sqrt{\frac{3n}{16}} \qquad \text{and} \qquad \sqrt{n} \approx 4(1.378) = 5.512.$$

Thus, n is approximately 30.4, and we round upward to 31. From either of the first two approximate equalities, we find that c is about equal to 10.9. Using $n = 31$ and $c = 10.9$ means that $K(1/2) = 0.05$ and $K(1/4) = 0.90$ are only approximate. In fact, since Y must be an integer, we could let $c = 10.5$. Then, with $n = 31$,

$$\begin{aligned} \alpha \;=\; K\left(\frac{1}{2}\right) &= P\left(Y \le 10.5;\, p = \frac{1}{2}\right) \;\approx\; 0.0362; \\ K\left(\frac{1}{4}\right) &= P\left(Y \le 10.5;\, p = \frac{1}{4}\right) \;\approx\; 0.8730. \end{aligned}$$

Or we could let $c = 11.5$ and $n = 32$, in which case

$$\begin{aligned} \alpha \;=\; K\left(\frac{1}{2}\right) &= P\left(Y \le 11.5;\, p = \frac{1}{2}\right) \;\approx\; 0.0558; \\ K\left(\frac{1}{4}\right) &= P\left(Y \le 11.5;\, p = \frac{1}{4}\right) \;\approx\; 0.9235. \end{aligned}$$

Exercises

8.5-1. A certain size of bag is designed to hold 25 pounds of potatoes. A farmer fills such bags in the field. Assume that the weight X of potatoes in a bag is $N(\mu, 9)$. We shall test the null hypothesis H_0: $\mu = 25$ against the alternative hypothesis H_1: $\mu < 25$. Let X_1, X_2, X_3, X_4 be a random sample of size 4 from this distribution, and let the critical region C for this test be defined by $\bar{x} \le 22.5$, where $\bar{x}$ is the observed value of $\overline{X}$.

(a) What is the power function $K(\mu)$ of this test? In particular, what is the significance level $\alpha = K(25)$ for your test?

(b) If the random sample of four bags of potatoes yielded the values $x_1 = 21.24$, $x_2 = 24.81$, $x_3 = 23.62$, and $x_4 = 26.82$, would your test lead you to accept or reject H_0?

(c) What is the p-value associated with $\bar{x}$ in part (b)?

8.5-2. Let X equal the number of milliliters of a liquid in a bottle that has a label volume of 350 ml. Assume that the distribution of X is $N(\mu, 4)$. To test the null hypothesis H_0: $\mu = 355$ against the alternative hypothesis H_1: $\mu < 355$, let the critical region be defined by $C = \{\bar{x} : \bar{x} \le 354.05\}$, where $\overline{x}$ is the sample mean of the contents of a random sample of $n = 12$ bottles.

(a) Find the power function $K(\mu)$ for this test.

(b) What is the (approximate) significance level of the test?

(c) Find the values of $K(354.05)$ and $K(353.1)$, and sketch the graph of the power function.

(d) Use the following 12 observations to state your conclusion from this test:

350	353	354	356	353	352
354	355	357	353	354	355

(e) What is the approximate p-value of the test?

8.5-3. Assume that SAT mathematics scores of students who attend small liberal arts colleges are $N(\mu, 8100)$. We shall test H_0: $\mu = 530$ against the alternative hypothesis H_1: $\mu < 530$. Given a random sample of size $n = 36$ SAT mathematics scores, let the critical region be defined by $C = \{\overline{x}: \overline{x} \le 510.77\}$, where $\overline{x}$ is the observed mean of the sample.

(a) Find the power function, $K(\mu)$, for this test.

(b) What is the value of the significance level of the test?

(c) What is the value of $K(510.77)$?

(d) Sketch the graph of the power function.

(e) What is the p-value associated with **(i)** $\overline{x} = 507.35$; **(ii)** $\overline{x} = 497.45$?

8.5-4. Let X be $N(\mu, 100)$. To test H_0: $\mu = 80$ against H_1: $\mu > 80$, let the critical region be defined by $C = \{(x_1, x_2, \ldots, x_{25}) : \overline{x} \ge 83\}$, where $\overline{x}$ is the sample mean of a random sample of size $n = 25$ from this distribution.

(a) What is the power function $K(\mu)$ for this test?

(b) What is the significance level of the test?

(c) What are the values of $K(80)$, $K(83)$, and $K(86)$?

(d) Sketch the graph of the power function.

(e) What is the p-value corresponding to $\overline{x} = 83.41$?

8.5-5. Let X equal the yield of alfalfa in tons per acre per year. Assume that X is $N(1.5, 0.09)$. It is hoped that a new fertilizer will increase the average yield. We shall test the null hypothesis H_0: $\mu = 1.5$ against the alternative hypothesis H_1: $\mu > 1.5$. Assume that the variance continues to equal $\sigma^2 = 0.09$ with the new fertilizer. Using $\overline{X}$, the mean of a random sample of size n, as the test statistic, reject H_0 if $\overline{x} \ge c$. Find n and c so that the power function $K(\mu) = P(\overline{X} \ge c : \mu)$ is such that $\alpha = K(1.5) = 0.05$ and $K(1.7) = 0.95$.

8.5-6. Let X equal the butterfat production (in pounds) of a Holstein cow during the 305-day milking period following the birth of a calf. Assume that the distribution of X is $N(\mu, 140^2)$. To test the null hypothesis H_0: $\mu = 715$ against the alternative hypothesis H_1: $\mu < 715$, let the critical region be defined by $C = \{\overline{x} : \overline{x} \le 668.94\}$, where $\overline{x}$ is the sample mean of $n = 25$ butterfat weights from 25 cows selected at random.

(a) Find the power function $K(\mu)$ for this test.

(b) What is the significance level of the test?

(c) What are the values of $K(668.94)$ and $K(622.88)$?

(d) Sketch a graph of the power function.

(e) What conclusion do you draw from the following 25 observations of X?

425	710	661	664	732	714	934	761	744
653	725	657	421	573	535	602	537	405
874	791	721	849	567	468	975		

(f) What is the approximate p-value of the test?

8.5-7. In Exercise 8.5-6, let $C = \{\overline{x} : \overline{x} \le c\}$ be the critical region. Find values for n and c so that the significance level of this test is $\alpha = 0.05$ and the power at $\mu = 650$ is 0.90.

8.5-8. Let X have a Bernoulli distribution with pmf

$$f(x; p) = p^x(1 - p)^{1-x}, \qquad x = 0, 1, \qquad 0 \le p \le 1.$$

We would like to test the null hypothesis H_0: $p \le 0.4$ against the alternative hypothesis H_1: $p > 0.4$. For the test statistic, use $Y = \sum_{i=1}^{n} X_i$, where $X_1, X_2, \ldots, X_n$ is a random sample of size n from this Bernoulli distribution. Let the critical region be of the form $C = \{y: y \ge c\}$.

(a) Let $n = 100$. On the same set of axes, sketch the graphs of the power functions corresponding to the three critical regions, $C_1 = \{y : y \ge 40\}$, $C_2 = \{y : y \ge 50\}$, and $C_3 = \{y : y \ge 60\}$. Use the normal approximation to compute the probabilities.

(b) Let $C = \{y : y \ge 0.45n\}$. On the same set of axes, sketch the graphs of the power functions corresponding to the three samples of sizes 10, 100, and 1000.

8.5-9. Let p denote the probability that, for a particular tennis player, the first serve is good. Since $p = 0.40$, this player decided to take lessons in order to increase p. When the lessons are completed, the hypothesis H_0: $p = 0.40$ will be tested against H_1: $p > 0.40$ on the basis of $n = 25$ trials. Let y equal the number of first serves that are good, and let the critical region be defined by $C = \{y: y \ge 14\}$.

(a) Find the power function $K(p)$ for this test.

(b) What is the value of the significance level, $\alpha = K(0.40)$? Use Table II in Appendix B.

(c) Evaluate $K(p)$ at $p = 0.45, 0.50, 0.60, 0.70, 0.80$, and 0.90. Use Table II.

(d) Sketch the graph of the power function.

(e) If $y = 15$ following the lessons, would H_0 be rejected?

(f) What is the p-value associated with $y = 15$?

8.5-10. Let $X_1, X_2, \ldots, X_8$ be a random sample of size $n = 8$ from a Poisson distribution with mean λ. Reject the simple null hypothesis H_0: $\lambda = 0.5$, and accept H_1: $\lambda > 0.5$, if the observed sum $\sum_{i=1}^{8} x_i \geq 8$.

(a) Compute the significance level α of the test.

(b) Find the power function $K(\lambda)$ of the test as a sum of Poisson probabilities.

(c) Using Table III in Appendix B, determine $K(0.75)$, $K(1)$, and $K(1.25)$.

8.5-11. Let p equal the fraction defective of a certain manufactured item. To test H_0: $p = 1/26$ against H_1: $p > 1/26$, we inspect n items selected at random and let Y be the number of defective items in this sample. We reject H_0 if the observed $y \geq c$. Find n and c so that $\alpha = K(1/26) \approx 0.05$ and $K(1/10) \approx 0.90$, where $K(p) = P(Y \geq c; p)$. Hint: Use either the normal or Poisson approximation to help solve this exercise.

8.5-12. Let X_1, X_2, X_3 be a random sample of size $n = 3$ from an exponential distribution with mean $\theta > 0$. Reject the simple null hypothesis H_0: $\theta = 2$, and accept the composite alternative hypothesis H_1: $\theta < 2$, if the observed sum $\sum_{i=1}^{3} x_i \leq 2$.

(a) What is the power function $K(\theta)$, written as an integral?

(b) Using integration by parts, define the power function as a summation.

(c) With the help of Table III in Appendix B, determine $\alpha = K(2)$, $K(1)$, $K(1/2)$, and $K(1/4)$.

8.6 BEST CRITICAL REGIONS

In this section, we consider the properties a satisfactory hypothesis test (or critical region) should possess. To introduce our investigation, we begin with a nonstatistical example.

Example 8.6-1

Say that you have α dollars with which to buy books. Further, suppose that you are not interested in the books themselves, but only in filling as much of your bookshelves as possible. How do you decide which books to buy? Does the following approach seem reasonable? First of all, take all the available free books. Then start choosing those books for which the cost of filling an inch of bookshelf is smallest. That is, choose those books for which the ratio c/w is a minimum, where w is the width of the book in inches and c is the cost of the book. Continue choosing books this way until you have spent the α dollars. ■

To see how Example 8.6-1 provides the background for selecting a good critical region of size α, let us consider a test of the simple hypothesis H_0: $\theta = \theta_0$ against a simple alternative hypothesis H_1: $\theta = \theta_1$. In this discussion, we assume that the random variables $X_1, X_2, \ldots, X_n$ under consideration have a joint pmf of the discrete type, which we here denote by $L(\theta; x_1, x_2, \ldots, x_n)$. That is,

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = L(\theta; x_1, x_2, \ldots, x_n).$$

A critical region C of size α is a set of points $(x_1, x_2, \ldots, x_n)$ with probability α when $\theta = \theta_0$. For a good test, this set C of points should have a large probability when $\theta = \theta_1$, because, under H_1: $\theta = \theta_1$, we wish to reject H_0: $\theta = \theta_0$. Accordingly, the first point we would place in the critical region C is the one with the smallest ratio:

$$\frac{L(\theta_0; x_1, x_2, \ldots, x_n)}{L(\theta_1; x_1, x_2, \ldots, x_n)}.$$

That is, the "cost" in terms of probability under $H_0\colon \theta = \theta_0$ is small compared with the probability that we can "buy" if $\theta = \theta_1$. The next point to add to C would be the one with the next-smallest ratio. We would continue to add points to C in this manner until the probability of C, under $H_0\colon \theta = \theta_0$, equals α. In this way, for the given significance level α, we have achieved the region C with the largest probability when $H_1\colon \theta = \theta_1$ is true. We now formalize this discussion by defining a best critical region and proving the well-known Neyman–Pearson lemma.

Definition 8.6-1
Consider the test of the simple null hypothesis $H_0\colon \theta = \theta_0$ against the simple alternative hypothesis $H_1\colon \theta = \theta_1$. Let C be a critical region of size α; that is, $\alpha = P(C;\theta_0)$. Then C is a **best critical region of size** α if, for every other critical region D of size $\alpha = P(D;\theta_0)$, we have

$$P(C;\theta_1) \geq P(D;\theta_1).$$

That is, when $H_1\colon \theta = \theta_1$ is true, the probability of rejecting $H_0\colon \theta = \theta_0$ with the use of the critical region C is at least as great as the corresponding probability with the use of any other critical region D of size α.

Thus, a best critical region of size α is the critical region that has the greatest power among all critical regions of size α. The Neyman–Pearson lemma gives sufficient conditions for a best critical region of size α.

Theorem 8.6-1 **(Neyman–Pearson Lemma)** Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with pdf or pmf $f(x;\theta)$, where θ_0 and θ_1 are two possible values of θ. Denote the joint pdf or pmf of $X_1, X_2, \ldots, X_n$ by the likelihood function

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta).$$

If there exist a positive constant k and a subset C of the sample space such that

(a) $P[(X_1, X_2, \ldots, X_n) \in C;\theta_0] = \alpha$,

(b) $\dfrac{L(\theta_0)}{L(\theta_1)} \leq k$ for $(x_1, x_2, \ldots, x_n) \in C$, and

(c) $\dfrac{L(\theta_0)}{L(\theta_1)} \geq k$ for $(x_1, x_2, \ldots, x_n) \in C'$,

then C is a best critical region of size α for testing the simple null hypothesis $H_0\colon \theta = \theta_0$ against the simple alternative hypothesis $H_1\colon \theta = \theta_1$.

Proof We prove the theorem when the random variables are of the continuous type; for discrete-type random variables, replace the integral signs by summation signs. To simplify the exposition, we shall use the following notation:

$$\int_B L(\theta) = \int \underset{B}{\cdots} \int L(\theta; x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n.$$

Assume that there exists another critical region of size α—say, D—such that, in this new notation,

$$\alpha = \int_C L(\theta_0) = \int_D L(\theta_0).$$

Then we have

$$\begin{aligned} 0 &= \int_C L(\theta_0) - \int_D L(\theta_0) \\ &= \int_{C\cap D'} L(\theta_0) + \int_{C\cap D} L(\theta_0) - \int_{C\cap D} L(\theta_0) - \int_{C'\cap D} L(\theta_0). \end{aligned}$$

Hence,

$$0 = \int_{C\cap D'} L(\theta_0) - \int_{C'\cap D} L(\theta_0).$$

By hypothesis (b), $kL(\theta_1) \ge L(\theta_0)$ at each point in C and therefore in $C \cap D'$; thus,

$$k \int_{C\cap D'} L(\theta_1) \ge \int_{C\cap D'} L(\theta_0).$$

By hypothesis (c), $kL(\theta_1) \le L(\theta_0)$ at each point in C' and therefore in $C' \cap D$; thus, we obtain

$$k \int_{C'\cap D} L(\theta_1) \le \int_{C'\cap D} L(\theta_0).$$

Consequently,

$$0 = \int_{C\cap D'} L(\theta_0) - \int_{C'\cap D} L(\theta_0) \le (k) \left\{ \int_{C\cap D'} L(\theta_1) - \int_{C'\cap D} L(\theta_1) \right\}.$$

That is,

$$0 \le (k) \left\{ \int_{C\cap D'} L(\theta_1) + \int_{C\cap D} L(\theta_1) - \int_{C\cap D} L(\theta_1) - \int_{C'\cap D} L(\theta_1) \right\}$$

or, equivalently,

$$0 \le (k) \left\{ \int_C L(\theta_1) - \int_D L(\theta_1) \right\}.$$

Thus,

$$\int_C L(\theta_1) \ge \int_D L(\theta_1);$$

that is, $P(C;\theta_1) \ge P(D;\theta_1)$. Since that is true for every critical region D of size α, C is a best critical region of size α. □

For a realistic application of the Neyman–Pearson lemma, consider the next example, in which the test is based on a random sample from a normal distribution.

Example 8.6-2 Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $N(\mu, 36)$. We shall find the best critical region for testing the simple hypothesis H_0: $\mu = 50$ against the simple alternative hypothesis H_1: $\mu = 55$. Using the ratio of the likelihood functions, namely, $L(50)/L(55)$, we shall find those points in the sample space for which this ratio is less than or equal to some positive constant k. That is, we shall solve the following inequality:

$$\frac{L(50)}{L(55)} = \frac{(72\pi)^{-n/2}\exp\left[-\left(\frac{1}{72}\right)\sum_{i=1}^{n}(x_i-50)^2\right]}{(72\pi)^{-n/2}\exp\left[-\left(\frac{1}{72}\right)\sum_{i=1}^{n}(x_i-55)^2\right]}$$

$$= \exp\left[-\left(\frac{1}{72}\right)\left(10\sum_{i=1}^{n}x_i + n50^2 - n55^2\right)\right] \le k.$$

If we take the natural logarithm of each member of the inequality, we find that

$$-10\sum_{i=1}^{n}x_i - n50^2 + n55^2 \le (72)\ln k.$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}x_i \ge -\frac{1}{10n}\,[n50^2 - n55^2 + (72)\ln k]$$

or, equivalently,

$$\bar{x} \ge c,$$

where $c = -(1/10n)[n50^2 - n55^2 + (72)\ln k]$. Hence, $L(50)/L(55) \le k$ is equivalent to $\bar{x} \ge c$. According to the Neyman–Pearson lemma, a best critical region is

$$C = \{(x_1, x_2, \ldots, x_n)\colon\ \bar{x} \ge c\},$$

where c is selected so that the size of the critical region is α. Say $n = 16$ and $c = 53$. Then, since $\overline{X}$ is $N(50, 36/16)$ under H_0, we have

$$\begin{aligned}\alpha &= P(\overline{X} \ge 53;\ \mu = 50)\\ &= P\left(\frac{\overline{X}-50}{6/4} \ge \frac{3}{6/4};\ \mu = 50\right) \;=\; 1 - \Phi(2) \;=\; 0.0228.\end{aligned}$$

This last example illustrates what is often true, namely, that the inequality

$$L(\theta_0)/L(\theta_1) \le k$$

can be expressed in terms of a function $u(x_1, x_2, \ldots, x_n)$, say,

$$u(x_1, \ldots, x_n) \le c_1$$

or

$$u(x_1, \ldots, x_n) \ge c_2,$$

where c_1 or c_2 is selected so that the size of the critical region is α. Thus, the test can be based on the statistic $u(x_1, \ldots, x_n)$. As an example, if we want α to be a given value—say, 0.05—we could then choose our c_1 or c_2. In Example 8.6-2, with $\alpha = 0.05$, we want

$$0.05 = P(\overline{X} \geq c;\ \mu = 50)$$
$$= P\left(\frac{\overline{X} - 50}{6/4} \geq \frac{c - 50}{6/4};\ \mu = 50\right) = 1 - \Phi\left(\frac{c - 50}{6/4}\right).$$

Hence, it must be true that $(c - 50)/(3/2) = 1.645$, or, equivalently,

$$c = 50 + \frac{3}{2}(1.645) \approx 52.47.$$

Example 8.6-3 Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a Poisson distribution with mean λ. A best critical region for testing H_0: $\lambda = 2$ against H_1: $\lambda = 5$ is given by

$$\frac{L(2)}{L(5)} = \frac{2^{\Sigma x_i} e^{-2n}}{x_1! x_2! \cdots x_n!} \frac{x_1! x_2! \cdots x_n!}{5^{\Sigma x_i} e^{-5n}} \leq k.$$

This inequality can be written as

$$\left(\frac{2}{5}\right)^{\Sigma x_i} e^{3n} \leq k, \qquad \text{or} \qquad (\Sigma x_i) \ln\left(\frac{2}{5}\right) + 3n \leq \ln k.$$

Since $\ln(2/5) < 0$, the latter inequality is the same as

$$\sum_{i=1}^{n} x_i \geq \frac{\ln k - 3n}{\ln(2/5)} = c.$$

If $n = 4$ and $c = 13$, then

$$\alpha = P\left(\sum_{i=1}^{4} X_i \geq 13;\ \lambda = 2\right) = 1 - 0.936 = 0.064,$$

from Table III in Appendix B, since $\sum_{i=1}^{4} X_i$ has a Poisson distribution with mean 8 when $\lambda = 2$. ■

When H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$ are both simple hypotheses, a critical region of size α is a best critical region if the probability of rejecting H_0 when H_1 is true is a maximum compared with all other critical regions of size α. The test using the best critical region is called a **most powerful test**, because it has the greatest value of the power function at $\theta = \theta_1$ compared with that of other tests with significance level α. If H_1 is a composite hypothesis, the power of a test depends on each simple alternative in H_1.

Definition 8.6-2
A test defined by a critical region C of size α is a **uniformly most powerful test** if it is a most powerful test against each simple alternative in H_1. The critical region C is called a **uniformly most powerful critical region of size** α.

Let us consider again Example 8.6-2 when the alternative hypothesis is composite.

Example 8.6-4

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, 36)$. We have seen that, in testing H_0: $\mu = 50$ against H_1: $\mu = 55$, a best critical region C is defined by $C = \{(x_1, x_2, \ldots, x_n) : \overline{x} \ge c\}$, where c is selected so that the significance level is α. Now consider testing H_0: $\mu = 50$ against the one-sided composite alternative hypothesis H_1: $\mu > 50$. For each simple hypothesis in H_1—say, $\mu = \mu_1$—the quotient of the likelihood functions is

$$\frac{L(50)}{L(\mu_1)} = \frac{(72\pi)^{-n/2} \exp\left[-\left(\frac{1}{72}\right)\sum_{i=1}^{n}(x_i - 50)^2\right]}{(72\pi)^{-n/2} \exp\left[-\left(\frac{1}{72}\right)\sum_{i=1}^{n}(x_i - \mu_1)^2\right]}$$

$$= \exp\left[-\frac{1}{72}\left\{2(\mu_1 - 50)\sum_{i=1}^{n} x_i + n(50^2 - \mu_1^2)\right\}\right].$$

Now, $L(50)/L(\mu_1) \le k$ if and only if

$$\overline{x} \ge \frac{(-72)\ln(k)}{2n(\mu_1 - 50)} + \frac{50 + \mu_1}{2} = c.$$

Thus, the best critical region of size α for testing H_0: $\mu = 50$ against H_1: $\mu = \mu_1$, where $\mu_1 > 50$, is given by $C = \{(x_1, x_2, \ldots, x_n) : \overline{x} \ge c\}$, where c is selected such that $P(\overline{X} \ge c;\, H_0\!: \mu = 50) = \alpha$. Note that the same value of c can be used for each $\mu_1 > 50$, but (of course) k does not remain the same. Since the critical region C defines a test that is most powerful against each simple alternative $\mu_1 > 50$, this is a uniformly most powerful test, and C is a uniformly most powerful critical region of size α. Again, if $\alpha = 0.05$, then $c \approx 52.47$. ■

Example 8.6-5

Let Y have the binomial distribution $b(n, p)$. To find a uniformly most powerful test of the simple null hypothesis H_0: $p = p_0$ against the one-sided alternative hypothesis H_1: $p > p_0$, consider, with $p_1 > p_0$,

$$\frac{L(p_0)}{L(p_1)} = \frac{\binom{n}{y} p_0^y (1 - p_0)^{n-y}}{\binom{n}{y} p_1^y (1 - p_1)^{n-y}} \le k.$$

This is equivalent to

$$\left[\frac{p_0(1 - p_1)}{p_1(1 - p_0)}\right]^y \left[\frac{1 - p_0}{1 - p_1}\right]^n \le k$$

and

$$y \ln\left[\frac{p_0(1 - p_1)}{p_1(1 - p_0)}\right] \le \ln k - n \ln\left[\frac{1 - p_0}{1 - p_1}\right].$$

Since $p_0 < p_1$, we have $p_0(1 - p_1) < p_1(1 - p_0)$. Thus, $\ln[p_0(1 - p_1)/p_1(1 - p_0)] < 0$. It follows that

$$\frac{y}{n} \ge \frac{\ln k - n \ln[(1 - p_0)/(1 - p_1)]}{n \ln[p_0(1 - p_1)/p_1(1 - p_0)]} = c$$

for each $p_1 > p_0$.

It is interesting to note that if the alternative hypothesis is the one-sided H_1: $p < p_0$, then a uniformly most powerful test is of the form $(y/n) \le c$. Thus, the tests of H_0: $p = p_0$ against the one-sided alternatives given in Table 8.3-1 are uniformly most powerful. ■

Exercise 8.6-5 will demonstrate that uniformly most powerful tests do not always exist; in particular, they usually do not exist when the composite alternative hypothesis is two sided.

REMARK We close this section with one easy but important observation: If a sufficient statistic $Y = u(X_1, X_2, \ldots, X_n)$ exists for θ, then, by the factorization theorem (Definition 6.7-1),

$$\frac{L(\theta_0)}{L(\theta_1)} = \frac{\phi[u(x_1, x_2, \ldots, x_n); \theta_0]\, h(x_1, x_2, \ldots, x_n)}{\phi[u(x_1, x_2, \ldots, x_n); \theta_1]\, h(x_1, x_2, \ldots, x_n)}$$

$$= \frac{\phi[u(x_1, x_2, \ldots, x_n); \theta_0)}{\phi[u(x_1, x_2, \ldots, x_n); \theta_1]}.$$

Thus, $L(\theta_0)/L(\theta_1) \le k$ provides a critical region that is a function of the observations $x_1, x_2, \ldots x_n$ only through the observed value of the sufficient statistic $y = u(x_1, x_2, \ldots, x_n)$. Hence, best critical and uniformly most powerful critical regions are based upon sufficient statistics when they exist. ■

Exercises

8.6-1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, 64)$.

(a) Show that $C = \{(x_1, x_2, \ldots, x_n) : \bar{x} \le c\}$ is a best critical region for testing H_0: $\mu = 80$ against H_1: $\mu = 76$.

(b) Find n and c so that $\alpha \approx 0.05$ and $\beta \approx 0.05$.

8.6-2. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(0, \sigma^2)$.

(a) Show that $C = \{(x_1, x_2, \ldots, x_n) : \sum_{i=1}^{n} x_i^2 \ge c\}$ is a best critical region for testing H_0: $\sigma^2 = 4$ against H_1: $\sigma^2 = 16$.

(b) If $n = 15$, find the value of c so that $\alpha = 0.05$. HINT: Recall that $\sum_{i=1}^{n} X_i^2/\sigma^2$ is $\chi^2(n)$.

(c) If $n = 15$ and c is the value found in part (b), find the approximate value of $\beta = P(\sum_{i=1}^{n} X_i^2 < c;\ \sigma^2 = 16)$.

8.6-3. Let X have an exponential distribution with a mean of θ; that is, the pdf of X is $f(x; \theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$. Let $X_1, X_2, \ldots, X_n$ be a random sample from this distribution.

(a) Show that a best critical region for testing H_0: $\theta = 3$ against H_1: $\theta = 5$ can be based on the statistic $\sum_{i=1}^{n} X_i$.

(b) If $n = 12$, use the fact that $(2/\theta)\sum_{i=1}^{12} X_i$ is $\chi^2(24)$ to find a best critical region of size $\alpha = 0.10$.

(c) If $n = 12$, find a best critical region of size $\alpha = 0.10$ for testing H_0: $\theta = 3$ against H_1: $\theta = 7$.

(d) If H_1: $\theta > 3$, is the common region found in parts (b) and (c) a uniformly most powerful critical region of size $\alpha = 0.10$?

8.6-4. Let $X_1, X_2, \ldots, X_n$ be a random sample of Bernoulli trials $b(1, p)$.

(a) Show that a best critical region for testing H_0: $p = 0.9$ against H_1: $p = 0.8$ can be based on the statistic $Y = \sum_{i=1}^{n} X_i$, which is $b(n, p)$.

(b) If $C = \{(x_1, x_2, \ldots, x_n) : \sum_{i=1}^{n} x_i \le n(0.85)\}$ and $Y = \sum_{i=1}^{n} X_i$, find the value of n such that $\alpha = P[\,Y \le n(0.85);\ p = 0.9\,] \approx 0.10$. HINT: Use the normal approximation for the binomial distribution.

(c) What is the approximate value of $\beta = P[\,Y > n(0.85);\ p = 0.8\,]$ for the test given in part (b)?

(d) Is the test of part (b) a uniformly most powerful test when the alternative hypothesis is $H_1\colon p < 0.9$?

8.6-5. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $N(\mu, 36)$.

(a) Show that a uniformly most powerful critical region for testing $H_0\colon \mu = 50$ against $H_1\colon \mu < 50$ is given by $C_2 = \{\bar{x}\colon \bar{x} \le c\}$.

(b) With this result and that of Example 8.6-4, argue that a uniformly most powerful test for testing $H_0\colon \mu = 50$ against $H_1\colon \mu \ne 50$ does not exist.

8.6-6. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $N(\mu, 9)$. To test the hypothesis H_0: $\mu = 80$ against $H_1\colon \mu \ne 80$, consider the following three critical regions: $C_1 = \{\bar{x}\colon \bar{x} \ge c_1\}$, $C_2 = \{\bar{x}\colon \bar{x} \le c_2\}$, and $C_3 = \{\bar{x}\colon |\bar{x} - 80| \ge c_3\}$.

(a) If $n = 16$, find the values of c_1, c_2, c_3 such that the size of each critical region is 0.05. That is, find c_1, c_2, c_3 such that

$$0.05 = P(\overline{X} \in C_1;\ \mu = 80) = P(\overline{X} \in C_2;\ \mu = 80)$$
$$= P(\overline{X} \in C_3;\ \mu = 80).$$

(b) On the same graph paper, sketch the power functions for these three critical regions.

8.6-7. Let $X_1, X_2, \ldots, X_{10}$ be a random sample of size 10 from a Poisson distribution with mean μ.

(a) Show that a uniformly most powerful critical region for testing $H_0\colon \mu = 0.5$ against $H_1\colon \mu > 0.5$ can be defined with the use of the statistic $\sum_{i=1}^{10} X_i$.

(b) What is a uniformly most powerful critical region of size $\alpha = 0.068$? Recall that $\sum_{i=1}^{10} X_i$ has a Poisson distribution with mean 10μ.

(c) Sketch the power function of this test.

8.6-8. Consider a random sample $X_1, X_2, \ldots, X_n$ from a distribution with pdf $f(x;\theta) = \theta(1-x)^{\theta-1}$, $0 < x < 1$, where $0 < \theta$. Find the form of the uniformly most powerful test of $H_0\colon \theta = 1$ against $H_1\colon \theta > 1$.

8.6-9. Let $X_1, X_2, \ldots, X_5$ be a random sample from the Bernoulli distribution $p(x;\theta) = \theta^x(1-\theta)^{1-x}$. We reject $H_0\colon \theta = 1/2$ and accept $H_1\colon \theta < 1/2$ if $Y = \sum_{i=1}^{5} X_i \le c$. Show that this is a uniformly most powerful test and find the power function $K(\theta)$ if $c = 1$.

8.7* LIKELIHOOD RATIO TESTS

In this section, we consider a general test-construction method that is applicable when either or both of the null and alternative hypotheses—say, H_0 and H_1—are composite. We continue to assume that the functional form of the pdf is known, but that it depends on one or more unknown parameters. That is, we assume that the pdf of X is $f(x;\theta)$, where θ represents one or more unknown parameters. We let Ω denote the total parameter space—that is, the set of all possible values of the parameter θ given by either H_0 or H_1. These hypotheses will be stated as

$$H_0: \theta \in \omega, \qquad H_1: \theta \in \omega',$$

where ω is a subset of Ω and ω' is the complement of ω with respect to Ω. The test will be constructed with the use of a ratio of likelihood functions that have been maximized in ω and Ω, respectively. In a sense, this is a natural generalization of the ratio appearing in the Neyman–Pearson lemma when the two hypotheses were simple.

Definition 8.7-1
The **likelihood ratio** is the quotient

$$\lambda = \frac{L(\widehat{\omega})}{L(\widehat{\Omega})},$$

where $L(\widehat{\omega})$ is the maximum of the likelihood function with respect to θ when $\theta \in \omega$ and $L(\widehat{\Omega})$ is the maximum of the likelihood function with respect to θ when $\theta \in \Omega$.

Because λ is the quotient of nonnegative functions, $\lambda \geq 0$. In addition, since $\omega \subset \Omega$, it follows that $L(\widehat{\omega}) \leq L(\widehat{\Omega})$ and hence $\lambda \leq 1$. Thus, $0 \leq \lambda \leq 1$. If the maximum of L in ω is much smaller than that in Ω, it would seem that the data $x_1, x_2, \ldots, x_n$ do not support the hypothesis H_0: $\theta \in \omega$. That is, a small value of the ratio $\lambda = L(\widehat{\omega})/L(\widehat{\Omega})$ would lead to the rejection of H_0. In contrast, a value of the ratio λ that is close to 1 would support the null hypothesis H_0. This reasoning leads us to the next definition.

Definition 8.7-2

To test H_0: $\theta \in \omega$ against H_1: $\theta \in \omega'$, the **critical region for the likelihood ratio test** is the set of points in the sample space for which

$$\lambda = \frac{L(\widehat{\omega})}{L(\widehat{\Omega})} \leq k,$$

where $0 < k < 1$ and k is selected so that the test has a desired significance level α.

The next example illustrates these definitions.

Example 8.7-1

Assume that the weight X in ounces of a "10-pound" bag of sugar is $N(\mu, 5)$. We shall test the hypothesis H_0: $\mu = 162$ against the alternative hypothesis H_1: $\mu \neq 162$. Thus, $\Omega = \{\mu : -\infty < \mu < \infty\}$ and $\omega = \{162\}$. To find the likelihood ratio, we need $L(\widehat{\omega})$ and $L(\widehat{\Omega})$. When H_0 is true, μ can take on only one value, namely, $\mu = 162$. Hence, $L(\widehat{\omega}) = L(162)$. To find $L(\widehat{\Omega})$, we must find the value of μ that maximizes $L(\mu)$. Recall that $\widehat{\mu} = \overline{x}$ is the maximum likelihood estimate of μ. Then $L(\widehat{\Omega}) = L(\overline{x})$, and the likelihood ratio $\lambda = L(\widehat{\omega})/L(\widehat{\Omega})$ is given by

$$\begin{aligned}
\lambda &= \frac{(10\pi)^{-n/2} \exp\left[-\left(\dfrac{1}{10}\right)\displaystyle\sum_{i=1}^{n}(x_i - 162)^2\right]}{(10\pi)^{-n/2} \exp\left[-\left(\dfrac{1}{10}\right)\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2\right]} \\
&= \frac{\exp\left[-\left(\dfrac{1}{10}\right)\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2 - \left(\dfrac{n}{10}\right)(\overline{x} - 162)^2\right]}{\exp\left[-\left(\dfrac{1}{10}\right)\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2\right]} \\
&= \exp\left[-\frac{n}{10}(\overline{x} - 162)^2\right].
\end{aligned}$$

On the one hand, a value of $\overline{x}$ close to 162 would tend to support H_0, and in that case λ is close to 1. On the other hand, an $\overline{x}$ that differs from 162 by too much would tend to support H_1. (See Figure 8.7-1 for the graph of this likelihood ratio when $n = 5$.)

A critical region for a likelihood ratio is given by $\lambda \leq k$, where k is selected so that the significance level of the test is α. Using this criterion and simplifying the

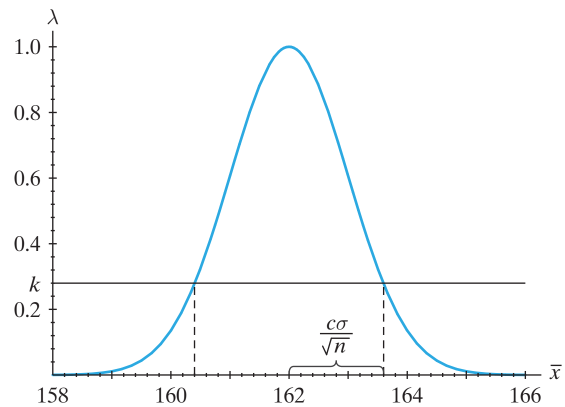

Figure 8.7-1 The likelihood ratio for testing $H_0: \mu = 162$

inequality as we do when we use the Neyman–Pearson lemma, we find that $\lambda \le k$ is equivalent to each of the following inequalities:

$$-\left(\frac{n}{10}\right)(\bar{x} - 162)^2 \le \ln k,$$

$$(\bar{x} - 162)^2 \ge -\left(\frac{10}{n}\right)\ln k,$$

$$\frac{|\bar{x} - 162|}{\sqrt{5}/\sqrt{n}} \ge \frac{\sqrt{-(10/n)\ln k}}{\sqrt{5}/\sqrt{n}} = c.$$

Since $Z = (\overline{X} - 162)/(\sqrt{5}/\sqrt{n})$ is $N(0, 1)$ when H_0: $\mu = 162$ is true, let $c = z_{\alpha/2}$. Thus, the critical region is

$$C = \left\{\bar{x}: \frac{|\bar{x} - 162|}{\sqrt{5}/\sqrt{n}} \ge z_{\alpha/2}\right\}.$$

To illustrate, if $\alpha = 0.05$, then $z_{0.025} = 1.96$. ■

As illustrated in Example 8.7-1, the inequality $\lambda \le k$ can often be expressed in terms of a statistic whose distribution is known. Also, note that although the likelihood ratio test is an intuitive test, it leads to the same critical region as that given by the Neyman–Pearson lemma when H_0 and H_1 are both simple hypotheses.

Suppose now that the random sample $X_1, X_2, \ldots, X_n$ arises from the normal population $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Let us consider the likelihood ratio test of the null hypothesis H_0: $\mu = \mu_0$ against the two-sided alternative hypothesis H_1: $\mu \ne \mu_0$. For this test,

$$\omega = \{(\mu, \sigma^2): \mu = \mu_0,\ 0 < \sigma^2 < \infty\}$$

and

$$\Omega = \{(\mu, \sigma^2): -\infty < \mu < \infty,\ 0 < \sigma^2 < \infty\}.$$

If $(\mu, \sigma^2) \in \Omega$, then the observed maximum likelihood estimates are $\widehat{\mu} = \overline{x}$ and $\widehat{\sigma^2} = (1/n)\sum_{i=1}^{n}(x_i - \overline{x})^2$.
Thus,

$$L(\widehat{\Omega}) = \left[\frac{1}{2\pi\left(\dfrac{1}{n}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]^{n/2} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\left(\dfrac{2}{n}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$$

$$= \left[\frac{ne^{-1}}{2\pi\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]^{n/2}.$$

Similarly, if $(\mu, \sigma^2) \in \omega$, then the observed maximum likelihood estimates are $\widehat{\mu} = \mu_0$ and $\widehat{\sigma^2} = (1/n)\sum_{i=1}^{n}(x_i - \mu_0)^2$. Hence,

$$L(\widehat{\omega}) = \left[\frac{1}{2\pi\left(\dfrac{1}{n}\right)\sum_{i=1}^{n}(x_i - \mu_0)^2}\right]^{n/2} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{\left(\dfrac{2}{n}\right)\sum_{i=1}^{n}(x_i - \mu_0)^2}\right]$$

$$= \left[\frac{ne^{-1}}{2\pi\sum_{i=1}^{n}(x_i - \mu_0)^2}\right]^{n/2}.$$

The likelihood ratio $\lambda = L(\widehat{\omega})/L(\widehat{\Omega})$ for this test is

$$\lambda = \frac{\left[\dfrac{ne^{-1}}{2\pi\sum_{i=1}^{n}(x_i - \mu_0)^2}\right]^{n/2}}{\left[\dfrac{ne^{-1}}{2\pi\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]^{n/2}} = \left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2}\right]^{n/2}.$$

However, note that

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2.$$

If this substitution is made in the denominator of λ, we have

$$\lambda = \left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2}\right]^{n/2}$$

$$= \left[\frac{1}{1 + \dfrac{n(\overline{x} - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right]^{n/2}.$$

Note that λ is close to 1 when $\overline{x}$ is close to μ_0 and λ is small when $\overline{x}$ and μ_0 differ by a great deal. The likelihood ratio test, given by the inequality $\lambda \le k$, is the same as

$$\frac{1}{1 + \dfrac{n(\overline{x} - \mu_0)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}} \le k^{2/n}$$

or, equivalently,

$$\frac{n(\overline{x} - \mu_0)^2}{\dfrac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2} \ge (n-1)(k^{-2/n} - 1).$$

When H_0 is true, $\sqrt{n}(\overline{X} - \mu_0)/\sigma$ is $N(0,1)$ and $\sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2$ has an independent chi-square distribution $\chi^2(n-1)$. Hence, under H_0,

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)/\sigma}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2/[\sigma^2(n-1)]}} = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)}} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

has a t distribution with $r = n - 1$ degrees of freedom. In accordance with the likelihood ratio test criterion, H_0 is rejected if the observed

$$T^2 \ge (n-1)(k^{-2/n} - 1).$$

That is, we reject H_0: $\mu = \mu_0$ and accept H_1: $\mu \ne \mu_0$ at the α level of significance if the observed $|T| \ge t_{\alpha/2}(n-1)$.

Note that this test is exactly the same as that listed in Table 8.1-2 for testing H_0: $\mu = \mu_0$ against H_1: $\mu \ne \mu_0$. That is, the test listed there is a likelihood ratio test. As a matter of fact, all six of the tests given in Tables 8.1-2 and 8.2-1 are likelihood ratio tests. Thus, the examples and exercises associated with those tables are illustrations of the use of such tests.

The final development of this section concerns a test about the variance of a normal population. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. We wish to test H_0: $\sigma^2 = \sigma_0^2$ against H_1: $\sigma^2 \ne \sigma_0^2$. For this purpose, we have

$$\omega = \{(\mu, \sigma^2) \colon -\infty < \mu < \infty,\ \sigma^2 = \sigma_0^2\}$$

and

$$\Omega = \{(\mu, \sigma^2) \colon -\infty < \mu < \infty,\ 0 < \sigma^2 < \infty\}.$$

As in the test concerning the mean, we obtain

$$L(\widehat{\Omega}) = \left[\frac{ne^{-1}}{2\pi \sum_{i=1}^{n} (x_i - \overline{x})^2}\right]^{n/2}.$$

If $(\mu, \sigma^2) \in \omega$, then $\widehat{\mu} = \overline{x}$ and $\widehat{\sigma^2} = \sigma_0^2$; thus,

$$L(\widehat{\omega}) = \left(\frac{1}{2\pi\sigma_0^2}\right)^{n/2} \exp\left[-\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{2\sigma_0^2}\right].$$

Accordingly, the likelihood ratio test $\lambda = L(\widehat{\omega})/L(\widehat{\Omega})$ is

$$\lambda = \left(\frac{w}{n}\right)^{n/2} \exp\left(-\frac{w}{2} + \frac{n}{2}\right) \le k,$$

where $w = \sum_{i=1}^{n}(x_i - \overline{x})^2/\sigma_0^2$. Solving this inequality for w, we obtain a solution of the form $w \le c_1$ or $w \ge c_2$, where the constants c_1 and c_2 are appropriate functions of the constants k and n so as to achieve the desired significance level α. However these values of c_1 and c_2 do not place probability $\alpha/2$ in each of the two regions $w \le c_1$ and $w \ge c_2$. Since $W = \sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma_0^2$ is $\chi^2(n-1)$ if H_0: $\sigma^2 = \sigma_0^2$ is true, most statisticians modify this test slightly by taking $c_1 = \chi^2_{1-\alpha/2}(n-1)$ and $c_2 = \chi^2_{\alpha/2}(n-1)$. As a matter of fact, most tests involving normal assumptions are likelihood ratio tests or modifications of them; included are tests involving regression and analysis of variance (see Chapter 9).

REMARK Note that likelihood ratio tests are based on sufficient statistics when they exist, as was also true of best critical and uniformly most powerful critical regions.

Exercises

8.7-1. In Example 8.7-1, suppose that $n = 20$ and $\overline{x} = 161.1$.

(a) Is H_0 accepted if $\alpha = 0.10$?

(b) Is H_0 accepted if $\alpha = 0.05$?

(c) What is the p-value of this test?

8.7-2. Assume that the weight X in ounces of a "10-ounce" box of cornflakes is $N(\mu, 0.03)$. Let $X_1, X_2, \ldots, X_n$ be a random sample from this distribution.

(a) To test the hypothesis H_0: $\mu \ge 10.35$ against the alternative hypothesis H_1: $\mu < 10.35$, what is the critical region of size $\alpha = 0.05$ specified by the likelihood ratio test criterion? HINT: Note that if $\mu \ge 10.35$ and $\overline{x} < 10.35$, then $\widehat{\mu} = 10.35$.

(b) If a random sample of $n = 50$ boxes yielded a sample mean of $\overline{x} = 10.31$, is H_0 rejected? HINT: Find the critical value z_α when H_0 is true by taking $\mu = 10.35$, which is the extreme value in $\mu \ge 10.35$.

(c) What is the p-value of this test?

8.7-3. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the normal distribution $N(\mu, 100)$.

(a) To test H_0: $\mu = 230$ against H_1: $\mu > 230$, what is the critical region specified by the likelihood ratio test criterion?

(b) Is this test uniformly most powerful?

(c) If a random sample of $n = 16$ yielded $\overline{x} = 232.6$, is H_0 accepted at a significance level of $\alpha = 0.10$?

(d) What is the p-value of this test?

8.7-4. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the normal distribution $N(\mu, \sigma_0^2)$, where σ_0^2 is known but μ is unknown.

(a) Find the likelihood ratio test for H_0: $\mu = \mu_0$ against H_1: $\mu \ne \mu_0$. Show that this critical region for a test with significance level α is given by $|\overline{X} - \mu_0| > z_{\alpha/2}\sigma_0/\sqrt{n}$.

(b) Test H_0: $\mu = 59$ against H_1: $\mu \ne 59$ when $\sigma^2 = 225$ and a sample of size $n = 100$ yielded $\overline{x} = 56.13$. Let $\alpha = 0.05$.

(c) What is the p-value of this test? Note that H_1 is a two-sided alternative.

8.7-5. It is desired to test the hypothesis H_0: $\mu = 30$ against the alternative hypothesis H_1: $\mu \ne 30$, where μ is the mean of a normal distribution and σ^2 is unknown. If a random sample of size $n = 9$ has $\overline{x} = 32.8$ and $s = 4$, is H_0 accepted at an $\alpha = 0.05$ significance level? What is the approximate p-value of this test?

8.7-6. To test H_0: $\mu = 335$ against H_1: $\mu < 335$ under normal assumptions, a random sample of size 17 yielded $\overline{x} = 324.8$ and $s = 40$. Is H_0 accepted at an $\alpha = 0.10$ significance level?

8.7-7. Let X have a normal distribution in which μ and σ^2 are both unknown. It is desired to test H_0: $\mu = 1.80$ against H_1: $\mu > 1.80$ at an $\alpha = 0.10$ significance level. If a random sample of size $n = 121$ yielded $\overline{x} = 1.84$ and $s = 0.20$, is H_0 accepted or rejected? What is the p-value of this test?

8.7-8. Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with mean θ. Show that the likelihood ratio test of H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$ has a critical region of the form $\sum_{i=1}^{n} x_i \leq c_1$ or $\sum_{i=1}^{n} x_i \geq c_2$. How would you modify this test so that chi-square tables can be used easily?

8.7-9. Let independent random samples of sizes n and m be taken respectively from two normal distributions with unknown means μ_X and μ_Y and unknown variances σ_X^2 and σ_Y^2.

(a) Show that when $\sigma_X^2 = \sigma_Y^2$, the likelihood ratio for testing H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X \neq \mu_Y$ is a function of the usual two-sample t statistic.

(b) Show that the likelihood ratio for testing H_0: $\sigma_X^2 = \sigma_Y^2$ against H_1: $\sigma_X^2 \neq \sigma_Y^2$ is a function of the usual two-sample F statistic.

8.7-10. Referring back to Exercise 6.4-19, find the likelihood ratio test of H_0: $\gamma = 1$, μ unspecified, against all alternatives.

8.7-11. Let $Y_1, Y_2, \ldots, Y_n$ be n independent random variables with normal distributions $N(\beta x_i, \sigma^2)$, where $x_1, x_2, \ldots, x_n$ are known and not all equal and β and σ^2 are unknown parameters.

(a) Find the likelihood ratio test for H_0: $\beta = 0$ against H_1: $\beta \neq 0$.

(b) Can this test be based on a statistic with a well-known distribution?

HISTORICAL COMMENTS Most of the tests presented in this section result from the use of methods found in the theories of Jerzy Neyman and Egon Pearson, a son of Karl Pearson. Neyman and Pearson formed a team, particularly in the 1920s and 1930s, which produced theoretical results that were important contributions to the area of testing statistical hypotheses.

Neyman and Pearson knew, in testing hypotheses, that they needed a critical region that had high probability when the alternative was true, but they did not have a procedure to find the best one. Neyman was thinking about this late one day when his wife told him they had to attend a concert. He kept thinking of this problem during the concert and finally, in the middle of the concert, the solution came to him: Select points in the critical region for which the ratio of the pdf under the alternative hypothesis to that under the null hypothesis is as large as possible. Hence, the Neyman–Pearson lemma was born. Sometimes solutions occur at the strangest times.

Shortly after Wilcoxon proposed his two-sample test, Mann and Whitney suggested a test based on the estimate of the probability $P(X < Y)$. In this test, they let U equal the number of times that $X_i < Y_j$, $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_2$. Using the data in Example 8.4-6, we find that the computed U is $u = 51$ among all $n_1 n_2 = (8)(8) = 64$ pairs of (X, Y). Thus, the estimate of $P(X < Y)$ is 51/64 or, in general, $u/n_1 n_2$. At the time of the Mann–Whitney suggestion, it was noted that U was just a linear function of Wilcoxon's W and hence really provided the same test. That relationship is

$$U = W - \frac{n_2(n_2 + 1)}{2},$$

which in our special case is

$$51 = 87 - \frac{8(9)}{2} = 87 - 36.$$

Thus, we often read about the test of Mann, Whitney, and Wilcoxon. From this observation, this test could be thought of as one testing H_0: $P(X < Y) = 1/2$ against the alternative H_1: $P(X < Y) > 1/2$ with critical region of the form $w \geq c$.

Note that the two-sample Wilcoxon test is much less sensitive to extreme values than is the Student's t test based on $\overline{X} - \overline{Y}$. Therefore, if there is considerable skewness or contamination, these proposed distribution-free tests are much safer. In particular, that of Wilcoxon is quite good and does not lose too much power in case the distributions are close to normal ones. It is important to note that the one-sample Wilcoxon test requires symmetry of the underlying distribution, but the two-sample Wilcoxon test does not and thus can be used for skewed distributions.

From theoretical developments beyond the scope of this text, the two Wilcoxon tests are strong competitors of the usual one- and two-sample tests based upon normality assumptions, so the Wilcoxon tests should be considered if those assumptions are questioned.

Computer programs, including Minitab, will calculate the value of the Wilcoxon or Mann–Whitney statistic. However, it is instructive to do these tests by hand so that you can see what is being calculated!