the basis of these data that transmitting underwater predator sounds is an effective technique for clearing fishing waters of unwanted whales?

(b) Calculate the P-value for these data. For what values of α would H_0 be rejected?

6.3.2. Efforts to find a genetic explanation for why certain people are right-handed and others left-handed have been largely unsuccessful. Reliable data are difficult to find because of environmental factors that also influence a child's "handedness." To avoid that complication, researchers often study the analogous problem of "pawedness" in animals, where both genotypes and the environment can be partially controlled. In one such experiment (27), mice were put into a cage having a feeding tube that was equally accessible from the right or the left. Each mouse was then carefully watched over a number of feedings. If it used its right paw more than half the time to activate the tube, it was defined to be "right-pawed." Observations of this sort showed that 67% of mice belonging to strain A/J are right-pawed. A similar protocol was followed on a sample of thirty-five mice belonging to strain A/HeJ. Of those thirty-five, a total of eighteen were eventually classified as right-pawed. Test whether the proportion of right-pawed mice found in the A/HeJ sample was significantly different from what was known about the A/J strain. Use a two-sided alternative and let 0.05 be the probability associated with the critical region.

6.3.3. Defeated in his most recent attempt to win a congressional seat because of a sizeable gender gap, a politician has spent the last two years speaking out in favor of women's rights issues. A newly released poll claims to have contacted a random sample of 120 of the politician's current supporters and found that 72 were men. In the election that he lost, exit polls indicated that 65% of those who voted for him were men. Using an $\alpha = 0.05$ level of significance, test the null hypothesis that the proportion of his male supporters has remained the same. Make the alternative hypothesis one-sided.

6.3.4. Suppose H_0: $p = 0.45$ is to be tested against H_1: $p > 0.45$ at the $\alpha = 0.14$ level of significance, where $p = P(i\text{th}$ trial ends in success). If the sample size is 200, what is the smallest number of successes that will cause H_0 to be rejected?

6.3.5. Recall the median test described in Example 5.3.2. Reformulate that analysis as a hypothesis test rather than a confidence interval. What P-value is associated with the outcomes listed in Table 5.3.3?

6.3.6. Among the early attempts to revisit the death postponement theory introduced in Case Study 6.3.2 was an examination of the birth dates and death dates of 348 U.S. celebrities (134). It was found that 16 of those individuals had died in the month preceding their birth month. Set up and test the appropriate H_0 against a one-sided H_1. Use the 0.05 level of significance.

6.3.7. What α levels are possible with a decision rule of the form "Reject H_0 if $k \geq k^*$" when H_0: $p = 0.5$ is to be tested against H_1: $p > 0.5$ using a random sample of size $n = 7$?

6.3.8. The following is a Minitab printout of the binomial pdf $p_X(k) = \binom{9}{k}(0.6)^k(0.4)^{9-k}$, $k = 0, 1, \ldots, 9$. Suppose H_0: $p = 0.6$ is to be tested against H_1: $p > 0.6$ and we wish the level of significance to be *exactly* 0.05. Use Theorem 2.4.1 to combine two different critical regions into a single *randomized decision rule* for which $\alpha = 0.05$.

```
MTB > pdf;
SUBC > binomial 9  0.6.
Probability Density Function
Binomial with n = 9 and p = 0.6
        x    P(X = x)
        0    0.000262
        1    0.003539
        2    0.021234
        3    0.074318
        4    0.167215
        5    0.250823
        6    0.250823
        7    0.161243
        8    0.060466
        9    0.010078
```

6.3.9. Suppose H_0: $p = 0.75$ is to be tested against H_1: $p < 0.75$ using a random sample of size $n = 7$ and the decision rule "Reject H_0 if $k \leq 3$."

(a) What is the test's level of significance?

(b) Graph the probability that H_0 will be rejected *as a function of p*.

6.4 Type I and Type II Errors

The possibility of drawing incorrect conclusions is an inevitable byproduct of hypothesis testing. No matter what sort of mathematical facade is laid atop the decision-making process, there is no way to guarantee that what the test tells us is the truth. One kind of error—rejecting H_0 when H_0 is true—figured prominently in Section 6.3: It was argued that critical regions should be defined so as to keep the probability of making such errors small, often on the order of 0.05.

In point of fact, there are two different kinds of errors that can be committed with any hypothesis test: (1) We can reject H_0 when H_0 is true and (2) we can fail to reject H_0 when H_0 is false. These are called *Type I* and *Type II* errors, respectively. At the same time, there are two kinds of correct decisions: (1) We can fail to reject a true H_0 and (2) we can reject a false H_0. Figure 6.4.1 shows these four possible "Decision/State of nature" combinations.

Figure 6.4.1

		True State of Nature	
		H_0 is true	H_1 is true
Our Decision	Fail to reject H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

Computing the Probability of Committing a Type I Error

Once an inference is made, there is no way to know whether the conclusion reached was correct. It *is* possible, though, to calculate the probability of having made an error, and the magnitude of that probability can help us better understand the "power" of the hypothesis test and its ability to distinguish between H_0 and H_1.

Recall the fuel additive example developed in Section 6.2: H_0: $\mu = 25.0$ was to be tested against H_1: $\mu > 25.0$ using a sample of size $n = 30$. The decision rule stated that H_0 should be rejected if $\overline{y}$, the average mpg with the new additive, equalled or exceeded 25.718. In that case, the probability of committing a Type I error is *0.05*:

$$\begin{aligned} P(\text{Type I error}) &= P(\text{Reject } H_0 \mid H_0 \text{ is true}) \\ &= P(\overline{Y} \geq 25.718 \mid \mu = 25.0) \\ &= P\left(\frac{\overline{Y} - 25.0}{2.4/\sqrt{30}} \geq \frac{25.718 - 25.0}{2.4/\sqrt{30}}\right) \\ &= P(Z \geq 1.64) = 0.05 \end{aligned}$$

Of course, the fact that the probability of committing a Type I error equals 0.05 should come as no surprise. In our earlier discussion of how "beyond reasonable doubt" should be interpreted numerically, we specifically chose the critical region so that the probability of the decision rule rejecting H_0 when H_0 is true *would* be 0.05.

In general, the probability of committing a Type I error is referred to as a test's *level of significance* and is denoted α (recall Definition 6.2.3). The concept is a crucial one: The level of significance is a single-number summary of the "rules" by which the decision process is being conducted. In essence, α reflects the amount of evidence the experimenter is demanding to see before abandoning the null hypothesis.

Computing the Probability of Committing a Type II Error

We just saw that calculating the probability of a Type I error is a nonproblem: There are no computations necessary, since the probability equals whatever value the experimenter sets a priori for α. A similar situation does not hold for Type

II errors. To begin with, Type II error probabilities are not specified explicitly by the experimenter; also, each hypothesis test has an infinite number of Type II error probabilities, one for each value of the parameter admissible under H_1.

As an example, suppose we want to find the probability of committing a Type II error in the gasoline experiment if the true μ (*with the additive*) were 25.750. By definition,

$$\begin{aligned} P(\text{Type II error} \mid \mu = 25.750) &= P(\text{We fail to reject } H_0 \mid \mu = 25.750) \\ &= P(\overline{Y} < 25.718 \mid \mu = 25.750) \\ &= P\left(\frac{\overline{Y} - 25.75}{2.4/\sqrt{30}} < \frac{25.718 - 25.75}{2.4/\sqrt{30}}\right) \\ &= P(Z < -0.07) = 0.4721 \end{aligned}$$

So, even if the new additive increased the fuel economy to 25.750 mpg (from 25 mpg), our decision rule would be "tricked" 47% of the time: that is, it would tell us on those occasions *not* to reject H_0.

The symbol for the probability of committing a Type II error is β. Figure 6.4.2 shows the sampling distribution of $\overline{Y}$ when $\mu = 25.0$ (i.e., when H_0 is true) and when $\mu = 25.750$ (H_1 is true); the areas corresponding to α and β are shaded.

Figure 6.4.2

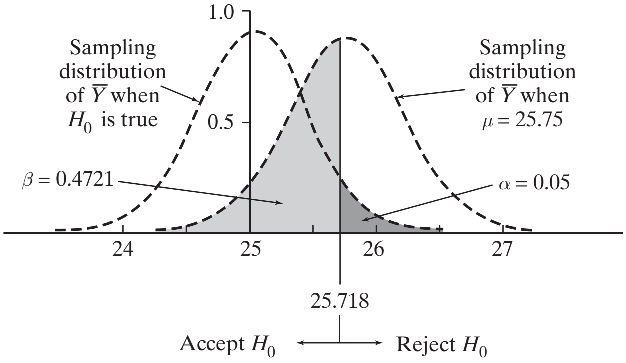
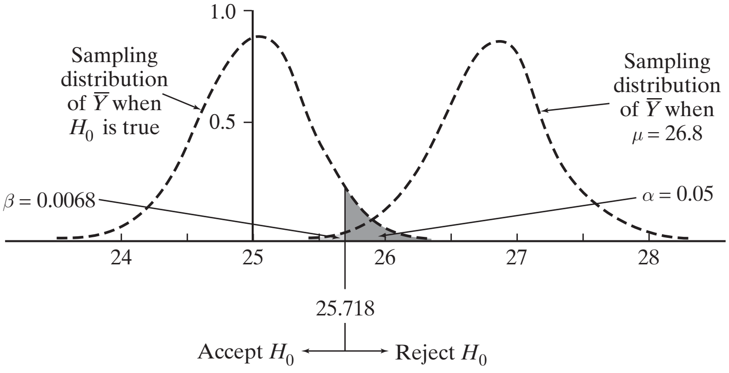


Figure 6.4.3



Clearly, the magnitude of β is a function of the presumed value for μ. If, for example, the gasoline additive is so effective as to raise fuel efficiency to 26.8 mpg, the probability that our decision rule would lead us to make a Type II error is a much smaller *0.0068*:

$$\begin{aligned} P(\text{Type II error} \mid \mu = 26.8) &= P(\text{We fail to reject } H_0 \mid \mu = 26.8) \\ &= P(\overline{Y} < 25.718 \mid \mu = 26.8) = P\left(\frac{\overline{Y} - 26.8}{2.4/\sqrt{30}} < \frac{25.718 - 26.8}{2.4/\sqrt{30}}\right) \\ &= P(Z < -2.47) = 0.0068 \end{aligned}$$

(See Figure 6.4.3.)

Power Curves

If β is the probability that we fail to reject H_0 when H_1 is true, then $1 - \beta$ is the probability of the complement—that we *reject* H_0 when H_1 is true. We call $1 - \beta$ the *power* of the test; it represents the ability of the decision rule to "recognize" (correctly) that H_0 is false.

The alternative hypothesis H_1 usually depends on a parameter, which makes $1 - \beta$ a function of that parameter. The relationship they share can be pictured by drawing a *power curve*, which is simply a graph of $1 - \beta$ versus the set of all possible parameter values.

Figure 6.4.4 shows the power curve for testing

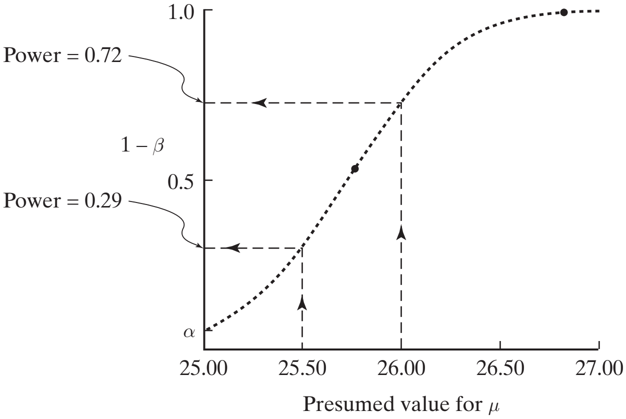
$$H_0\colon \mu = 25.0$$

$$\text{versus}$$

$$H_1\colon \mu > 25.0$$

where μ is the mean of a normal distribution with $\sigma = 2.4$, and the decision rule is "Reject H_0 if $\overline{y} \geq 25.718$." The two marked points on the curve represent the $(\mu, 1 - \beta)$ pairs just determined, (25.75, 0.5297) and (26.8, 0.9932). One other point can be gotten for every power curve, without doing any calculations: When $\mu = \mu_0$ (the value specified by H_0), $1 - \beta = \alpha$. Of course, as the true mean gets further and further away from the H_0 mean, the power will converge to 1.

Figure 6.4.4



Power curves serve two different purposes. On the one hand, they completely characterize the performance that can be expected from a hypothesis test. In

Figure 6.4.5

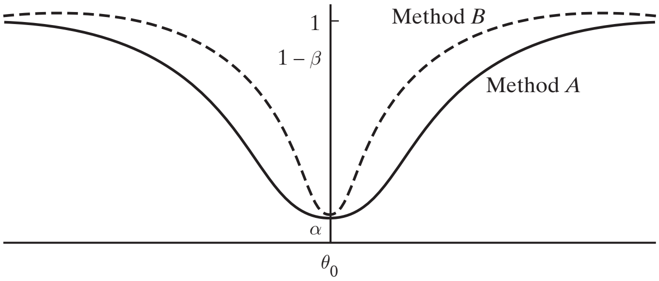


Figure 6.4.4, for example, the two arrows show that the probability of rejecting $H_0 : \mu = 25$ in favor of $H_1 : \mu > 25$ when $\mu = 26.0$ is approximately *0.72*. (Or, equivalently, Type II errors will be committed roughly *28%* of the time when $\mu = 26.0$.) As the true mean moves closer to μ_o (and becomes more difficult to distinguish) the power of the test understandably diminishes. If $\mu = 25.5$, for example, the graph shows that $1 - \beta$ falls to 0.29.

Power curves are also useful for *comparing* one inference procedure with another. For every conceivable hypothesis testing situation, a variety of procedures for choosing between H_0 and H_1 will be available. How do we know which to use?

The answer to that question is not always simple. Some procedures will be computationally more convenient or easier to explain than others; some will make slightly different assumptions about the pdf being sampled. Associated with each of them, though, is a power curve. If the selection of a hypothesis test is to hinge solely on its ability to distinguish H_0 from H_1, then the procedure to choose is the one having the *steepest* power curve.

Figure 6.4.5 shows the power curves for two hypothetical methods A and B, each of which is testing $H_0: \theta = \theta_o$ versus $H_1: \theta \neq \theta_o$ at the α level of significance. From the standpoint of power, Method B is clearly the better of the two—it always has a higher probability of correctly rejecting H_0 when the parameter θ is not equal to θ_o.

Factors That Influence the Power of a Test

The ability of a test procedure to reject H_0 when H_0 is false is clearly of prime importance, a fact that raises an obvious question: What can an experimenter do to influence the value of $1 - \beta$? In the case of the Z test described in Theorem 6.2.1, $1 - \beta$ is a function of α, σ, and n. By appropriately raising or lowering the values of those parameters, the power of the test against any given μ can be made to equal any desired level.

The Effect of α on $1 - \beta$

Consider again the test of

$$H_0: \mu = 25.0$$

$$\text{versus}$$

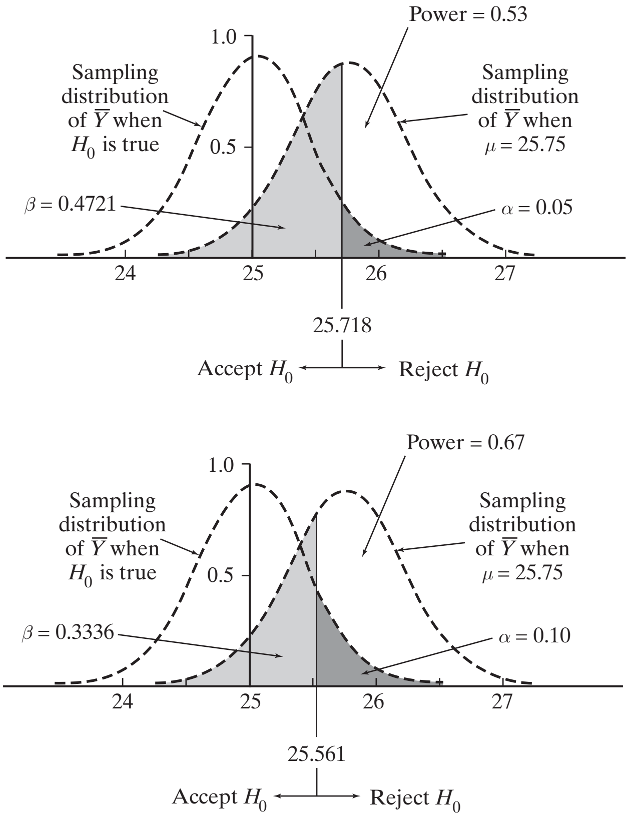
$$H_1: \mu > 25.0$$

discussed earlier in this section. In its original form, $\alpha = 0.05$, $\sigma = 2.4$, $n = 30$, and the decision rule called for H_0 to be rejected if $\overline{y} \geq 25.718$.

Figure 6.4.6 shows what happens to $1 - \beta$ (when $\mu = 25.75$) if σ, n, and μ are held constant but α is increased to 0.10. The top pair of distributions shows the configuration that appears in Figure 6.4.2; the power in this case is $1 - 0.4721$, or *0.53*. The bottom portion of the graph illustrates what happens when α is set at 0.10 instead of 0.05—the decision rule changes from "Reject H_0 if $\overline{y} \geq 25.718$" to "Reject H_0 if $\overline{y} \geq 25.561$" (see Question 6.4.2) and the power increases from 0.53 to *0.67*:

$$\begin{aligned}
1 - \beta &= P(\text{Reject } H_0 \mid H_1 \text{ is true}) \\
&= P(\overline{Y} \geq 25.561 \mid \mu = 25.75) \\
&= P\left(\frac{\overline{Y} - 25.75}{2.4/\sqrt{30}} \geq \frac{25.561 - 25.75}{2.4/\sqrt{30}}\right) \\
&= P(Z \geq -0.43) \\
&= 0.6664
\end{aligned}$$

Figure 6.4.6



The specifics of Figure 6.4.6 accurately reflect what is true in general: *Increasing α decreases β and increases the power*. That said, it does not follow *in practice* that experimenters should manipulate α to achieve a desired $1 - \beta$. For all the reasons cited in Section 6.2, α should typically be set equal to a number somewhere in the neighborhood of 0.05. If the corresponding $1 - \beta$ against a particular μ is deemed to be inappropriate, adjustments should be made in the values of σ and/or n.

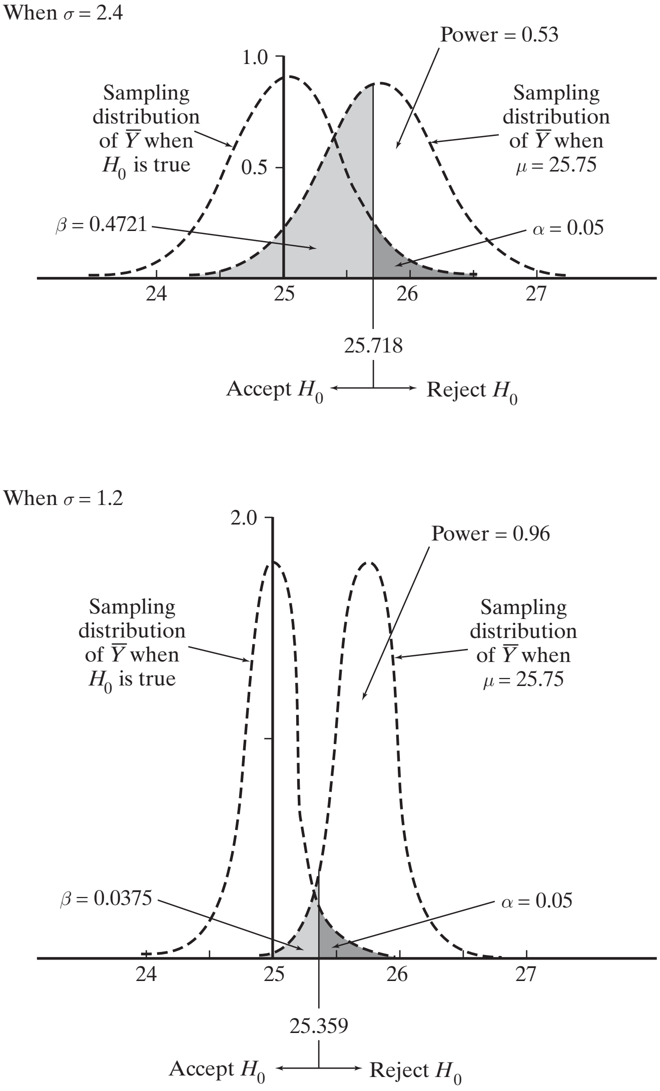
The Effects of σ and n on $1-\beta$

Although it may not always be feasible (or even possible), *decreasing* σ will necessarily *increase* $1-\beta$. In the gasoline additive example, σ is assumed to be 2.4 mpg, the latter being a measure of the variation in gas mileages from driver to driver achieved in a cross-country road trip from Boston to Los Angeles (recall p. 351). Intuitively, the environmental differences inherent in a trip of that magnitude would be considerable. Different drivers would encounter different weather conditions and varying amounts of traffic, and would perhaps take alternate routes.

Suppose, instead, that the drivers simply did laps around a test track rather than drive on actual highways. Conditions from driver to driver would then be much more uniform and the value of σ would surely be smaller. What would be the effect on $1-\beta$ when $\mu=25.75$ (and $\alpha=0.05$) if σ could be reduced from 2.4 mpg to 1.2 mpg?

As Figure 6.4.7 shows, reducing σ has the effect of making the H_0 distribution of $\overline{Y}$ more concentrated around $\mu_o(=25)$ and the H_1 distribution of $\overline{Y}$ more concentrated around $\mu(=25.75)$. Substituting into Equation 6.2.1 (with 1.2 for σ in place

Figure 6.4.7



of 2.4), we find that the critical value $\bar{y}^*$ moves closer to μ_o [from 25.718 to *25.359* $\left(=25+1.64\cdot\frac{1.2}{\sqrt{30}}\right)$] and the proportion of the H_1 distribution above the rejection region (i.e., the power) *increases* from 0.53 to *0.96*:

$$\begin{aligned}1-\beta &= P(\overline{Y}\geq 25.359\mid\mu=25.75)\\ &= P\left(Z\geq\frac{25.359-25.75}{1.2/\sqrt{30}}\right)=P(Z\geq -1.78)=0.9625\end{aligned}$$

In theory, reducing σ can be a very effective way of increasing the power of a test, as Figure 6.4.7 makes abundantly clear. In practice, though, refinements in the way data are collected that would have a substantial impact on the magnitude of σ are often either difficult to identify or prohibitively expensive. More typically, experimenters achieve the same effect by simply increasing the sample size.

Look again at the two sets of distributions in Figure 6.4.7. The increase in $1-\beta$ from 0.53 to 0.96 was accomplished by cutting the denominator of the test statistic $\left(z=\frac{\bar{y}-25}{\sigma/\sqrt{30}}\right)$ in half by reducing the standard deviation from 2.4 to 1.2. The same numerical effect would be produced if σ were left unchanged but n was increased from 30 to 120—that is, $\frac{1.2}{\sqrt{30}}=\frac{2.4}{\sqrt{120}}$. Because it can easily be increased or decreased, the sample size is the parameter that researchers almost invariably turn to as the mechanism for ensuring that a hypothesis test will have a sufficiently high power against a given alternative.

Example 6.4.1

Suppose an experimenter wishes to test

$$H_0\colon \mu=100$$

versus

$$H_1\colon \mu>100$$

at the $\alpha=0.05$ level of significance and wants $1-\beta$ to equal 0.60 when $\mu=103$. What is the smallest (i.e., cheapest) sample size that will achieve that objective? Assume that the variable being measured is normally distributed with $\sigma=14$.

Finding n, given values for $\alpha, 1-\beta, \sigma$, and μ, requires that two simultaneous equations be written for the critical value $\bar{y}^*$, one in terms of the H_0 distribution and the other in terms of the H_1 distribution. Setting the two equal will yield the minimum sample size that achieves the desired α and $1-\beta$.

Consider, first, the consequences of the level of significance being equal to 0.05. By definition,

$$\begin{aligned}\alpha &= P(\text{We reject } H_0\mid H_0 \text{ is true})\\ &= P(\overline{Y}\geq\bar{y}^*\mid\mu=100)\\ &= P\left(\frac{\overline{Y}-100}{14/\sqrt{n}}\geq\frac{\bar{y}^*-100}{14/\sqrt{n}}\right)\\ &= P\left(Z\geq\frac{\bar{y}^*-100}{14/\sqrt{n}}\right)\\ &= 0.05\end{aligned}$$

But $P(Z\geq 1.64)=0.05$, so

$$\frac{\bar{y}^*-100}{14/\sqrt{n}}=1.64$$

or, equivalently,

$$\bar{y}^* = 100 + 1.64 \cdot \frac{14}{\sqrt{n}} \tag{6.4.1}$$

Similarly,

$$\begin{aligned} 1 - \beta &= P(\text{We reject } H_0 \mid H_1 \text{ is true}) = P(\bar{Y} \geq \bar{y}^* \mid \mu = 103) \\ &= P\left(\frac{\bar{Y} - 103}{14/\sqrt{n}} \geq \frac{\bar{y}^* - 103}{14/\sqrt{n}}\right) = 0.60 \end{aligned}$$

From Appendix Table A.1, though, $P(Z \geq -0.25) = 0.5987 \doteq 0.60$, so

$$\frac{\bar{y}^* - 103}{14/\sqrt{n}} = -0.25$$

which implies that

$$\bar{y}^* = 103 - 0.25 \cdot \frac{14}{\sqrt{n}} \tag{6.4.2}$$

It follows, then, from Equations 6.4.1 and 6.4.2 that

$$100 + 1.64 \cdot \frac{14}{\sqrt{n}} = 103 - 0.25 \cdot \frac{14}{\sqrt{n}}$$

Solving for n shows that a minimum of *seventy-eight* observations must be taken to guarantee that the hypothesis test will have the desired precision. ■

Decision Rules for Nonnormal Data

Our discussion of hypothesis testing thus far has been confined to inferences involving either binomial data or normal data. Decision rules for other types of probability functions are rooted in the same basic principles.

In general, to test $H_0: \theta = \theta_o$, where θ is the unknown parameter in a pdf $f_Y(y; \theta)$, we initially define the decision rule in terms of $\hat{\theta}$, where the latter is a sufficient statistic for θ. The corresponding critical region is the set of values of $\hat{\theta}$ least compatible with θ_o (but admissible under H_1) whose total probability when H_0 is true is α. In the case of testing $H_0: \mu = \mu_o$ versus $H_1: \mu > \mu_o$, for example, where the data are normally distributed, $\bar{Y}$ is a sufficient statistic for μ, and the least likely values for the sample mean that are admissible under H_1 are those for which $\bar{y} \geq \bar{y}^*$, where $P(\bar{Y} \geq \bar{y}^* \mid H_0 \text{ is true}) = \alpha$.

Example 6.4.2

A random sample of size $n = 8$ is drawn from the uniform pdf, $f_Y(y; \theta) = 1/\theta, 0 \leq y \leq \theta$, for the purpose of testing

$$H_0: \theta = 2.0$$

versus

$$H_1: \theta < 2.0$$

at the $\alpha = 0.10$ level of significance. Suppose the decision rule is to be based on Y_8', the largest order statistic. What would be the probability of committing a Type II error when $\theta = 1.7$?