8.7 THE POWER OF A TEST

In the initial discussion of hypothesis testing, the two types of risks that are taken when decisions are made about population parameters based only on sample evidence were defined. Recall from section 8.1 that α represents the probability that the null hypothesis is rejected when in fact it is true and should not be rejected, and β represents the probability that the null hypothesis is not rejected when in fact it is false and should be rejected. The power of the test, which is $1 - \beta$ (that is, the complement of β), indicates the sensitivity of the statistical test in detecting changes that have occurred by measuring the probability of rejecting the null hypothesis when in fact it is false and should be rejected. The power of the statistical test depends on how different the actual population mean really is from the value being hypothesized (under H_0), the value of α used, and the sample size. If there is a large difference between the actual population mean and the hypothesized mean, the power of the test will be much greater than if the difference between the actual population mean and the hypothesized mean is small. Selecting a larger value of α makes it easier to reject H_0 and therefore increases the power of a test. Increasing the sample size increases the precision in the estimate and therefore increases the ability to detect a difference from in the hypothesized parameters and increases the power of a test.

In this section, the cereal-box-filling process is examined in order to further develop the concept of the power of a statistical test. Suppose that the filling process is subject to periodic inspection from a representative of the local office of consumer affairs. It is this representative's job to detect the possible "short weighting" of boxes, a situation in which cereal boxes are sold at less than the specified 368 grams. Thus, the representative is interested in determining whether there is evidence that the cereal boxes have an average amount that is less than 368 grams. The null and alternative hypotheses are set up as follows:

$$H_0: \mu \geq 368 \text{ (filling process is working properly)}$$

$$H_1: \mu < 368 \text{ (filling process is not working properly)}$$

The representative of the office of consumer affairs is willing to accept the company's claim that the standard deviation σ over the entire packaging process is equal to 15 grams. Therefore, the Z test is appropriate. If the level of significance α of 0.05 is selected and a random sample of 25 boxes is obtained, the value of $\bar{X}$ that enables you to reject the null hypothesis is found from Equation (8.1) as follows with $\bar{X}_L$ used in place of $\bar{X}$:

$$Z = \frac{\bar{X}_L - \mu}{\dfrac{\sigma}{\sqrt{n}}}$$

$$Z\frac{\sigma}{\sqrt{n}} = \bar{X}_L - \mu$$

$$\bar{X}_L = \mu + Z\frac{\sigma}{\sqrt{n}}$$

Because this is a one-tail test with a level of significance of 0.05, the value of Z equal to 1.645 standard deviations below the hypothesized mean is obtained from Table E.2 (see Figure 8.16). Therefore, $Z = -1.645$

$$\bar{X}_L = 368 + (-1.645)\frac{(15)}{\sqrt{25}} = 368 - 4.935 = 363.065$$

The decision rule for this one-tail test is

Reject H_0 if $\bar{X} < 363.065$;
otherwise do not reject H_0.

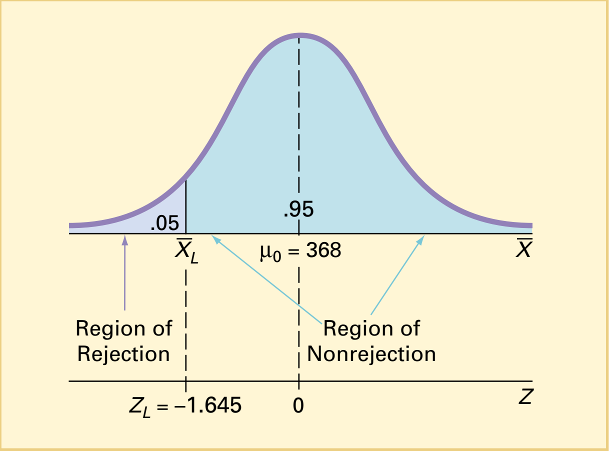


FIGURE 8.16
Determining the lower critical value for a one-tail Z test for a population mean at the 0.05 level of significance

The decision rule states that if a random sample of 25 boxes reveals a sample mean of less than 363.065 grams, the null hypothesis is rejected and the representative concludes that the process is not working properly. The power of the test measures the probability of concluding that the process is not working properly for differing values of the true population mean.

Suppose that you want to determine the chance of rejecting the null hypothesis when the population mean is actually 360 grams. On the basis of the decision rule, the probability or area under the normal curve below 363.065 grams needs to be determined. From the central limit theorem and the assumption of normality in the population, you can assume that the sampling distribution of the mean follows a normal distribution. Therefore, the area under the normal curve below 363.065 grams can be expressed in standard deviation units, because you are finding the probability of rejecting the null hypothesis when the true population mean has shifted to 360 grams. Using Equation (8.1),

$$Z = \frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}}$$

where μ_1 is the actual population mean. Thus,

$$Z = \frac{363.065 - 360}{\frac{15}{\sqrt{25}}} = 1.02$$

From Table E.2, there is an 84.61% chance of observing a Z value less than +1.02. This is the power of the test or area below 363.065 (see Figure 8.17). The probability (β) that the null hypothesis (μ = 368) will not be rejected is 1 – 0.8461 = 0.1539. Thus, the probability of committing a Type II error is 15.39%.

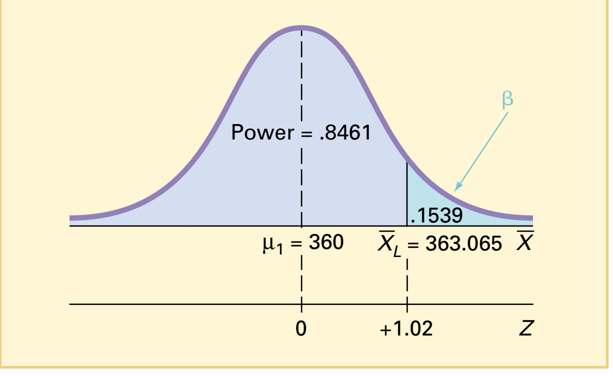


FIGURE 8.17
Determining the power of the test and the probability of a Type II error when μ_1 = 360 grams

Now that you have determined the power of the test if the population mean were really equal to 360, the power for any other value that μ could attain can be calculated. For example, what is the power of the test if the population mean is equal to 352 grams? Assuming the same standard deviation, sample size, and level of significance, the decision rule is

$$\text{Reject } H_0 \text{ if } \bar{X} < 363.065;$$

$$\text{otherwise do not reject } H_0.$$

Once again, because you are testing a hypothesis for a mean, from Equation (8.1)

$$Z = \frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}}$$

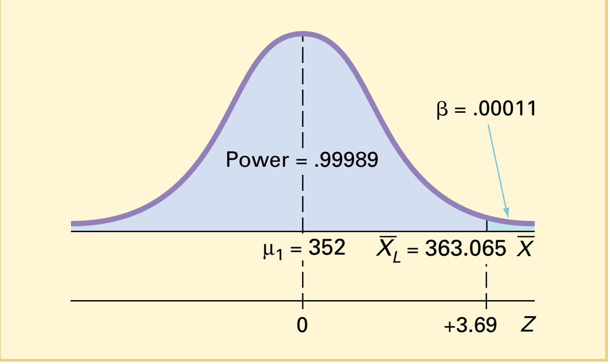
If the population mean shifts down to 352 grams (see Figure 8.18), then

$$Z = \frac{363.065 - 352}{\frac{15}{\sqrt{25}}} = 3.69$$

From Table E.2, there is a 99.989% chance of observing a Z value less than +3.69. This is the power of the test or area below 363.065. The probability (β) that the null hypothesis (μ = 368) will not be rejected is 1 – 0.99989 = 0.00011. Thus, the probability of committing a Type II error is only 0.011%.

FIGURE 8.18

Determining the power of the test and the probability of a Type II error when $\mu_1 = 352$ grams



In the preceding two cases the power of the test was quite high, whereas, conversely, the chance of committing a Type II error was quite low. In the next example, the power of the test is computed for the case in which the population mean is equal to 367 grams—a value that is very close to the hypothesized mean of 368 grams.

Once again, from Equation (8.1),

$$Z = \frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}}$$

If the population mean is equal to 367 grams (see Figure 8.19), then

$$Z = \frac{363.065 - 367}{\frac{15}{\sqrt{25}}} = -1.31$$

From Table E.2, observe that the probability (area under the curve) less than $Z = -1.31$ is 0.0951 (or 9.51%). Because in this instance the rejection region is in the lower tail of the distribution, the power of the test is 9.51% and the chance of making a Type II error is 90.49%.

Figure 8.20 illustrates the power of the test for various possible values of μ_1 (including the three cases examined). This is called a **power curve**. The computations for the three cases are summarized in Figure 8.21.

FIGURE 8.19

Determining the power of the test and the probability of a Type II error when $\mu_1 = 367$ grams

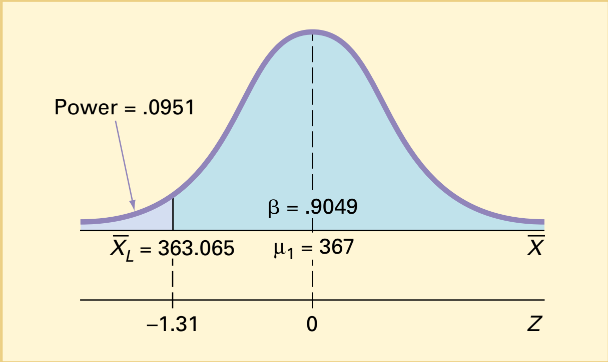
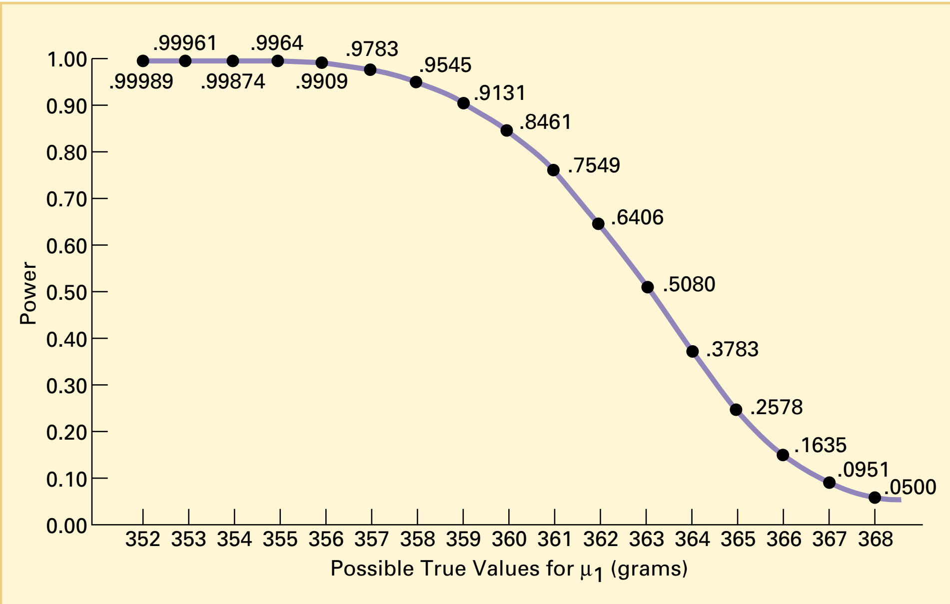


FIGURE 8.20

Power curve of the cereal-box-filling process for the alternative hypothesis H_1: $\mu < 368$ grams



From Figure 8.21, observe that the power of this one-tail test increases sharply (and approaches 100%) as the actual population mean takes on values farther below the hypothesized mean of 368 grams. Clearly, for this one-tail test, the smaller the actual mean μ_1 is when compared with the hypothesized mean, the greater will be the power to detect this disparity.* On the other hand, for values of μ_1 close to 368 grams, the power is rather small because the test cannot effectively detect small differences between the actual population mean and the hypothesized value of 368 grams.

You can observe the drastic changes in the power of the test for differing values of the actual population means by reviewing the different panels of Figure 8.21. From panels A and B you can see that when the population mean does not greatly differ from 368 grams, the chance of rejecting the null hypothesis, based on the decision rule involved, is not large. However, once the actual population mean shifts substantially below the hypothesized 368 grams, the power of the test greatly increases, approaching its maximum value of 1 (or 100%).

In the discussion of the power of a statistical test, a one-tail test, a level of significance of 0.05, and a sample size of 25 boxes have been used. With this in mind, you can determine the effect on the power of the test by varying, one at a time,

- the type of statistical test—one-tail versus two-tail.
- the level of significance α.
- the sample size n.

While these exercises are left to the reader (see problems 8.95–8.101), three basic conclusions regarding the power of the test are summarized in Exhibit 8.6.

*For situations involving one-tail tests in which the actual mean μ_1 really exceeds the hypothesized mean, the converse would be true. The larger the actual mean μ_1 compared with the hypothesized mean, the greater is the power. On the other hand, for two-tail tests, the greater the distance between the actual mean μ_1 and the hypothesized mean, the greater the power of the test.

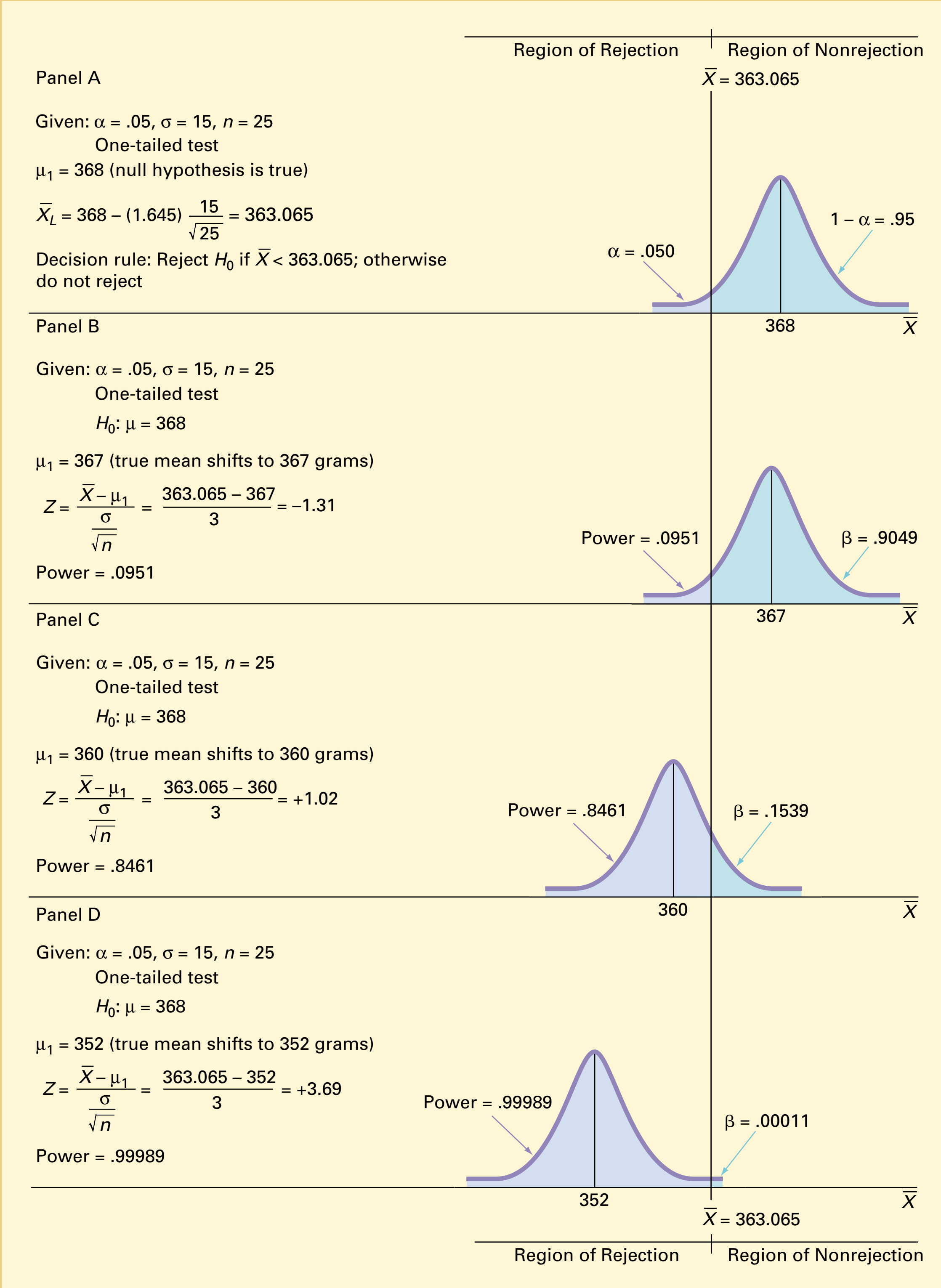


FIGURE 8.21
Determining statistical power for varying values of the actual population mean

EXHIBIT 8.6 THE POWER OF A TEST

There are three basic conclusions involved in understanding the power of the test.

1. A one-tail test is more powerful than a two-tail test and should be used whenever it is appropriate to specify the direction of the alternative hypothesis.
2. Because the probability of committing a Type I error (α) and the probability of committing a Type II error (β) have an inverse relationship and the latter is the complement of the power of the test ($1 - \beta$), then α and the power of the test vary directly. An increase in the value of the level of significance (α) results in an increase in power, and a decrease in α results in a decrease in power.
3. An increase in the size of the sample *n* chosen results in an increase in power. A decrease in the size of the sample selected results in a decrease in power.

PROBLEMS FOR SECTION 8.7

Applying the Concepts

8.95 A coin-operated soft-drink machine is designed to discharge, when it is operating properly, at least 7 ounces of beverage per cup with a standard deviation of 0.2 ounce. If a random sample of 16 cupfuls is selected by a statistician for a consumer testing service and the statistician is willing to take a risk of $\alpha = 0.05$ of committing a Type I error, compute the power of the test and the probability of a Type II error (β) if the population average amount dispensed is actually

a. 6.9 ounces per cup.
b. 6.8 ounces per cup.

8.96 Refer to problem 8.95. If the statistician is willing to take a risk of only $\alpha = 0.01$ of committing a Type I error, compute the power of the test and the probability of a Type II error (β) if the population average amount dispensed is actually

a. 6.9 ounces per cup.
b. 6.8 ounces per cup.
c. Compare the results in (a) and (b) of this problem and in problem 8.95. What conclusion can you reach?

8.97 Refer to problem 8.95. If the statistician selects a random sample of 25 cupfuls and is willing to take a risk of $\alpha = 0.05$ of committing a Type I error, compute the power of the test and the probability of a Type II error (β) if the population average amount dispensed is actually

a. 6.9 ounces per cup.
b. 6.8 ounces per cup.
c. Compare the results in (a) and (b) of this problem and in problem 8.95. What conclusion can you draw?

8.98 A tire manufacturer produces tires that last, on average, at least 25,000 miles when the production process is working properly. Based on past experience, the standard deviation of the tires is assumed to be 3,500 miles. The operations manager will stop the production process if there is evidence that the average tire life is below 25,000 miles. If a random sample of 100 tires is selected (to be subjected to *destructive testing*) and the operations manager is willing to take a risk of $\alpha = 0.05$ of committing a Type I error, compute the power of the test and the probability of a Type II error (β) if the population average life is actually

a. 24,000 miles.
b. 24,900 miles.

8.99 Refer to problem 8.98. If the operations manager is willing to take a risk of $\alpha = 0.01$ of committing a Type I error, compute the power of the test and the probability of a Type II error (β) if the population average life is actually

a. 24,000 miles.
b. 24,900 miles.
c. Compare the results in (a) and (b) of this problem and (a) and (b) in problem 8.98. What conclusion can you draw?

8.100 Refer to problem 8.98. If the operations manager selects a random sample of 25 tires and is willing to take a risk of $\alpha = 0.05$ of committing a Type I error, compute the power of the test and the probability of a Type II error (β) if the population average life is actually

a. 24,000 miles.
b. 24,900 miles.
c. Compare the results in (a) and (b) of this problem and (a) and (b) in problem 8.98. What conclusion can you draw?

8.101 Refer to problem 8.98. If the operations manager will stop the process when there is evidence that the average life is different from 25,000 miles (either less than or greater than) and a random sample of 100 tires is selected along with a level of significance of $\alpha = 0.05$, compute the power of the test and the probability of a Type II error (β) if the population average life is actually

a. 24,000 miles.
b. 24,900 miles.
c. Compare the results in (a) and (b) of this problem and (a) and (b) in problem 8.98. What conclusion can you draw?