6.61 Refer to the previous problem. A researcher has performed 12 tests of significance and wants to apply the Bonferroni procedure with $\alpha = 0.05$. The calculated P-values are 0.041, 0.569, 0.050, 0.416, 0.001, 0.004, 0.256, 0.041 0.888, 0.010, 0.002, 0.433. Which of the null hypotheses are rejected with this procedure?

6.62 The text cites an example in which researchers carried out 77 separate significance tests, of which 2 were significant at the 5% level. Suppose that these tests are independent of each other. (In fact they were not independent, because all involved the same subjects.) If all of the null hypotheses are true, each test has probability 0.05 of being significant at the 5% level.

(a) What is the distribution of the number X of tests that are significant?

(b) Find the probability that 2 or more of the tests are significant.

6.4 Power and Inference as a Decision*

Although we prefer to use P-values rather than the reject-or-not view of the fixed α significance test, the latter view is very important for planning studies and for understanding statistical decision theory. We will discuss these two topics in this section.

Power

In examining the usefulness of a confidence interval, we are concerned with both the level of confidence and the margin of error. The confidence level tells us how reliable the method is in repeated use. The margin of error tells us how sensitive the method is, that is, how closely the interval pins down the parameter being estimated. Fixed level α significance tests are closely related to confidence intervals—in fact, we saw that a two-sided test can be carried out directly from a confidence interval. The significance level, like the confidence level, says how reliable the method is in repeated use. If we use 5% significance tests repeatedly when H_0 is in fact true, we will be wrong (the test will reject H_0) 5% of the time and right (the test will fail to reject H_0) 95% of the time.

High confidence is of little value if the interval is so wide that few values of the parameter are excluded. Similarly, a test with a small level of α is of little value if it almost never rejects H_0 even when the true parameter value is far from the hypothesized value. We must be concerned with the ability of a test to detect that H_0 is false, just as we are concerned with the margin of error of a confidence interval. This ability is measured by the probability that the test will reject H_0 when an alternative is true. The higher this probability is, the more sensitive the test is.

*Although the topics in this section are important in planning and interpreting significance tests, they can be omitted without loss of continuity.

Power

The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test against that alternative.

EXAMPLE 6.17 Can a 6-month exercise program increase the total body bone mineral content (TBBMC) of young women? A team of researchers is planning a study to examine this question. Based on the results of a previous study, they are willing to assume that $\sigma = 2$ for the percent change in TBBMC over the 6-month period. A change in TBBMC of 1% would be considered important, and the researchers would like to have a reasonable chance of detecting a change this large or larger. Is 25 subjects a large enough sample for this project?

We will answer this question by calculating the power of the significance test that will be used to evaluate the data to be collected. The calculation consists of three steps:

1. State H_0, H_a, the particular alternative we want to detect, and the significance level α.
2. Find the values of $\bar{x}$ that will lead us to reject H_0.
3. Calculate the probability of observing these values of $\bar{x}$ when the alternative is true.

Step 1 The null hypothesis is that the exercise program has no effect on TBBMC. In other words, the mean percent change is zero. The alternative is that exercise is beneficial; that is, the mean change is positive. Formally, we have

$$H_0: \mu = 0$$
$$H_a: \mu > 0$$

The alternative of interest is $\mu = 1\%$. A 5% test of significance will be used.

Step 2 The z test rejects H_0 at the $\alpha = 0.05$ level whenever

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0}{2/\sqrt{25}} \geq 1.645$$

Be sure you understand why we use 1.645. Rewrite this in terms of $\bar{x}$:

$$\bar{x} \geq 1.645\frac{2}{\sqrt{25}}$$
$$\bar{x} \geq 0.658$$

Because the significance level is $\alpha = 0.05$, this event has probability 0.05 of occurring *when the population mean* μ *is* 0.

Step 3 The power against the alternative $\mu = 1\%$ increase in TBBMC is the probability that H_0 will be rejected *when in fact* $\mu = 1$. We calculate this probability by standardizing $\bar{x}$, using the value $\mu = 1$, the population standard deviation $\sigma = 2$, and the sample size $n = 25$. The power is

$$P(\bar{x} \geq 0.658 \text{ when } \mu = 1) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.658 - 1}{2/\sqrt{25}}\right)$$
$$= P(Z \geq -0.855) = 0.80$$

Figure 6.13 illustrates the power with the sampling distribution of $\bar{x}$ when $\mu = 1$. This significance test rejects the null hypothesis that exercise has no effect on TBBMC 80% of the time if the true effect of exercise is a 1% increase in TBBMC. If the true effect of exercise is a greater percent increase, the test will have greater power; it will reject with a higher probability.

High power is desirable. Along with 95% confidence intervals and 5% significance tests, 80% power is becoming a standard. Many U.S. government agencies that provide research funds require that the sample size for the funded studies be sufficient to detect important results 80% of the time using a 5% test of significance.

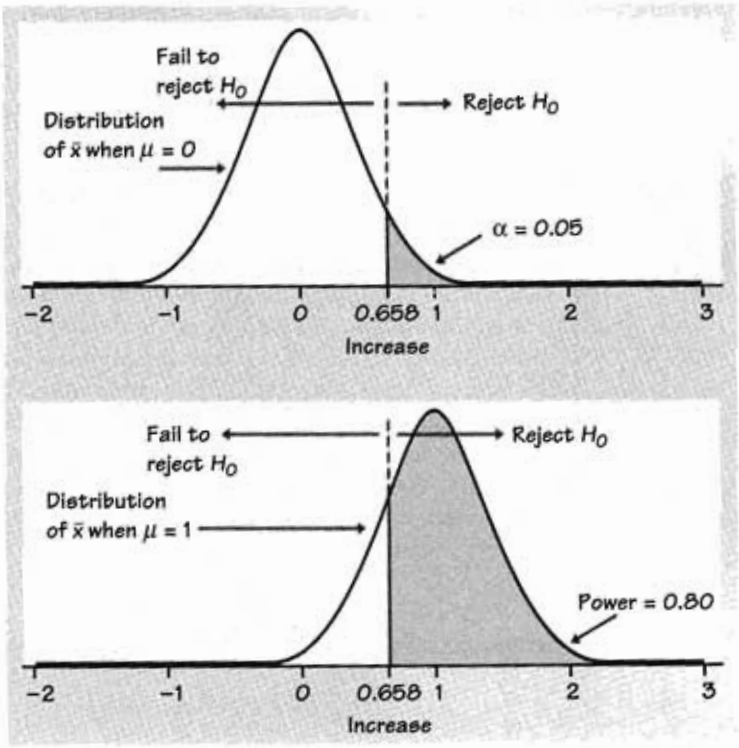

FIGURE 6.13 The sampling distributions of $\bar{x}$ when $\mu = 0$ and when $\mu = 1$ with the α and the power. Power is the probability that the test rejects H_0 when the alternative is true.

Increasing the power

Suppose you have performed a power calculation and found that the power is too small. What can you do to increase it? Here are four ways:

- Increase α. A 5% test of significance will have a greater chance of rejecting the alternative than a 1% test because the strength of evidence required for rejection is less.
- Consider a particular alternative that is farther away from μ_0. Values of μ that are in H_a but lie close to the hypothesized value μ_0 are harder to detect (lower power) than values of μ that are far from μ_0.
- Increase the sample size. More data will provide more information about $\bar{x}$ so we have a better chance of distinguishing values of μ.
- Decrease σ. This has the same effect as increasing the sample size: more information about μ. Improving the measurement process and restricting attention to a subpopulation are two common ways to decrease σ.

Power calculations are important in planning studies. Using a significance test with low power makes it unlikely that you will find a significant effect even if the truth is far from the null hypothesis. A null hypothesis that is in fact false can become widely believed if repeated attempts to find evidence against it fail because of low power. The following example illustrates this point.

EXAMPLE 6.18 The "efficient market hypothesis" for the time series of stock prices says that future stock prices (when adjusted for inflation) show only random variation. No information available now will help us predict stock prices in the future, because the efficient working of the market has already incorporated all available information in the present price. Many studies have tested the claim that one or another kind of information is helpful. In these studies, the efficient market hypothesis is H_0, and the claim that prediction is possible is H_a. Almost all the studies have failed to find good evidence against H_0. As a result, the efficient market theory is quite popular. But an examination of the significance tests employed finds that the power is generally low. Failure to reject H_0 when using tests of low power is not evidence that H_0 is true. As one expert says, "The widespread impression that there is strong evidence for market efficiency may be due just to a lack of appreciation of the low power of many statistical tests."[6]

Here is another example of a power calculation, this time for a two-sided z test.

EXAMPLE 6.19 Example 6.13 presented a test of

$$H_0: \mu = 0.86$$
$$H_a: \mu \neq 0.86$$

at the 1% level of significance. What is the power of this test against the specific alternative $\mu = 0.845$?

The test rejects H_0 when $|z| \geq 2.576$. The test statistic is

$$z = \frac{\bar{x} - 0.86}{0.0068/\sqrt{3}}$$

Some arithmetic shows that the test rejects when either of the following is true:

$$z \geq 2.576 \quad \text{(in other words, } \bar{x} \geq 0.870\text{)}$$
$$z \leq -2.576 \quad \text{(in other words, } \bar{x} \leq 0.850\text{)}$$

These are disjoint events, so the power is the sum of their probabilities, *computed assuming that the alternative* $\mu = 0.845$ *is true.* We find that

$$P(\bar{x} \geq 0.87) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.87 - 0.845}{0.0068/\sqrt{3}}\right)$$
$$= P(Z \geq 6.37) \doteq 0$$
$$P(\bar{x} \leq 0.85) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.85 - 0.845}{0.0068/\sqrt{3}}\right)$$
$$= P(Z \leq 1.27) = 0.8980$$

Figure 6.14 illustrates this calculation. Because the power is about 0.9, we are quite confident that the test will reject H_0 when this alternative is true.

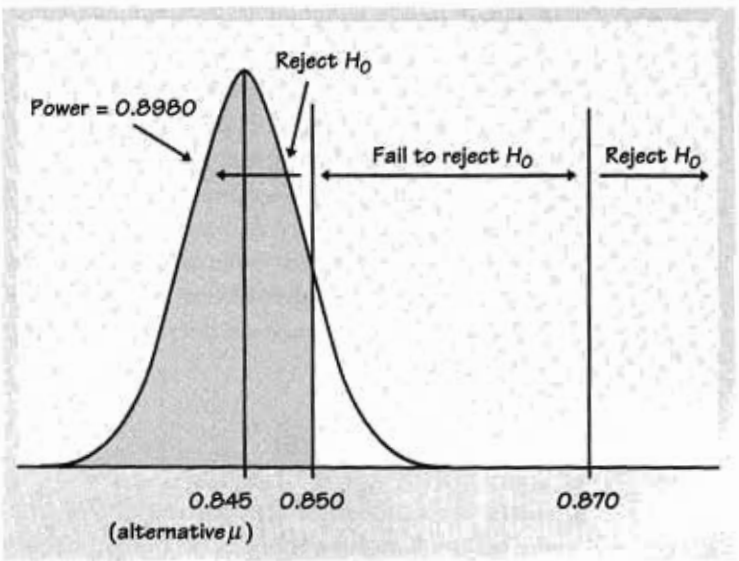

FIGURE 6.14 The power for Example 6.19.