**EXAMPLE 8.11** The true average voltage drop from collector to emitter of insulated gate bipolar transistors of a certain type is supposed to be at most 2.5 volts. An investigator selects a sample of $n = 10$ such transistors and uses the resulting voltages as a basis for testing $H_0$: $\mu = 2.5$ versus $H_a$: $\mu > 2.5$ using a $t$ test with significance level $\alpha = .05$. If the standard deviation of the voltage distribution is $\sigma = .100$, how likely is it that $H_0$ will not be rejected when in fact $\mu = 2.6$? With $d = |2.5 - 2.6|/.100 = 1.0$, the point on the $\beta$ curve at 9 df for a one-tailed test with $\alpha = .05$ above 1.0 has a height of approximately .1, so $\beta \approx .1$. The investigator might think that this is too large a value of $\beta$ for such a substantial departure from $H_0$ and may wish to have $\beta = .05$ for this alternative value of $\mu$. Since $d = 1.0$, the point $(d, \beta) = (1.0, .05)$ must be located. This point is very close to the 14 df curve, so using $n = 15$ will give both $\alpha = .05$ and $\beta = .05$ when the value of $\mu$ is 2.6 and $\sigma = .10$. A larger value of $\sigma$ would give a larger $\beta$ for this alternative, and an alternative value of $\mu$ closer to 2.5 would also result in an increased value of $\beta$. ■

Most of the widely used statistical software packages are capable of calculating type II error probabilities. They generally work in terms of **power**, which is simply $1 - \beta$. A small value of $\beta$ (close to 0) is equivalent to large power (near 1). A *powerful* test is one that has high power and therefore good ability to detect when the null hypothesis is false.

As an example, we asked Minitab to determine the power of the upper-tailed test in Example 8.11 for the three sample sizes 5, 10, and 15 when $\alpha = .05$, $\sigma = .10$, and the value of $\mu$ is actually 2.6 rather than the null value 2.5—a "difference" of. 2.6−2.5 = .1. We also asked the software to determine the necessary sample size for a power of .9 ($\beta = .1$) and also .95. Here is the resulting output:

```
Power and Sample Size
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 0.1
```

```
Power and Sample Size
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 0.1

             Sample
Difference     Size     Power
0.1               5     0.579737
0.1              10     0.897517
0.1              15     0.978916

                Sample Target
                                          Actual
Difference       Size       Power        Power
0.1                11        0.90        0.924489
0.1                13        0.95        0.959703
```

The power for the sample size $n = 10$ is a bit smaller than .9. So if we insist that the power be at least .9, a sample size of 11 is required and the actual power for that $n$ is roughly .92. The software says that for a target power of .95, a sample size of $n = 13$ is required, whereas eyeballing our $\beta$ curves gave 15. When available, this type of software is more reliable than the curves. Finally, Minitab now also provides power curves for the specified sample sizes, as shown in Figure 8.9. Such curves illustrate how the power increases for each sample size as the actual value of $\mu$ moves farther and farther away from the null value.

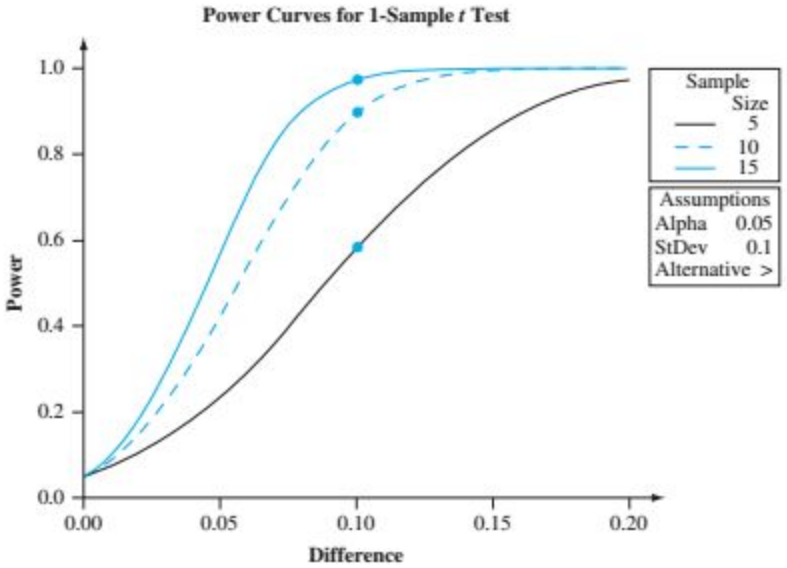


**Figure 8.9** Power curves from Minitab for the *t* test of Example 8.11

**EXAMPLE 8.12** The fuel efficiency (mpg) of any particular new vehicle under specified driving conditions may not be identical to the EPA figure that appears on the vehicle's sticker. Suppose that four different vehicles of a particular type are to be selected and driven over a certain course, after which the fuel efficiency of each one is to be determined.

Let $\mu$ denote the true average fuel efficiency under these conditions. Consider testing $H_0$: $\mu = 20$ versus $H_0$: $\mu > 20$ using the one-sample $t$ test based on the resulting sample. Since the test is based on $n - 1 = 3$ degrees of freedom, the $P$-value for an upper-tailed test is the area under the $t$ curve with 3 df to the right of the calculated $t$.

Let's first suppose that the null hypothesis is true. We asked Minitab to generate 10,000 different samples, each containing 4 observations, from a normal population distribution with mean value $\mu = 20$ and standard deviation $\sigma = 2$. The first sample and resulting summary quantities were

$$x_1 = 20.830,\ x_2 = 22.232,\ x_3 = 20.276,\ x_4 = 17.718$$

$$\bar{x} = 20.264 \quad s = 1.8864 \quad t = \frac{20.264 - 20}{.1.8864/\sqrt{4}} = .2799$$

The $P$-value is the area under the 3-df $t$ curve to the right of .2799, which according to Minitab is .3989. Using a significance level of .05, the null hypothesis would of course not be rejected. The values of $t$ for the next four samples were $-1.7591$, .6082, $-.7020$, and 3.1053, with corresponding $P$-values .912, .293, .733, and .0265.

Figure 8.10(a) shows a histogram of the 10,000 $P$-values from this simulation experiment. About 4.5% of these $P$-values are in the first class interval from 0 to .05. Thus when using a significance level of .05, the null hypothesis is rejected in roughly 4.5% of these 10,000 tests. If we continued to generate samples and carry out the test for each sample at significance level .05, in the long run 5% of the $P$-values would be in the first class interval. This is because when $H_0$ is true and a test with significance level .05 is used, by definition the probability of rejecting $H_0$ is .05.

Looking at the histogram, it appears that the distribution of $P$-values is relatively flat. In fact, it can be shown that when $H_0$ is true, the probability distribution of the $P$-value is a uniform distribution on the interval from 0 to 1. That is, the density curve is completely flat on this interval, and thus must have a height of 1 if the total area under the curve is to be 1. Since the area under such a curve to the left of .05 is $(.05)(1) = .05$, we again have that the probability of rejecting $H_0$ when it is true that it is .05, the chosen significance level.

it is .05, the chosen significance level.

Now consider what happens when $H_0$ is false because $\mu = 21$. We again had Minitab generate 10,000 different samples of size 4 (each from a normal distribution with $\mu = 21$ and $\sigma = 2$), calculate $t = (\bar{x} - 20)/(s/\sqrt{4})$ for each one, and then determine the $P$-value. The first such sample resulted in $\bar{x} = 20.6411$, $s = .49637$, $t = 2.5832$, $P$-value $= .0408$. Figure 8.10(b) gives a histogram of the resulting $P$-values. The shape of this histogram is quite different from that of Figure 8.10(a)—there is a much greater tendency for the $P$-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$. Again $H_0$ is rejected at significance level .05 whenever the $P$-value is at most .05 (in the first class interval). Unfortunately, this is the case for only about 19% of the $P$-values. So only about 19% of the 10,000 tests correctly reject the null hypothesis; for the other 81%, a type II error is committed. The difficulty is that the sample size is quite small and 21 is not very different from the value asserted by the null hypothesis.

Figure 8.10(c) illustrates what happens to the $P$-value when $H_0$ is false because $\mu = 22$ (still with $n = 4$ and $\sigma = 2$). The histogram is even more concentrated toward values close to 0 than was the case when $\mu = 21$. In general, as $\mu$ moves farther to the right of the null value 20, the distribution of the $P$-value will become more and more concentrated on values close to 0. Even here a bit fewer than 50% of the $P$-values are smaller than .05. So it is still slightly more likely than

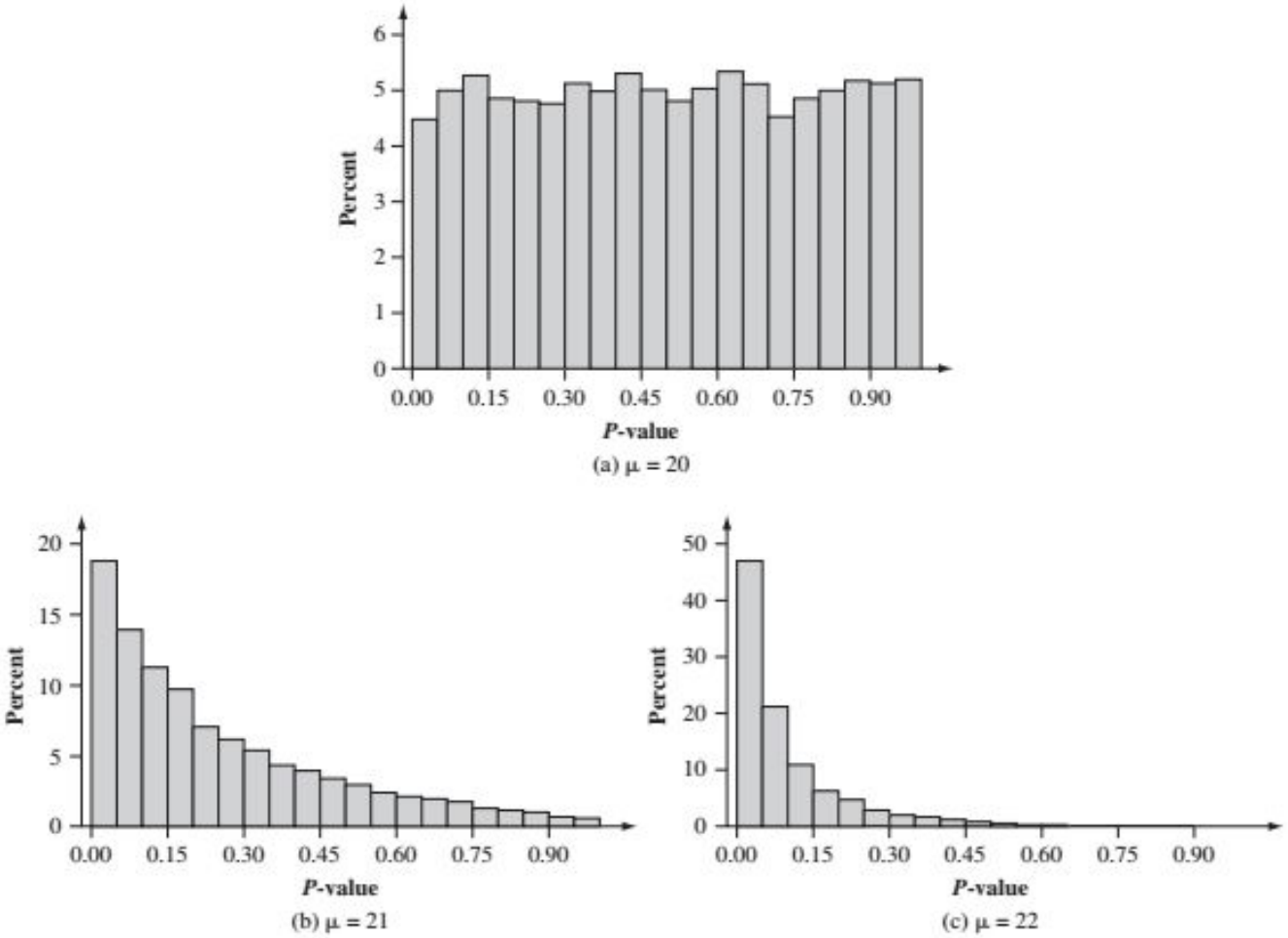


Figure 8.10 *P*-value simulation results for Example 8.12