8 Tests of Hypotheses Based on a Single Sample

INTRODUCTION

A parameter can be estimated from sample data either by a single number (a point estimate) or an entire interval of plausible values (a confidence interval). Frequently, however, the objective of an investigation is not to estimate a parameter but to decide which of two contradictory claims about the parameter is correct. Methods for accomplishing this comprise the part of statistical inference called *hypothesis testing*. In this chapter, we first discuss some of the basic concepts and terminology in hypothesis testing and then develop decision procedures for the most frequently encountered testing problems based on a sample from a single population.

8.1 Hypotheses and Test Procedures

A **statistical hypothesis**, or just *hypothesis,* is a claim or assertion either about the value of a single parameter (population characteristic or characteristic of a probability distribution), about the values of several parameters, or about the form of an entire probability distribution. One example of a hypothesis is the claim $\mu = .75$, where μ is the true average inside diameter of a certain type of PVC pipe. Another example is the statement $p < .10$, where p is the proportion of defective circuit boards among all circuit boards produced by a certain manufacturer. If μ_1 and μ_2 denote the true average breaking strengths of two different types of twine, one hypothesis is the assertion that $\mu_1 - \mu_2 = 0$, and another is the statement $\mu_1 - \mu_2 > 5$. Yet another example of a hypothesis is the assertion that vehicle braking distance under particular conditions has a normal distribution. Hypotheses of this latter sort will be considered in Chapter 14. In this and the next several chapters, we concentrate on hypotheses about parameters.

In any hypothesis-testing problem, there are two contradictory hypotheses under consideration. One hypothesis might be the claim $\mu = .75$ and the other $\mu \neq .75$, or the two contradictory statements might be $p \geq .10$ and $p < .10$. The objective is to decide, based on sample information, which of the two hypotheses is correct. There is a familiar analogy to this in a criminal trial. One claim is the assertion that the accused individual is innocent. In the U.S. judicial system, this is the claim that is initially believed to be true. Only in the face of strong evidence to the contrary should the jury reject this claim in favor of the alternative assertion that the accused is guilty. In this sense, the claim of innocence is the favored or protected hypothesis, and the burden of proof is placed on those who believe in the alternative claim.

Similarly, in testing statistical hypotheses, the problem will be formulated so that one of the claims is initially favored. This initially favored claim will not be rejected in favor of the alternative claim unless sample evidence contradicts it and provides strong support for the alternative assertion.

DEFINITION

The **null hypothesis,** denoted by H_0, is the claim that is initially assumed to be true (the "prior belief" claim). The **alternative hypothesis,** denoted by H_a, is the assertion that is contradictory to H_0.

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis. The two possible conclusions from a hypothesis-testing analysis are then *reject* H_0 or *fail to reject* H_0.

A **test of hypotheses** is a method for using sample data to decide whether the null hypothesis should be rejected. Thus we might test H_0: $\mu = .75$ against the alternative H_a: $\mu \neq .75$. Only if sample data strongly suggests that μ is something other than .75 should the null hypothesis be rejected. In the absence of such evidence, H_0 should not be rejected, since it is still quite plausible.

Sometimes an investigator does not want to accept a particular assertion unless and until data can provide strong support for the assertion. As an example, suppose a company is considering putting a new type of coating on bearings that it produces.

The true average wear life with the current coating is known to be 1000 hours. With μ denoting the true average life for the new coating, the company would not want to make a change unless evidence strongly suggested that μ exceeds 1000. An appropriate problem formulation would involve testing H_0: $\mu = 1000$ against H_a: $\mu > 1000$. The conclusion that a change is justified is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced by a more plausible and satisfactory explanation of the phenomenon under investigation. A conservative approach is to identify the current theory with H_0 and the researcher's alternative explanation with H_a. Rejection of the current theory will then occur only when evidence is much more consistent with the new theory. In many situations, H_a is referred to as the "researcher's hypothesis," since it is the claim that the researcher would really like to validate. The word *null* means "of no value, effect, or consequence," which suggests that H_0 should be identified with the hypothesis of no change (from current opinion), no difference, no improvement, and so on. Suppose, for example, that 10% of all circuit boards produced by a certain manufacturer during a recent period were defective. An engineer has suggested a change in the production process in the belief that it will result in a reduced defective rate. Let p denote the true proportion of defective boards resulting from the changed process. Then the research hypothesis, on which the burden of proof is placed, is the assertion that $p < .10$. Thus the alternative hypothesis is H_a: $p < .10$.

In our treatment of hypothesis testing, H_0 will generally be stated as an equality claim. If θ denotes the parameter of interest, the null hypothesis will have the form H_0: $\theta = \theta_0$, where θ_0 is a specified number called the **null value** of the parameter (value claimed for θ by the null hypothesis). As an example, consider the circuit board situation just discussed. The suggested alternative hypothesis was H_a: $p < .10$, the claim that the defective rate is reduced by the process modification. A natural choice of H_0 in this situation is the claim that $p \geq .10$, according to which the new process is either no better *or* worse than the one currently used. We will instead consider H_0: $p = .10$ versus H_a: $p < .10$. The rationale for using this simplified null hypothesis is that any reasonable decision procedure for deciding between H_0: $p = .10$ and H_a: $p < .10$ will also be reasonable for deciding between the claim that $p \geq .10$ and H_a. The use of a simplified H_0 is preferred because it has certain technical benefits, which will be apparent shortly.

The alternative to the null hypothesis H_0: $\theta = \theta_0$ will look like one of the following three assertions:

1. H_a: $\theta > \theta_0$ (in which case the implicit null hypothesis is $\theta \leq \theta_0$),
2. H_a: $\theta < \theta_0$ (in which case the implicit null hypothesis is $\theta \geq \theta_0$), or
3. H_a: $\theta \neq \theta_0$

For example, let σ denote the standard deviation of the distribution of inside diameters (inches) for a certain type of metal sleeve. If the decision was made to use the sleeve unless sample evidence conclusively demonstrated that $\sigma > .001$, the appropriate hypotheses would be H_0: $\sigma = .001$ versus H_a: $\sigma > .001$. The number θ_0 that appears in both H_0 and H_a (separates the alternative from the null) is the null value.

Test Procedures and *P*-Values

A test procedure is a rule, based on sample data, for deciding whether H_0 should be rejected. The key issue will be the following: Suppose that H_0 is in fact true. Then

how likely is it that a (random) sample at least as contradictory to this hypothesis as our sample would result? Consider the following two scenarios:

1. There is only a .1% chance (a probability of .001) of getting a sample at least as contradictory to H_0 as what we obtained assuming that H_0 is true.
2. There is a 25% chance (a probability of .25) of getting a sample at least as contradictory to H_0 as what we obtained when H_0 is true.

In the first scenario, something as extreme as our sample is very unlikely to have occurred when H_0 is true—in the long run only 1 in 1000 samples would be at least as contradictory to the null hypothesis as the one we ended up selecting. In contrast, for the second scenario, in the long run 25 out of every 100 samples would be at least as contradictory to H_0 as what we obtained assuming that the null hypothesis is true. So our sample is quite consistent with H_0, and there is no reason to reject it.

We must now flesh out this reasoning by being more specific as to what is meant by "at least as contradictory to H_0 as the sample we obtained when H_0 is true." Before doing so in a general way, let's consider several examples.

EXAMPLE 8.1 The company that manufactures brand D Greek-style yogurt is anxious to increase its market share, and in particular persuade those who currently prefer brand C to switch brands. So the marketing department has devised the following blind taste experiment. Each of 100 brand C consumers will be asked to taste yogurt from two bowls, one containing brand C and the other brand D, and then say which one he or she prefers. The bowls are marked with a code so that the experimenters know which bowl contains which yogurt, but the experimental subjects do not have this information (Note: Such an experiment involving beers was actually carried out several decades ago, with the now defunct Schlitz beer playing the role of brand D and Michelob being the target beer).

Let p denote the proportion of all brand C consumers who would prefer C to D in such circumstances. Let's consider testing the hypotheses H_0: $p = .5$ versus H_a: $p < .5$. The alternative hypothesis says that a majority of brand C consumers actually prefer brand D. Of course the brand D company would like to have H_0 rejected so that H_a is judged the more plausible hypothesis. If the null hypothesis is true, then whether a single randomly selected brand C consumer prefers C or D is analogous to the result of flipping a fair coin.

The sample data will consist of a sequence of 100 preferences, each one a C or a D. Let X = the number among the 100 selected individuals who prefer C to D. This random variable will serve as our *test statistic*, the function of sample data on which we'll base our conclusion. Now X is a binomial random variable (the number of successes in an experiment with a fixed number of independent trials having constant success probability p). When H_0 is true, this test statistic has a binomial distribution with $p = .5$, in which case $E(X) = np = 100(.5) = 50$.

Intuitively, a value of X "considerably" smaller than 50 argues for rejection of H_0 in favor of H_a. Suppose the observed value of X is $x = 37$. How contradictory is this value to the null hypothesis? To answer this question, let's first identify values of X that are even more contradictory to H_0 than is 37 itself. Clearly 35 is one such value, and 30 is another; in fact, any number smaller than 37 is a value of X more contradictory to the null hypothesis than is the value we actually observed. Now consider the probability, computed assuming that the null hypothesis is true, of obtaining a value of X at least as contradictory to H_0 as is our observed value:

$$\begin{aligned} P(X \le 37 \text{ when } H_0 \text{ is true}) &= P(X \le 37 \text{ when } X \sim \text{Bin}(100, .5)) \\ &= B(37; 100, .5) = .006 \end{aligned}$$

(from software). Thus if the null hypothesis is true, there is less than a 1% chance of seeing 37 or fewer successes amongst the 100 trials. This suggests that $x = 37$ is much more consistent with the alternative hypothesis than with the null, and that rejection of H_0 in favor of H_a is a sensible conclusion. In addition, note that $\sigma_X = \sqrt{npq} = \sqrt{100(.5)(.5)} = 5$ when H_0 is true. It follows that 37 is more than 2.5 standard deviations smaller than what we'd expect to see were H_0 true.

Now suppose that 45 of the 100 individuals in the experiment prefer C (45 successes). Let's again calculate the probability, assuming H_0 true, of getting a test statistic value at least as contradictory to H_0 as this:

$$\begin{aligned} P(X \le 45 \text{ when } H_0 \text{ is true}) &= P(X \le 45 \text{ when } X \sim \text{Bin}(100, .5)) \\ &= B(45; 100, .5) = .184 \end{aligned}$$

So if in fact $p = .5$, it would not be surprising to see 45 or fewer successes. For this reason, the value 45 does not seem very contradictory to H_0 (it is only one standard deviation smaller than what we'd expect were H_0 true). Rejection of H_0 in this case does not seem sensible. ■

EXAMPLE 8.2 According to the article **"Freshman 15: Fact or Fiction"** (***Obesity*, 2006: 1438–1443**), "A common belief among the lay public is that body weight increases after entry into college, and the phrase 'freshman 15' has been coined to describe the 15 pounds that students presumably gain over their freshman year." Let μ denote the true average weight gain of women over the course of their first year in college. The foregoing quote suggests that we should test the hypotheses H_0: $\mu = 15$ versus H_a: $\mu \neq 15$. For this purpose, suppose that a random sample of n such individuals is selected and the weight gain of each one is determined, resulting in a sample mean weight gain $\bar{x}$ and a sample standard deviation s (Note: The data here is actually *paired*, with each weight gain resulting from obtaining a (beginning, ending) weight pair and then subtracting to determine the difference; more will be said about such data in Section 9.3). Before data is obtained, the sample mean weight gain is a random variable $\overline{X}$ and the sample standard deviation is also a random variable S.

A natural test statistic (function of the data on which the decision will be based) is the sample mean $\overline{X}$ itself; if H_0 is true, then $E(\overline{X}) = \mu = 15$, whereas if μ differs considerably from 15, then the sample mean weight gain should do the same. But there is a more convenient test statistic that has appealing intuitive and technical properties: the sample mean standardized assuming that H_0 is true. Recall that the standard deviation (standard error) of $\overline{X}$ is $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Supposing that the population distribution of weight gains is normal, it follows that the sampling distribution of $\overline{X}$ itself is normal. Now standardizing a normally distributed variable gives a variable having a standard normal distribution (the z curve):

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

If the value of σ were known, we could obtain a test statistic simply by replacing μ by the null value $\mu_0 = 15$:

$$Z = \frac{\overline{X} - 15}{\sigma/\sqrt{n}}$$

If substitution of $\bar{x}$, σ, and n results in $z = 3$, the interpretation is that the observed value of the sample mean is three standard deviations larger than what we would have expected it to be were the null hypothesis true. Of course in "normal land" such an occurrence is exceedingly rare. Alternatively, if $z = -1$, then the sample mean is

only one standard deviation less than what would be expected under H_0, a result not surprising enough to cast substantial doubt on H_0.

A practical glitch in the foregoing development is that the value of σ is virtually never available to an investigator. However, as discussed in the previous chapter, substitution of S for σ in Z typically introduces very little extra variability when n is large ($n > 40$ was our earlier rule of thumb). In this case the resulting variable still has *approximately* a standard normal distribution. The implied large-sample test statistic for our weight-gain scenario is

$$Z = \frac{\bar{X} - 15}{S/\sqrt{n}}$$

Thus when H_0 is true, Z has approximately a standard normal distribution.

Suppose that $\bar{x} = 13.7$, and that substitution of this along with s and n gives $z = -2.80$. Which values of the test statistic are at least as contradictory to H_0 as -2.80 itself? To answer this, let's first determine values of $\bar{x}$ that are at least as contradictory to H_0 as 13.7. One such value is 13.5, another is 13.0, and in fact *any* value smaller than 13.7 is more contradictory to H_0 than 13.7.

But that is not the whole story. Recall that the alternative hypothesis says that the value of μ is something other than 15. In light of this, the value 16.3 is just as contradictory to H_0 as is 13.7; it falls the same distance above the null value 15 as 13.7 does below 15—and the resulting z value is 3.0, just as extreme as -3.0. And any particular $\bar{x}$ that exceeds 16.3 is just as contradictory to H_0 as is a value the same distance below 15—e.g., 16.8 and 14.2, 17.0 and 13.0, and so on.

Just as values of $\bar{x}$ that are at most 13.7 correspond to $z \leq -2.80$, values of $\bar{x}$ that are at least 16.3 correspond to $z \geq 2.80$. Thus values of the test statistic that are at least as contradictory to H_0 as the value -2.80 actually obtained are $\{z: z \leq -2.80 \text{ or } z \geq 2.80\}$. We can now calculate the probability, assuming H_0 true, of obtaining a test statistic value at least as contradictory to H_0 as what our sample yielded:

$$\begin{aligned} &P(Z \leq -2.80 \text{ or } Z \geq 2.80 \text{ assuming } H_0 \text{ true}) \\ &\quad \approx 2 \cdot (\text{area under the } z \text{ curve to the right of } 2.80) \\ &\quad = 2[1 - \Phi(2.80)] = 2[1 - .9974] = .0052 \end{aligned}$$

That is, if the null hypothesis is in fact true, only about one half of one percent of all samples would result in a test statistic value at least as contradictory to the null hypothesis as is our value. Clearly -2.80 is among the possible test statistic values that are most contradictory to H_0. It would therefore make sense to reject H_0 in favor of H_a.

Suppose we had instead obtained the test statistic value $z = .89$, which is less than one standard deviation larger than what we'd expect if H_0 were true. The foregoing probability would then be

$$\begin{aligned} &P(Z \leq -0.89 \text{ or } Z \geq 0.89 \text{ assuming } H_0 \text{ true}) \\ &\quad \approx 2 \cdot (\text{area under the } z \text{ curve to the right of } .89) \\ &\quad = 2[1 - \Phi(.89)] = 2[1 - .8133] = .3734 \end{aligned}$$

More than 1/3 of all samples would give a test statistic value at least as contradictory to H_0 as is .89 when H_0 is true. So the data is quite consistent with the null hypothesis; it remains plausible that $\mu = 15$.

The article cited at the outset of this example reported that for a sample of 137 students, the sample mean weight gain was only 2.42 lb with a sample standard deviation of 5.72 lb (some students lost weight). This gives $z = (2.42 - 15)/(5.72/\sqrt{137}) = -25.7$! The probability of observing a value at least

this extreme in either direction is essentially 0. The data very strongly contradicts the null hypothesis, and there is substantial evidence that true average weight gain is much closer to 0 than to 15. ■

The type of probability calculated in Examples 8.1 and 8.2 will now provide the basis for obtaining general test procedures.

DEFINITIONS

A **test statistic** is a function of the sample data used as a basis for deciding whether H_0 should be rejected. The selected test statistic should discriminate effectively between the two hypotheses. That is, values of the statistic that tend to result when H_0 is true should be quite different from those typically observed when H_0 is not true.

The ***P*-value** is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample data. A conclusion is reached in a hypothesis testing analysis by selecting a number α, called the **significance level** (alternatively, *level of significance*) of the test, that is reasonably close to 0. Then H_0 will be rejected in favor of H_a if $P\text{-value} \leq \alpha$, whereas H_0 will not be rejected (still considered to be plausible) if $P\text{-value} > \alpha$. The significance levels used most frequently in practice are (in order) $\alpha = .05, .01, .001$, and .10.

For example, if we select a significance level of .05 and then compute $P\text{-value} = .0032$, H_0 would be rejected because $.0032 \leq .05$. With this same P-value, the null hypothesis would also be rejected at the smaller significance level of .01 because $.0032 \leq .01$. However, at a significance level of .001 we would not be able to reject H_0 since $.0032 > .001$. Figure 8.1 illustrates the comparison of the P-value with the significance level in order to reach a conclusion.

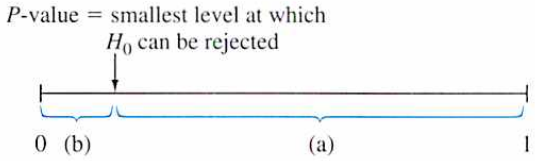


Figure 8.1 Comparing α and the P-value: (a) reject H_0 when α lies here; (b) do not reject H_0 when α lies here

We will shortly consider in some detail the consequences of selecting a smaller significance level rather than a larger one. For the moment, note that the smaller the significance level, the more protection is being given to the null hypothesis and the harder it is for that hypothesis to be rejected.

The definition of a P-value is obviously somewhat complicated, and it doesn't roll off the tongue very smoothly without a good deal of practice. In fact, many users of statistical methodology use the specified decision rule repeatedly to test hypotheses, but would be hard put to say what a P-value is! Here are some important points:

- The P-value is a probability.
- This probability is calculated assuming that the null hypothesis is true.
- To determine the P-value, we must first decide which values of the test statistic are at least as contradictory to H_0 as the value obtained from our sample.

- The smaller the P-value, the stronger is the evidence against H_0 and in favor of H_a.
- The P-value is not the probability that the null hypothesis is true or that it is false, nor is it the probability that an erroneous conclusion is reached.

EXAMPLE 8.3 Urban storm water can be contaminated by many sources, including discarded batteries. When ruptured, these batteries release metals of environmental significance. The article **"Urban Battery Litter" (*J. Environ. Engr.*, 2009: 46–57)** presented summary data for characteristics of a variety of batteries found in urban areas around Cleveland. A random sample of 51 Panasonic AAA batteries gave a sample mean zinc mass of 2.06 g. and a sample standard deviation of .141 g. Does this data provide compelling evidence for concluding that the population mean zinc mass exceeds 2.0 g.? Let's employ a significance level of .01 to reach a conclusion.

With μ denoting the true average zinc mass for such batteries, the relevant hypotheses are

$$H_0: \mu = 2.0 \text{ versus } H_a: \mu > 2.0.$$

The reasonably large sample size allows us to invoke the Central Limit Theorem, according to which the sample mean $\overline{X}$ has approximately a normal distribution. Furthermore, the standardized variable $Z = (\overline{X} - \mu)/(S/\sqrt{n})$ has approximately a standard normal distribution (the z curve). The test statistic results from standardizing $\overline{X}$ assuming that H_0 is true:

$$\text{Test statistic: } Z = \frac{\overline{X} - 2.0}{S/\sqrt{n}}$$

Substituting $n = 51$, $\bar{x} = 2.06$, and $s = .141$ gives $z = .06/.0197 = 3.04$. The sample mean here is roughly three (estimated) standard errors larger than would be expected were H_0 true (it does not appear to exceed 2 by very much, but there is only a small amount of variability in the 51 sample observations).

Any value of $\bar{x}$ larger than 2.06 is more contradictory to H_0 than 2.06 itself, and values of $\bar{x}$ that exceed 2.06 correspond to values of z that exceed 3.04. So any $z \geq 3.04$ is at least as contradictory to H_0. Since the test statistic has approximately a standard normal distribution when H_0 is true, we have

$$P\text{-value} \approx P(\text{a standard normal rv is} \geq 3.04) = 1 - \Phi(3.04) = 1 - .9988 = .0012$$

Because P-value $= .0012 \leq .01 = \alpha$, the null hypothesis should be rejected at the chosen significance level. It appears that true average zinc mass does indeed exceed 2. ■

Errors in Hypothesis Testing

The basis for choosing a particular significance level α lies in consideration of the errors that one might be faced with in drawing a conclusion. Recall the judicial scenario in which the null hypothesis is that the individual accused of committing a crime is in fact innocent. In rendering a verdict, the jury must consider the possibility of committing one of two different kinds of errors. One of these involves convicting an innocent person, and the other involves letting a guilty person go free. Similarly, there are two different types of errors that might be made in the course of a statistical hypothesis testing analysis.

DEFINITIONS

A **type I error** consists of rejecting the null hypothesis H_0 when it is true.

A **type II error** involves not rejecting H_0 when it is false.

As an example, a cereal manufacturer claims that a serving of one of its brands provides 100 calories (calorie content used to be determined by a destructive testing method, but the requirement that nutritional information appear on packages has led to more straightforward techniques). Of course the actual calorie content will vary somewhat from serving to serving (of the specified size), so 100 should be interpreted as an average. It could be distressing to consumers of this cereal if the true average calorie content exceeded the asserted value. So an appropriate formulation of hypotheses is to test H_0: $\mu = 100$ versus H_a: $\mu > 100$. The alternative hypothesis says that consumers are ingesting on average a greater amount of calories than what the company claims. A type I error here consists of rejecting the manufacturer's claim that $\mu = 100$ when it is actually true. A type II error results from not rejecting the manufacturer's claim when it is actually the case that $\mu > 100$.

Suppose μ_1 and μ_2 represent the true average lifetimes for two different brands of rollerball pen under controlled experimental conditions (utilizing a machine that writes continuously until a pen fails). It is natural to test the hypotheses H_0: $\mu_1 - \mu_2 = 0$ (i.e., $\mu_1 = \mu_2$) versus H_a: $\mu_1 - \mu_2 \neq 0$ (i.e., $\mu_1 \neq \mu_2$). A type I error would be to conclude that the true average lifetimes are different when in fact they are identical. A type II error involves deciding that the true average lifetimes may be the same when in fact they really differ from one another.

In the best of all possible worlds, we'd have a judicial system that never convicted an innocent person and never let a guilty person go free. This gold standard for judicial decisions has proven to be extremely elusive. Similarly, we would like to find test procedures for which neither type of error is ever committed. However, this ideal can be achieved only by basing a conclusion on an examination of the entire population. The difficulty with using a procedure based on sample data is that because of sampling variability, a sample unrepresentative of the population may result. In the calorie content scenario, even if the manufacturer's assertion is correct, an unusually large value of $\bar{X}$ may result in a P-value smaller than the chosen significance level and the consequent commission of a type I error. Alternatively, the true average calorie content may exceed what the manufacturer claims, yet a sample of servings may yield a relatively large P-value for which the null hypothesis cannot be rejected.

Instead of demanding error-free test procedures, we must seek procedures for which either type of error is unlikely to be committed. That is, a good procedure is one for which the probability of making a type I error is small and the probability of making a type II error is also small.

EXAMPLE 8.4 An automobile model is known to sustain no visible damage 25% of the time in 10-mph crash tests. A modified bumper design has been proposed in an effort to increase this percentage. Let p denote the proportion of all 10-mph crashes with this new bumper that result in no visible damage. The hypotheses to be tested are H_0: $p = .25$ (no improvement) versus H_a: $p > .25$. The test will be based on an experiment involving $n = 20$ independent crashes with prototypes of the new design. The natural test statistic here is X = the number of crashes with no visible damage. If H_0 is true, $E(X) = np_0 = (20)(.25) = 5$. Intuition suggests that an observed value x much larger than this would provide strong evidence against H_0 and in support of H_a.

Consider using a significance level of .10. The P-value is $P(X \geq x$ when X has a binomial distribution with $n = 20$ and $p = .25) = 1 - B(x - 1; 20, .25)$ for $x > 0$.

Appendix Table A.1 shows that in this case,

$$P(X \geq 7) = 1 - B(6; 20, .25) = 1 - .786 = .214$$
$$P(X \geq 8) = 1 - .898 = .102 \approx .10, P(X \geq 9) = 1 - .959 = .041$$

Thus rejecting H_0 when P-value $\leq .10$ is equivalent to rejecting H_0 when $X \geq 8$. Therefore

$$\begin{aligned} P(\text{committing a type I error}) &= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) \\ &= P(X \geq 8 \text{ when } X \text{ has a binomial distribution with } n = 20 \text{ and } p = .25) \\ &= .102 \\ &\approx .10 \end{aligned}$$

That is, *the probability of a type I error is just the significance level* α. If the null hypothesis is true here and the test procedure is used over and over again, each time in conjunction with a group of 20 crashes, in the long run the null hypothesis will be incorrectly rejected in favor of the alternative hypothesis about 10% of the time. So our test procedure offers reasonably good protection against committing a type I error.

There is only one type I error probability because there is only one value of the parameter for which H_0 is true (this is one benefit of simplifying the null hypothesis to a claim of equality). Let β denote the probability of committing a type II error. Unfortunately there is not a single value of β, because there are a multitude of ways for H_0 to be false—it could be false because $p = .30$, because $p = .37$, because $p = .5$, and so on. There is in fact a different value of β for each different value of p that exceeds .25. At the chosen significance level .10, H_0 will be rejected if and only if $X \geq 8$, so H_0 will not be rejected if and only if $X \leq 7$. Thus

$$\begin{aligned} \beta(.3) &= P(\text{type II error when } p = .3) \\ &= P(H_0 \text{ is not rejected when } p = .3) \\ &= P[X \leq 7 \text{ when } X \sim \text{Bin}(20, .3)] \\ &= B(7; 20, .3) = .772 \end{aligned}$$

When p is actually .3 rather than .25 (a "small" departure from H_0), roughly 77% of all experiments of this type would result in H_0 being incorrectly not rejected!

The accompanying table displays β for selected values of p (each calculated as we just did for $\beta(.3)$). Clearly, β decreases as the value of p moves farther to the right of the null value .25. Intuitively, the greater the departure from H_0, the more likely it is that such a departure will be detected.

p	.3	.4	.5	.6	.7	.8
$\beta(p)$	.772	.416	.132	.021	.001	.000

The probability of committing a type II error here is quite large when $p = .3$ or .4. This is because those values are quite close to what H_0 asserts and the sample size of 20 is too small to permit accurate discrimination between .25 and those values of p.

The proposed test procedure is still reasonable for testing the more realistic null hypothesis that $p \leq .25$. In this case, there is no longer a single type I error probability α, but instead there is an α for each p that is at most .25: $\alpha(.25)$, $\alpha(.23)$, $\alpha(.20)$, $\alpha(.15)$, and so on. It is easily verified, though, that $\alpha(p) < \alpha(.25) = .102$ if $p < .25$. That is, the largest type I error probability occurs for the boundary value .25 between H_0 and H_a. Thus if α is small for the simplified null hypothesis, it will also be as small as or smaller for the more realistic H_0. ■

EXAMPLE 8.5 The drying time of a type of paint under specified test conditions is known to be normally distributed with mean value 75 min and standard deviation 9 min. Chemists have proposed a new additive designed to decrease average drying time. It is believed that drying times with this additive will remain normally distributed with $\sigma = 9$. Because of the expense associated with the additive, evidence should strongly

suggest an improvement in average drying time before such a conclusion is adopted. Let μ denote the true average drying time when the additive is used. The appropriate hypotheses are H_0: $\mu = 75$ versus H_a: $\mu < 75$. Only if H_0 can be rejected will the additive be declared successful and used.

Experimental data is to consist of drying times from $n = 25$ test specimens. Let $X_1,\ldots,X_{25}$ denote the 25 drying times—a random sample of size 25 from a normal distribution with mean value μ and standard deviation $\sigma = 9$ (although the assumption of a known value of σ is generally unrealistic in practice, it considerably simplifies calculation of type II error probabilities). The sample mean drying time $\overline{X}$ then has a normal distribution with expected value $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 9/\sqrt{25} = 1.8$. When H_0 is true, we expect $\overline{X}$ to be 75; a sample mean much smaller than this would be contradictory to H_0 and supportive of H_a.

Our test statistic here will be $\overline{X}$ standardized assuming that H_0 is true:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X} - 75}{1.8}$$

The sampling distribution of $\overline{X}$ is normal because the population distribution is normal, which implies that Z has a standard normal distribution when H_0 is true (in contrast to Examples 8.2 and 8.3, we are assuming and using a known value of σ here).

Consider carrying out the test using a significance level of .01, i.e., H_0 will be rejected if P-value $\leq .01$. For a given value $\bar{x}$ of the sample mean and corresponding calculated value z, the form of the alternative hypothesis implies that values more contradictory to H_0 than this are values less than $\bar{x}$ and, correspondingly, values of the test statistic that are less than z. Thus the P-value is

$$\begin{aligned} P\text{-value} &= P(\text{obtaining a value of } Z \text{ at least as contradictory to} \\ &\qquad H_0 \text{ as } z \text{ when } H_0 \text{ is true}) \\ &= P(Z \leq z \text{ when } H_0 \text{ is true}) \\ &= \text{area under the standard normal curve to the left of } z \\ &= \Phi(z) \end{aligned}$$

So the P-value will equal .01 when z captures lower-tail area .01 under the z curve. From Appendix Table A.3, this happens when $z = -2.33$ [verify that $\Phi(-2.33) = .01$]. As illustrated in Figure 8.2, the P-value will therefore be at most .01 when $z \leq -2.33$. This in turn implies that

$$\begin{aligned} P(\text{type I error}) &= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) \\ &= P(P\text{-value} \leq .01 \text{ when } H_0 \text{ is true}) \\ &= P(Z \leq -2.33 \text{ when } Z \text{ has a standard normal distribution}) \\ &= .01 \end{aligned}$$

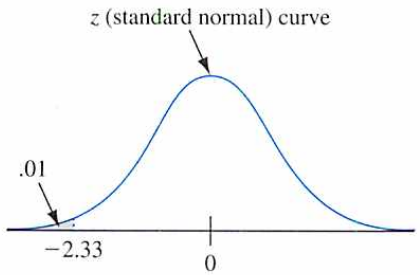


Figure 8.2 P-value $\leq .01$ if and only if $z \leq -2.33$

As in the previous example, the chosen significance level α is in fact the probability of committing a type I error. If the above test procedure [test statistic Z, reject H_0 if

P-value $\leq .01$] is used repeatedly on sample after sample, in the long run the null hypothesis will be incorrectly rejected only 1% of the time. Our proposed test procedure offers excellent protection against the commission of a type I error. Note that if the more realistic null hypothesis H_0: $\mu \geq 75$ is considered, it can be shown that P(type I error) $\leq$ .01; the maximum occurs at the null value 75, which is the boundary between H_0 and H_a.

The calculation of P(type I error) in this example relied on the fact that P-value $\leq$ .01 is equivalent to $Z = (\bar{X} - 75)/1.8 \leq -2.33$. Multiplying both sides of this latter inequality by 1.8 and then adding 75 to both sides results in $\bar{X} \leq 70.8$. Thus rejecting H_0 at significance level .01 [if P-value $\leq .01$] is equivalent to rejecting H_0 if $\bar{X} \leq 70.8$; H_0 will not be rejected if $\bar{X} > 70.8$. The probability of committing a type II error when $\mu = 72$ is now

$$\begin{aligned}\beta(72) &= P(\text{not rejecting } H_0 \text{ when } \mu = 72)\\ &= P(\bar{X} > 70.8 \text{ when } \bar{X} \sim \text{ normal with } \mu_{\bar{X}} = 72, \sigma_{\bar{X}} = 1.8)\\ &= 1 - \Phi[(70.8 - 72)/1.8] = 1 - \Phi(-.67) = 1 - .2514 = .7486\end{aligned}$$

This is an awfully large error probability. If the test with $\alpha = .01$ is used repeatedly on sample after sample and the actual value of μ is 72, almost 75% of the time the null hypothesis will not be rejected. The difficulty is that 72 is too close to the null value for a test with this sample size and value of α to have a good chance of detecting such a departure from H_0.

Similar calculations give

$$\beta(70) = 1 - \Phi[(70.8 - 70)/1.8] = .3300, \quad \beta(67) = .0174$$

These type II error probabilities are much smaller than $\beta(72)$ because 70 and 67 are both farther away from the null value than is 72. Figure 8.3 illustrates α and the first two type II error probabilities.

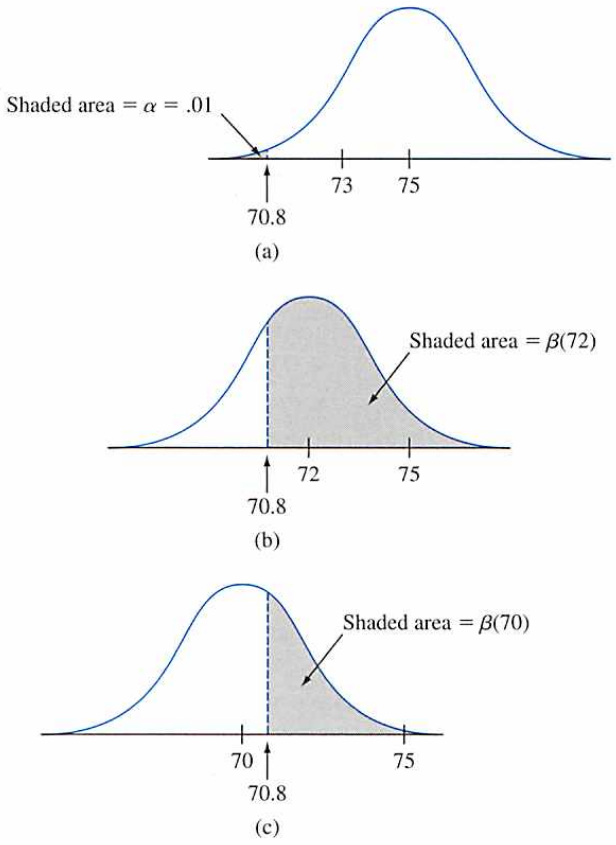


Figure 8.3 α and β illustrated for Example 8.5: (a) the distribution of $\bar{X}$ when $\mu = 75$ (H_0 true); (b) the distribution of $\bar{X}$ when $\mu = 72$ (H_0 false); (c) the distribution of $\bar{X}$ when $\mu = 70$ (H_0 false)

The investigators might regard $\mu = 72$ as an important departure from the null hypothesis, in which case $\beta(72) = .7486$ is intolerably large. Consider changing the significance level (type I error probability from .01 to .05; that is, we now propose rejecting H_0 if P-value $\leq .05$. Appendix Table A.3 shows that the z critical value -1.645 captures a lower-tail z curve area of .05. Using the same reasoning that we previously applied when $\alpha = .01$, rejecting H_0 when P-value $\leq .05$ is equivalent to rejecting when $Z \leq -1.645$. This in turn is equivalent to rejecting when $\bar{X} \leq 72$ (notice that by increasing the significance level, we have made it easier for the null hypothesis to get rejected). Proceeding as in the previous calculations, we find that

$$\beta(72) = .5, \quad \beta(70) = .1335, \quad \beta(67) = .0027$$

These type II error probabilities are all smaller than their counterparts for the test with $\alpha = .01$. The important message here is that if a larger significance level (type I error probability) can be tolerated, then the resulting test will have better ability to detect when the null hypothesis is false. ■

It is no accident that in the two foregoing examples, the significance level α turned out to be the probability of a type I error.

PROPOSITION

The test procedure that rejects H_0 if P-value $\leq \alpha$ and otherwise does not reject H_0 has $P(\text{type I error}) = \alpha$. That is, the significance level employed in the test procedure is the probability of a type I error.

A partial proof of this proposition is sketched out at the end of the section.

The inverse relationship between the significance level α and type II error probabilities in Example 8.5 can be generalized in the following manner:

PROPOSITION

Suppose an experiment or sampling procedure is selected, a sample size is specified, and a test statistic is chosen. Then increasing the significance level α, i.e., employing a larger type I error probability, results in a smaller value of β for any particular parameter value consistent with H_a.

This result is intuitively obvious because when α is increased, it becomes more likely that we'll have P-value $\leq \alpha$ and therefore less likely that P-value $> \alpha$.

The proposition implies that once the test statistic and n are fixed, it is not possible to make both α and any values of β that might be of interest arbitrarily small. Deciding on an appropriate significance level involves compromising between small α and small β's. In Example 8.5, the type II error probability for a test with $\alpha = .01$ was quite large for a value of μ close to the value in H_0. A strategy that is sometimes (but perhaps not often enough) used in practice is to specify α and also β for some alternative value of the parameter that is of particular importance to the investigator. Then the sample size n can be determined to satisfy these two conditions. For example, the article **"Cognitive Treatment of Illness Perceptions in Patients with Chronic Low Back Pain: A Randomized Controlled Trial" (*Physical Therapy*, 2013: 435–438)** contains the following passage: "A decrease of 18 to 24 mm on the PSC was determined as being a clinically relevant change in patients with low back pain. The sample size was calculated with a minimum change of 18 mm, a

2-sided α of .05, a $1 - \beta$ of .90, and a standard deviation of 26.01.... The sample size calculation resulted in a total of 135 participants." We'll consider such sample size determinations in subsequent sections and chapters.

In practice it is usually the case that the hypotheses of interest can be formulated so that a type I error is more serious than a type II error. The approach adhered to by most statistical practitioners is to reflect on the relative seriousness of a type I error compared to a type II error and then use the largest value of α that can be tolerated. This amounts to doing the best we can with respect to type II error probabilities while ensuring that the type I error probability is sufficiently small. For example, if $\alpha = .05$ is the largest significance level that can be tolerated, it would be better to use that rather than $\alpha = .01$, because all β's for the former α will be smaller than those for the latter one. As previously mentioned, the most frequently employed significance levels are $\alpha = .05, .01, .001$, and .10. However, there are exceptions. Here is one example from particle physics: according to the article **"Discovery or Fluke: Statistics in Particle Physics" (*Physics Today*, July 2012: 45–50),** "the usual choice of alpha is 3×10^{-7}, corresponding to the 5σ of a Gaussian [i.e., normal] H_0 distribution.... Why so stringent? For one thing, recent history offers many cautionary examples of exciting 3σ and 4σ signals that went away when more data arrived."

If the distribution of the test statistic is continuous (e.g., if the test statistic has the standard normal distribution or a particular t distribution when H_0 is true), then any significance level α between 0 and 1 can be employed—for example, reject H_0 if P-value $\leq .035$. However, this is not necessarily the case if the distribution of the test statistic is discrete. As an example, consider again the bumper design scenario of Example 8.4 in which the hypotheses of interest were H_0: $p = .25$ versus H_a: $p > .25$. The test statistic X had a binomial distribution and

$$P\text{-value} = P(X \geq x \text{ when } n = 20 \text{ and } p = .25)$$

Appendix Table A.1 shows that in this case, $P(X \geq 8) = .102$ and $P(X \geq 9) = .041$. Thus if we want the significance level to be .05, the closest achievable level is actually .041: reject H_0 if P-value $\leq .041$.

Some Further Comments on the *P*-Value

Suppose that the P-value is calculated to be .038. The null hypothesis will then be rejected if $.038 \leq \alpha$ and not rejected otherwise. So H_0 can be rejected if $\alpha = .10$ or .05 but not if $\alpha = .01$ or .001. In fact, H_0 would be rejected for any significance level that is at least .038 but not for any level smaller than .038. For this reason, the P-value is often referred to as the **observed significance level** (OSL): it is the smallest value of α for which H_0 can be rejected.

One very appealing aspect of basing a conclusion from a hypothesis testing analysis on the P-value is that all widely used statistical software packages will calculate and output the P-value for any of the commonly used test procedures. Once the P-value is available, the investigator need only compare it to the selected significance level to decide whether H_0 should be rejected. This explains how an investigator can forget the definition of a P-value and still use it to reach a conclusion!

Sometimes a situation is encountered in which various individuals are interested in testing the same pair of hypotheses but may wish to use different significance levels. For example, suppose the true average time to pain relief for the current best-selling pain reliever is known to be 15 minutes. A new formulation has been developed that it is hoped will reduce this time. The relevant hypotheses are H_0: $\mu = 15$ versus H_a: $\mu < 15$, where μ is the true average time to relief using the new formulation. You may be quite satisfied with the current product and therefore wish to use a small significance

level such as .01. I on the other hand may be less satisfied and thus more willing to switch, in which case a larger level such as .10 may be sensible. In using the larger α, I am giving less protection to H_0 than you are. Once the P-value is available, each of us can employ our own significance level irrespective of what the other person is using. Thus when medical journals report a P-value, a significance level is not mandated; instead it is left to the reader to select his or her own level and conclude accordingly. Furthermore, if someone else carried out the test and simply reported that H_0 was rejected at significance level .05 without revealing the P-value, then anyone wishing to use a smaller significance level would not know which conclusion is appropriate. That individual would have imposed his or her own significance level on other decision makers. Access to the P-value prevents such an imposition.

A final point concerning the utility of the P-value is that it allows one to distinguish between a close call and a very clear-cut conclusion at any particular significance level. For example, suppose you are told that H_0 was rejected at significance level .05. This conclusion is consistent with a P-value of .0498 and also with a P-value of .0003, since in each case P-value $\le \alpha = .05$. But of course with a P-value of .0498, the null hypothesis is barely rejected, whereas with P-value = .0003, the null hypothesis is rejected by a country mile. So it is always preferable to report the P-value rather than just stating the conclusion at a particular significance level.

Unfortunately most journal articles containing summaries of hypothesis testing analyses do not report exact P-values. Instead what typically appears is one of the following statements: "$P < .05$" if the P-value is between .05 and .01, "$P < .01$" if it is between .01 and .001, and "$P < .001$" if the P-value really is smaller than .001. In a tabular summary, you will often see *, **, and *** corresponding to these three cases.

Proof of the proposition stating that P(type I error) = the significance level α:

Denote the test statistic by Y, and let $F(\cdot)$ be the cumulative distribution function of Y when H_0 is true (e.g., F might be the standard normal cdf Φ or the cdf of an rv having a t distribution with some specified number of df). Suppose the distribution of Y is continuous over some interval (often infinite in extent) so that F is a strictly increasing function over this interval. Then F has a well-defined inverse function F^{-1}. Consider the case in which only values of the test statistic smaller than the calculated value y are more contradictory to H_0 than y itself. This implies that

$$P\text{-value} = P(\text{obtaining a test statistic value at least as contradictory to } H_0 \text{ when } H_0 \text{ is true}) = F(y)$$

Now before the sample data is available, the value of the test statistic is a random variable Y, and so the P-value itself is a random variable. Thus

$$P(\text{type I error}) = P(P\text{-value} \le \alpha \text{ when } H_0 \text{ is true}) = P(F(Y) \le \alpha)$$

Let's now apply F^{-1} to both sides of the inequality inside the last set of parentheses:

$$P(\text{type I error}) = P[F^{-1}(F(Y)) \le F^{-1}(\alpha)] = P(Y \le F^{-1}(\alpha)) = F(F^{-1}(\alpha)) = \alpha$$

The argument in the case in which only values of Y larger than y are more contradictory to H_0 than y itself is similar to what we have just shown. The case in which either large or small Y values are more contradictory to H_0 than y itself is a bit trickier. And when the test statistic has a discrete distribution, the inverse function F^{-1} is not uniquely defined, so extra care is needed to make the argument valid.

EXERCISES Section 8.1 (1–14)

1. For each of the following assertions, state whether it is a legitimate statistical hypothesis and why:
   a. $H: \sigma > 100$ **b.** $H: \tilde{x} = 45$
   c. $H: s \leq .20$ **d.** $H: \sigma_1/\sigma_2 < 1$
   e. $H: \overline{X} - \overline{Y} = 5$
   f. $H: \lambda \leq .01$, where λ is the parameter of an exponential distribution used to model component lifetime

2. For the following pairs of assertions, indicate which do not comply with our rules for setting up hypotheses and why (the subscripts 1 and 2 differentiate between quantities for two different populations or samples):
   a. $H_0: \mu = 100, H_a: \mu > 100$
   b. $H_0: \sigma = 20, H_a: \sigma \leq 20$
   c. $H_0: p \neq .25, H_a: p = .25$
   d. $H_0: \mu_1 - \mu_2 = 25, H_a: \mu_1 - \mu_2 > 100$
   e. $H_0: S_1^2 = S_2^2, H_a: S_1^2 \neq S_2^2$
   f. $H_0: \mu = 120, H_a: \mu = 150$
   g. $H_0: \sigma_1/\sigma_2 = 1, H_a: \sigma_1/\sigma_2 \neq 1$
   h. $H_0: p_1 - p_2 = -.1, H_a: p_1 - p_2 < -.1$

3. For which of the given P-values would the null hypothesis be rejected when performing a level .05 test?
   a. .001 **b.** .021 **c.** .078
   d. .047 **e.** .148

4. Pairs of P-values and significance levels, α, are given. For each pair, state whether the observed P-value would lead to rejection of H_0 at the given significance level.
   a. P-value = .084, α = .05
   b. P-value = .003, α = .001
   c. P-value = .498, α = .05
   d. P-value = .084, α = .10
   e. P-value = .039, α = .01
   f. P-value = .218, α = .10

5. To determine whether the pipe welds in a nuclear power plant meet specifications, a random sample of welds is selected, and tests are conducted on each weld in the sample. Weld strength is measured as the force required to break the weld. Suppose the specifications state that mean strength of welds should exceed 100 lb/in^2; the inspection team decides to test $H_0: \mu = 100$ versus $H_a: \mu > 100$. Explain why it might be preferable to use this H_a rather than $\mu < 100$.

6. Let μ denote the true average radioactivity level (picocuries per liter). The value 5 pCi/L is considered the dividing line between safe and unsafe water. Would you recommend testing $H_0: \mu = 5$ versus $H_a: \mu > 5$ or $H_0: \mu = 5$ versus $H_a: \mu < 5$? Explain your reasoning. [*Hint:* Think about the consequences of a type I and type II error for each possibility.]

7. Before agreeing to purchase a large order of polyethylene sheaths for a particular type of high-pressure oil-filled submarine power cable, a company wants to see conclusive evidence that the true standard deviation of sheath thickness is less than .05 mm. What hypotheses should be tested, and why? In this context, what are the type I and type II errors?

8. Many older homes have electrical systems that use fuses rather than circuit breakers. A manufacturer of 40-amp fuses wants to make sure that the mean amperage at which its fuses burn out is in fact 40. If the mean amperage is lower than 40, customers will complain because the fuses require replacement too often. If the mean amperage is higher than 40, the manufacturer might be liable for damage to an electrical system due to fuse malfunction. To verify the amperage of the fuses, a sample of fuses is to be selected and inspected. If a hypothesis test were to be performed on the resulting data, what null and alternative hypotheses would be of interest to the manufacturer? Describe type I and type II errors in the context of this problem situation.

9. Water samples are taken from water used for cooling as it is being discharged from a power plant into a river. It has been determined that as long as the mean temperature of the discharged water is at most 150°F, there will be no negative effects on the river's ecosystem. To investigate whether the plant is in compliance with regulations that prohibit a mean discharge water temperature above 150°, 50 water samples will be taken at randomly selected times and the temperature of each sample recorded. The resulting data will be used to test the hypotheses $H_0: \mu = 150°$ versus $H_a: \mu > 150°$. In the context of this situation, describe type I and type II errors. Which type of error would you consider more serious? Explain.

10. A regular type of laminate is currently being used by a manufacturer of circuit boards. A special laminate has been developed to reduce warpage. The regular laminate will be used on one sample of specimens and the special laminate on another sample, and the amount of warpage will then be determined for each specimen. The manufacturer will then switch to the special laminate only if it can be demonstrated that the true average amount of warpage for that laminate is less than for the regular laminate. State the relevant hypotheses, and describe the type I and type II errors in the context of this situation.

11. Two different companies have applied to provide cable television service in a certain region. Let p denote the proportion of all potential subscribers who favor the first company over the second. Consider testing $H_0: p = .5$ versus $H_a: p \neq .5$ based on a random sample of 25 individuals. Let the test statistic X be the number in the

sample who favor the first company and x represent the observed value of X.

a. Describe type I and II errors in the context of this problem situation.

b. Suppose that $x = 6$. Which values of X are at least as contradictory to H_0 as this one?

c. What is the probability distribution of the test statistic X when H_0 is true? Use it to compute the P-value when $x = 6$.

d. If H_0 is to be rejected when P-value $\leq .044$, compute the probability of a type II error when $p = .4$, again when $p = .3$, and also when $p = .6$ and $p = .7$. [*Hint:* P-value $> .044$ is equivalent to what inequalities involving x (see Example 8.4)?]

e. Using the test procedure of (d), what would you conclude if 6 of the 25 queried favored company 1?

12. A mixture of pulverized fuel ash and Portland cement to be used for grouting should have a compressive strength of more than 1300 KN/m^2. The mixture will not be used unless experimental evidence indicates conclusively that the strength specification has been met. Suppose compressive strength for specimens of this mixture is normally distributed with $\sigma = 60$. Let μ denote the true average compressive strength.

a. What are the appropriate null and alternative hypotheses?

b. Let $\overline{X}$ denote the sample average compressive strength for $n = 10$ randomly selected specimens. Consider the test procedure with test statistic $\overline{X}$ itself (not standardized). If $\overline{x} = 1340$, should H_0 be rejected using a significance level of .01? [*Hint:* What is the probability distribution of the test statistic when H_0 is true?]

c. What is the probability distribution of the test statistic when $\mu = 1350$? For a test with $\alpha = .01$, what is the probability that the mixture will be judged unsatisfactory when in fact $\mu = 1350$ (a type II error)?

13. The calibration of a scale is to be checked by weighing a 10-kg test specimen 25 times. Suppose that the results of different weighings are independent of one another and that the weight on each trial is normally distributed with $\sigma = .200$ kg. Let μ denote the true average weight reading on the scale.

a. What hypotheses should be tested?

b. With the sample mean itself as the test statistic, what is the P-value when $\overline{x} = 9.85$, and what would you conclude at significance level .01?

c. For a test with $\alpha = .01$, what is the probability that recalibration is judged unnecessary when in fact $\mu = 10.1$? When $\mu = 9.8$?

14. A new design for the braking system on a certain type of car has been proposed. For the current system, the true average braking distance at 40 mph under specified conditions is known to be 120 ft. It is proposed that the new design be implemented only if sample data strongly indicates a reduction in true average braking distance for the new design.

a. Define the parameter of interest and state the relevant hypotheses.

b. Suppose braking distance for the new system is normally distributed with $\sigma = 10$. Let $\overline{X}$ denote the sample average braking distance for a random sample of 36 observations. Which values of $\overline{x}$ are more contradictory to H_0 than 117.2, what is the P-value in this case, and what conclusion is appropriate if $\alpha = .10$?

c. What is the probability that the new design is not implemented when its true average braking distance is actually 115 ft and the test from part (b) is used?

8.2 z Tests for Hypotheses about a Population Mean

Recall from the previous section that a conclusion in a hypothesis testing analysis is reached by proceeding as follows:

i. Compute the value of an appropriate test statistic.

ii. Then determine the P-value—the probability, calculated assuming that the null hypothesis H_0 true, of observing a test statistic value at least as contradictory to H_0 as what resulted from the available data.

iii. Reject the null hypothesis if P-value $\leq \alpha$, where α is the specified or chosen significance level, i.e., the probability of a type I error (rejecting H_0 when it is true); if P-value $> \alpha$, there is not enough evidence to justify rejecting H_0 (it is still deemed plausible).

Determination of the P-value depends on the distribution of the test statistic when H_0 is true. In this section we describe z tests for testing hypotheses about a single population mean μ. By "z test," we mean that the test statistic has at least approximately a

standard normal distribution when H_0 is true. The P-value will then be a z curve area which depends on whether the inequality in H_a is $>$, $<$, or $\neq$.

In the development of confidence intervals for μ in Chapter 7, we first considered the case in which the population distribution is normal with known σ, then relaxed the normality and known σ assumptions when the sample size n is large, and finally described the one-sample t CI for the mean of a normal population. In this section we discuss the first two cases, and then present the one-sample t test in Section 8.3.

A Normal Population Distribution with Known σ

Although the assumption that the value of σ is known is rarely met in practice, this case provides a good starting point because of the ease with which general procedures and their properties can be developed. The null hypothesis in all three cases will state that μ has a particular numerical value, the *null value*. We denote this value by the symbol μ_0, so the null hypothesis has the form H_0: $\mu = \mu_0$. Let $X_1, \ldots, X_n$ represent a random sample of size n from the normal population. Then the sample mean $\overline{X}$ has a normal distribution with expected value $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. When H_0 is true, $\mu_{\overline{X}} = \mu_0$. Consider now the statistic Z obtained by standardizing $\overline{X}$ under the assumption that H_0 is true:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Substitution of the computed sample mean $\bar{x}$ gives z, the distance between $\bar{x}$ and μ_0 expressed in "standard deviation units." For example, if the null hypothesis is H_0: $\mu = 100$, $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 10/\sqrt{25} = 2.0$, and $\bar{x} = 103$, then the test statistic value is $z = (103 - 100)/2.0 = 1.5$. That is, the observed value of $\bar{x}$ is 1.5 standard deviations (of $\overline{X}$) larger than what we expect it to be when H_0 is true. The statistic Z is a natural measure of the distance between $\overline{X}$, the estimator of μ, and its expected value when H_0 is true. If this distance is too great in a direction consistent with H_a, there is substantial evidence that H_0 is false.

Suppose first that the alternative hypothesis is of the form H_a: $\mu > \mu_0$. Then an $\bar{x}$ value that considerably exceeds μ_0 provides evidence against H_0. Such an $\bar{x}$ value corresponds to a large positive value z. This in turn implies that any value *exceeding* the calculated z is more contradictory to H_0 than is z itself. It follows that

$$P\text{-value} = P(Z \geq z \text{ when } H_0 \text{ is true})$$

Now here is the key point: when H_0 is true, the test statistic Z has a standard normal distribution—because we created Z by standardizing $\overline{X}$ assuming that H_0 is true (i.e., by subtracting μ_0). The implication is that in this case, the P-value is just the area under the standard normal curve to the right of z. Because of this, the test is referred to as *upper-tailed*. For example, in the previous paragraph we calculated $z = 1.5$. If in the alternative hypothesis there is H_a: $\mu > 100$, then P-value = area under the z curve to the right of $1.5 = 1 - \Phi(1.50) = .0668$. At significance level .05 we would not be able to reject the null hypothesis because the P-value exceeds α.

Now consider an alternative hypothesis of the form H_a: $\mu < \mu_0$. In this case any value of the sample mean smaller than our $\bar{x}$ is even more contradictory to the null hypothesis. Thus any test statistic value *smaller* than the calculated z is more contradictory to H_0 than is z itself. It follows that

$$\begin{aligned} P\text{-value} &= P(Z \leq z \text{ when } H_0 \text{ is true}) \\ &= \text{area under the standard normal curve to the left of } z = \Phi(z) \end{aligned}$$

The test in this case is customarily referred to as *lower-tailed.* If, for example, the alternative hypothesis is H_a: $\mu < 100$ and $z = -2.75$, then P-value $= \Phi(-2.75) =$ .0030. This is small enough to justify rejection of H_0 at a significance level of either .05 or .01, but not .001.

The third possible alternative, H_a: $\mu \neq \mu_0$, requires a bit more careful thought. Suppose, for example, that the null value is 100 and that $\bar{x} = 103$ results in $z = 1.5$. Then any $\bar{x}$ value exceeding 103 is more contradictory to H_0 than is 103 itself. So any z exceeding 1.5 is likewise more contradictory to H_0 than is 1.5. However, 97 is just as contradictory to the null hypothesis as is 103, since it is the same distance below 100 as 103 is above 100. Thus $z = -1.5$ is just as contradictory to H_0 as is $z = 1.5$. Therefore any z smaller than -1.5 is more contradictory to H_0 than is 1.5 or -1.5. It follows that

$$\begin{aligned} P\text{-value} &= P(Z \text{ either } \geq 1.5 \text{ or } \leq -1.5 \text{ when } H_0 \text{ is true}) \\ &= (\text{area under the } z \text{ curve to the right of } 1.5) \\ &\quad + (\text{area under the } z \text{ curve to the left of } -1.5) \\ &= 1 - \Phi(1.5) + \Phi(-1.5) = 2[1 - \Phi(1.5)] \\ &= 2(.0668) = .1336 \end{aligned}$$

This would also be the P-value if $\bar{x} = 97$ results in $z = -1.5$. The important point is that because of the inequality $\neq$ in H_a, the P-value is the sum of an upper-tail area and a lower-tail area. By symmetry of the standard normal distribution, this becomes twice the area captured in the tail in which z falls. Equivalently, it is twice the area captured in the upper tail by $|z|$, i.e., $2[1 - \Phi(|z|)]$. It is natural to refer to this test as being *two-tailed* because z values far out in either tail of the z curve argue for rejection of H_0.

The test procedure is summarized in the accompanying box, and the P-value for each of the possible alternative hypotheses is illustrated in Figure 8.4.

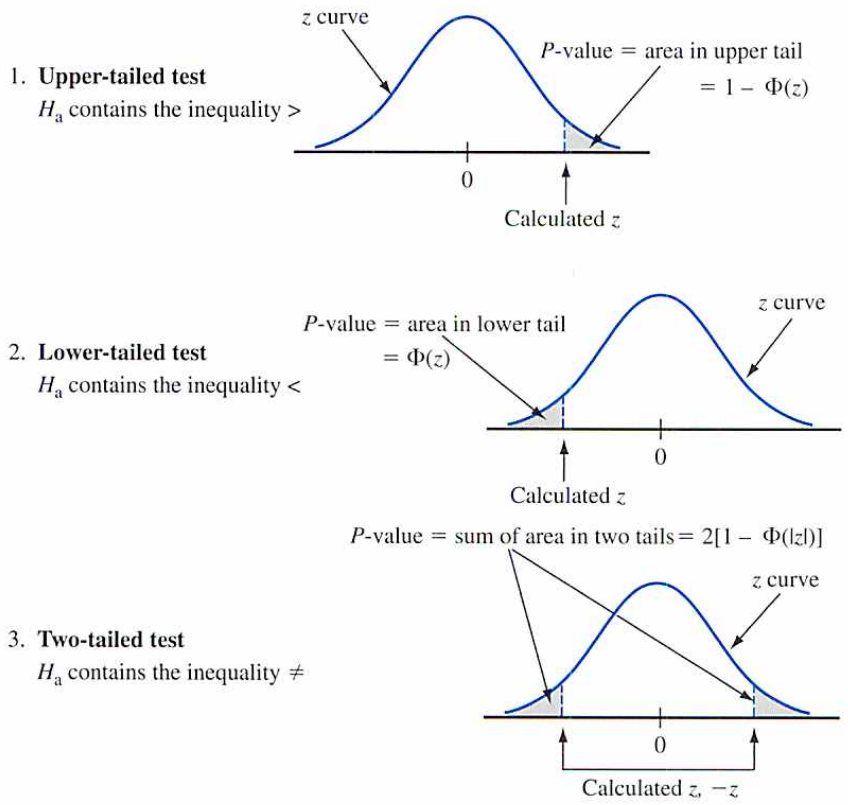


Figure 8.4 Determination of the P-value for a z test

Null hypothesis: H_0: $\mu = \mu_0$

Test statistic: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	P-Value Determination
H_a: $\mu > \mu_0$	Area under the standard normal curve to the right of z
H_a: $\mu < \mu_0$	Area under the standard normal curve to the left of z
H_a: $\mu \neq \mu_0$	$2 \cdot$ (area under the standard normal curve to the right of $\lvert z \rvert$)

Assumptions: A normal population distribution with known value of σ.

Use of the following sequence of steps is recommended when testing hypotheses about a parameter. The plausibility of any assumptions underlying use of the selected test procedure should of course be checked before carrying out the test.

1. Identify the parameter of interest and describe it in the context of the problem situation.
2. Determine the null value and state the null hypothesis.
3. State the appropriate alternative hypothesis.
4. Give the formula for the computed value of the test statistic (substituting the null value and the known values of any other parameters, but *not* those of any sample-based quantities).
5. Compute any necessary sample quantities, substitute into the formula for the test statistic value, and compute that value.
6. Determine the P-value.
7. Compare the selected or specified significance level to the P-value to decide whether H_0 should be rejected, and state this conclusion in the problem context.

The formulation of hypotheses (Steps 2 and 3) should be done before examining the data, and the significance level α should be chosen prior to determination of the P-value.

EXAMPLE 8.6 A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°. A sample of $n = 9$ systems, when tested, yields a sample average activation temperature of 131.08°F. If the distribution of activation times is normal with standard deviation 1.5°F, does the data contradict the manufacturer's claim at significance level $\alpha = .01$?

1. Parameter of interest: μ = true average activation temperature.
2. Null hypothesis: H_0: $\mu = 130$ (null value = $\mu_0 = 130$).
3. Alternative hypothesis: H_a: $\mu \neq 130$ (a departure from the claimed value in *either* direction is of concern).
4. Test statistic value:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{n}}$$

5. Substituting $n = 9$ and $\bar{x} = 131.08$,

$$z = \frac{131.08 - 130}{1.5/\sqrt{9}} = \frac{1.08}{.5} = 2.16$$

That is, the observed sample mean is a bit more than 2 standard deviations above what would have been expected were H_0 true.

6. The inequality in H_a implies that the test is two-tailed, so the P-value results from doubling the captured tail area:

$$P\text{-value} = 2[1 - \Phi(2.16)] = 2(.0154) = .0308$$

7. Because P-value $= .0308 > .01 = \alpha$, H_0 cannot be rejected at significance level .01. The data does not give strong support to the claim that the true average differs from the design value of 130. ■

β and Sample Size Determination The z tests with known σ are among the few in statistics for which there are simple formulas available for β, the probability of a type II error. Consider first the alternative H_a: $\mu > \mu_0$. The null hypothesis is rejected if P-value $\leq \alpha$, and the P-value is the area under the standard normal curve to the right of z. Suppose that $\alpha = .05$. The z critical value that captures an upper-tail area of .05 is $z_{.05} = 1.645$ (look for a cumulative area of .95 in Table A.3). Thus if the calculated test statistic value z is smaller than 1.645, the area to the right of z will be larger than .05 and the null hypothesis will then *not* be rejected. Now substitute $(\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ in place of z in the inequality $z < 1.645$ and manipulate to isolate $\bar{x}$ on the left (multiply both sides by $\sigma/\sqrt{n}$ and then add μ_0 to both sides). This gives the equivalent inequality $\bar{x} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$. Now let μ' denote a particular value of μ that exceeds the null value μ_0. Then,

$$\begin{aligned}
\beta(\mu') &= P(H_0 \text{ is not rejected when } \mu = \mu') \\
&= P(\bar{X} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n} \text{ when } \mu = \mu') \\
&= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu = \mu'\right) \\
&= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

As μ' increases, $\mu_0 - \mu'$ becomes more negative, so $\beta(\mu')$ will be small when μ' greatly exceeds μ_0 (because the value at which Φ is evaluated will then be quite negative). Error probabilities for the lower-tailed and two-tailed tests are derived in an analogous manner.

If σ is large, the probability of a type II error can be large at an alternative value μ' that is of particular concern to an investigator. Suppose we fix α and also specify β for such an alternative value. In the sprinkler example, company officials might view $\mu' = 132$ as a very substantial departure from H_0: $\mu = 130$ and therefore wish $\beta(132) = .10$ in addition to $\alpha = .01$. More generally, consider the two restrictions $P(\text{type I error}) = \alpha$ and $\beta(\mu') = \beta$ for specified α, μ', and β. Then for an upper-tailed test, the sample size n should be chosen to satisfy

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

This implies that

$$-z_\beta = \begin{matrix} z \text{ critical value that} \\ \text{captures lower-tail area } \beta \end{matrix} = z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}$$

This equation is easily solved for the desired n. A parallel argument yields the necessary sample size for lower- and two-tailed tests as summarized in the next box.

Alternative Hypothesis	Type II Error Probability $\beta(\mu')$ for a Level α Test
H_a: $\mu > \mu_0$	$\Phi\left(z_\alpha + \dfrac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
H_a: $\mu < \mu_0$	$1 - \Phi\left(-z_\alpha + \dfrac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
H_a: $\mu \neq \mu_0$	$\Phi\left(z_{\alpha/2} + \dfrac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \dfrac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$

where $\Phi(z) =$ the standard normal cdf.

The sample size n for which a level α test also has $\beta(\mu') = \beta$ at the alternative value μ' is

$$n = \begin{cases} \left[\dfrac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{for a one-tailed (upper or lower) test} \\ \left[\dfrac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{for a two-tailed test (an approximate solution)} \end{cases}$$

EXAMPLE 8.7 Let μ denote the true average tread life of a certain type of tire. Consider testing H_0: $\mu = 30{,}000$ versus H_a: $\mu > 30{,}000$ based on a sample of size $n = 16$ from a normal population distribution with $\sigma = 1500$. A test with $\alpha = .01$ requires $z_\alpha = z_{.01} = 2.33$. The probability of making a type II error when $\mu = 31{,}000$ is

$$\beta(31{,}000) = \Phi\left(2.33 + \frac{30{,}000 - 31{,}000}{1500/\sqrt{16}}\right) = \Phi(-.34) = .3669$$

Since $z_{.1} = 1.28$, the requirement that the level .01 test also have $\beta(31{,}000) = .1$ necessitates

$$n = \left[\frac{1500(2.33 + 1.28)}{30{,}000 - 31{,}000}\right]^2 = (-5.42)^2 = 29.32$$

The sample size must be an integer, so $n = 30$ tires should be used. ■

Large-Sample Tests

When the sample size is large, the foregoing z tests are easily modified to yield valid test procedures without requiring either a normal population distribution or known σ. The key result was used in Chapter 7 to justify large-sample confidence intervals: A large n implies that the standardized variable

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has *approximately* a standard normal distribution. Substitution of the null value μ_0 in place of μ yields the test statistic

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

which has approximately a standard normal distribution when H_0 is true. The P-value is then determined exactly as was previously described in this section (e.g., $\Phi(z)$ when the alternative hypothesis is H_a: $\mu < \mu_0$). Rejecting H_0 when P-value $\leq \alpha$ gives a test with *approximate* significance level α. The rule of thumb $n > 40$ will again be used to characterize a large sample size.

EXAMPLE 8.8 A dynamic cone penetrometer (DCP) is used for measuring material resistance to penetration (mm/blow) as a cone is driven into pavement or subgrade. Suppose that for a particular application it is required that the true average DCP value for a certain type of pavement be less than 30. The pavement will not be used unless there is conclusive evidence that the specification has been met. Let's state and test the appropriate hypotheses using the following data **("Probabilistic Model for the Analysis of Dynamic Cone Penetrometer Test Values in Pavement Structure Evaluation," *J. of Testing and Evaluation*, 1999: 7–14):**

14.1	14.5	15.5	16.0	16.0	16.7	16.9	17.1	17.5	17.8
17.8	18.1	18.2	18.3	18.3	19.0	19.2	19.4	20.0	20.0
20.8	20.8	21.0	21.5	23.5	27.5	27.5	28.0	28.3	30.0
30.0	31.6	31.7	31.7	32.5	33.5	33.9	35.0	35.0	35.0
36.7	40.0	40.0	41.3	41.7	47.5	50.0	51.0	51.8	54.4
55.0	57.0								

Figure 8.5 shows a descriptive summary obtained from Minitab. The sample mean DCP is less than 30. However, there is a substantial amount of variation in the data (sample coefficient of variation = $s/\bar{x} = .4265$), so the fact that the mean is less than the design specification cutoff may be a consequence just of sampling variability. Notice that the histogram does not resemble at all a normal curve (and a normal probability plot does not exhibit a linear pattern). However, the large-sample z tests do not require a normal population distribution.

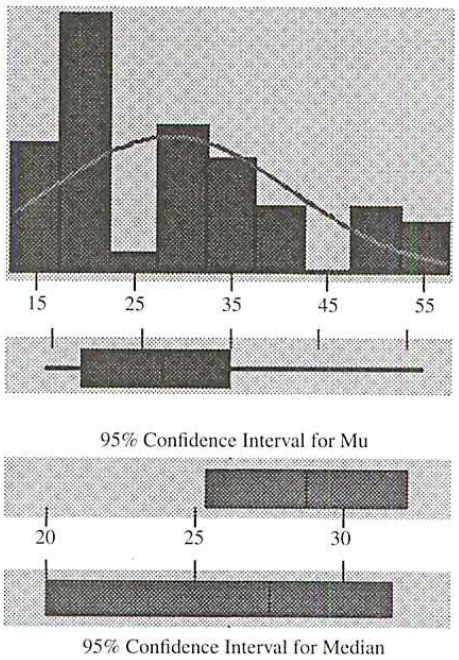


Figure 8.5 Minitab descriptive summary for the DCP data of Example 8.8

1. μ = true average DCP value
2. H_0: $\mu = 30$
3. H_a: $\mu < 30$ (so the pavement will not be used unless the null hypothesis is rejected)
4. $z = \dfrac{\bar{x} - 30}{s/\sqrt{n}}$
5. With $n = 52$, $\bar{x} = 28.76$, and $s = 12.2647$,

$$z = \frac{28.76 - 30}{12.2647/\sqrt{52}} = \frac{-1.24}{1.701} = -.73$$

6. The P-value for this lower-tailed z test is $\Phi(-.73) = .2327$.
7. Since $.2327 > .05$, H_0 cannot be rejected. We do not have compelling evidence for concluding that $\mu < 30$; use of the pavement is not justified. Note that in not rejecting H_0, we might possibly have committed a type II error. ■

Determination of β and the necessary sample size for these large-sample tests can be based either on specifying a plausible value of σ and using the previous formulas (even though s is used in the test) or on using the methodology to be introduced in connection with the one-sample t tests discussed in Section 8.3.

EXERCISES Section 8.2 (15–28)

15. Let μ denote the true average reaction time to a certain stimulus. For a z test of H_0: $\mu = 5$ versus H_a: $\mu > 5$, determine the P-value for each of the following values of the z test statistic.

a. 1.42 **b.** .90 **c.** 1.96 **d.** 2.48 **e.** −.11

16. Newly purchased tires of a particular type are supposed to be filled to a pressure of 30 psi. Let μ denote the true average pressure. A test is to be carried out to decide whether μ differs from the target value. Determine the P-value for each of the following z test statistic values.

a. 2.10 **b.** −1.75 **c.** −.55 **d.** 1.41 **e.** −5.3

17. Answer the following questions for the tire problem in Example 8.7.

a. If $\bar{x} = 30{,}960$ and a level $\alpha = .01$ test is used, what is the decision?

b. If a level .01 test is used, what is $\beta(30{,}500)$?

c. If a level .01 test is used and it is also required that $\beta(30{,}500) = .05$, what sample size n is necessary?

d. If $\bar{x} = 30{,}960$, what is the smallest α at which H_0 can be rejected (based on $n = 16$)?

18. Reconsider the paint-drying situation of Example 8.5, in which drying time for a test specimen is normally distributed with $\sigma = 9$. The hypotheses H_0: $\mu = 75$ versus H_a: $\mu < 75$ are to be tested using a random sample of $n = 25$ observations.

a. How many standard deviations (of $\bar{X}$) below the null value is $\bar{x} = 72.3$?

b. If $\bar{x} = 72.3$, what is the conclusion using $\alpha = .002$?

c. For the test procedure with $\alpha = .002$, what is $\beta(70)$?

d. If the test procedure with $\alpha = .002$ is used, what n is necessary to ensure that $\beta(70) = .01$?

e. If a level .01 test is used with $n = 100$, what is the probability of a type I error when $\mu = 76$?

19. The melting point of each of 16 samples of a certain brand of hydrogenated vegetable oil was determined, resulting in $\bar{x} = 94.32$. Assume that the distribution of the melting point is normal with $\sigma = 1.20$.

a. Test H_0: $\mu = 95$ versus H_a: $\mu \neq 95$ using a two-tailed level .01 test.

b. If a level .01 test is used, what is $\beta(94)$, the probability of a type II error when $\mu = 94$?

c. What value of n is necessary to ensure that $\beta(94) = .1$ when $\alpha = .01$?

20. Lightbulbs of a certain type are advertised as having an average lifetime of 750 hours. The price of these bulbs is very favorable, so a potential customer has decided to go ahead with a purchase arrangement unless it can be conclusively demonstrated that the true average lifetime is smaller than what is advertised. A random sample of 50 bulbs was selected, the lifetime of each bulb determined,

and the appropriate hypotheses were tested using Minitab, resulting in the accompanying output.

```
Variable   N    Mean  StDev  SE Mean      Z  P-Value
lifetime  50  738.44  38.20     5.40  -2.14    0.016
```

What conclusion would be appropriate for a significance level of .05? A significance level of .01? What significance level and conclusion would you recommend?

21. The desired percentage of SiO_2 in a certain type of aluminous cement is 5.5. To test whether the true average percentage is 5.5 for a particular production facility, 16 independently obtained samples are analyzed. Suppose that the percentage of SiO_2 in a sample is normally distributed with $\sigma = .3$ and that $\bar{x} = 5.25$.

a. Does this indicate conclusively that the true average percentage differs from 5.5?

b. If the true average percentage is $\mu = 5.6$ and a level $\alpha = .01$ test based on $n = 16$ is used, what is the probability of detecting this departure from H_0?

c. What value of n is required to satisfy $\alpha = .01$ and $\beta(5.6) = .01$?

22. To obtain information on the corrosion-resistance properties of a certain type of steel conduit, 45 specimens are buried in soil for a 2-year period. The maximum penetration (in mils) for each specimen is then measured, yielding a sample average penetration of $\bar{x} = 52.7$ and a sample standard deviation of $s = 4.8$. The conduits were manufactured with the specification that true average penetration be at most 50 mils. They will be used unless it can be demonstrated conclusively that the specification has not been met. What would you conclude?

23. Automatic identification of the boundaries of significant structures within a medical image is an area of ongoing research. The paper **"Automatic Segmentation of Medical Images Using Image Registration: Diagnostic and Simulation Applications" (*J. of Medical Engr. and Tech.*, 2005: 53–63)** discussed a new technique for such identification. A measure of the accuracy of the automatic region is the average linear displacement (ALD). The paper gave the following ALD observations for a sample of 49 kidneys (units of pixel dimensions).

1.38	0.44	1.09	0.75	0.66	1.28	0.51
0.39	0.70	0.46	0.54	0.83	0.58	0.64
1.30	0.57	0.43	0.62	1.00	1.05	0.82
1.10	0.65	0.99	0.56	0.56	0.64	0.45
0.82	1.06	0.41	0.58	0.66	0.54	0.83
0.59	0.51	1.04	0.85	0.45	0.52	0.58
1.11	0.34	1.25	0.38	1.44	1.28	0.51

a. Summarize/describe the data.

b. Is it plausible that ALD is at least approximately normally distributed? Must normality be assumed prior to calculating a CI for true average ALD or testing hypotheses about true average ALD? Explain.

c. The authors commented that in most cases the ALD is better than or of the order of 1.0. Does the data in fact provide strong evidence for concluding that true average ALD under these circumstances is less than 1.0? Carry out an appropriate test of hypotheses.

d. Calculate an upper confidence bound for true average ALD using a confidence level of 95%, and interpret this bound.

24. Unlike most packaged food products, alcohol beverage container labels are not required to show calorie or nutrient content. The article **"What Am I Drinking? The Effects of Serving Facts Information on Alcohol Beverage Containers" (*J. of Consumer Affairs*, 2008: 81–99)** reported on a pilot study in which each of 58 individuals in a sample was asked to estimate the calorie content of a 12-oz can of beer known to contain 153 calories. The resulting sample mean estimated calorie level was 191 and the sample standard deviation was 89. Does this data suggest that the true average estimated calorie content in the population sampled exceeds the actual content? Test the appropriate hypotheses at significance level .001.

25. Body armor provides critical protection for law enforcement personnel, but it does affect balance and mobility. The article **"Impact of Police Body Armour and Equipment on Mobility" (*Applied Ergonomics*, 2013: 957–961)** reported that for a sample of 52 male enforcement officers who underwent an acceleration task that simulated exiting a vehicle while wearing armor, the sample mean was 1.95 sec, and the sample standard deviation was .20 sec. Does it appear that true average task time is less than 2 sec? Carry out a test of appropriate hypotheses using a significance level of .01.

26. The recommended daily dietary allowance for zinc among males older than age 50 years is 15 mg/day. The article **"Nutrient Intakes and Dietary Patterns of Older Americans: A National Study" (*J. of Gerontology*, 1992: M145–150)** reports the following summary data on intake for a sample of males age 65–74 years: $n = 115$, $\bar{x} = 11.3$, and $s = 6.43$. Does this data indicate that average daily zinc intake in the population of all males ages 65–74 falls below the recommended allowance?

27. Show that for any $\Delta > 0$, when the population distribution is normal and σ is known, the two-tailed test satisfies $\beta(\mu_0 - \Delta) = \beta(\mu_0 + \Delta)$, so that $\beta(\mu')$ is symmetric about μ_0.

28. For a fixed alternative value μ', show that $\beta(\mu') \to 0$ as $n \to \infty$ for either a one-tailed or a two-tailed z test in the case of a normal population distribution with known σ.

8.3 The One-Sample t Test

When n is small, the Central Limit Theorem (CLT) can no longer be invoked to justify the use of a large-sample test. We faced this same difficulty in obtaining a small-sample confidence interval (CI) for μ in Chapter 7. Our approach here will be the same one used there: We will assume that the population distribution is at least approximately normal and describe test procedures whose validity rests on this assumption. If an investigator has good reason to believe that the population distribution is quite nonnormal, a distribution-free test from Chapter 15 may be appropriate. Alternatively, a statistician can be consulted regarding procedures valid for specific families of population distributions other than the normal family. Or a bootstrap procedure can be developed.

The key result on which tests for a normal population mean are based was used in Chapter 7 to derive the one-sample t CI: If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution, the standardized variable

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

has a t distribution with $n - 1$ degrees of freedom (df). Consider testing H_0: $\mu = \mu_0$ using the test statistic $T = (\overline{X} - \mu_0)/(S/\sqrt{n})$. That is, the test statistic results from standardizing $\overline{X}$ under the assumption that H_0 is true (using $S/\sqrt{n}$, the estimated standard deviation of $\overline{X}$, rather than $\sigma/\sqrt{n}$). When H_0 is true, this test statistic has a t distribution with $n - 1$ df. Knowledge of the test statistic's distribution when H_0 is true (the "null distribution") allows us to determine the P-value.

The test statistic is really the same here as in the large-sample case but is labeled T to emphasize that the reference distribution for P-value determination is a t distribution with $n - 1$ df rather than the standard normal (z) distribution. Instead of being a z curve area as was the case for large-sample tests, the P-value will now be an area under the t_{n-1} curve (see Figure 8.6).

The One-Sample t Test

Null hypothesis: H_0: $\mu = \mu_0$

Test statistic value: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	P-Value Determination
H_a: $\mu > \mu_0$	Area under the t_{n-1} curve to the right of t
H_a: $\mu < \mu_0$	Area under the t_{n-1} curve to the left of t
H_a: $\mu \neq \mu_0$	$2 \cdot$ (Area under the t_{n-1} curve to the right of $\lvert t \rvert$)

Assumption: The data consists of a random sample from a normal population distribution.

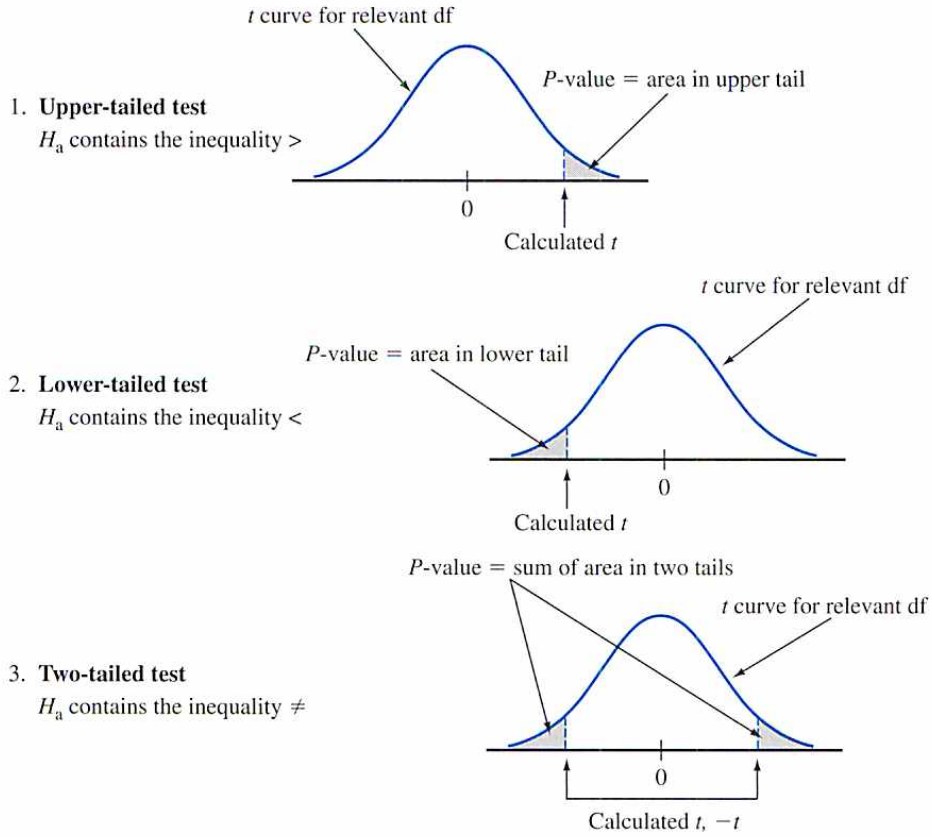


Figure 8.6 *P*-values for *t* tests

Unfortunately the table of *t* critical values that we used for confidence and prediction interval calculations in Chapter 7 does not provide much information about *t* curve tail areas. This is because for each *t* distribution there are values for only the seven most commonly used tail areas: .10, .05, .025, .01, .005, .001, and .0005. *P*-value determination would be straightforward if we had a table of tail areas (or alternatively, cumulative areas) that resembled our *z* table: for each different *t* distribution, the area under the corresponding curve to the right (or the left) of values 0.00, 0.01, 0.02, 0.03, … , 3.97, 3.98, 3.99, and finally 4.00. But this would necessitate an entire page of text for each different *t* distribution.

So we have included another *t* table in Appendix Table A.8. It contains a tabulation of upper-tail *t* curve areas but with less decimal accuracy than what the *z* table provides. Each different column of the table is for a different number of df, and the rows are for calculated values of the test statistic *t* ranging from 0.0 to 4.0 in increments of .1. For example, the number .074 appears at the intersection of the 1.6 row and the 8 df column. Thus the area under the 8 df curve to the right of 1.6 (an upper-tail area) is .074. Because *t* curves are symmetric about 0, .074 is also the area under the 8 df curve to the left of −1.6.

Suppose, for example, that a test of H_0: $\mu = 100$ versus H_a: $\mu > 100$ is based on the 8 df *t* distribution. If the calculated value of the test statistic is $t = 1.6$, then the *P*-value for this upper-tailed test is .074. Because .074 exceeds .05, we would not be able to reject H_0 at a significance level of .05. If the alternative hypothesis is H_a: $\mu < 100$ and a test based on 20 df yields $t = -3.2$, then Appendix Table A.7 shows that the *P*-value is the captured lower-tail area .002. The null hypothesis can be rejected at either level .05 or .01. In the next chapter, we will present a *t* test for hypotheses about a difference between two population means. Suppose the relevant hypotheses are H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 \neq 0$; the null hypothesis states that the means of the two

populations are identical, whereas the alternative hypothesis states that they are different without specifying a direction of departure from H_0. If a t test is based on 20 df and $t = 3.2$, then the P-value for this two-tailed test is $2(.002) = .004$. This would also be the P-value for $t = -3.2$. The tail area is doubled because values both larger than 3.2 and smaller than -3.2 are more contradictory to H_0 than what was calculated (values farther out in *either* tail of the t curve).

EXAMPLE 8.9 Carbon nanofibers have potential application as heat-management materials, for composite reinforcement, and as components for nanoelectronics and photonics. The accompanying data on failure stress (MPa) of fiber specimens was read from a graph in the article **"Mechanical and Structural Characterization of Electrospun PAN-Derived Carbon Nanofibers" (*Carbon*, 2005: 2175–2185).**

300	312	327	368	400	425	470	556	573	575
580	589	626	637	690	715	757	891	900	

Summary quantities include $n = 19, \bar{x} = 562.68, s = 180.874, s/\sqrt{n} = 41.495$. Does the data provide compelling evidence for concluding that true average failure stress exceeds 500 MPa?

Figure 8.7 shows a normal probability plot of the data; the substantial linear pattern indicates that a normal population distribution of failure stress is quite plausible, giving us license to employ the one-sample t test (the box to the right of the plot gives information about a formal test of the hypothesis that the population distribution is normal; this will be discussed in Chapter 14).

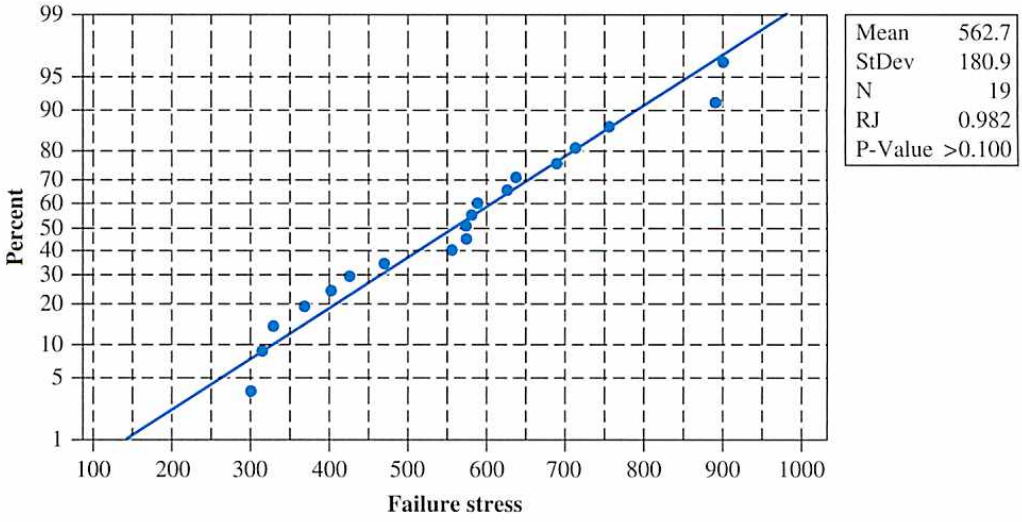


Figure 8.7 Normal probability plot of the failure stress data ■

Let's carry out a test of the relevant hypotheses using a significance level of .05.

1. The parameter of interest is μ = the true average failure stress
2. The null hypothesis is H_0: $\mu = 500$
3. The appropriate alternative hypothesis is H_a: $\mu > 500$ (so we'll believe that true average failure stress exceeds 500 only if the null hypothesis can be rejected).
4. The one-sample t test statistic is $T = (\bar{X} - 500)/(S/\sqrt{n})$. Its value t for the given data results from replacing $\bar{X}$ by $\bar{x}$ and S by s.
5. The test-statistic value is $t = (562.68 - 500)/41.495 = 1.51$

6. The test is based on $19 - 1 = 18$ df. The entry in that column and the 1.5 row of Appendix Table A.8 is .075. Since the test is upper-tailed (because $>$ appears in H_a), it follows that P-value $\approx .075$ (Minitab says .074).
7. Because $.075 > .05$, there is not enough evidence to justify rejecting the null hypothesis at significance level .05. Rather than conclude that the true average failure stress exceeds 500, it appears that sampling variability provides a plausible explanation for the fact that the sample mean exceeds 500 by a rather substantial amount. ■

EXAMPLE 8.10 Many deleterious effects of smoking on health have been well documented. The article **"Smoking Abstinence Impairs Time Estimation Accuracy in Cigarette Smokers" (*Psychopharmacology Bull.*, 2003: 90–95)** described an investigation into whether time perception, an indicator of a person's ability to concentrate, is impaired during nicotine withdrawal. After a 24-hour smoking abstinence, each of 20 smokers was asked to estimate how much time had elapsed during a 45-second period. The following data on perceived elapsed time is consistent with summary quantities given in the cited article.

69	65	72	73	59	55	39	52	67	57
56	50	70	47	56	45	70	64	67	53

A normal probability plot of this data shows a very substantial linear pattern. Let's carry out a test of hypotheses at significance level .05 to decide whether true average perceived elapsed time differs from the known time 45.

1. μ = true average perceived elapsed time for all smokers exposed to the described experimental regimen
2. H_0: $\mu = 45$
3. H_0: $\mu \neq 45$
4. $t = (\bar{x} - 45)/(s/\sqrt{n})$
5. With $\bar{x} = 59.30$ and $s/\sqrt{n} = 9.84/\sqrt{20} = 2.200$, the test statistic value is $t = 14.3/2.200 = 6.50$.
6. The P-value for a two-tailed test is twice the area under the 19 df t curve to the right of 6.50. Since Table A.8 shows that the area under this t curve to the right of 4.0 is 0, the area to the right of 6.50 is certainly 0. The P-value is then $2(0) = 0$ (.00000 according to software).
7. A P-value as small as what we obtained argues very strongly for rejection of H_0 at any reasonable significance level, and in particular at significance level .05. The difference between the sample mean and its expected value when H_0 is true cannot plausibly be explained simply by chance variation. The true average perceived elapsed time is evidently something other than 45, so nicotine withdrawal does appear to impair perception of time. ■

β and Sample Size Determination

The calculation of β at the alternative value μ' for a normal population distribution with known σ was carried out by converting the inequality P-value $> \alpha$ to a statement about $\bar{x}$ (e.g., $\bar{x} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$) and then subtracting μ' to standardize correctly. An equivalent approach involves noting that when $\mu = \mu'$, the test statistic $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ still has a normal distribution with variance 1, but now the mean value of Z is given by $(\mu' - \mu_0)/(\sigma/\sqrt{n})$. That is, when $\mu = \mu'$, the test

statistic still has a normal distribution though not the standard normal distribution. Because of this, $\beta(\mu')$ is an area under the normal curve corresponding to mean value $(\mu' - \mu_0)/(\sigma/\sqrt{n})$ and variance 1. Both α and β involve working with normally distributed variables.

The calculation of $\beta(\mu')$ for the t test is much less straightforward. This is because the distribution of the test statistic $T = (\overline{X} - \mu_0)/(S/\sqrt{n})$ is quite complicated when H_0 is false and H_a is true. Thus, for an upper-tailed test, determining

$$\beta(\mu') = P(T < t_{\alpha, n-1} \text{ when } \mu = \mu' \text{ rather than } \mu_0)$$

involves integrating a very unpleasant density function. This must be done numerically. The results are summarized in graphs of β that appear in Appendix Table A.17. There are four sets of graphs, corresponding to one-tailed tests at level .05 and level .01 and two-tailed tests at the same levels.

To understand how these graphs are used, note first that both β and the necessary sample size n are as before functions not just of the absolute difference $|\mu_0 - \mu'|$ but of $d = |\mu_0 - \mu'|/\sigma$. Suppose, for example, that $|\mu_0 - \mu'| = 10$. This departure from H_0 will be much easier to detect (smaller β) when $\sigma = 2$, in which case μ_0 and μ' are 5 population standard deviations apart, than when $\sigma = 10$. The fact that β for the t test depends on d rather than just $|\mu_0 - \mu'|$ is unfortunate, since to use the graphs one must have some idea of the true value of σ. A conservative (large) guess for σ will yield a conservative (large) value of $\beta(\mu')$ and a conservative estimate of the sample size necessary for prescribed α and $\beta(\mu')$.

Once the alternative μ' and value of σ are selected, d is calculated and its value located on the horizontal axis of the relevant set of curves. The value of β is the height of the $n - 1$ df curve above the value of d (visual interpolation is necessary if $n - 1$ is not a value for which the corresponding curve appears), as illustrated in Figure 8.8.

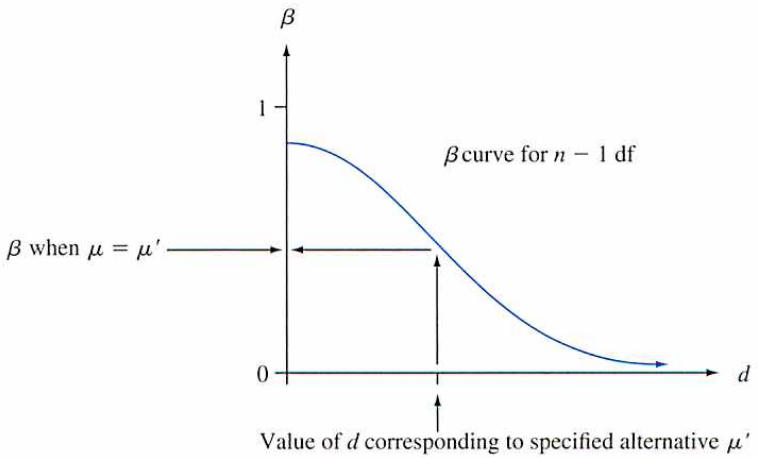


Figure 8.8 A typical β curve for the t test

Rather than fixing n (i.e., $n - 1$, and thus the particular curve from which β is read), one might prescribe both α (.05 or .01 here) and a value of β for the chosen μ' and σ. After computing d, the point (d, β) is located on the relevant set of graphs. The curve below and closest to this point gives $n - 1$ and thus n (again, interpolation is often necessary).

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 β 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 β for such a substantial departure from H_0 and may wish to have $\beta = .05$ for this alternative value of μ. 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 μ is 2.6 and $\sigma = .10$. A larger value of σ would give a larger β for this alternative, and an alternative value of μ closer to 2.5 would also result in an increased value of β. ■

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 β (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 μ 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

              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 β 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 μ moves farther and farther away from the null value.

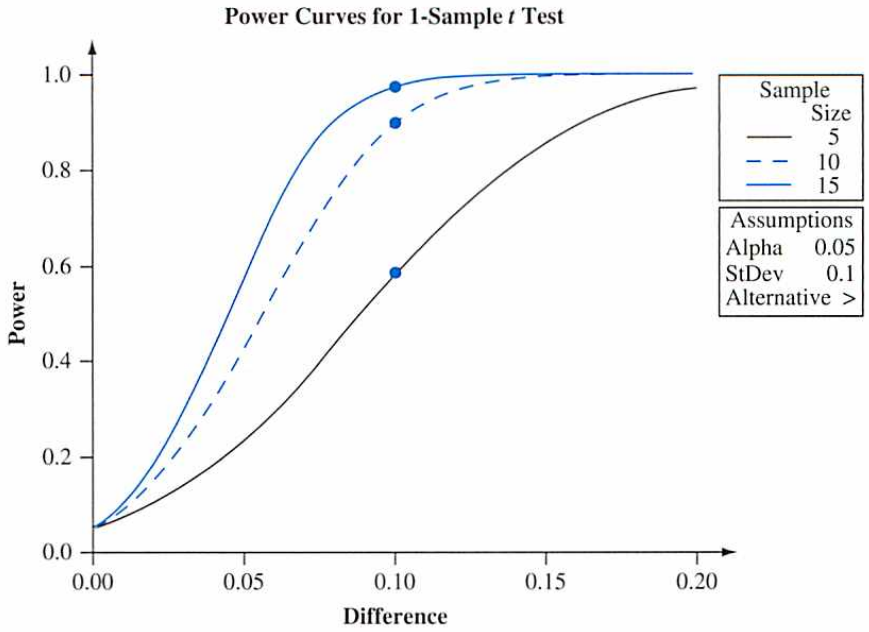


Figure 8.9 Power curves from Minitab for the t test of Example 8.11

Variation in *P*-values

The P-value resulting from carrying out a test on a selected sample is *not* the probability that H_0 is true, nor is it the probability of rejecting the null hypothesis. Once again, it is the probability, calculated assuming that H_0 is true, of obtaining a test statistic value at least as contradictory to the null hypothesis as the value that actually resulted. For example, consider testing H_0: $\mu = 50$ against H_0: $\mu < 50$ using a lower-tailed t test based on 20 df. If the calculated value of the test statistic is $t = -2.00$, then

$$\begin{aligned} P\text{-value} &= P(T < -2.00 \text{ when } \mu = 50) \\ &= \text{area under the } t_{20} \text{ to the left of } -2.00 = .030 \end{aligned}$$

But if a second sample is selected, the resulting value of t will almost surely be different from -2.00, so the corresponding P-value will also likely differ from .030. Because the test statistic value itself varies from one sample to another, the P-value will also vary from one sample to another. That is, the test statistic is a random variable, and so the P-value will also be a random variable. A first sample may give a P-value of .030, a second sample may result in a P-value of .117, a third may yield .061 as the P-value, and so on.

If H_0 is false, we hope the P-value will be close to 0 so that the null hypothesis can be rejected. On the other hand, when H_0 is true, we'd like the P-value to exceed the selected significance level so that the correct decision to not reject H_0 is made. The next example presents simulations to show how the P-value behaves both when the null hypothesis is true and when it is false.

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 μ 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.

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 μ 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

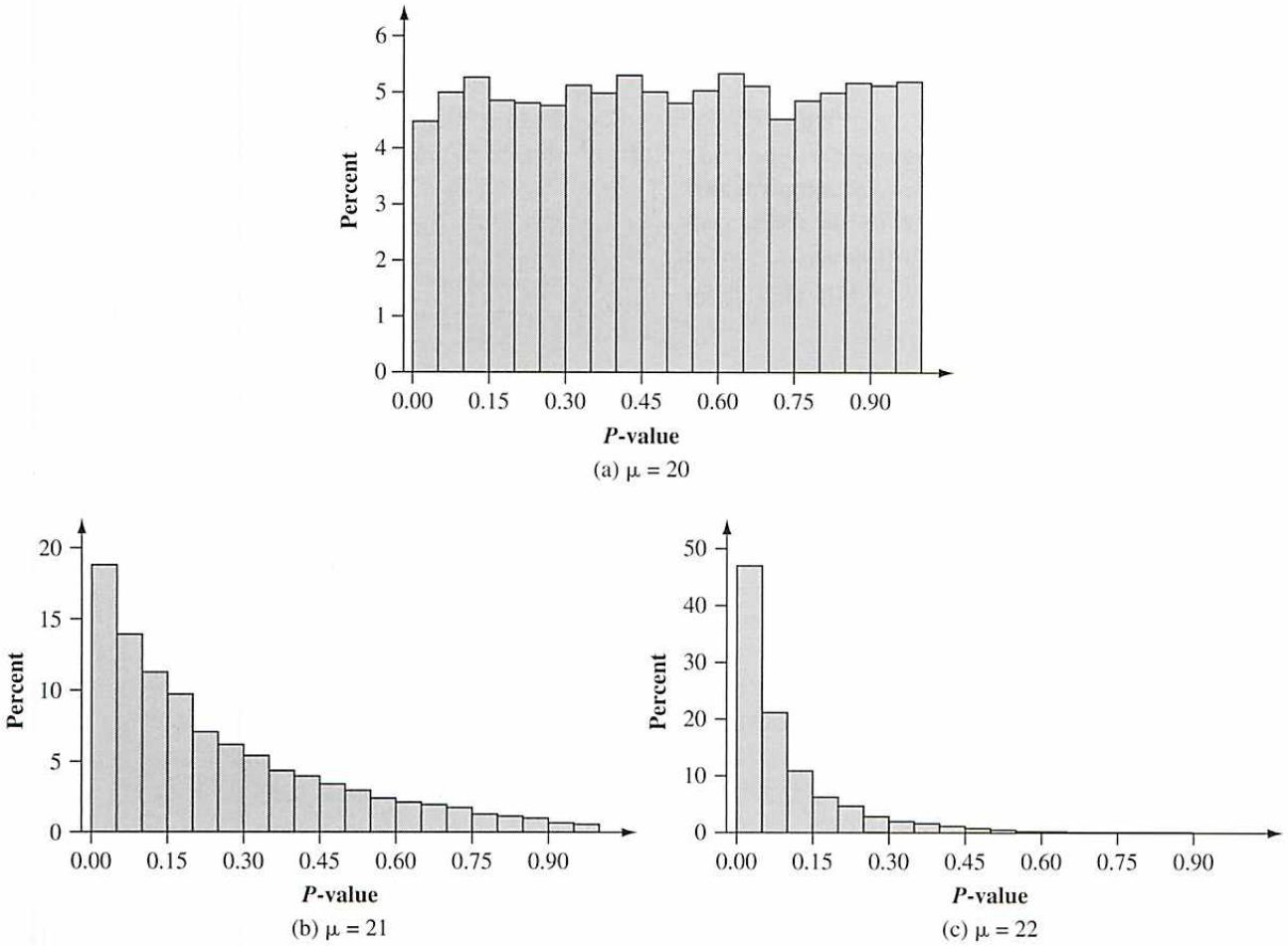


Figure 8.10 *P*-value simulation results for Example 8.12

not that the null hypothesis is incorrectly not rejected. Only for values of μ much larger than 20 (e.g., at least 24 or 25) is it highly likely that the P-value will be smaller than .05 and thus give the correct conclusion.

The big idea of this example is that because the value of any test statistic is random, the P-value will also be a random variable and thus have a distribution. The farther the actual value of the parameter is from the value specified by the null hypothesis, the more the distribution of the P-value will be concentrated on values close to 0 and the greater the chance that the test will correctly reject H_0 (corresponding to smaller β). ■

Whenever the observed value of a statistic such as $\overline{X}$ or $\hat{p}$ is reported, it is good statistical practice to include a quantitative measure of the statistic's precision, e.g., that the estimated standard error of $\overline{X}$ is $s/\sqrt{n}$. The P-value itself is a statistic—its value can be calculated once sample data is available and a particular test procedure is selected, and before such data is in hand, the P-value is subject to randomness. So it would be nice to have available σ_P or an estimate of this standard deviation.

Unfortunately the sampling distribution of a P-value is in general quite complicated. The simulation results of Example 8.12 suggest that the sampling distribution is quite skewed when H_0 is false (it is uniformly distributed on (0,1) when H_0 is true and the test statistic has a continuous distribution, e.g., a t distribution). A standard deviation is not as easy to interpret and use when there is substantial non-normality. The statisticians Dennis Boos and Leonard Stefanski investigated the random behavior of

the P-value in their article **"*P*-Value Precision and Reproducibility" (*The American Statistician*, 2011: 213–221).** To address non-normality, they focused on the quantity –log(P-value). The log-transformed P-value does for many test procedures have approximately a normal distribution when n is large.

Suppose application of a particular test procedure to sample data results in a P-value of .001. Then H_0 would be rejected using either a significance level of .05 or .01. If a new sample from the same population distribution is then selected, how likely is it that the P-value for this new data will lead to rejection of H_0 at a significance level of .05 or .01? This is what the authors of the foregoing article meant by "reproducibility": How likely is it that a new sample will lead to the same conclusion as that reached using the original sample? The answer to this question depends on the population distribution, the sample size, and the test procedure used. Nevertheless, based on their investigations, the authors suggested the following general guidelines:

> If the P-value for the original data is .0001, then P(new P-value $\leq .05) \approx .97$, whereas this probability is roughly .91 if the original P-value is .001 and it is roughly .73 when the original P-value is .01.

Particularly when the original P-value is around .01, there is a reasonably good chance that a new sample will not lead to rejection of H_0 at the 5% significance level. Thus unless the original P-value is really small, it would not be surprising to have a new sample contradict the inference drawn from the original data. A P-value not too much smaller than a chosen significance level such as .05 or .01 should be viewed with some caution!

EXERCISES Section 8.3 (29–41)

29. The true average diameter of ball bearings of a certain type is supposed to be .5 in. A one-sample t test will be carried out to see whether this is the case. What conclusion is appropriate in each of the following situations?

a. $n = 13, t = 1.6, \alpha = .05$

b. $n = 13, t = -1.6, \alpha = .05$

c. $n = 25, t = -2.6, \alpha = .01$

d. $n = 25, t = -3.9$

30. A sample of n sludge specimens is selected and the pH of each one is determined. The one-sample t test will then be used to see if there is compelling evidence for concluding that true average pH is less than 7.0. What conclusion is appropriate in each of the following situations?

a. $n = 6, t = -2.3, \alpha = .05$

b. $n = 15, t = -3.1, \alpha = .01$

c. $n = 12, t = -1.3, \alpha = .05$

d. $n = 6, t = .7, \alpha = .05$

e. $n = 6, \bar{x} = 6.68, s/\sqrt{n} = .0820$

31. The paint used to make lines on roads must reflect enough light to be clearly visible at night. Let μ denote the true average reflectometer reading for a new type of paint under consideration. A test of H_0: $\mu = 20$ versus H_a: $\mu > 20$ will be based on a random sample of size n from a normal population distribution. What conclusion is appropriate in each of the following situations?

a. $n = 15, t = 3.2, \alpha = .05$

b. $n = 9, t = 1.8, \alpha = .01$

c. $n = 24, t = -.2$

32. The relative conductivity of a semiconductor device is determined by the amount of impurity "doped" into the device during its manufacture. A silicon diode to be used for a specific purpose requires an average cut-on voltage of .60 V, and if this is not achieved, the amount of impurity must be adjusted. A sample of diodes was selected and the cut-on voltage was determined. The accompanying SAS output resulted from a request to test the appropriate hypotheses.

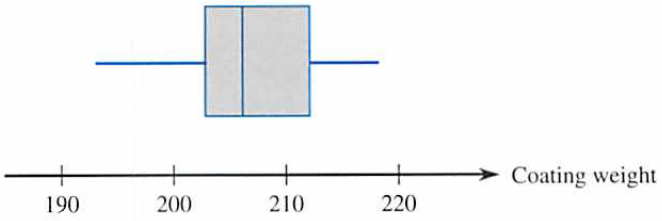
```
N          Mean      Std Dev           T   Prob. > |T|
15   0.0453333   0.0899100   1.9527887        0.0711
```

[*Note:* SAS explicitly tests H_0: $\mu = 0$, so to test H_0: $\mu = .60$, the null value .60 must be subtracted from each x_i; the reported mean is then the average of the $(x_i - .60)$ values. Also, SAS's P-value is always for a two-tailed test.] What would be concluded for a significance level of .01? .05? .10?

33. The article **"The Foreman's View of Quality Control" (*Quality Engr.*, 1990: 257–280)** described an investigation into the coating weights for large pipes resulting from a galvanized coating process. Production standards call for a true average weight of 200 lb per pipe. The accompanying descriptive summary and boxplot are from Minitab.

```
Variable   N    Mean  Median  TrMean  StDev  SEMean
ctg wt    30  206.73  206.00  206.81   6.35    1.16

Variable      Min      Max      Q1      Q3
ctg wt     193.00   218.00   202.75  212.00
```



a. What does the boxplot suggest about the status of the specification for true average coating weight?

b. A normal probability plot of the data was quite straight. Use the descriptive output to test the appropriate hypotheses.

34. The following observations are on stopping distance (ft) of a particular truck at 20 mph under specified experimental conditions **("Experimental Measurement of the Stopping Performance of a Tractor-Semitrailer from Multiple Speeds," NHTSA, DOT HS 811 488, June 2011):**

32.1 30.6 31.4 30.4 31.0 31.9

The cited report states that under these conditions, the maximum allowable stopping distance is 30. A normal probability plot validates the assumption that stopping distance is normally distributed.

a. Does the data suggest that true average stopping distance exceeds this maximum value? Test the appropriate hypotheses using $\alpha = .01$.

b. Determine the probability of a type II error when $\alpha = .01$, $\sigma = .65$, and the actual value of μ is 31. Repeat this for $\mu = 32$ (use either statistical software or Table A.17).

c. Repeat (b) using $\sigma = .80$ and compare to the results of (b).

d. What sample size would be necessary to have $\alpha = .01$ and $\beta = .10$ when $\mu = 31$ and $\sigma = .65$?

35. The article **"Uncertainty Estimation in Railway Track Life-Cycle Cost" (*J. of Rail and Rapid Transit*, 2009)** presented the following data on time to repair (min) a rail break in the high rail on a curved track of a certain railway line.

159 120 480 149 270 547 340 43 228 202 240 218

A normal probability plot of the data shows a reasonably linear pattern, so it is plausible that the population distribution of repair time is at least approximately normal. The sample mean and standard deviation are 249.7 and 145.1, respectively.

a. Is there compelling evidence for concluding that true average repair time exceeds 200 min? Carry out a test of hypotheses using a significance level of .05.

b. Using $\sigma = 150$, what is the type II error probability of the test used in (a) when true average repair time is actually 300 min? That is, what is $\beta(300)$?

36. Have you ever been frustrated because you could not get a container of some sort to release the last bit of its contents? The article **"Shake, Rattle, and Squeeze: How Much Is Left in That Container?" (*Consumer Reports*, May 2009: 8)** reported on an investigation of this issue for various consumer products. Suppose five 6.0 oz tubes of toothpaste of a particular brand are randomly selected and squeezed until no more toothpaste will come out. Then each tube is cut open and the amount remaining is weighed, resulting in the following data (consistent with what the cited article reported): .53, .65, .46, .50, .37. Does it appear that the true average amount left is less than 10% of the advertised net contents?

a. Check the validity of any assumptions necessary for testing the appropriate hypotheses.

b. Carry out a test of the appropriate hypotheses using a significance level of .05. Would your conclusion change if a significance level of .01 had been used?

c. Describe in context type I and II errors, and say which error might have been made in reaching a conclusion.

37. The accompanying data on cube compressive strength (MPa) of concrete specimens appeared in the article **"Experimental Study of Recycled Rubber-Filled High-Strength Concrete" (*Magazine of Concrete Res.*, 2009: 549–556):**

112.3 97.0 92.7 86.0 102.0
99.2 95.8 103.5 89.0 86.7

a. Is it plausible that the compressive strength for this type of concrete is normally distributed?

b. Suppose the concrete will be used for a particular application unless there is strong evidence that true average strength is less than 100 MPa. Should the concrete be used? Carry out a test of appropriate hypotheses.

38. A random sample of soil specimens was obtained, and the amount of organic matter (%) in the soil was determined for each specimen, resulting in the accompanying data (from **"Engineering Properties of Soil," *Soil Science*, 1998: 93–102).**

1.10 5.09 0.97 1.59 4.60 0.32 0.55 1.45
0.14 4.47 1.20 3.50 5.02 4.67 5.22 2.69
3.98 3.17 3.03 2.21 0.69 4.47 3.31 1.17
0.76 1.17 1.57 2.62 1.66 2.05

The values of the sample mean, sample standard deviation, and (estimated) standard error of the mean are 2.481, 1.616, and .295, respectively. Does this data suggest that the true average percentage of organic matter in such soil is something other than 3%? Carry out a test of the appropriate hypotheses at significance level .10. Would your conclusion be different if $\alpha = .05$ had been used? [*Note:* A normal probability plot of the data shows an acceptable pattern in light of the reasonably large sample size.]

39. Reconsider the accompanying sample data on expense ratio (%) for large-cap growth mutual funds first introduced in Exercise 1.53.

0.52	1.06	1.26	2.17	1.55	0.99	1.10	1.07	1.81	2.05
0.91	0.79	1.39	0.62	1.52	1.02	1.10	1.78	1.01	1.15

A normal probability plot shows a reasonably linear pattern.

a. Is there compelling evidence for concluding that the population mean expense ratio exceeds 1%? Carry out a test of the relevant hypotheses using a significance level of .01.

b. Referring back to (a), describe in context type I and II errors and say which error you might have made in reaching your conclusion. The source from which the data was obtained reported that $\mu = 1.33$ for the population of all 762 such funds. So did you actually commit an error in reaching your conclusion?

c. Supposing that $\sigma = .5$, determine and interpret the power of the test in (a) for the actual value of μ stated in (b).

40. Polymer composite materials have gained popularity because they have high strength to weight ratios and are relatively easy and inexpensive to manufacture. However, their nondegradable nature has prompted development of environmentally friendly composites using natural materials. The article **"Properties of Waste Silk Short Fiber/Cellulose Green Composite Films" (*J. of Composite Materials*, 2012: 123–127)** reported that for a sample of 10 specimens with 2% fiber content, the sample mean tensile strength (MPa) was 51.3 and the sample standard deviation was 1.2. Suppose the true average strength for 0% fibers (pure cellulose) is known to be 48 MPa. Does the data provide compelling evidence for concluding that true average strength for the WSF/cellulose composite exceeds this value?

41. A spectrophotometer used for measuring CO concentration [ppm (parts per million) by volume] is checked for accuracy by taking readings on a manufactured gas (called span gas) in which the CO concentration is very precisely controlled at 70 ppm. If the readings suggest that the spectrophotometer is not working properly, it will have to be recalibrated. Assume that if it is properly calibrated, measured concentration for span gas samples is normally distributed. On the basis of the six readings—85, 77, 82, 68, 72, and 69—is recalibration necessary? Carry out a test of the relevant hypotheses using $\alpha = .05$.

8.4 Tests Concerning a Population Proportion

Let p denote the proportion of individuals or objects in a population who possess a specified property (e.g., college students who graduate without any debt, or computers that do not need service during the warranty period). If an individual or object with the property is labeled a success (S), then p is the population proportion of successes. Tests concerning p will be based on a random sample of size n from the population. Provided that n is small relative to the population size, X (the number of S's in the sample) has (approximately) a binomial distribution. Furthermore, if n itself is large [$np \geq 10$ and $n(1 - p) \geq 10$], both X and the estimator $\hat{p} = X/n$ are approximately normally distributed. We first consider large-sample tests based on this latter fact and then turn to the small-sample case that directly uses the binomial distribution.

Large-Sample Tests

Large-sample tests concerning p are a special case of the more general large-sample procedures for a parameter θ. Let $\hat{\theta}$ be an estimator of θ that is (at least approximately) unbiased and has approximately a normal distribution. The null hypothesis has the form H_0: $\theta = \theta_0$ where θ_0 denotes a number

(the null value) appropriate to the problem context. Suppose that when H_0 is true, the standard deviation of $\hat{\theta}$, $\sigma_{\hat{\theta}}$, involves no unknown parameters. For example, if $\theta = \mu$ and $\hat{\theta} = \overline{X}$, $\sigma_{\hat{\theta}} = \sigma_{\overline{X}} = \sigma/\sqrt{n}$, which involves no unknown parameters only if the value of σ is known. A large-sample test statistic results from standardizing $\hat{\theta}$ under the assumption that H_0 is true (so that $E(\hat{\theta}) = \theta_0$):

$$\text{Test statistic: } Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

If the alternative hypothesis is H_a: $\theta > \theta_0$, an upper-tailed test whose significance level is approximately α has P-value $= 1 - \Phi(z)$. The other two alternatives, H_a: $\theta < \theta_0$ and H_a: $\theta \neq \theta_0$, are tested using a lower-tailed z test and a two-tailed z test, respectively.

In the case $\theta = p$, $\sigma_{\hat{\theta}}$ will not involve any unknown parameters when H_0 is true, but this is atypical. When $\sigma_{\hat{\theta}}$ does involve unknown parameters, it is often possible to use an estimated standard deviation $S_{\hat{\theta}}$ in place of $\sigma_{\hat{\theta}}$ and still have Z approximately normally distributed when H_0 is true (because this substitution does not increase variability in Z by very much). The large-sample test of the previous section furnishes an example of this: Because σ is usually unknown, we use $s_{\hat{\theta}} = s_{\overline{X}} = s/\sqrt{n}$ in place of $\sigma/\sqrt{n}$ in the denominator of z.

The estimator $\hat{p} = X/n$ is unbiased ($E(\hat{p}) = p$) and its standard deviation is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$. These facts along with approximate normality were used in Section 7.2 to obtain a confidence interval for p. When H_0 is true, $E(\hat{p}) = p_0$ and $\sigma_{\hat{p}} = \sqrt{p_0(1-p_0)/n}$, so $\sigma_{\hat{p}}$ does not involve any unknown parameters. It then follows that when n is large and H_0 is true, the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

has approximately a standard normal distribution. The P-value for the test is then a z curve area, just as it was in the case of large-sample z tests concerning μ. Its calculation depends on which of the three inequalities in H_a is under consideration.

Null hypothesis: H_0: $p = p_0$

Test statistic value: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

Alternative Hypothesis	P-Value Determination
H_a: $p > p_0$	Area under the standard normal curve to the right of z
H_a: $p < p_0$	Area under the standard normal curve to the left of z
H_a: $p \neq p_0$	2·(Area under the standard normal curve to the right of $\lvert z \rvert$)

These test procedures are valid provided that $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

They are referred to as *upper-tailed*, *lower-tailed*, and *two-tailed*, respectively, for the three different alternative hypotheses.

EXAMPLE 8.13 Student use of cell phones during class is perceived by many faculty to be an annoying but perhaps harmless distraction. However, the use of a phone to text during an exam is a serious breach of conduct. The article **"The Use and Abuse of Cell Phones and Text Messaging During Class: A Survey of College Students" (*College Teaching*, 2012: 1–9)** reported that 27 of the 267 students in a sample admitted to doing this. Can it be concluded at significance level .001 that more than 5% of all students in the population sampled had texted during an exam?

1. The parameter of interest is the proportion p of the sampled population that has texted during an exam.
2. The null hypothesis is H_0: $p = .05$
3. The alternative hypothesis is H_a: $p > .05$
4. Since $np_0 = 267(.05) = 13.35 \geq 10$ and $nq_0 = 267(.95) = 253.65 \geq 10$, the large-sample z test can be used. The test statistic value is $z = (\hat{p} - .05)/\sqrt{(.05)(.95)/n}$.
5. $\hat{p} = 27/267 = .1011$, from which $z = (.1011 - .05)/\sqrt{(.05)(.95)/267} = .0511/.0133 = 3.84$
6. The P-value for this upper-tailed z test is $1 - \Phi(3.84) < 1 - \Phi(3.49) = .0003$ (software gives .000062).
7. The null hypothesis is resoundingly rejected because P-value $< .0003 \leq .001 = \alpha$. The evidence for concluding that the population percentage of students who text during an exam exceeds 5% is very compelling. The cited article's abstract contained the following comment: "The majority of the students surveyed believe that instructors are largely unaware of the extent to which texting and other cell phone activities engage students in the classroom." Maybe it is time for instructors, administrators, and student leaders to become proactive about this issue ■

β and Sample Size Determination When H_0 is true, the test statistic Z has approximately a standard normal distribution. Now suppose that H_0 is *not* true and that $p = p'$. Then Z still has approximately a normal distribution (because it is a linear function of $\hat{p}$), but its mean value and variance are no longer 0 and 1, respectively. Instead,

$$E(Z) = \frac{p' - p_0}{\sqrt{p_0(1 - p_0)/n}} \qquad V(Z) = \frac{p'(1 - p')/n}{p_0(1 - p_0)/n}$$

The null hypothesis will not be rejected if P-value $> \alpha$. For an upper-tailed z test (inequality $>$ in H_a), we argued previously that this is equivalent to $z < z_\alpha$. The probability of a type II error (not rejecting H_0 when it is false) is $\beta(p') = P(Z < z_\alpha \text{ when } p = p')$. This can be computed by using the given mean and variance to standardize and then referring to the standard normal cdf. In addition, if it is desired that the level α test also have $\beta(p') = \beta$ for a specified value of β, this equation can be solved for the necessary n as in Section 8.2. General expressions for $\beta(p')$ and n are given in the accompanying box.

Alternative Hypothesis	$\beta(p')$
H_a: $p > p_0$	$\Phi\left[\dfrac{p_0 - p' + z_\alpha\sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$
H_a: $p < p_0$	$1 - \Phi\left[\dfrac{p_0 - p' - z_\alpha\sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$

H_a: $p \neq p_0$

$$\Phi\left[\frac{p_0 - p' + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right] - \Phi\left[\frac{p_0 - p' - z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right]$$

The sample size n for which the level α test also satisfies $\beta(p') = \beta$ is

$$n = \begin{cases} \left[\dfrac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p'(1-p')}}{p' - p_0}\right]^2 & \text{one-tailed test} \\ \left[\dfrac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p'(1-p')}}{p' - p_0}\right]^2 & \text{two-tailed test (an approximate solution)} \end{cases}$$

EXAMPLE 8.14 A package-delivery service advertises that at least 90% of all packages brought to its office by 9 A.M. for delivery in the same city are delivered by noon that day. Let p denote the true proportion of such packages that are delivered as advertised and consider the hypotheses H_0: $p = .9$ versus H_a: $p < .9$. If only 80% of the packages are delivered as advertised, how likely is it that a level .01 test based on $n = 225$ packages will detect such a departure from H_0? What should the sample size be to ensure that $\beta(.8) = .01$? With $\alpha = .01$, $p_0 = .9$, $p' = .8$, and $n = 225$,

$$\begin{aligned}\beta(.8) &= 1 - \Phi\left(\frac{.9 - .8 - 2.33\sqrt{(.9)(.1)/225}}{\sqrt{(.8)(.2)/225}}\right) \\ &= 1 - \Phi(2.00) = .0228\end{aligned}$$

Thus the probability that H_0 will be rejected using the test when $p = .8$ is .9772; roughly 98% of all samples will result in correct rejection of H_0.

Using $z_\alpha = z_\beta = 2.33$ in the sample size formula yields

$$n = \left[\frac{2.33\sqrt{(.9)(.1)} + 2.33\sqrt{(.8)(.2)}}{.8 - .9}\right]^2 \approx 266$$

■

Small-Sample Tests

Test procedures when the sample size n is small are based directly on the binomial distribution rather than the normal approximation. Consider the alternative hypothesis H_a: $p > p_0$ and again let X be the number of successes in the sample. Then X is the test statistic. When H_0 is true, X has a binomial distribution with parameters n and p_0, so

$$\begin{aligned}P\text{-value} &= P(X \geq x \text{ when } H_0 \text{ is true}) \\ &= P(X \geq x \text{ when } X \sim \text{Bin}(n, p_0)) \\ &= 1 - P(X \leq x - 1 \text{ when } X \sim \text{Bin}(n, p_0)) \\ &= 1 - B(x - 1; n, p_0)\end{aligned}$$

Because X has a discrete probability distribution, it is usually not possible to obtain a test for which P(type I error) is exactly the desired significance level α (e.g., .05 or .01; refer back to middle of page 323 for an example).

Let p' denote an alternative value of $p(p' > p_0)$. When $p = p'$, $X \sim \text{Bin}(n, p')$. The probability of a type II error is then calculated by expressing the condition P-value $> \alpha$ in the equivalent form $x < c_\alpha$. Then

$$\begin{aligned}\beta(p') &= P(\text{type II error when } p = p') \\ &= P(X < c_\alpha \text{ when } X \sim \text{Bin}(n, p')) = B(c_\alpha - 1; n, p')\end{aligned}$$

That is, $\beta(p')$ is the result of a straightforward binomial probability calculation. The sample size n necessary to ensure that a level α test also has specified β at a particular alternative value p' must be determined by trial and error using the binomial cdf.

Test procedures for H_a: $p < p_0$ and for H_a: $p \neq p_0$ are constructed in a similar manner. In the former case, the P-value is $B(x; n, p_0)$. The P-value when the alternative hypothesis is H_a: $p \neq p_0$ is twice the smaller of the two probabilities $B(x; n, p_0)$ and $1 - B(x - 1; n, p_0)$.

EXAMPLE 8.15 A plastics manufacturer has developed a new type of plastic trash can and proposes to sell them with an unconditional 6-year warranty. To see whether this is economically feasible, 20 prototype cans are subjected to an accelerated life test to simulate 6 years of use. The proposed warranty will be modified only if the sample data strongly suggests that fewer than 90% of such cans would survive the 6-year period. Let p denote the proportion of all cans that survive the accelerated test. The relevant hypotheses are H_0: $p = .9$ versus H_a: $p < .9$. A decision will be based on the test statistic X, the number among the 20 that survive. Because of the inequality in H_a, any value smaller than the observed value x is more contradictory to H_0 than is x itself. Therefore

$$P\text{-value} = P(X \leq x \text{ when } H_0 \text{ is true}) = B(x; 20, .9)$$

From Appendix Table A.1, $B(15; 20, .9) = .043$, whereas $B(16; 20, .9) = .133$. The closest achievable significance level to .05 is therefore .043. Since $B(14; 20, .9) = .011$, H_0 would be rejected at this significance level if the accelerated test results in $x = 14$. It would then be appropriate to modify the proposed warranty. Because P-value $\leq .043$ is equivalent to $x \leq 15$, the probability of a type II error for the alternative value $p' = .8$ is

$$\begin{aligned}\beta(.8) &= P(H_0 \text{ is not rejected when } X \sim \text{Bin}(20, .8)) \\ &= P(X \geq 16 \text{ when } X \sim \text{Bin}(20, .8)) \\ &= 1 - B(15; 20, .8) = 1 - .370 = .630\end{aligned}$$

That is, when $p = .8$, 63% of all samples consisting of $n = 20$ cans would result in H_0 being incorrectly not rejected. This error probability is high because 20 is a small sample size and $p' = .8$ is close to the null value $p_0 = .9$. ■

EXERCISES Section 8.4 (42–52)

42. Consider using a z test to test H_0: $p = .6$. Determine the P-value in each of the following situations.

a. H_a: $p > .6$, $z = 1.47$

b. H_a: $p < .6$, $z = -2.70$

c. H_a: $p \neq .6$, $z = -2.70$

d. H_a: $p < .6$, $z = .25$

43. A common characterization of obese individuals is that their body mass index is at least 30 [BMI = weight/(height)^2, where height is in meters and weight is in kilograms]. The article **"The Impact of Obesity on Illness Absence and Productivity in an Industrial Population of Petrochemical Workers"** (*Annals of Epidemiology*, **2008: 8–14**) reported

that in a sample of female workers, 262 had BMIs of less than 25, 159 had BMIs that were at least 25 but less than 30, and 120 had BMIs exceeding 30. Is there compelling evidence for concluding that more than 20% of the individuals in the sampled population are obese?

a. State and test appropriate hypotheses with a significance level of .05.

b. Explain in the context of this scenario what constitutes type I and II errors.

c. What is the probability of not concluding that more than 20% of the population is obese when the actual percentage of obese individuals is 25%?

44. A manufacturer of nickel-hydrogen batteries randomly selects 100 nickel plates for test cells, cycles them a specified number of times, and determines that 14 of the plates have blistered.

a. Does this provide compelling evidence for concluding that more than 10% of all plates blister under such circumstances? State and test the appropriate hypotheses using a significance level of .05. In reaching your conclusion, what type of error might you have committed?

b. If it is really the case that 15% of all plates blister under these circumstances and a sample size of 100 is used, how likely is it that the null hypothesis of part (a) will not be rejected by the level .05 test? Answer this question for a sample size of 200.

c. How many plates would have to be tested to have $\beta(.15) = .10$ for the test of part (a)?

45. A random sample of 150 recent donations at a certain blood bank reveals that 82 were type A blood. Does this suggest that the actual percentage of type A donations differs from 40%, the percentage of the population having type A blood? Carry out a test of the appropriate hypotheses using a significance level of .01. Would your conclusion have been different if a significance level of .05 had been used?

46. It is known that roughly 2/3 of all human beings have a dominant right foot or eye. Is there also right-sided dominance in kissing behavior? The article **"Human Behavior: Adult Persistence of Head-Turning Asymmetry" (*Nature*, 2003: 771)** reported that in a random sample of 124 kissing couples, both people in 80 of the couples tended to lean more to the right than to the left.

a. If 2/3 of all kissing couples exhibit this right-leaning behavior, what is the probability that the number in a sample of 124 who do so differs from the expected value by at least as much as what was actually observed?

b. Does the result of the experiment suggest that the 2/3 figure is implausible for kissing behavior? State and test the appropriate hypotheses.

47. The article **"Effects of Bottle Closure Type on Consumer Perception of Wine Quality" (*Amer. J. of Enology and Viticulture*, 2007: 182–191)** reported that in a sample of 106 wine consumers, 22 (20.8%) thought that screw tops were an acceptable substitute for natural corks. Suppose a particular winery decided to use screw tops for one of its wines unless there was strong evidence to suggest that fewer than 25% of wine consumers found this acceptable.

a. Using a significance level of .10, what would you recommend to the winery?

b. For the hypotheses tested in (a), describe in context what the type I and II errors would be, and say which type of error might have been committed.

48. With domestic sources of building supplies running low several years ago, roughly 60,000 homes were built with imported Chinese drywall. According to the article **"Report Links Chinese Drywall to Home Problems" (*New York Times*, Nov. 24, 2009)**, federal investigators identified a strong association between chemicals in the drywall and electrical problems, and there is also strong evidence of respiratory difficulties due to the emission of hydrogen sulfide gas. An extensive examination of 51 homes found that 41 had such problems. Suppose these 51 were randomly sampled from the population of all homes having Chinese drywall.

a. Does the data provide strong evidence for concluding that more than 50% of all homes with Chinese drywall have electrical/environmental problems? Carry out a test of hypotheses using $\alpha = .01$.

b. Calculate a lower confidence bound using a confidence level of 99% for the percentage of all such homes that have electrical/environmental problems.

c. If it is actually the case that 80% of all such homes have problems, how likely is it that the test of (a) would not conclude that more than 50% do?

49. A plan for an executive travelers' club has been developed by an airline on the premise that 5% of its current customers would qualify for membership. A random sample of 500 customers yielded 40 who would qualify.

a. Using this data, test at level .01 the null hypothesis that the company's premise is correct against the alternative that it is not correct.

b. What is the probability that when the test of part (a) is used, the company's premise will be judged correct when in fact 10% of all current customers qualify?

50. Each of a group of 20 intermediate tennis players is given two rackets, one having nylon strings and the other synthetic gut strings. After several weeks of playing with the two rackets, each player will be asked to state a preference for one of the two types of strings. Let p denote the proportion of all such players who would prefer gut to nylon, and let X be the number of players in the sample who prefer gut. Because gut strings are more expensive, consider the null hypothesis that at most 50% of all such players prefer gut. We simplify this to $H_0: p = .5$, planning to reject H_0 only if sample evidence strongly favors gut strings.

a. Is a significance level of exactly .05 achievable? If not, what is the largest α smaller than .05 that is achievable?

b. If 60% of all enthusiasts prefer gut, calculate the probability of a type II error using the significance level from part (a). Repeat if 80% of all enthusiasts prefer gut.

c. If 13 out of the 20 players prefer gut, should H_0 be rejected using the significance level of (a)?

51. A manufacturer of plumbing fixtures has developed a new type of washerless faucet. Let $p = P$(a randomly selected faucet of this type will develop a leak within 2 years under normal use). The manufacturer has decided to proceed with production unless it can be determined that p is too large; the borderline acceptable value of p is specified as .10. The manufacturer decides to subject n of these faucets to accelerated testing (approximating 2 years of normal use). With $X =$ the number among the n faucets that leak before the test concludes, production will commence unless the observed X is too large. It is decided that if $p = .10$, the probability of not proceeding should be at most .10, whereas if $p = .30$ the probability of proceeding should be at most .10. Can $n = 10$ be used? $n = 20$? $n = 25$? What are the actual error probabilities for the chosen n?

52. In a sample of 171 students at an Australian university that introduced the use of plagiarism-detection software in a number of courses, 58 students indicated a belief that such software unfairly targets students (**"Student and Staff Perceptions of the Effectiveness of Plagiarism Detection Software,"** ***Australian J. of Educ. Tech.*****, 2008: 222–240).** Does this suggest that a majority of students at the university do not share this belief? Test appropriate hypotheses.

8.5 Further Aspects of Hypothesis Testing

We close this introductory chapter on hypothesis testing by briefly considering a variety of issues involving the use of test procedures: the distinction between statistical and practical significance, the relationship between tests and confidence intervals, the implications of multiple testing, and a general method for deriving test statistics.

Statistical Versus Practical Significance

Statistical significance means simply that the null hypothesis was rejected at the selected significance level. That is, in the judgment of the investigator, any observed discrepancy between the data and what would be expected were H_0 true cannot be explained solely by chance variation. However, a small P-value, which would ordinarily indicate statistical significance, may be the result of a large sample size in combination with a departure from H_0 that has little **practical significance**. In many experimental situations, only departures from H_0 of large magnitude would be worthy of detection, whereas a small departure from H_0 would have little practical significance.

As an example, let μ denote the true average IQ of all children in the very large city of Euphoria. Consider testing H_0: $\mu = 100$ versus H_a: $\mu > 100$ assuming a normal IQ distribution with $\sigma = 15$ (100 is conventionally believed to be the average IQ for all individuals, so parents of Euphorian children might be euphoric to have the null hypothesis rejected). But one IQ point is no big deal, so the value $\mu = 101$ certainly does not represent a departure from H_0 that has practical significance. For a reasonably large sample size n, this μ would lead to an $\bar{x}$ value near 101, so we would not want this sample evidence to argue strongly for rejection of H_0 when $\bar{x} = 101$ is observed. For various sample sizes, Table 8.1 records both the P-value when $\bar{x} = 101$ and also the probability of not rejecting H_0 at level .01 when $\mu = 101$.

The second column in Table 8.1 shows that even for moderately large sample sizes, the P-value resulting from $\bar{x} = 101$ argues very strongly for rejection of H_0, whereas the observed $\bar{x}$ itself suggests that in practical terms the true value of μ differs

Table 8.1 An Illustration of the Effect of Sample Size on *P*-values and β

n	P-Value When $\bar{x} = 101$	$\beta(101)$ for Level .01 Test
25	.3707	.9772
100	.2514	.9525
400	.0918	.8413
900	.0228	.6293
1600	.0038	.3707
2500	.0004	.1587
5000	.0000012	.0087
10,000	.0000000	.0000075

little from the null value $\mu_0 = 100$. The third column points out that even when there is little practical difference between the true μ and the null value, for a fixed level of significance a large sample size will frequently lead to rejection of the null hypothesis at that level. To summarize, *one must be especially careful in interpreting evidence when the sample size is large, since any small departure from H_0 will almost surely be detected by a test, yet such a departure may have little practical significance.*

The Relationship between Confidence Intervals and Hypothesis Tests

Suppose the standardized variable $Z = (\hat{\theta} - \theta)/\hat{\sigma}_{\hat{\theta}}$ has (at least approximately) a standard normal distribution. The central z curve area captured between -1.96 and 1.96 is .95 (and the remaining area .05 is split equally between the two tails, giving area .025 in each one). This implies that a confidence interval for θ with confidence level 95% is $\hat{\theta} \pm 1.96\hat{\sigma}_{\hat{\theta}}$.

Now consider testing H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$ at significance level .05 using the test statistic $Z = (\hat{\theta} - \theta_0)/\hat{\sigma}_{\hat{\theta}}$. The phrase "$z$ test" implies that when the null hypothesis is true, Z has (at least approximately) a standard normal distribution. So the P-value will be twice the area under the z curve to the right of $|z|$. This P-value will be less than or equal to .05, allowing for rejection of the null hypothesis, if and only if either $z \geq 1.96$ or $z \leq -1.96$. The null hypothesis will therefore not be rejected if $-1.96 < z < 1.96$.

Substituting the formula for z into this latter system of inequalities and manipulating them to isolate θ_0 gives the equivalent system $\hat{\theta} - 1.96\hat{\sigma}_{\hat{\theta}} < \theta_0 < \hat{\theta} + 1.96\hat{\sigma}_{\hat{\theta}}$. The lower limit in this system is just the left endpoint of the 95% confidence interval, and the upper limit is the right endpoint of the interval. What this says is that the null hypothesis will not be rejected if and only if the null value θ_0 lies in the confidence interval. Suppose, for example, that sample data yields the 95% CI (68.6, 72.0). Then the null hypothesis H_0: $\theta = 70$ cannot be rejected at significance level .05 because 70 lies in the CI. But the null hypothesis H_0: $\theta = 65$ can be rejected because 65 does not lie in the CI. There is an analogous relationship between a 99% CI and a test with significance level .01—the null hypothesis cannot be rejected if the null value lies in the CI and should be rejected if the null value is outside the CI. There is a duality between a two-sided confidence interval with confidence level $100(1 - \alpha)\%$ and the conclusion from a two-tailed test with significance level α.

Now consider testing H_0: $\theta = \theta_0$ against the alternative H_a: $\theta > \theta_0$ at significance level .01. Because of the inequality in H_a, the P-value is the area under the z curve to the right of the calculated z. The z critical value 2.33 captures upper-tail area .01. Therefore the P-value (captured upper-tail area) will be at most .01 if and only if

$z \geq 2.33$; we will not be able to reject the null hypothesis if and only if $z < 2.33$. Again substituting the formula for z into this inequality and manipulating to isolate θ_0 gives the equivalent inequality $\hat{\theta} - 2.33\hat{\sigma}_{\hat{\theta}} < \theta_0$. The lower limit of this inequality is the lower confidence bound for θ with a confidence level of 99%. So the null hypothesis won't be rejected at significance level .01 if and only if the null value exceeds the lower confidence bound. Thus there is a duality between a lower confidence bound and the conclusion from an upper-tailed test. This is why the Minitab software package will output a lower confidence bound when an upper-tailed test is performed. If, for example, the 90% lower confidence bound is 25.3, i.e., $25.3 < \theta$ with confidence level 90%, then we would not be able to reject H_0: $\theta = 26$ versus H_a: $\theta > 26$ at significance level .10 but would be able to reject H_0: $\theta = 24$ in favor of H_a: $\theta > 24$. There is an analogous duality between an upper confidence bound and the conclusion from a lower-tailed test. And there are analogous relationships for t tests and t confidence intervals or bounds.

PROPOSITION

Let $(\hat{\theta}_L, \hat{\theta}_U)$ be a confidence interval for θ with confidence level $100(1 - \alpha)\%$. Then a test of H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$ with significance level α rejects the null hypothesis if the null value θ_0 is not included in the CI and does not reject H_0 if the null value does lie in the CI. There is an analogous relationship between a lower confidence bound and an upper-tailed test, and also between an upper confidence bound and a lower-tailed test.

In light of these relationships, it is tempting to carry out a test of hypotheses by calculating the corresponding CI or CB. Don't yield to temptation! Instead carry out a more informative analysis by determining and reporting the P-value.

Simultaneous Testing of Several Hypotheses

Many published articles report the results of more than just a single test of hypotheses. For example, the article **"Distributions of Compressive Strength Obtained from Various Diameter Cores" (*ACI Materials J.*, 2012: 597–606)** considered the plausibility of Weibull, normal, and lognormal distributions as models for compressive strength distributions under various experimental conditions. Table 3 of the cited article reported exact P-values for a total of 71 different tests.

Consider two different tests, one for a pair of hypotheses about a population mean and another for a pair of hypotheses about a population proportion—e.g., the mean wing length for adult Monarch butterflies and the proportion of schoolchildren in a particular state who are obese. Assume that the sample used to test the first pair of hypotheses is selected independently of that used to test the second pair. Then if each test is carried out at significance level .05 (type I error probability .05),

$$\begin{aligned} P(\text{at least one type I error is committed}) &= 1 - P(\text{no type I errors are committed}) \\ &= 1 - P(\text{no type I error in the 1st test}) \cdot P(\text{no type I error in the 2nd test}) \\ &= 1 - (.95)^2 = 1 - .9025 = .0975 \end{aligned}$$

Thus the probability of committing at least one type I error when two independent tests are carried out is much higher than the probability that a type I error will result from a single test. If three tests are independently carried out, each at significance level .05, then the probability that at least one type I error is committed is $1 - (.95)^3 = .1426$. Clearly as the number of tests increases, the probability of committing at least one type I error gets larger and in fact will approach 1.

Suppose we want the probability of committing at least one type I error in two independent tests to be .05—an *experimentwise* error rate of .05. Then the significance level α for each test must be smaller than .05:

$$.05 = 1 - (1 - \alpha)^2 \Rightarrow 1 - \alpha = \sqrt{.95} = .975 \Rightarrow \alpha = .025$$

If the probability of committing at least one type I error in three independent tests is to be .05, the significance level for each one must be .017 (replace the square root by the cube root in the foregoing argument). As the number of tests increases, the significance level for each one must decrease to 0 in order to maintain an experimentwise error rate of .05.

Often it is not reasonable to assume that the various tests are independent of one another. In the example cited at the beginning of this subsection, four different tests were carried out based on the same sample involving one particular type of concrete in combination with a specified core diameter and length-to-diameter ratio. It is then no longer clear how the experimentwise error rate relates to the significance level for each individual test. Let A_i denote the event that the *ith* test results in a type I error. Then in the case of k tests,

$$\begin{aligned} &P(\text{at least one type I error}) \\ &\quad = P(A_1 \cup A_2 \cup \ldots \cup A_k) \le P(A_1) + \cdots + P(A_k) = k\alpha \end{aligned}$$

(the inequality in the last line is called the *Bonferroni inequality*; it can be proved by induction on k). Thus a significance level of $.05/k$ for each test will ensure that the experimentwise significance level is at most .05.

Again, the central idea here is that in order for the probability of at least one type I error among k tests to be small, the significance level for each individual test must be quite small. If the significance level for each individual test is .05, for even a moderate number of tests it is rather likely that at least one type I error will be committed. That is, with $\alpha = .05$ for each test, when each null hypothesis is actually true, it is rather likely that at least one of the tests will yield a statistically significant result. This is why one should view a statistically significant result with skepticism when many tests are carried out using one of the traditional significance levels.

The Likelihood Ratio Principle

The test procedures presented in this and subsequent chapters will (at least for the most part) be intuitively sensible. But there are many situations that arise in practice where intuition is not a reliable guide to obtaining a test statistic. We now describe a general strategy for this purpose. Let $x_1, x_2, \ldots, x_n$ be the observations in a random sample of size n from a probability distribution $f(x; \theta)$. The joint distribution evaluated at these sample values is the product $f(x_1; \theta) \cdot f(x_2; \theta) \cdot \cdots \cdot f(x_n; \theta)$. As in the discussion of maximum likelihood estimation, the *likelihood function* is this joint distribution, regarded as a function of θ. Consider testing H_0: θ is in Ω_0 versus H_a: θ is in Ω_a, where Ω_0 and Ω_a are disjoint (for example, H_0: $\theta \le 100$ versus H_a: $\theta > 100$). The **likelihood ratio principle** for test construction proceeds as follows:

1. Find the largest value of the likelihood for any θ in Ω_0 (by finding the maximum likelihood estimate within Ω_0 and substituting back into the likelihood function).
2. Find the largest value of the likelihood for any θ in Ω_a.
3. Form the ratio

$$\lambda(x_1, \ldots, x_n) = \frac{\text{maximum likelihood for } \theta \text{ in } \Omega_0}{\text{maximum likelihood for } \theta \text{ in } \Omega_a}$$

The ratio $\lambda(x_1, \ldots, x_n)$ is called the *likelihood ratio statistic value*. Intuitively, the smaller the value of λ, the stronger is the evidence against H_0. It can, for example, be shown that for testing H_0: $\mu \leq \mu_0$ versus H_a: $\mu > \mu_0$ in the case of population normality, a small value of λ is equivalent to a large value of t. Thus the one-sample t test comes from applying the likelihood ratio principle. We emphasize that once a test statistic has been selected, its distribution when H_0 is true is required for P-value determination; statistical theory must again come to the rescue!

The likelihood ratio principle can also be applied when the X_i's have different distributions and even when they are dependent, though the likelihood function can be complicated in such cases. Many of the test procedures to be presented in subsequent chapters are obtained from the likelihood ratio principle. These tests often turn out to minimize β among all tests that have the desired α, so are truly best tests. For more details and some worked examples, refer to one of the references listed in the Chapter 6 bibliography.

A practical limitation is that, to construct the likelihood ratio test statistic, the form of the probability distribution from which the sample comes must be specified. Derivation of the t test from the likelihood ratio principle requires assuming a normal pdf. If an investigator is willing to assume that the distribution is symmetric but does not want to be specific about its exact form (such as normal, uniform, or Cauchy), then the principle fails because there is no way to write a joint pdf simultaneously valid for all symmetric distributions. In Chapter 15, we will present several **distribution-free** test procedures, so called because the probability of a type I error is controlled simultaneously for many different underlying distributions. These procedures are useful when the investigator has limited knowledge of the underlying distribution. We shall also consider criteria for selection of a test procedure when several sensible candidates are available, and comment on the performance of several procedures when an underlying assumption such as normality is violated.

EXERCISES Section 8.5 (53–56)

53. Reconsider the paint-drying problem discussed in Example 8.5. The hypotheses were H_0: $\mu = 75$ versus H_a: $\mu < 75$, with σ assumed to have value 9.0. Consider the alternative value $\mu = 74$, which in the context of the problem would presumably not be a practically significant departure from H_0.

- **a.** For a level .01 test, compute β at this alternative for sample sizes $n = 100$, 900, and 2500.
- **b.** If the observed value of $\bar{X}$ is $\bar{x} = 74$, what can you say about the resulting P-value when $n = 2500$? Is the data statistically significant at any of the standard values of α?
- **c.** Would you really want to use a sample size of 2500 along with a level .01 test (disregarding the cost of such an experiment)? Explain.

54. Consider the large-sample level .01 test in Section 8.4 for testing H_0: $p = .2$ against H_a: $p > .2$.

- **a.** For the alternative value $p = .21$, compute $\beta(.21)$ for sample sizes $n = 100$, 2500, 10,000, 40,000, and 90,000.
- **b.** For $\hat{p} = x/n = .21$, compute the P-value when $n = 100$, 2500, 10,000, and 40,000.
- **c.** In most situations, would it be reasonable to use a level .01 test in conjunction with a sample size of 40,000? Why or why not?

55. Consider carrying out m tests of hypotheses based on independent samples, each at significance level (exactly) .01.

- **a.** What is the probability of committing at least one type I error when $m = 5$? When $m = 10$?
- **b.** How many such tests would it take for the probability of committing at least one type I error to be at least .5?

56. A 95% CI for true average amount of warpage (mm) of laminate sheets under specified conditions was calculated as (1.81, 1.95), based on a sample size of $n = 15$ and the assumption that amount of warpage is normally distributed.

- **a.** Suppose you want to test H_0: $\mu = 2$ versus H_a: $\mu \neq 2$ using $\alpha = .05$. What conclusion would be appropriate, and why?
- **b.** If you wanted to use a significance level of .01 for the test in (a), what conclusion would be appropriate?

SUPPLEMENTARY EXERCISES (57–80)

57. A sample of 50 lenses used in eyeglasses yields a sample mean thickness of 3.05 mm and a sample standard deviation of .34 mm. The desired true average thickness of such lenses is 3.20 mm. Does the data strongly suggest that the true average thickness of such lenses is something other than what is desired? Test using $\alpha = .05$.

58. In Exercise 57, suppose the experimenter had believed before collecting the data that the value of σ was approximately .30. If the experimenter wished the probability of a type II error to be .05 when $\mu = 3.00$, was a sample size 50 unnecessarily large?

59. It is specified that a certain type of iron should contain .85 g of silicon per 100 g of iron (.85%). The silicon content of each of 25 randomly selected iron specimens was determined, and the accompanying Minitab output resulted from a test of the appropriate hypotheses.

```
Variable   N    Mean   StDev  SE Mean     T     P
sil cont  25  0.8880  0.1807   0.0361  1.05  0.30
```

a. What hypotheses were tested?

b. What conclusion would be reached for a significance level of .05, and why? Answer the same question for a significance level of .10.

60. One method for straightening wire before coiling it to make a spring is called "roller straightening." The article **"The Effect of Roller and Spinner Wire Straightening on Coiling Performance and Wire Properties" (*Springs*, 1987: 27–28)** reports on the tensile properties of wire. Suppose a sample of 16 wires is selected and each is tested to determine tensile strength (N/mm^2). The resulting sample mean and standard deviation are 2160 and 30, respectively.

a. The mean tensile strength for springs made using spinner straightening is 2150 N/mm^2. What hypotheses should be tested to determine whether the mean tensile strength for the roller method exceeds 2150?

b. Assuming that the tensile strength distribution is approximately normal, what test statistic would you use to test the hypotheses in part (a)?

c. What is the value of the test statistic for this data?

d. What is the P-value for the value of the test statistic computed in part (c)?

e. For a level .05 test, what conclusion would you reach?

61. Contamination of mine soils in China is a serious environmental problem. The article **"Heavy Metal Contamination in Soils and Phytoaccumulation in a Manganese Mine Wasteland, South China" (*Air, Soil, and Water Res.*, 2008: 31–41)** reported that, for a sample of 3 soil specimens from a certain restored mining area, the sample mean concentration of Total Cu was 45.31 mg/kg with a corresponding (estimated) standard error of the mean of 5.26. It was also stated that the China background value for this concentration was 20. The results of various statistical tests described in the article were predicated on assuming normality.

a. Does the data provide strong evidence for concluding that the true average concentration in the sampled region exceeds the stated background value? Carry out a test at significance level .01. Does the result surprise you? Explain.

b. Referring back to the test of (a), how likely is it that the P-value would be at least .01 when the true average concentration is 50 and the true standard deviation of concentration is 10?

62. The article **"Orchard Floor Management Utilizing Soil-Applied Coal Dust for Frost Protection" (*Agri. and Forest Meteorology*, 1988: 71–82)** reports the following values for soil heat flux of eight plots covered with coal dust.

34.7 35.4 34.7 37.7 32.5 28.0 18.4 24.9

The mean soil heat flux for plots covered only with grass is 29.0. Assuming that the heat-flux distribution is approximately normal, does the data suggest that the coal dust is effective in increasing the mean heat flux over that for grass? Test the appropriate hypotheses using $\alpha = .05$.

63. The article **"Caffeine Knowledge, Attitudes, and Consumption in Adult Women" (*J. of Nutrition Educ.*, 1992: 179–184)** reports the following summary data on daily caffeine consumption for a sample of adult women: $n = 47$, $\bar{x} = 215$ mg, $s = 235$ mg, and range = 5−1176.

a. Does it appear plausible that the population distribution of daily caffeine consumption is normal? Is it necessary to assume a normal population distribution to test hypotheses about the value of the population mean consumption? Explain your reasoning.

b. Suppose it had previously been believed that mean consumption was at most 200 mg. Does the given data contradict this prior belief? Test the appropriate hypotheses at significance level .10.

64. Annual holdings turnover for a mutual fund is the percentage of a fund's assets that are sold during a particular year. Generally speaking, a fund with a low value of turnover is more stable and risk averse, whereas a high value of turnover indicates a substantial amount of buying and selling in an attempt to take advantage of short-term market fluctuations. Here are values of turnover for a sample of 20 large-cap blended funds (refer to Exercise 1.53 for a bit more information) extracted from **Morningstar.com:**

1.03 1.23 1.10 1.64 1.30 1.27 1.25 0.78 1.05 0.64
0.94 2.86 1.05 0.75 0.09 0.79 1.61 1.26 0.93 0.84

a. Would you use the one-sample t test to decide whether there is compelling evidence for concluding that the population mean turnover is less than 100%? Explain.

b. A normal probability plot of the 20 ln(turnover) values shows a very pronounced linear pattern, suggesting it is reasonable to assume that the turnover distribution is lognormal. Recall that X has a lognormal distribution if $\ln(X)$ is normally distributed with mean value μ and variance σ^2. Because μ is also the median of the $\ln(X)$ distribution, e^{μ} is the median of the X distribution. Use this information to decide whether there is compelling evidence for concluding that the median of the turnover population distribution is less than 100%.

65. The true average breaking strength of ceramic insulators of a certain type is supposed to be at least 10 psi. They will be used for a particular application unless sample data indicates conclusively that this specification has not been met. A test of hypotheses using $\alpha = .01$ is to be based on a random sample of ten insulators. Assume that the breaking-strength distribution is normal with unknown standard deviation.

a. If the true standard deviation is .80, how likely is it that insulators will be judged satisfactory when true average breaking strength is actually only 9.5? Only 9.0?

b. What sample size would be necessary to have a 75% chance of detecting that the true average breaking strength is 9.5 when the true standard deviation is .80?

66. The accompanying observations on residual flame time (sec) for strips of treated children's nightwear were given in the article **"An Introduction to Some Precision and Accuracy of Measurement Problems" (*J. of Testing and Eval.*, 1982: 132–140).** Suppose a true average flame time of at most 9.75 had been mandated. Does the data suggest that this condition has not been met? Carry out an appropriate test after first investigating the plausibility of assumptions that underlie your method of inference.

9.85	9.93	9.75	9.77	9.67	9.87	9.67
9.94	9.85	9.75	9.83	9.92	9.74	9.99
9.88	9.95	9.95	9.93	9.92	9.89	

67. The incidence of a certain type of chromosome defect in the U.S. adult male population is believed to be 1 in 75. A random sample of 800 individuals in U.S. penal institutions reveals 16 who have such defects. Can it be concluded that the incidence rate of this defect among prisoners differs from the presumed rate for the entire adult male population?

a. State and test the relevant hypotheses using $\alpha = .05$. What type of error might you have made in reaching a conclusion?

b. Based on the P-value calculated in (a), could H_0 be rejected at significance level .20?

68. In an investigation of the toxin produced by a certain poisonous snake, a researcher prepared 26 different vials, each containing 1 g of the toxin, and then determined the amount of antitoxin needed to neutralize the toxin. The sample average amount of antitoxin necessary was found to be 1.89 mg, and the sample standard deviation was .42. Previous research had indicated that the true average neutralizing amount was 1.75 mg/g of toxin. Does the new data contradict the value suggested by prior research? Test the relevant hypotheses. Does the validity of your analysis depend on any assumptions about the population distribution of neutralizing amount? Explain.

69. The sample average unrestrained compressive strength for 45 specimens of a particular type of brick was computed to be 3107 psi, and the sample standard deviation was 188. The distribution of unrestrained compressive strength may be somewhat skewed. Does the data strongly indicate that the true average unrestrained compressive strength is less than the design value of 3200? Test using $\alpha = .001$.

70. The Dec. 30, 2009, the ***NewYork Times*** reported that in a survey of 948 American adults who said they were at least somewhat interested in college football, 597 said the current Bowl Championship System should be replace by a playoff similar to that used in college basketball. Does this provide compelling evidence for concluding that a majority of all such individuals favor replacing the B.C.S. with a playoff? Test the appropriate hypotheses using a significant level of .001.

71. When $X_1, X_2, \ldots, X_n$ are independent Poisson variables, each with parameter μ, and n is large, the sample mean $\bar{X}$ has approximately a normal distribution with $\mu = E(\bar{X})$ and $V(\bar{X}) = \mu/n$. This implies that

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

has approximately a standard normal distribution. For testing H_0: $\mu = \mu_0$, we can replace μ by μ_0 in the equation for Z to obtain a test statistic. This statistic is actually preferred to the large-sample statistic with denominator $S/\sqrt{n}$ (when the X_i's are Poisson) because it is tailored explicitly to the Poisson assumption. If the number of requests for consulting received by a certain statistician during a 5-day work week has a Poisson distribution and the total number of consulting requests during a 36-week period is 160, does this suggest that the true average number of weekly requests exceeds 4.0? Test using $\alpha = .02$.

72. An article in the Nov. 11, 2005, issue of the ***San Luis Obispo Tribune*** reported that researchers making random purchases at California Wal-Mart stores found scanners coming up with the wrong price 8.3% of the time. Suppose this was based on 200 purchases. The National Institute for Standards and Technology says that in the long run at most two out of every 100 items should have incorrectly scanned prices.

a. Develop a test procedure with a significance level of (approximately) .05, and then carry out the test to decide whether the NIST benchmark is not satisfied.

b. For the test procedure you employed in (a), what is the probability of deciding that the NIST benchmark has been satisfied when in fact the mistake rate is 5%?

73. The article **"Heavy Drinking and Polydrug Use Among College Students" (*J. of Drug Issues*, 2008: 445–466)** stated that 51 of the 462 college students in a sample had a lifetime abstinence from alcohol. Does this provide strong evidence for concluding that more than 10% of the population sampled had completely abstained from alcohol use? Test the appropriate hypotheses. [*Note:* The article used more advanced statistical methods to study the use of various drugs among students characterized as light, moderate, and heavy drinkers.]

74. The article **"Analysis of Reserve and Regular Bottlings: Why Pay for a Difference Only the Critics Claim to Notice?" (*Chance*, Summer 2005, pp. 9–15)** reported on an experiment to investigate whether wine tasters could distinguish between more expensive reserve wines and their regular counterparts. Wine was presented to tasters in four containers labeled A, B, C, and D, with two of these containing the reserve wine and the other two the regular wine. Each taster randomly selected three of the containers, tasted the selected wines, and indicated which of the three he/she believed was different from the other two. Of the $n = 855$ tasting trials, 346 resulted in correct distinctions (either the one reserve that differed from the two regular wines or the one regular wine that differed from the two reserves). Does this provide compelling evidence for concluding that tasters of this type have some ability to distinguish between reserve and regular wines? State and test the relevant hypotheses. Are you particularly impressed with the ability of tasters to distinguish between the two types of wine?

75. The American Academy of Pediatrics recommends a vitamin D level of at least 20 ng/ml for infants. The article **"Vitamin D and Parathormone Levels of Late-Preterm Formula Fed Infants During the First Year of Life" (*European J. of Clinical Nutr.*, 2012: 224–230)** reported that for a sample of 102 preterm infants judged to be of appropriate weight for their gestational age, the sample mean vitamin D level at 2 weeks was 21 with a sample standard deviation of 11. Does this provide convincing evidence that the population mean vitamin D level for such infants exceeds 20? Test the relevant hypotheses using a significance level of .10.

76. Chapter 7 presented a CI for the variance σ^2 of a normal population distribution. The key result there was that the rv $\chi^2 = (n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ df. Consider the null hypothesis H_0: $\sigma^2 = \sigma_0^2$ (equivalently, $\sigma = \sigma_0$). Then when H_0 is true, the test statistic $\chi^2 = (n-1)S^2/\sigma_0^2$ has a chi-squared distribution with $n-1$ df. If the relevant alternative is H_a: $\sigma^2 > \sigma_0^2$ the P-value is the area under the χ^2 curve with $n-1$ df to the right of the calculated χ^2 value. To ensure reasonably uniform characteristics for a particular application, it is desired that the true standard deviation of the softening point of a certain type of petroleum pitch be at most .50°C. The softening points of ten different specimens were determined, yielding a sample standard deviation of .58°C. Does this strongly contradict the uniformity specification? Test the appropriate hypotheses using $\alpha = .01$. [*Hint:* Consult Table A.11.]

77. Referring to Exercise 76, suppose an investigator wishes to test H_0: $\sigma^2 = .04$ versus H_a: $\sigma^2 < .04$ based on a sample of 21 observations. The computed value of $20s^2/.04$ is 8.58. Place bounds on the P-value and then reach a conclusion at level .01. [*Hint:* Consult Table A.7.]

78. When the population distribution is normal and n is large, the sample standard deviation S has approximately a normal distribution with $E(S) \approx \sigma$ and $V(S) \approx \sigma^2/(2n)$. We already know that in this case, for any n, $\overline{X}$ is normal with $E(\overline{X}) = \mu$ and $V(\overline{X}) = \sigma^2/n$.

a. Assuming that the underlying distribution is normal, what is an approximately unbiased estimator of the 99th percentile $\theta = \mu + 2.33\sigma$?

b. When the X_i's are normal, it can be shown that $\overline{X}$ and S are independent rv's (one measures location whereas the other measures spread). Use this to compute $V(\hat{\theta})$ and $\sigma_{\hat{\theta}}$ for the estimator $\hat{\theta}$ of part (a). What is the estimated standard error $\hat{\sigma}_{\hat{\theta}}$?

c. Write a test statistic for testing H_0: $\theta = \theta_0$ that has approximately a standard normal distribution when H_0 is true. If soil pH is normally distributed in a certain region and 64 soil samples yield $\bar{x} = 6.33$, $s = .16$, does this provide strong evidence for concluding that at most 99% of all possible samples would have a pH of less than 6.75? Test using $\alpha = .01$.

79. Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with parameter λ. Then it can be shown that $2\lambda\Sigma X_i$ has a chi-squared distribution with $\nu = 2n$ (by first showing that $2\lambda X_i$ has a chi-squared distribution with $\nu = 2$).

a. Use this fact to obtain a test statistic for testing H_0: $\mu = \mu_0$. Then explain how you would determine the P-value when the alternative hypothesis is H_a: $\mu < \mu_0$. [*Hint:* $E(X_i) = \mu = 1/\lambda$, so $\mu = \mu_0$ is equivalent to $\lambda = 1/\mu_0$.]

b. Suppose that ten identical components, each having exponentially distributed time until failure, are tested. The resulting failure times are

95 16 11 3 42 71 225 64 87 123

Use the test procedure of part (a) to decide whether the data strongly suggests that the true average lifetime is less than the previously claimed value of 75. [*Hint:* Consult Table A.7.]

80. Because of variability in the manufacturing process, the actual yielding point of a sample of mild steel subjected to increasing stress will usually differ from the theoretical yielding point. Let p denote the true proportion of samples that yield before their theoretical yielding point. If on the basis of a sample it can be concluded that more than 20% of all specimens yield before the theoretical point, the production process will have to be modified.

a. If 15 of 60 specimens yield before the theoretical point, what is the P-value when the appropriate test is used, and what would you advise the company to do?

b. If the true percentage of "early yields" is actually 50% (so that the theoretical point is the median of the yield distribution) and a level .01 test is used, what is the probability that the company concludes a modification of the process is necessary?

BIBLIOGRAPHY

See the bibliographies at the ends of Chapter 6 and Chapter 7.