INTRODUCTION TO THE PRACTICE OF STATISTICS

FOURTH EDITION

David S. Moore
George P. McCabe
Purdue University

W. H. Freeman and Company
New York

Power

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

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

Power

The probability that a fixed level α significance test will reject H_0 particular alternative value of the parameter is true is called the p the test to detect that alternative.

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

We will answer this question by calculating the power of the s test that will be used to evaluate the data to be collected. The

Step 1

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

$$H_0: \mu = 0$$

$$H_a: \mu > 0$$

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

Step 2

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

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

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

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

$$\bar{x} \geq 0.658$$

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

Step 3

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

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

$$= P(Z \geq -0.855) = 0.80$$

Figure 6.13 illustrates the power with the sampling distribution of $\bar{x}$ when

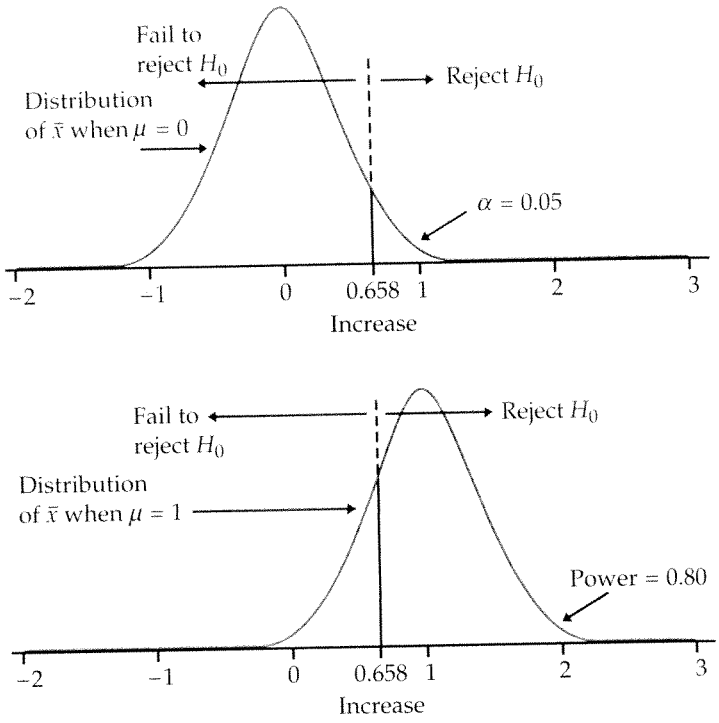

FIGURE 6.13 The sampling distributions of $\bar{x}$ when $\mu = 0$ and when $\mu = 1$ with the α and the power. Power is the probability that the test rejects H_0 when the alternative is true.

the funded studies be sufficient to detect important results 80% of the time using a 5% test of significance.

Increasing the power

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

Increase α. A 5% test of significance will have a greater chance of rejecting the alternative than a 1% test because the strength of evidence required for rejection is less.

Consider a particular alternative that is farther away from μ_0. Values of μ that are in H_a but lie close to the hypothesized value μ_0 are harder to detect (lower power) than values of μ that are far from μ_0.

Increase the sample size. More data will provide more information about $\bar{x}$ so we have a better chance of distinguishing values of μ.

Decrease σ. This has the same effect as increasing the sample size: more information about μ. Improving the measurement process and restricting attention to a subpopulation are two common ways to decrease σ.

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

EXAMPLE 6.18

The "efficient market hypothesis" for the time series of stock prices says that future stock prices (when adjusted for inflation) show only random variation. No information available now will help us predict stock prices in the future, because the efficient working of the market has already incorporated all available information in the present price. Many studies have tested the claim that one or another kind of information is helpful. In these studies, the efficient market hypothesis is H_0, and the claim that prediction is possible is H_a. Almost all the studies have failed to find good evidence against H_0. As a result, the efficient market theory is quite popular. But an examination of the significance tests employed finds that the power is generally low. Failure to reject H_0 when using tests of low power is not evidence that H_0 is true. As one expert says, "The widespread impression that there is strong evidence for market efficiency may be due just to a lack of appreciation of the low power of many statistical tests."[13]

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

EXAMPLE 6.19

Example 6.13 (page 448) presented a test of

$$H_0: \mu = 0.86$$

$$H_a: \mu \neq 0.86$$

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

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

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

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

$$z \geq 2.576 \qquad \text{(in other words, } \bar{x} \geq 0.870\text{)}$$

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

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

$$P(\bar{x} \geq 0.87) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.87 - 0.845}{0.0068/\sqrt{3}}\right)$$

$$= P(Z \geq 6.37) = 0$$

$$P(\bar{x} \leq 0.85) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.85 - 0.845}{0.0068/\sqrt{3}}\right)$$

$$= P(Z \leq 1.27) = 0.8980$$

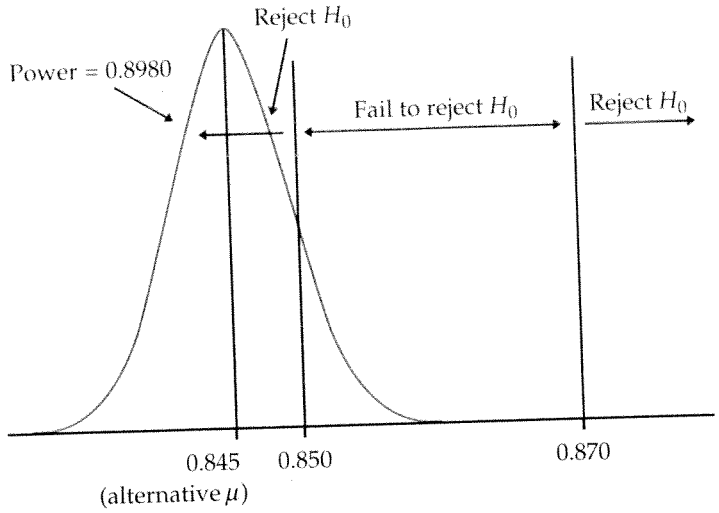

FIGURE 6.14 The power for Example 6.19.

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

Inference as decision*

We have presented tests of significance as methods for assessing the strengt of evidence against the null hypothesis. This assessment is made by the P value, which is a probability computed under the assumption that H_0 is tru The alternative hypothesis (the statement we seek evidence for) enters the tes only to help us see what outcomes count against the null hypothesis. Such i the reasoning of tests of significance as advocated by Fisher and as practice by many users of statistics.

But signs of another way of thinking were present in the discussion c significance tests with fixed level α. A level of significance α chosen in advanc points to the outcome of the test as a *decision*. If the P-value is less than α, w reject H_0 in favor of H_a. Otherwise we fail to reject H_0. The transition fro measuring the strength of evidence to making a decision is not a small ste Many statisticians agree with Fisher that making decisions is too grand a goa especially in scientific inference. A decision is reached only after the evidenc of many studies is weighed. Indeed, the goal of research is not "decision" b a gradually evolving understanding. Statistical inference should content itse with confidence intervals and tests of significance. Many users of statistics a content with such methods. It is rare to set up a level α in advance as a ru

*The purpose of this section is to clarify the reasoning of significance tests by contrast with a related type of reasoning. It can be omitted without loss of continuity.

for making a decision in a scientific problem. More commonly, users think of significance at level 0.05 as a description of good evidence. This is made clearer by giving the *P*-value.

Yet there are circumstances that call for a decision or action as the end result of inference. **Acceptance sampling** is one such circumstance. A producer of bearings and the consumer of the bearings agree that each carload lot shall meet certain quality standards. When a carload arrives, the consumer chooses a sample of bearings to be inspected. On the basis of the sample outcome, the consumer will either accept or reject the carload. Fisher agreed that this is a genuine decision problem. But he insisted that acceptance sampling is completely different from scientific inference. Other eminent statisticians have argued that if "decision" is given a broad meaning, almost all problems of statistical inference can be posed as problems of making decisions in the presence of uncertainty. We will not venture further into the arguments over how we ought to think about inference. We do want to show how a different concept—inference as decision—changes the reasoning used in tests of significance.

Two types of error

Tests of significance concentrate on H_0, the null hypothesis. If a decision is called for, however, there is no reason to single out H_0. There are simply two hypotheses, and we must accept one and reject the other. It is convenient to call the two hypotheses H_0 and H_a, but H_0 no longer has the special status (the statement we try to find evidence against) that it had in tests of significance. In the acceptance sampling problem, we must decide between

H_0: the lot of bearings meets standards

H_a: the lot does not meet standards

on the basis of a sample of bearings.

We hope that our decision will be correct, but sometimes it will be wrong. There are two types of incorrect decisions. We can accept a bad lot of bearings, or we can reject a good lot. Accepting a bad lot injures the consumer, while rejecting a good lot hurts the producer. To help distinguish these two types of error, we give them specific names.

Type I and Type II Errors

If we reject H_0 (accept H_a) when in fact H_0 is true, this is a **Type I error.** If we accept H_0 (reject H_a) when in fact H_a is true, this is a **Type II error.**

The possibilities are summed up in Figure 6.15. If H_0 is true, our decision either is correct (if we accept H_0) or is a Type I error. If H_a is true, our decision either is correct or is a Type II error. Only one error is possible at one time. Figure 6.16 applies these ideas to the acceptance sampling example.

Decision based on sample	Truth about the population: H_0 true	Truth about the population: H_a true
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

FIGURE 6.15 The two types of error in testing hypotheses.

Decision based on sample	Truth about the lot: Does meet standards	Truth about the lot: Does not meet standards
Reject the lot	Type I error	Correct decision
Accept the lot	Correct decision	Type II error

FIGURE 6.16 The two types of error in the acceptance sampling setting.

Error probabilities

Any rule for making decisions is assessed in terms of the probabilities of th two types of error. This is in keeping with the idea that statistical inference i based on probability. We cannot (short of inspecting the whole lot) guarante that good lots of bearings will never be rejected and bad lots never be acceptec But by random sampling and the laws of probability, we can say what th probabilities of both kinds of error are.

Significance tests with fixed level α give a rule for making decisions, bc cause the test either rejects H_0 or fails to reject it. If we adopt the decisior making way of thought, failing to reject H_0 means deciding that H_0 is tru We can then describe the performance of a test by the probabilities of Type and Type II errors.

EXAMPLE 6.20 The mean diameter of a type of bearing is supposed to be 2.000 centimeters (cm) The bearing diameters vary normally with standard deviation $\sigma = 0.010$ cm. Wher a lot of the bearings arrives, the consumer takes an SRS of 5 bearings from the lo and measures their diameters. The consumer rejects the bearings if the sample mear diameter is significantly different from 2 at the 5% significance level.

This is a test of the hypotheses

$$H_0: \mu = 2$$

$$H_a: \mu \neq 2$$

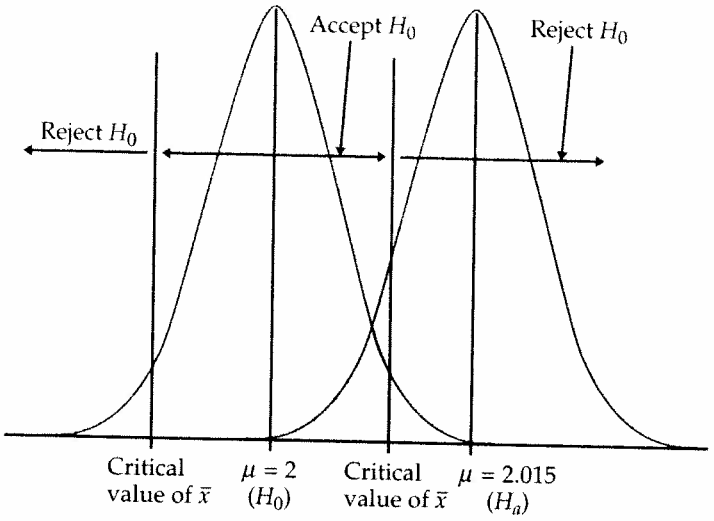

FIGURE 6.17 The two error probabilities for Example 6.20. The probability of a Type I error (*light tan area*) is the probability of rejecting H_0: $\mu = 2$ when in fact $\mu = 2$. The probability of a Type II error (*dark tan area*) is the probability of accepting H_0 when in fact $\mu = 2.015$.

To carry out the test, the consumer computes the z statistic:

$$z = \frac{\bar{x} - 2}{0.01/\sqrt{5}}$$

and rejects H_0 if

$$z < -1.96 \text{ or } z > 1.96$$

A Type I error is to reject H_0 when in fact $\mu = 2$.

What about Type II errors? Because there are many values of μ in H_a, we will concentrate on one value. The producer and the consumer agree that a lot of bearings with mean 0.015 cm away from the desired mean 2.000 should be rejected. So a particular Type II error is to accept H_0 when in fact $\mu = 2.015$.

Figure 6.17 shows how the two probabilities of error are obtained from the two sampling distributions of $\bar{x}$, for $\mu = 2$ and for $\mu = 2.015$. When $\mu = 2$, H_0 is true and to reject H_0 is a Type I error. When $\mu = 2.015$, accepting H_0 is a Type II error. We will now calculate these error probabilities.

The probability of a Type I error is the probability of rejecting H_0 when it is really true. In Example 6.20, this is the probability that $|z| \geq 1.96$ when $\mu = 2$. But this is exactly the significance level of the test. The critical value 1.96 was chosen to make this probability 0.05, so we do not have to compute it again. The definition of "significant at level 0.05" is that sample outcomes this extreme will occur with probability 0.05 when H_0 is true.

Significance and Type I Error

The significance level α of any fixed level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is in fact true.

The probability of a Type II error for the particular alternative $\mu = 2.01$ in Example 6.20 is the probability that the test will fail to reject H_0 when μ ha this alternative value. The *power* of the test against the alternative $\mu = 2.01$ is just the probability that the test *does* reject H_0. By following the method Example 6.19, we can calculate that the power is about 0.92. The probabili of a Type II error is therefore $1 - 0.92$, or 0.08.

Power and Type II Error

The power of a fixed level test against a particular alternative is 1 minus the probability of a Type II error for that alternative.

The two types of error and their probabilities give another interpretatic of the significance level and power of a test. The distinction between tests significance and tests as rules for deciding between two hypotheses does n lie in the calculations but in the reasoning that motivates the calculation In a test of significance we focus on a single hypothesis (H_0) and a sing probability (the P-value). The goal is to measure the strength of the samp evidence against H_0. Calculations of power are done to check the sensitivi of the test. If we cannot reject H_0, we conclude only that there is not sufficie evidence against H_0, not that H_0 is actually true. If the same inference proble is thought of as a decision problem, we focus on two hypotheses and give rule for deciding between them based on the sample evidence. We therefc must focus equally on two probabilities, the probabilities of the two typ of error. We must choose one or the other hypothesis and cannot abstain grounds of insufficient evidence.

The common practice of testing hypotheses

Such a clear distinction between the two ways of thinking is helpful for und standing. In practice, the two approaches often merge. We continued to call c of the hypotheses in a decision problem H_0. The common practice of *testing potheses* mixes the reasoning of significance tests and decision rules as follo

1. State H_0 and H_a just as in a test of significance.
2. Think of the problem as a decision problem, so that the probabilities Type I and Type II errors are relevant.
3. Because of Step 1, Type I errors are more serious. So choose an α (significance level) and consider only tests with probability of Type I error no greater than α.

4. Among these tests, select one that makes the probability of a Type II error as small as possible (that is, power as large as possible). If this probability is too large, you will have to take a larger sample to reduce the chance of an error.

Testing hypotheses may seem to be a hybrid approach. It was, historically, the effective beginning of decision-oriented ideas in statistics. An impressive mathematical theory of hypothesis testing was developed between 1928 and 1938 by Jerzy Neyman and the English statistician Egon Pearson. The decision-making approach came later (1940s). Because decision theory in its pure form leaves you with two error probabilities and no simple rule on how to balance them, it has been used less often than either tests of significance or tests of hypotheses. Decision ideas have been applied in testing problems mainly by way of the Neyman-Pearson hypothesis-testing theory. That theory asks you first to choose α, and the influence of Fisher often has led users of hypothesis testing comfortably back to $\alpha = 0.05$ or $\alpha = 0.01$. Fisher, who was exceedingly argumentative, violently attacked the Neyman-Pearson decision-oriented ideas, and the argument still continues.

SUMMARY

The **power** of a significance test measures its ability to detect an alternative hypothesis. Power against a specific alternative is calculated as the probability that the test will reject H_0 when that alternative is true. This calculation requires knowledge of the sampling distribution of the test statistic under the alternative hypothesis. Increasing the size of the sample increases the power when the significance level remains fixed.

An alternative to significance testing regards H_0 and H_a as two statements of equal status that we must decide between. This **decision theory** point of view regards statistical inference in general as giving rules for making decisions in the presence of uncertainty.

In the case of testing H_0 versus H_a, decision analysis chooses a decision rule on the basis of the probabilities of two types of error. A **Type I error** occurs if H_0 is rejected when it is in fact true. A **Type II error** occurs if H_0 is accepted when in fact H_a is true.

In a fixed level α significance test, the significance level α is the probability of a Type I error, and the power against a specific alternative is 1 minus the probability of a Type II error for that alternative.

SECTION 6.4 EXERCISES

82 You want to see if a redesign of the cover of a mail-order catalog will increase sales. A very large number of customers will receive the original catalog, and a random sample of customers will receive the one with the

new cover. For planning purposes, you are willing to assume that the sales from the new catalog will be approximately normal with $\sigma = 50$ dollars and that the mean for the original catalog will be $\mu = 25$ dollars. You decide to use a sample size of $n = 900$. You wish to test

$$H_0: \mu = 25$$

$$H_a: \mu > 25$$

You decide to reject H_0 if $\bar{x} > 26$ and to accept H_0 otherwise.

(a) Find the probability of a Type I error, that is, the probability that your test rejects H_0 when in fact $\mu = 25$ dollars.

(b) Find the probability of a Type II error when $\mu = 28$ dollars. This is the probability that your test accepts H_0 when in fact $\mu = 28$.

(c) Find the probability of a Type II error when $\mu = 30$.

(d) The distribution of sales is not normal because many customers buy nothing. Why is it nonetheless reasonable in this circumstance to assume that the mean will be approximately normal?

6.83 Example 6.12 gives a test of a hypothesis about the SAT scores of Californi high school students based on an SRS of 500 students. The hypotheses are

$$H_0: \mu = 450$$

$$H_a: \mu > 450$$

Assume that the population standard deviation is $\sigma = 100$. The test rejects H_0 at the 1% level of significance when $z \geq 2.326$, where

$$z = \frac{\bar{x} - 450}{100/\sqrt{500}}$$

Is this test sufficiently sensitive to usually detect an increase of 12 points in the population mean SAT score? Answer this question by calculating th power of the test against the alternative $\mu = 462$.

6.84 Example 6.16 discusses a test about the mean contents of cola bottles. The hypotheses are

$$H_0: \mu = 300$$

$$H_a: \mu < 300$$

The sample size is $n = 6$, and the population is assumed to have a norma distribution with $\sigma = 3$. A 5% significance test rejects H_0 if $z \leq -1.645$, where the test statistic z is

$$z = \frac{\bar{x} - 300}{3/\sqrt{6}}$$

Power calculations help us see how large a shortfall in the bottle contents the test can be expected to detect.

(a) Find the power of this test against the alternative $\mu = 298$.

(b) Find the power against the alternative $\mu = 294$.

(c) Is the power against $\mu = 296$ higher or lower than the value you found in (b)? Explain why this result makes sense.

6.85 Increasing the sample size increases the power of a test when the level α is unchanged. Suppose that in the previous exercise a sample of n bottles had been measured. In that exercise, $n = 6$. The 5% significance test still rejects H_0 when $z \leq -1.645$, but the z statistic is now

$$z = \frac{\bar{x} - 300}{3/\sqrt{n}}$$

where we substitute the sample size for n.

(a) Find the power of this test against the alternative $\mu = 298$ when $n = 30$.

(b) Find the power against $\mu = 298$ when $n = 120$.

6.86 In Example 6.11, a company medical director failed to find significant evidence that the mean blood pressure of a population of executives differed from the national mean $\mu = 128$. The medical director now wonders if the test used would detect an important difference if one were present. For the SRS of size 72 from a population with standard deviation $\sigma = 15$, the z statistic is

$$z = \frac{\bar{x} - 128}{15/\sqrt{72}}$$

The two-sided test rejects

$$H_0: \mu = 128$$

at the 5% level of significance when $|z| \geq 1.96$.

(a) Find the power of the test against the alternative $\mu = 135$.

(b) Find the power of the test against $\mu = 121$. Can the test be relied on to detect a mean that differs from 128 by 7?

(c) If the alternative were farther from H_0, say $\mu = 138$, would the power be higher or lower than the values calculated in (a) and (b)?

6.87 You have an SRS of size $n = 16$ from a normal distribution with $\sigma = 1$. You wish to test

$$H_0: \mu = 0$$

$$H_a: \mu > 0$$

You decide to reject H_0 if $\bar{x} > 0$ and to accept H_0 otherwise.

(a) Find the probability of a Type I error, that is, the probability that your test rejects H_0 when in fact $\mu = 0$.

(b) Find the probability of a Type II error when $\mu = 0.2$. This is the probability that your test accepts H_0 when in fact $\mu = 0.2$.

(c) Find the probability of a Type II error when $\mu = 0.6$.

6.88 Use the result of Exercise 6.84 to give the probabilities of Type I and Type II errors for the test discussed there. Take the alternative hypothesis to be $\mu = 294$.

6.89 Use the result of Exercise 6.83 to give the probability of a Type I error and the probability of a Type II error for the test in that exercise when the alternative is $\mu = 462$.

6.90 You must decide which of two discrete distributions a random variable X has. We will call the distributions p_0 and p_1. Here are the probabilities the assign to the values x of X:

x	0	1	2	3	4	5	6
p_0	0.1	0.1	0.1	0.1	0.2	0.1	0.3
p_1	0.2	0.1	0.1	0.2	0.2	0.1	0.1

You have a single observation on X and wish to test

H_0: p_0 is correct

H_a: p_1 is correct

One possible decision procedure is to accept H_0 if $X = 4$ or $X = 6$ and reject H_0 otherwise.

(a) Find the probability of a Type I error, that is, the probability that you reject H_0 when p_0 is the correct distribution.

(b) Find the probability of a Type II error.

6.91 You are designing a computerized medical diagnostic program. The program will scan the results of routine medical tests (pulse rate, blood pressure, urinalysis, etc.) and either clear the patient or refer the case to doctor. The program will be used as part of a preventive-medicine systen to screen many thousands of persons who do not have specific medical complaints. The program makes a decision about each patient.

(a) What are the two hypotheses and the two types of error that the program can make? Describe the two types of error in terms of "fals positive" and "false-negative" test results.

(b) The program can be adjusted to decrease one error probability, at the cost of an increase in the other error probability. Which error probability would you choose to make smaller, and why? (This is a matter of judgment. There is no single correct answer.)

6.92 **(Optional)** The acceptance sampling test in Example 6.20 has probabili 0.05 of rejecting a good lot of bearings and probability 0.08 of accepting bad lot. The consumer of the bearings may imagine that acceptance

sampling guarantees that most accepted lots are good. Alas, it is not so. Suppose that 90% of all lots shipped by the producer are bad.

(a) Draw a tree diagram for shipping a lot (the branches are "bad" and "good") and then inspecting it (the branches at this stage are "accept" and "reject").

(b) Write the appropriate probabilities on the branches, and find the probability that a lot shipped is accepted.

(c) Use the definition of conditional probability or Bayes's formula (page 350) to find the probability that a lot is bad, given that the lot is accepted. This is the proportion of bad lots among the lots that the sampling plan accepts.

CHAPTER 6 EXERCISES

6.93 Patients with chronic kidney failure may be treated by dialysis, using a machine that removes toxic wastes from the blood, a function normally performed by the kidneys. Kidney failure and dialysis can cause other changes, such as retention of phosphorus, that must be corrected by changes in diet. A study of the nutrition of dialysis patients measured the level of phosphorus in the blood of several patients on six occasions. Here are the data for one patient (in milligrams of phosphorus per deciliter of blood):[14]

5.6 5.1 4.6 4.8 5.7 6.4

The measurements are separated in time and can be considered an SRS of the patient's blood phosphorus level. Assuming that this level varies normally with $\sigma = 0.9$ mg/dl, give a 90% confidence interval for the mean blood phosphorus level.

6.94 The normal range of phosphorus in the blood is considered to be 2.6 to 4.8 mg/dl. Is there strong evidence that the patient in the previous exercise has a mean phosphorus level that exceeds 4.8?

6.95 Because sulfur compounds cause "off-odors" in wine, oenologists (wine experts) have determined the odor threshold, the lowest concentration of a compound that the human nose can detect. For example, the odor threshold for dimethyl sulfide (DMS) is given in the oenology literature as 25 micrograms per liter of wine (μg/l). Untrained noses may be less sensitive, however. Here are the DMS odor thresholds for 10 beginning students of oenology:

31 31 43 36 23 34 32 30 20 24

Assume (this is not realistic) that the standard deviation of the odor threshold for untrained noses is known to be $\sigma = 7$ μg/l.

(b) Calculate the P-value.

(c) Is the result significant at the $\alpha = 0.05$ level? Do you think the study gives strong evidence that the mean compensation of all CEOs went up?

6.102 Statisticians prefer large samples. Describe briefly the effect of increasing the size of a sample (or the number of subjects in an experiment) on each the following:

(a) The width of a level C confidence interval.

(b) The P-value of a test, when H_0 is false and all facts about the population remain unchanged as n increases.

(c) The power of a fixed level α test, when α, the alternative hypothesis, and all facts about the population remain unchanged.

6.103 A roulette wheel has 18 red slots among its 38 slots. You observe many s and record the number of times that red occurs. Now you want to use th data to test whether the probability of a red has the value that is correct a fair roulette wheel. State the hypotheses H_0 and H_a that you will test. (will describe the test for this situation in Chapter 8.)

6.104 When asked to explain the meaning of "statistically significant at the α 0.05 level," a student says, "This means there is only probability 0.05 tha the null hypothesis is true." Is this an essentially correct explanation of statistical significance? Explain your answer.

6.105 Another student, when asked why statistical significance appears so often in research reports, says, "Because saying that results are signific tells us that they cannot easily be explained by chance variation alone. Do you think that this statement is essentially correct? Explain your answer.

6.106 Use a computer to generate $n = 9$ observations from a normal distribu with mean 15 and standard deviation 6: $N(15, 6)$. Find the 95% confide interval for μ. Repeat this process 100 times and then count the numbe times that the confidence interval includes the value $\mu = 15$. Explain y results.

6.107 Use a computer to generate $n = 9$ observations from a normal distribu with mean 15 and standard deviation 6: $N(15, 6)$. Test the null hypothe that $\mu = 15$ using a two-sided significance test. Repeat this process 10 times and then count the number of times that you reject H_0. Explain y results.

6.108 Use the same procedure for generating data as in the previous exercise Now test the null hypothesis that $\mu = 18$. Explain your results.

6.109 Figure 6.2 (page 420) demonstrates the behavior of a confidence interv repeated sampling by showing the results of 25 samples from the

same population. Now you will do a similar demonstration. Suppose that (unknown to the researcher) the mean SATM score of all California high school seniors is $\mu = 475$, and that the standard deviation is known to be $\sigma = 100$. The scores vary normally.

(a) Simulate the drawing of 50 SRSs of size $n = 100$ from this population.

(b) The 95% confidence interval for the population mean μ has the form $\bar{x} \pm m$. What is the margin of error m? (Remember that we know $\sigma = 100$.)

(c) Use your software to calculate the 95% confidence interval for μ when $\sigma = 100$ for each of your 50 samples. Verify the computer's calculations by checking the interval given for the first sample against your result in (b). Use the $\bar{x}$ reported by the software.

(d) How many of the 50 confidence intervals contain the true mean $\mu = 475$? If you repeated the simulation, would you expect exactly the same number of intervals to contain μ? In a very large number of samples, what percent of the confidence intervals would contain μ?

6.110 In the previous exercise you simulated the SATM scores of 50 SRSs of 100 California seniors. Now use these samples to demonstrate the behavior of a significance test. We know that the population of all SATM scores is normal with standard deviation $\sigma = 100$.

(a) Use your software to carry out a test of

$$H_0: \mu = 475$$

$$H_a: \mu \neq 475$$

for each of the 50 samples.

(b) Verify the computer's calculations by using Table A to find the P-value of the test for the first of your samples. Use the $\bar{x}$ reported by your software.

(c) How many of your 50 tests reject the null hypothesis at the $\alpha = 0.05$ significance level? (That is, how many have P-value 0.05 or smaller?) Because the simulation was done with $\mu = 475$, samples that lead to rejecting H_0 produce the wrong conclusion. In a very large number of samples, what percent would falsely reject the hypothesis?

6.111 Suppose that in fact the mean SATM score of California high school seniors is $\mu = 500$. Would the test in the previous exercise usually detect a mean this far from the hypothesized value? This is a question about the power of the test.

(a) Simulate the drawing of 50 SRSs from a normal population with mean $\mu = 500$ and $\sigma = 100$. These represent the results of sampling when in fact the alternative $\mu = 500$ is true.

(b) Repeat on these new data the test of

$$H_0: \mu = 475$$

$$H_a: \mu \neq 475$$

that you did in the previous exercise. How many of the 50 tests have P-values 0.05 or smaller? These tests reject the null hypothesis at the $\alpha = 0.05$ significance level, which is the correct conclusion.

(c) The power of the test against the alternative $\mu = 500$ is the probabili that the test will reject H_0: $\mu = 475$ when in fact $\mu = 500$. Calculate this power. In a very large number of samples from a population with mean 500, what percent would reject H_0?

6.112 Persons aged 55 and over represented 21.3% of the U.S. population in the year 2000. This group is expected to increase to 30.5% by 2025. In terms of actual numbers of people, the increase is from 58.6 million to 101.4 million. Restauranteurs have found this market to be important and would like to make their businesses attractive to older customers. C study used a questionnaire to collect data from people aged 50 and over For one part of the analysis, individuals were classified into two age groups: 50–64 and 65–79. There were 267 people in the first group and 263 in the second. One set of items concerned ambience, menu design, and service. A series of questions were rated on a 1 to 5 scale with 1 representing "strongly disagree" and 5 representing "strongly agree." I some cases the wording of questions has been shortened in the table be Here are the means:

Question	50–64	65–79
Ambience		
Most restaurants are too dark	2.75	2.93
Most restaurants are too noisy	3.33	3.43
Background music is often too loud	3.27	3.55
Restaurants are too smoky	3.17	3.12
Tables are too small	3.00	3.19
Tables are too close together	3.79	3.81
Menu design		
Print size is not large enough	3.68	3.77
Glare makes menus difficult to read	2.81	3.01
Colors of menus make them difficult to read	2.53	2.72
Service		
It is difficult to hear the service staff	2.65	3.00
I would rather be served than serve myself	4.23	4.14
I would rather pay the server than a cashier	3.88	3.48
Service is too slow	3.13	3.10

First examine the means of the people who are 50 to 64. Order the statements according to the means and describe the results. Then do the same for the older group. For each question compute the z statistic and the associated P-value for the comparison between the two groups. For these calculations you can assume that the standard deviation of the difference is 0.08, so z is simply the difference in the means divided by 0.08. Note that you are performing 13 significance tests in this exercise. Keep this in mind when you interpret your results. Write a report summarizing your work.