AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

Minitab is a statistical package, a computer program that performs many statistical procedures. The versions of Minitab now available for use on personal computers are menu-driven and much easier to use than the main-frame version originally discussed in this text. Those sections are not included in this online edition of the text. At this time, the most recent version is Minitab 15, available at very reasonable prices for purchase or rental from:

www.e-academy.com/minitab

System Requirements

Processor:	PC with a 1 GHz 32- or 64-bit processor
Memory:	512 MB or more of available RAM
Disk Space:	125 MB free space available
Operating System:	Microsoft Windows 2000, XP, or Vista.
Display:	A display capable of 1024 X 768 or higher resolution
Software:	Adobe Acrobat Reader 5.0 or higher for Meet Minitab

CHAPTER 9

Basic Ideas in Statistics

IN THIS CHAPTER

Population, parameter
Sample, random sample
Statistical inference
Experimental design
Hypothesis testing, null and alternative hypotheses
Test statistic
Observed significance level, *p*-value
Significance level, power
Point estimate, interval estimate, confidence interval
Parametric, nonparametric

The realm of statistics includes both data analysis and statistical inference. In data analysis, we use graphical tools and descriptive statistics to explore a data set. We want to learn as much as we can about variables and relationships between variables, within a particular sample. With statistical inference, we move beyond the goals of data analysis. We want to use a sample to learn about (make inferences about) the population from which the sample came.

We use the techniques of statistical inference to make probability statements based on our observations. The validity of these statements depends on certain assumptions about the observations (regarding independence or probability distributions, for instance). To assure the reasonableness of these assumptions, we must take care in how we obtain the observations. Collecting samples that allow valid inferences is the realm of experimental design. We should not talk about statistical inference without discussing design of experiments.

Consider, for example, a taste test of Coke and Pepsi. What are the goals of a taste test? How can we conduct the test in order to meet these goals? Should we give some tasters Coke and some other tasters Pepsi and try to determine which group is more satisfied? Should we hand each taster a can of Coke and a can of Pepsi and ask him to decide which he likes best? Or should we do something else? Our willingness to assess the results of the taste test will depend on whether we think the experimental design allowed a fair comparison of the two soft drinks. We will see how one group of students approached this problem, in Example 9-1.

Now consider this question: Will you pedal farther in 15 minutes on an exercise bicycle while listening to music than if you do not listen to music? One student designed an experiment to answer this question for herself. As we will see in Example 9-2, she tried to control as many extraneous variables (such as clothing type and time of day) as possible to allow a valid assessment of her experimental results.

As another example, the engineer in charge of an electroplating process is concerned that too much gold is being plated onto components. Thousands of components are plated each day and she cannot examine all of them. So she decides to use a sample of the components plated one day to learn about the population of all components plated that day. She would like to use the sample to estimate the average thickness of gold plated on all components that day. She would also like to use the sample to decide if the average gold thickness on components plated that day was above the target value. How should she select components for her sample to give her the most confidence in her inferences? We will consider this problem in Example 9-3.

These three examples, introduced in Section 9-2, form the basis of our introductory discussion of statistical inference. Before considering them, however, we need to define some basic terms that will make the discussion clearer.

9-1

Some Definitions Related to Statistical Inference

In *statistical inference,* we use a sample to learn about a population:

> **Statistical inference** refers to the process of drawing conclusions about a population based on a sample from that population.

The *population* is the group or collection we want to learn about. For our purposes, we assume the population is very large compared with the size of the sample.

> The **population** is the set or collection we are interested in learning about. We learn about the population by studying a sample of the population.

Sometimes the population is hypothetical. For instance, to get a feel for whether a coin is fair, we might flip it 200 times and count the number of heads. We make inferences about the fairness of the coin (or the probability of heads) based on the sample of 200 tosses. The population is the hypothetical collection of results we would get if we could toss the coin indefinitely.

A similar situation occurs when an investigator carries out a controlled experiment. Say, he divides some animals into two groups. He subjects all the animals in one group to one experimental condition or treatment, and subjects all the animals in the other group to another condition. He then compares the responses across the two experimental groups. The sample consists of all the animals involved in the experiment. The population is the hypothetical collection of similar animals that might be subjected to the same experimental conditions, if time and money permitted unlimited investigation.

Other times the population has substance. In a quality control setting, a sample of items is used to draw inferences about the acceptability of a large collection of items produced. Or, a sample of 1-year-old babies may be examined so that inferences can be made about physical characteristics (such as height and weight) of a large population of 1-year-old babies.

When we can study the entire population, we are in the realm of data analysis (as when we studied the World Bank data set). Statistical inference is appropriate when we cannot observe the entire population, but can observe only a subset or *sample* of the population. We use this sample to learn about the population.

> A **sample** is a subset of the population. We use the observations in the sample to make inferences about the population.

In formal statistical analysis, the subject of the remaining chapters, we want our sample to be a *random sample in the probability sense:*

> By a **random sample in the probability sense,** we mean a sample with observations that are independent and have the same probability distribution.

We want the probability distribution of the observations in the sample to be the same as that of the population. That is, we want a sample that is *representative* of the population.

> We say a sample is **representative** of the population if it is similar to the population with respect to characteristics we want to study.

The best way to ensure getting a random sample in the probability sense is to collect a *random sample in the experimental sense:*

> When sampling from a population, we say we have a **random sample in the experimental sense** if each member of the population had an equal and independent chance of being included in the sample.

Sometimes, as in a quality control setting, a true random sample from the population may not be hard to get. Other times, as in examining 1-year-old babies, a true random sample from the population may be difficult, or costly, or even impossible to obtain. In such cases, we must be especially careful when interpreting results.

In formal statistical analysis, we select a sample from a population. We make assumptions about the distribution of the observations, then use the sample to make probability statements about the population. Sometimes our statistical inference is *parametric:*

> In **parametric statistical inference,** we assume a specific type of probability distribution (for example, Gaussian) for the observations.

Other times our inference is *nonparametric:*

> In **nonparametric statistical inference,** our assumptions do not specify a particular type of probability distribution for the observations.

We will use both parametric and nonparametric statistical analyses in the chapters that follow.

The descriptions "parametric" and "nonparametric" come from the idea of a *parameter:*

> A number that characterizes the distribution of a population is a **parameter.**

The mean and median of a population are parameters that describe the center of the population values. The standard deviation of a population is a parameter that describes the spread in the population values. One goal of statistical inference is to use a sample to estimate population parameters.

One way to estimate a population parameter is with a *point estimate:*

> A **point estimate** is a single number calculated from the sample to estimate a population parameter.

We often use the sample mean as a point estimate of the population mean. We may use the sample standard deviation as a point estimate of the population standard deviation.

Because of sampling variation, we would not expect to get the same point

estimate of a parameter from different samples. Point estimates have variation associated with them. So in addition to a point estimate, we might like an *interval estimate:*

> An **interval estimate** for a population parameter is a range of reasonable values for the parameter.

Our interval estimates will take the form of *confidence intervals,* with probability interpretations. We will calculate a confidence interval for a population mean (average gold thickness) in Example 9-3. Confidence intervals have different forms depending on the parameter we estimate and the assumptions we make about the sample observations. We will see a number of different forms of confidence intervals in the chapters that follow.

> A **confidence interval** is an interval estimate of a parameter, with a probability interpretation.

In *hypothesis testing,* we formalize the way we ask questions about one or more populations. We ask whether the sample is consistent with a specific hypothesis about the population(s).

> **Hypothesis testing** is a formal strategy for comparing two statements about the state of nature in an experimental situation.

We want to compare two statements about the state of nature. The first statement is generally a "status quo" or "no difference" statement. We call it the *null hypothesis,* or H_0.

> The **null hypothesis** is a statement about the state of nature in an experimental situation, generally a "status quo" or "no difference" statement. We often denote the null hypothesis by H_0.

The other statement is an alternative to the null hypothesis. We call it the *alternative hypothesis,* or H_a.

> The **alternative hypothesis** is a statement about the state of nature, providing an alternative to that specified in the null hypothesis. We often denote the alternative hypothesis by H_a.

As we will see when we discuss the general strategy of hypothesis testing in Section 9-3, we build a probability model for our experiment under the null hypothesis and use this model to make probability statements based on our sample.

The general strategy of hypothesis testing is the same for all hypothesis testing situations. Before describing this general strategy, let's consider three examples.

9-2 Three Examples

We will refer to the following examples later in the chapter as we discuss experimental design and formal statistical analysis.

EXAMPLE 9-1

On the first day of a statistics course, I walked into the classroom with a package of paper cups and several cans of the diet, caffeine-free versions of Coke and Pepsi. I asked the class to design and carry out a taste test comparing these two soft drinks. In particular, I asked the class to:

- State a specific goal. Write this goal in terms of hypotheses to be compared.
- Design an experiment to meet this goal.
- State assumptions and describe how the experiment will be analyzed.
- Carry out the experiment.
- Analyze and interpret the results.

The class enthusiastically accepted my challenge. The following is a summary of their discussion and results.

State a specific goal. The class discussed two goals. The first was to determine whether tasters could tell a difference between Coke and Pepsi. The second was to see, more specifically, if they had a preference for one of these soft drinks. The students decided the second goal was more interesting and relevant for a taste test.

Goal: To see if tasters have a preference for Coke or Pepsi.

They then wrote this goal in terms of two statements to be compared. One statement, called the null hypothesis, says there is no preference. The other statement, called the alternative hypothesis, says there is a preference. (As noted earlier, we sometimes denote the null hypothesis by H_0 and the alternative hypothesis by H_a.)

Null hypothesis: Tasters have no preference for Coke or Pepsi.
Alternative hypothesis: Tasters have a preference for either Coke or Pepsi.

Design an experiment to meet this goal. One student, Karen, could not taste either soft drink for dietary reasons, so she carried out the mechanics of the taste test. The class decided that each of the other 12 students should taste both Coke and Pepsi. The students did not want any taster to know the brand of beverage tasted at any stage. (Why?) They also did not want Karen to know the identity of beverages tasted. (Why?) So the students chose a **double-blind experiment.**

> A **double-blind experiment** is one in which neither the participants nor the experimenter know the identity of the treatments as they are administered and evaluated.

The students wrote a letter C on the bottoms of 12 paper cups and a letter P on the bottoms of 12 more. They filled all the cups about three-quarters full—those labeled C, with Coke, and those labeled P, with Pepsi. Then they paired the cups, one C cup and one P cup per pair. They switched the cups within a pair around, so no one knew which was the C cup and which was the P cup.

Everyone except Karen left the room. Karen placed one pair of cups on

each desk, then the others returned. A taster sampled from each cup as often as desired, leaving enough beverage in a cup so that the letter C or P on the bottom was not visible. Each of the 12 tasters chose a preferred beverage. If a taster really felt no preference, he or she had to choose one of the cups anyway. After everyone had made a selection, each taster gave Karen the cup of the preferred drink; she checked the letter on the bottom of the cup and wrote it down.

State assumptions and describe how the experiment will be analyzed. All the students had already had a beginning statistics course, so they knew they needed to make some simplifying assumptions in order to analyze the results of the experiment. The assumptions they made were:

Experimental assumptions
The tasters make choices independently of one another.
Each taster has the same probability p of choosing Coke.

The students believed the first assumption was reasonable. However, they thought the second assumption was most likely a great simplification of reality; we will discuss this later in the chapter.

Let Y denote the number of tasters who choose Coke. Under the two assumptions listed above, Y has a Binomial(12, p) distribution. We can rewrite the null and alternative hypotheses as

$$H_0: \ p = \frac{1}{2} \qquad \text{and} \qquad H_a: \ p \neq \frac{1}{2}$$

The null hypothesis says that a taster is just as likely to select Pepsi as Coke. This seems reasonable when a taster really has no preference, but must select one of the cups anyway. Under the null hypothesis, the random variable Y has a Binomial(12, $\frac{1}{2}$) distribution. We found this probability distribution in Example 7-5. The probabilities, listed in Table 7-5, are shown again in Table 9-1.

Extreme values of Y, near 0 or 12, cause us to doubt the null hypothesis, suggesting that there is a preference for Pepsi or Coke. Moderate values of Y, not too far from 6, are consistent with the null hypothesis of no preference. With this in mind, the students chose the following decision rule:

Decision rule
If $3 \leq Y \leq 9$, say the experimental results are consistent with the no-preference null hypothesis.
If $Y \leq 2$ or $Y \geq 10$, say the experimental results are inconsistent with the no-preference null hypothesis, suggesting there is a preference.

Using Table 9-1, we can calculate the probability of saying the results are inconsistent with the null hypothesis, when the null hypothesis is really true:

$$P(Y \leq 2 \text{ or } Y \geq 10 \text{ when } H_0 \text{ is true}) = .038$$

We call .038 the *significance level* of the test. This is the probability the students will say the experimental results are inconsistent with what they would expect

TABLE 9-1 Probabilities for a random variable Y having a Binomial(12, $\frac{1}{2}$) distribution. The probabilities have been rounded to 4 decimal places; this is why they do not sum exactly to 1.

k	$P(Y = k)$
0	.0002
1	.0029
2	.0161
3	.0537
4	.1208
5	.1934
6	.2256
7	.1934
8	.1208
9	.0537
10	.0161
11	.0029
12	.0002

under the null hypothesis, when the null hypothesis is really true. With this decision rule, there is about a 4% chance the students will say there is a preference for Coke or Pepsi when there really is none.

Carry out the experiment. The class carried out the experiment as described above. Nine of the students selected Coke as their preferred beverage and three selected Pepsi.

Analyze and interpret the results. Since nine students selected Coke, $Y = 9$. Using the decision rule given above, the class decided the experimental results were consistent with the no-preference null hypothesis.

The students then calculated the probability under the null hypothesis of seeing results as extreme as or more extreme than the observed result. This probability is called the *observed significance level* or *p-value*. We can find the p-value using the probabilities in Table 9-1:

$$p\text{-value} = P(Y \leq 3 \text{ or } Y \geq 9 \text{ when } H_0 \text{ is true}) = .146$$

Note that $Y \leq 3$ is just as extreme in the direction of a Pepsi preference as $Y \geq 9$ is in the direction of a Coke preference. If there were really no preference, there would be slightly better than a 14% chance of seeing a result at least as extreme as the one observed.

The class decided that seeing 9 of 12 students prefer Coke was not extreme enough to discredit the no-preference null hypothesis; they were unwilling to say there was a preference. In addition, they noted that the 12 tasters (graduate students at a northeastern university) were unlikely to be representative of the population of American soft drink consumers. Therefore, they could not interpret the experimental results in terms of any population beyond the confines of the class.

EXAMPLE 9-2 Will you pedal farther in 15 minutes on an exercise bicycle while listening to a Walkman or not? As part of a class project, a student tried to answer this question for herself (Walsh, 1988). The student, Michele, used an exercise bicycle and a Walkman equipped with a tape of five "pop rock" songs. She planned to use her results to compare two statements:

Null hypothesis: The median distance pedaled is the same with and without the Walkman.

Alternative hypothesis: The median distance pedaled is different with than without the Walkman.

Michele carried out her experiment on four consecutive weekdays, between 8:00 and 8:30 A.M. Each morning she pedaled for 15 minutes, timed by a clock with an alarm. The odometer on the exercise bicycle was covered for each trial, so she could not see the distance pedaled. Two of the four runs were with the Walkman and two without. All other conditions (clothing, warm-up, room temperature) were constant. By a random process such as drawing numbers from a hat, Michele determined that she would make the runs on the first and fourth days wearing the Walkman. She made the runs on the second and third days without the Walkman.

This experiment was not double-blind as the taste test was. Michele was both the experimenter and the participant in her investigation. She obviously knew which treatment (Walkman or no Walkman) she used in each trial. Since she knew why she was conducting the experiment, her preconceptions could influence her results. This is a major drawback of a study that is not blinded. Can you think of some ways Michele could have improved her experimental design?

The results of Michele's experiment are shown in Table 9-2. Let's use her results to compare her null and alternative hypotheses.

We should always start by plotting data in any way that seems reasonable. Figure 9-1 shows a plot of Michele's experimental results. We see that she pedaled farther in both runs with the Walkman than she did in either run without the Walkman.

These results suggest that Michele tends to pedal farther wearing the Walkman. Let's do a formal statistical analysis of the results. As in many experimental situations, there are several approaches we could take. For this illustration, we will keep things simple, using a method of analysis known as the median test (see Section 11-5).

The median of Michele's four observations is 3.2. We classify each observation by treatment and by position relative to the overall median. This information is displayed in Table 9-3.

We know that half the experimental runs were made with the Walkman and half without. We also know that half the observations are above the overall median and half below. Would we be surprised, if the null hypothesis were true, to see both the Walkman results above the overall median and both the no-Walkman results below? To answer this question, we need to develop a probability model for the experiment under the null hypothesis.

TABLE 9-2 Results of an experiment to determine the effect of wearing a Walkman on distance pedaled on an exercise bicycle

Day	Experimental condition	Miles pedaled
Tuesday	Walkman	3.5
Wednesday	No Walkman	2.9
Thursday	No Walkman	2.7
Friday	Walkman	3.8

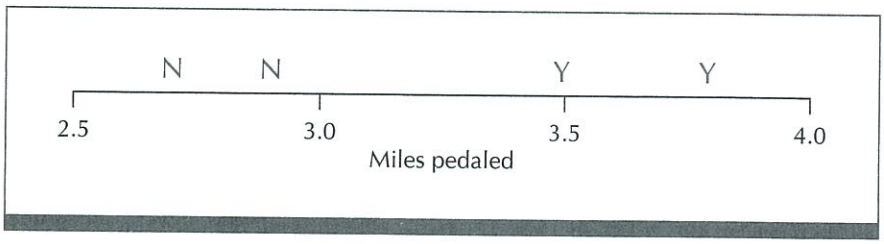

FIGURE 9-1 Plot of the results of the exercise bicycle experiment in Example 9-2: Y = Walkman, N = No Walkman

TABLE 9-3 The observations in Example 9-2 classified by treatment and position relative to the overall median

	Treatment		
	Walkman	No Walkman	Total
Above median	2	0	2
Below median	0	2	2
Total	2	2	4

Let's assume that the results of the four trials are independent of one another:

Experimental assumptions
The results of the four trials are independent.

Under the null hypothesis, we expect a random distribution of the two below-median results across the four trials, regardless of treatment. Let X equal the number of Walkman results below the overall median. Then under the null hypothesis, X has a hypergeometric distribution:

$$P(X = k \text{ when } H_0 \text{ is true}) = \frac{\binom{2}{k}\binom{2}{2-k}}{\binom{4}{2}} \quad \text{for } k = 0, 1, 2$$

The hypergeometric probability $P(X = k$ when H_0 is true) equals the number of ways we could select k Walkman trials and $2 - k$ no-Walkman trials to be below the overall median, divided by the number of ways we could select two of the four trials to be below the overall median. (We discussed hypergeometric distributions in Section 7-3.)

The observed value of X is 0. This value is as extreme as possible in supporting the statement that the median distance pedaled is greater with the Walkman. A value of 2 is just as extreme in supporting the statement that the median distance pedaled is greater without the Walkman. The alternative hypothesis covers both these possibilities. Therefore, the observed significance level or p-value is

$$P(X = 0 \text{ or } X = 2 \text{ when } H_0 \text{ is true}) = \frac{\binom{2}{0}\binom{2}{2}}{\binom{4}{2}} + \frac{\binom{2}{2}\binom{2}{0}}{\binom{4}{2}} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

If the null hypothesis were true, there would be 1 chance in 3 of seeing results at least as extreme as those actually observed. This is a pretty good chance; based on the p-value, we would say the experimental results are consistent with the null hypothesis.

This conclusion seems to contradict the observations. Both distances pedaled with the Walkman are greater than both distances pedaled without the Walkman. An explanation of this seeming contradiction is that we have a very small sample size. Stronger evidence in favor of either hypothesis requires larger sample sizes. Also, in our analysis, we ignored the actual distances pedaled, noting only whether each observation was above or below the overall median. In Chapter 11 we discuss ways to use the relative magnitudes (ranks) as well as the actual values of the observations to analyze the results of this type of experiment.

EXAMPLE 9-3

In a high-technology factory, workers make hardware for personal computers. At one stage in the production process, gold is electroplated onto small components somewhat resembling sewing needles. The target thickness for the gold is 1.000 unit. The minimum acceptable thickness is .980 unit. While thicknesses greater than 1.000 unit are acceptable for use, managers want to avoid the use of too much gold.

Emily is in charge of the electroplating process. She knows that the process is controlled to keep thicknesses above the minimum acceptable value of .980 unit, but she is concerned that perhaps too much gold is being used. Emily decides to design an experiment to answer the question: Is too much gold being plated onto these components?

Out of a day's production lot consisting of many thousands of these components, Emily selects 100 at random for inspection. (She selects a *random sample,* meaning that each component in the lot has an equal and independent

chance of being included in the sample.) She plans to use her results to compare two statements:

Null hypothesis: The average thickness of gold plated onto the components is 1.000 unit.

Alternative hypothesis: The average thickness of gold plated onto the components is greater than 1.000 unit.

The alternative hypothesis is one-sided, allowing for an average thickness greater than 1.000 unit, but not for an average less than 1.000 unit. This one-sided alternative corresponds to Emily's concern that too much gold is being used.

Let μ denote the average thickness of gold on components in the day's production lot. We can write the null and alternative hypotheses as

$$H_0: \ \mu = 1.000 \qquad \text{and} \qquad H_a: \ \mu > 1.000$$

Let $\bar{X}$ and s denote the sample mean and standard deviation, respectively, of the 100 thicknesses observed in the sample. The estimated standard deviation of $\bar{X}$ is called the *standard error* (SE) of $\bar{X}$. The standard error of $\bar{X}$ for this example is $\text{SE} = s/\sqrt{100}$.

> The **standard error** of a variable quantity is an estimate of the standard deviation of the quantity.
>
> For example, if $\bar{X}$ and s represent the sample average and standard deviation of a random sample of size n, then the standard error of $\bar{X}$, called the **standard error of the mean,** is $\text{SE} = s/\sqrt{n}$.

To carry out a statistical analysis, Emily makes three assumptions about her experiment:

Experimental assumptions

The observations are independent.

The observations come from the same distribution.

The sample size of 100 is large enough that $(\bar{X} - \mu)/\text{SE}$ has approximately the standard Gaussian distribution (see the Central Limit Theorem results in Section 8-3).

Because she selects 100 components at random from a large production lot, Emily believes these assumptions are reasonable for her experiment.

Under the null hypothesis, $\mu = 1.000$ and $(\bar{X} - 1.000)/\text{SE}$ has approximately the standard Gaussian distribution. Emily decides that values of this quantity near 0 are consistent with the null hypothesis. Values much greater than 0 are inconsistent with her null hypothesis, suggesting that too much gold is being electroplated. With this in mind, she chooses the following decision rule:

Decision rule

If $(\bar{X} - 1.000)/\text{SE} < 1.65$, say the results are consistent with the null hypothesis.

If $(\bar{X} - 1.000)/\text{SE} \geq 1.65$, say the results are inconsistent with the null hypothesis.

Using a table of probabilities for the standard Gaussian distribution (such as Table B), we see that

$$P\left(\frac{\bar{X} - 1.000}{\text{SE}} \geq 1.65 \text{ when } H_0 \text{ is true}\right) \doteq .0495$$

or about .05. (The symbol $\doteq$ means two quantities are close in value or approximately equal.) There is about a 5% chance that Emily will say her experimental results are inconsistent with the null hypothesis, when the null hypothesis is really true. We call this probability the *significance level* of her test of hypothesis.

Besides asking whether too much gold is being used, Emily wants to estimate the mean thickness μ of gold on components in the production lot. She will use the sample mean $\bar{X}$ as a point estimate of μ. But, because it is based on a sample, she does not expect $\bar{X}$ to equal μ exactly. She would like to use her sample to find a range of reasonable values for the average thickness of gold plated onto the components that day. Like the test of hypotheses, her range of reasonable values for μ is based on the standard Gaussian distribution.

Referring to a table of probabilities for the standard Gaussian distribution (such as Table B), Emily makes the following approximate probability statement:

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\text{SE}} \leq 1.96\right) \doteq .95$$

Using some algebra, she rewrites this as

$$P(\bar{X} - 1.96\text{SE} \leq \mu \leq \bar{X} + 1.96\text{SE}) \doteq .95$$

From this probability statement, Emily decides to use the interval from $\bar{X} - 1.96\text{SE}$ to $\bar{X} + 1.96\text{SE}$ as a range of reasonable values (or interval estimate) of μ.

We say that the interval $(\bar{X} - 1.96\text{SE}, \bar{X} + 1.96\text{SE})$ is an approximate 95% *confidence interval* for μ. The correct probability interpretation of this confidence interval is: Suppose Emily selects 100 components at random from the production lot. Based on the 100 measured thicknesses, she constructs a confidence interval as above. She puts these 100 components back into the lot and then selects another 100 components at random, again constructing a confidence interval based on the 100 measured thicknesses. If she repeats this experiment many times, she obtains many confidence intervals for μ, one from each experiment. She would expect about 95% of these confidence intervals to contain μ, about 5% not to contain μ. The probability interpretation of a confidence interval is based on this idea of repeated sampling. Since Emily actually takes only one sample, she calculates just one confidence interval for μ; this interval either contains μ or it does not.

If Emily had started with a different probability statement, she would have obtained a different confidence interval. For instance, $(\bar{X} - 2.58\text{SE}, \bar{X} + 2.58\text{SE})$ is an approximate 99% confidence interval for μ.

FIGURE 9-2 Stem-and-leaf plot of the measured thickness of gold for 100 components in Example 9-3. The stem is the thickness to one-hundredth of a unit; the leaf gives the nearest thousandth of a unit.

Stem	Leaf
.97	8 9
.98	4 5 6 6 7 9 9
.99	4 4 5 6 6 7 7 7 7 8 8 9 9 9
1.00	0 0 0 1 1 1 1 2 2 3 3 3 4 5 5 6 6 8 9
1.01	0 0 0 0 0 1 1 1 2 2 2 3 3 3 4 5
1.02	1 1 3 4 5 5 6 6 8 9
1.03	2 3 4 7 8
1.04	5 6 9
1.05	
1.06	0 3
1.07	4 6 8
1.08	1 2 5 6
1.09	6 7
1.10	4 6 7
1.11	
1.12	5
1.13	1 4
1.14	
1.15	2 5 9
1.16	5
1.17	
1.18	3
1.19	
1.20	1
1.21	
1.22	7

Having decided on her analysis, Emily goes ahead and randomly selects 100 components from the day's production lot. The 100 measured thicknesses are displayed in the stem-and-leaf plot in Figure 9-2. The distribution is positively skewed. We see that two of the 100 observations are below the acceptable limit of .980 unit, most thicknesses are clustered around the target value of 1.000 unit, and there is a long tail toward larger values. This plot supports Emily's concern that too much gold is being used in the plating process. (Note, however, that 2% of units with gold thickness below the .980 unit specification may be unacceptably large. This is a consideration in quality control that we are not addressing here.)

Emily calculates the sample mean, standard deviation, and standard error based on her 100 observations:

$$\bar{X} = 1.0345 \text{ units} \qquad s = .0537 \text{ unit} \qquad \text{SE} = .00537 \text{ unit}$$

To compare her hypotheses, Emily then calculates

$$\frac{\bar{X} - 1.000}{\text{SE}} = \frac{1.0345 - 1.000}{.00537} = 6.42$$

Since 6.42 is larger than 1.65, she decides the results are inconsistent with the null hypothesis; the average thickness μ of gold on the components seems to be greater than 1.000 unit.

Emily decides to calculate the probability of seeing a value of $(\bar{X} - 1.000)/\text{SE}$ greater than or equal to 6.42 under the null hypothesis:

$$P\left(\frac{\bar{X} - 1.000}{\text{SE}} \geq 6.42 \text{ when } H_0 \text{ is true}\right)$$

This probability is the *p-value* or *observed significance level* of her experimental results. This p-value approximately equals $P(Z \geq 6.42)$, where Z has the standard Gaussian distribution. From Table B, we see that $P(Z \geq 6.42)$ is less than .0002. If the average gold thickness were really 1.000 unit, the probability is less than .0002 that a random sample of size 100 would result in a value of $(\bar{X} - 1.000)/\text{SE}$ greater than or equal to 6.42. Such a small p-value strongly discredits the null hypothesis; we say the experimental results are extremely inconsistent with the null hypothesis.

Emily estimates the average thickness μ of gold plated onto components that day with the sample mean, $\bar{X} = 1.0345$ units. She calculates a 95% confidence interval for μ:

$$(\bar{X} - 1.96\text{SE}, \bar{X} + 1.96\text{SE})$$
$$= (1.0345 - 1.96 \times .00537, 1.0345 + 1.96 \times .00537) = (1.0240, 1.0450)$$

Emily uses the interval 1.024 to 1.045 units as her range of reasonable values for μ. This interval is illustrated in Figure 9-3, along with a box plot of the observations. Because of the large sample size, the interval estimate of the mean is narrow compared with the range of the observations.

Emily estimates that the average gold thickness on components in her lot is .024 to .045 unit greater than the target value. Her next step is to decide, with other members of management, whether this excess justifies the expense of modifications to the gold plating process. If so, she will try to modify the process so that less gold is used while still achieving acceptable minimum

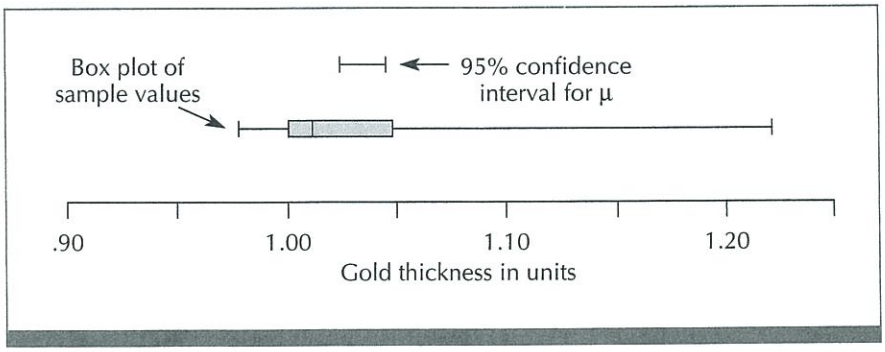

FIGURE 9-3 An approximate 95% confidence interval for the mean thickness μ of gold plated onto components in Example 9-3. Also shown is a box plot summarizing the 100 sample values.

thickness. Do you think, looking at Figure 9-2, that the plating process has more problems with *variability* than with *average* thickness of gold plated? (We will discuss problems in variability in Chapter 14.)

We now leave these three examples briefly to outline the general strategy of hypothesis testing, in Section 9-3. We then return to the examples in Section 9-4, discussing them in terms of ideas in hypothesis testing and estimation.

9-3.

The General Strategy of Hypothesis Testing

We want to use the sample to compare two statements about the state of nature. The first statement is generally a "status quo" or "no difference" statement, called the *null hypothesis*. The other statement is an alternative to the null hypothesis, called the *alternative hypothesis*. We will discuss two approaches to hypothesis testing: the *significance level approach* and the *p-value approach*. Let's outline the significance level approach first.

The significance level approach to hypothesis testing

1. State a null hypothesis and an alternative hypothesis to be compared.
2. Formulate a *test statistic* that measures how far the observations differ from what we would expect under the null hypothesis.
3. Make assumptions about the sample, specifying the probability distribution the test statistic would have if the null hypothesis were true.
4. Select a number α, called the *significance level* of the test. The significance level is a probability. Typical values of α are .05 and .01.
5. Define a set of possible values of the test statistic called the *acceptance region*. Values in the acceptance region are consistent with the null hypothesis, and

   $$P(\text{test statistic is in the acceptance region when } H_0 \text{ is true}) = 1 - \alpha$$

   The remaining possible values of the test statistic make up the *rejection region*. Values in the rejection region are inconsistent with the null hypothesis, and

   $$P(\text{test statistic is in the rejection region when } H_0 \text{ is true}) = \alpha$$

6. Based on the regions defined in step 5, formulate a *decision rule:*

   If the test statistic is in the acceptance region, say the results are consistent with the null hypothesis.

   If the test statistic is in the rejection region, say the results are inconsistent with the null hypothesis.

7. Collect a random sample in the probability sense that will allow us to compare the hypotheses in step 1. Calculate the test statistic based on the sample. Use the decision rule in step 6 to decide whether the observations are consistent with the null hypothesis. Draw conclusions based on the experimental results.

Using the formal strategy of hypothesis testing, we use the test statistic to decide whether the experimental results seem to be consistent with the null hypothesis.

> A **test statistic** is a measure of how much the sample observations differ from what we would expect if the null hypothesis were true.

When our test statistic falls in the acceptance region, we say the experimental results are consistent with the null hypothesis; using the terminology of classical statistics, we fail to reject the null hypothesis. When the test statistic falls in the rejection region, we say the experimental results are inconsistent with the null hypothesis; in the words of classical statistics, we reject the null hypothesis. The classical terminology implies a finality to a hypothesis testing situation that seldom exists. A single experiment rarely decides an issue conclusively. Rather, it adds to our body of knowledge in some way. We make decisions and choose courses of action based on experimental results, doing the best we can with the information available, aware that later results may lead us to different conclusions. We should keep this in mind whenever we interpret the results of an experiment.

We can see from steps 4, 5, and 6 above that the significance level is the probability of saying the results are inconsistent with the null hypothesis, when the null hypothesis is really true.

> Using the significance level approach to hypothesis testing, the **significance level** (or level) is the probability of saying the observations are inconsistent with the null hypothesis, when the null hypothesis is really true.

The significance level is the probability of making what we call a *Type I error.*

> In hypothesis testing, we make a **Type I error** if we say the experimental results are inconsistent with the null hypothesis, when the null hypothesis is really true.

We make a *Type II error* if we say the results are consistent with the null hypothesis when the alternative is really true.

> In hypothesis testing, we make a **Type II error** if we say the experimental results are consistent with the null hypothesis, when the alternative is really true.

Whether we make an error depends on the decision we make and on which hypothesis is true, as summarized in Table 9-4.

Related to significance level is the *power of the test:*

> Using the significance level approach to hypothesis testing, the **power of the test** is the probability of saying the results are inconsistent with the null hypothesis, when an alternative to the null hypothesis is really true.

For a specific alternative, the power of the test is 1 minus the probability of making a Type II error. We will discuss how to calculate the power of a test in Section 9-4.

TABLE 9-4 Whether we make an error in hypothesis testing depends on the decision we make and on the state of nature.

	State of nature	
Decision	Null hypothesis is true	Alternative hypothesis is true
Results consistent with null hypothesis	No error	Type II error
Results inconsistent with null hypothesis	Type I error	No error

Another approach to hypothesis testing is the p-value approach, outlined below.

The p-value approach to hypothesis testing

1. State a null hypothesis and an alternative hypothesis to be compared.
2. Formulate a test statistic that measures how far the observations differ from what we would expect under the null hypothesis.
3. Make assumptions about the sample, specifying the probability distribution the test statistic would have if the null hypothesis were true.
4. Collect a random sample in the probability sense that will allow us to compare the hypotheses stated in step 1. Calculate the test statistic based on the sample.
5. Calculate the *observed significance level* or *p-value*. The p-value is the probability under the null hypothesis of seeing a test statistic as extreme as or more extreme (in the direction of the alternative) than the one actually observed.
6. If the p-value is large, say the results are consistent with the null hypothesis. If the p-value is small, say the results are inconsistent with the null hypothesis.

A large p-value means a test statistic at least as extreme as that observed would not be surprising if the null hypothesis were true; so, we say the results are consistent with the null hypothesis. A small p-value means we would not be likely to see a test statistic as extreme as or more extreme than the one observed, if the null hypothesis were true. Therefore, we say the results are inconsistent with the null hypothesis.

> The ***p*-value,** or observed significance level, is the probability of seeing a test statistic as extreme as or more extreme (in the direction of the alternative) than the one observed, if the null hypothesis were really true.

There are many questions we can ask at this point. How do we decide what the hypotheses should be? What assumptions do we make about the sample? What if the assumptions we make do not apply to the actual sampling process? How do we select a test statistic? Should we use the significance level

or p-value approach to hypothesis testing? When is a p-value "small" and when is it "large"? As we will see in Section 9-4, the answer to all of these questions is: It depends on the situation.

9-4 Some Comments on Hypothesis Testing

Let's address the questions we just posed in terms of the examples in Section 9-2. Then we will make some additional observations on hypothesis testing.

How do we decide what the hypotheses should be? The null hypothesis is generally a statement of no difference or status quo. In Example 9-1, the null hypothesis states that the students have no preference for Coke or Pepsi. The null hypothesis in Example 9-2 is that the median distance pedaled is the same with and without the Walkman. In Example 9-3, the null hypothesis says that the average thickness of gold plated onto the components is equal to the target value of 1.000 unit.

The alternative hypothesis is an alternative to the null hypothesis. In Example 9-1, the alternative states that the students have a preference for either Coke or Pepsi. This alternative allows for deviations from the no-preference null hypothesis in the direction of either a preference for Coke or a preference for Pepsi. Such a two-sided alternative is appropriate when we do not know what to expect from the experiment. We also use a two-sided alternative when we want to allow for extremes on either side of the null hypothesis, even when we do have some prior expectation of how the experiment will turn out.

> A **two-sided alternative hypothesis** allows for extremes in two directions, on either side of the state of nature specified in the null hypothesis.

The alternative hypothesis in Example 9-2 is two-sided. It states that the median distance pedaled is different with than without the Walkman. This alternative allows the possibility that the median distance pedaled is greater with the Walkman, as well as the possibility that the median distance pedaled is greater without the Walkman.

A one-sided alternative makes sense in Example 9-3. The manager is concerned that too much gold is being used. Her alternative hypothesis says that the average thickness of gold plated onto the components is greater than the target value of 1.000 unit.

> A **one-sided alternative hypothesis** allows for extremes in just one direction, on just one side of the state of nature specified in the null hypothesis.

Null and alternative hypotheses must be developed specially for each experimental situation. They should be written down *before* the data are examined. The reason is that after looking at the sample, we may be able to state hypotheses that fit what we see, defeating the point of hypothesis testing. In

fact, null and alternative hypotheses should be developed during the experimental design stage, before the sample is selected. With clearly stated hypotheses, we can then determine whether our planned experiment will be adequate to test the hypotheses that interest us.

What assumptions do we make about the sample? The assumptions we make depend on what we know about the experimental design, on how the observations were obtained. They also depend on what we know or are willing to assume about the distribution of values in the population. The major assumption that is common to all the hypothesis testing situations we will discuss is that our observations were made independently and came from a common population. Our aim is to use this sample of independent observations to infer characteristics of the larger population.

The students in Example 9-1 assumed that the tasters made independent beverage selections. They also assumed that each taster had the same probability p of choosing Coke. Such assumptions are simplifying, perhaps unrealistic. There may be biological differences that make some tasters more likely to prefer Coke, others more likely to prefer Pepsi. We might think of p as the proportion of tasters who will choose Coke. (In a large-scale test, this is the most reasonable interpretation of p. Then the null hypothesis states that half the population prefers Coke; the alternative states that either more than half the population prefers Coke or more than half prefers Pepsi.) We then simplify the situation by supposing that each taster has probability p of choosing Coke. Under the no-preference null hypothesis, this probability p is equal to $\frac{1}{2}$.

For the bicycling experiment in Example 9-2, we assumed the four observations were independent. Under the null hypothesis, we assumed a random distribution of results across the four trials, regardless of treatment (Walkman or no Walkman).

In the sampling inspection problem of Example 9-3, the experimenter assumed the observations were independent, with the same distribution. These assumptions seemed reasonable because the components in the sample were selected at random from a large production lot. She made an additional assumption—that the sample size was large enough to use a Gaussian approximation to the distribution of the sample mean.

In the chapters that follow, we will see how different sets of assumptions lead to different probability models. For a given experiment, the assumptions we make about the sampling process determine the hypothesis testing procedure we should use.

What if the assumptions we make do not apply to the actual sampling process? In most hypothesis testing situations, we must make some simplifying assumptions. We then build a probability model we think will be good enough to make reasonable inferences about the state of nature. If our assumptions provide too poor a model of reality, the inferences we make may be incorrect; how incorrect depends on how badly our model approximates reality.

For instance, in Example 9-1, there were 12 tasters. We assumed the tasters made independent selections, each taster having probability p of choosing Coke. There were men and women in the class. Students represented several racial backgrounds and national origins. If some factors (such as sex, race, or cultural background) make some students tend to have similar responses, the model assumptions break down. Then it is not clear how the results of the test of hypotheses apply to the actual experiment.

How do we select a test statistic? The test statistic measures how far the observed sample differs from what we would expect if the null hypothesis were true. It must have a known probability distribution under the null hypothesis. This probability distribution allows us to make probability statements based on the experimental results.

The test statistic Y in Example 9-1 equals the number of tasters choosing Coke. With the assumptions we make, Y has a Binomial(12, $\frac{1}{2}$) distribution under the null hypothesis.

In Example 9-2, the test statistic X equals the number of Walkman results below the overall median. With independent observations, X has a hypergeometric distribution under the null hypothesis.

The test statistic is $(\bar{X} - 1.000)/\text{SE}$ in Example 9-3. With the assumptions we make, this test statistic has approximately the standard Gaussian distribution under the null hypothesis.

In the chapters that follow, we will discuss tests of hypotheses for a number of different experimental situations. We will see how the choice of test statistic depends on the experimental design, on the type of observations (such as continuous or categorical), and on the assumptions we make about the sampling process.

How do we define an acceptance region and a rejection region when using the significance level approach? The acceptance region contains possible values of the test statistic near what is expected under the null hypothesis. The rejection region contains values far from what is expected under the null hypothesis.

In Example 9-1, the test statistic Y is the number of tasters choosing Coke. We expect values of Y near 6 under the null hypothesis. Values close to 0 or 12 are far from what we expect under the null hypothesis. Three reasonable choices of acceptance and rejection regions are shown in Table 9-5, along with the significance level associated with each choice.

TABLE 9-5 Three possible choices of acceptance and rejection regions in Example 9-1

Choice	Acceptance region	Rejection region	Significance level, α
1	{2, 3, 4, 5, 6, 7, 8, 9, 10}	{0, 1, 11, 12}	.006
2	{3, 4, 5, 6, 7, 8, 9}	{0, 1, 2, 10, 11, 12}	.038
3	{4, 5, 6, 7, 8}	{0, 1, 2, 3, 9, 10, 11, 12}	.146

The students in Example 9-1 selected choice 2, with significance level .038. With this choice, the probability of saying the results are inconsistent with the null hypothesis, when the null hypothesis is really true, is .038.

The test statistic in Example 9-3 is $(\bar{X} - 1.000)/\text{SE}$, where $\bar{X}$ is the sample mean gold thickness. We expect values of the test statistic near 0 under the null hypothesis, values greater than 0 under the alternative. The acceptance region is $(-\infty, c)$ and the rejection region is $[c, \infty)$. Here, c is a constant that gives the desired significance level.

The manager wanted a significance level α close to .05. She knew that if Z is a standard Gaussian random variable, then $P(Z \geq 1.65) = .0495$. She decided to use $c = 1.65$, so her acceptance region was $(-\infty, 1.65)$. Her rejection region was $[1.65, \infty)$.

For a different significance level, she would choose a different value of c. If she wanted significance level $\alpha = .004$, she would choose $c = 2.65$. Then she would say values of the test statistic less than 2.65 are consistent with the null hypothesis and values greater than or equal to 2.65 are inconsistent with the null hypothesis.

Why do we consider values of the test statistic more extreme than the one actually observed when calculating a p-value? The p-value is the probability of observing a test statistic as extreme as or more extreme (in the direction of the alternative hypothesis) than the one actually observed, if the null hypothesis were really true. The p-value is a measure of our surprise at seeing such a test statistic, if the null hypothesis were true. The "as extreme" part of the definition of p-value makes sense, but why also consider values of the test statistic more extreme than that observed? One way to answer this question is with an example.

Suppose we are in charge of deciding whether the coin that will be used to determine the kick-off in the next Super Bowl appears to be fair. We want to compare the following hypotheses:

Null hypothesis: The coin is fair.
Alternative hypothesis: The coin is not fair.

We toss the coin 100 times and observe 53 heads. Is this result consistent with the null hypothesis or not? Most of us would not be surprised at seeing 53 heads in 100 tosses of a fair coin. How does this lack of surprise relate to a test of hypothesis?

We must formulate a probability model. Let's suppose that the tosses are independent. Under the null hypothesis, the probability of a head on each toss is $\frac{1}{2}$. Then the number of heads observed in 100 tosses has a Binomial$(100, \frac{1}{2})$ distribution.

What is the probability of observing exactly 53 heads if the coin is fair? This is the probability of seeing 53 successes in 100 independent trials, with probability $\frac{1}{2}$ of success on each trial:

$$P(53 \text{ successes when } H_0 \text{ is true}) = \binom{100}{53}\left(\frac{1}{2}\right)^{100} = .067$$

Even though seeing 53 heads in 100 tosses of a fair coin does not surprise us, the probability of its happening is just .067. (In fact, the probability of seeing exactly 50 heads is only slightly larger, .08. If we repeated this 100-toss experiment many times, we would be surprised to see exactly 50 heads and 50 tails in a large percentage of these repetitions. We would expect about 92% of our repetitions to result in something other than an exact 50-50 split between heads and tails.)

Looking at the probability of 53 heads in 100 tosses is not enough. So we ask: If the null hypothesis were true, what results would be more surprising than 53 heads? Fifty-four or more heads would be more surprising than 53 heads. If we add up the probabilities of 53 or more heads, we have

$$P(53 \text{ or more heads when } H_0 \text{ is true}) = \sum_{k=53}^{100} \binom{100}{k}\left(\frac{1}{2}\right)^{100} = .31$$

If the coin were fair, we would expect about 31% of all 100-toss experiments to end in 53 or more heads.

Forty-seven or fewer heads would be just as surprising as 53 or more heads. The chance of 47 or fewer heads under the null hypothesis is also .31. Therefore, the p-value associated with our two-sided alternative is equal to .31 + .31 = .62. This large p-value corroborates our lack of surprise at seeing 53 heads in 100 tosses, under the fair-coin hypothesis. Exactly 53 heads in 100 tosses has relatively small probability. But 53 is close to what we would expect if the null hypothesis were true.

In general, we consider the probability of seeing a test statistic at least as extreme as the observed value, under the null hypothesis. If this probability is large, we consider the test statistic to be close to what is expected under the null hypothesis; then we say the results are consistent with the null hypothesis. If this probability is small, then we consider the test statistic to be far from what is expected under the null hypothesis; then we say the results are inconsistent with the null hypothesis.

When is a p-value "small" and when is it "large"? Large p-values are consistent with the null hypothesis and small p-values are inconsistent with the null hypothesis, but what is "large" and what is "small"? This is a difficult question to answer. Our interpretation of experimental results depends on considerations such as sample size and practical importance.

Statistical significance is associated with a small p-value. *Practical significance* is associated with a deviation from the null hypothesis that we consider to be important. In Example 9-3, for instance, we obtained a very small p-value. The sample average gold thickness was 1.0345 units, compared to the target value of 1.000 unit, and this difference was highly statistically significant. It could be, however, that this difference is not large enough to justify expensive modifications of the electroplating process. Also, such a difference may be necessary to maintain minimum acceptable gold thicknesses. These are practical considerations, distinct from the issue of statistical significance.

In Example 9-2, we had a large p-value because the sample size was so

small. However, the experimental results suggested that there was a real difference in median distance pedaled under the two experimental conditions (Walkman and no Walkman).

We will consider the issue of statistical significance versus practical importance again in later chapters.

Should we use the significance level or p-value approach to hypothesis testing? This is a matter of personal preference. Many investigators prefer the significance level approach; they interpret their experimental results relative to a specified significance level and decision rule. Other workers prefer the p-value approach; they report a p-value when they write up their results. We can use a combination of the two approaches, as illustrated in Examples 9-1 and 9-3. We can specify a decision rule as outlined in the significance level approach. Then, after collecting our observations, we can calculate a p-value as well.

How do we evaluate the power of a test? The *power* of a test is the likelihood that we will reject the null hypothesis, when a specific alternative is really true. Power depends on how far the particular alternative is from the null hypothesis. The farther the true state of nature is from that specified in the null hypothesis, the greater the likelihood of deciding in favor of the alternative. Power also depends on the sample size.

Let's illustrate how the power of the test can be determined, for a taste test such as the one in Example 9-1. Investigators want to conduct a taste test comparison of Pepsi and Coke. They assume that tasters choose independently of one another and they wish to compare the hypotheses

$$H_0: \ p = \frac{1}{2} \qquad \text{and} \qquad H_a: \ p \neq \frac{1}{2}$$

where p is the probability that a single taster selects Coke. These investigators are making the same simplifying assumptions that the students made in Example 9-1.

The investigators plan to select a random sample of tasters from some larger population. During the process of planning their experiment, they come to us and ask how many tasters they should include. "What sample size do we need?" is a common and extremely sensible question asked by experimenters. Our answer is:

The experimental sample size needed depends on two considerations:
The particular alternative to the null hypothesis that is of interest.
The power of the test desired under this alternative.

We show the investigators what we mean with the following calculations. We will compare the power of the test under three specific alternatives, for sample sizes of 10 and 20.

Consider first a sample size of 10. The test statistic is Y, the number of

Coke selections. Under the no-preference null hypothesis, Y has the Binomial(10, $\frac{1}{2}$) distribution. Suppose we use the following decision rule:

Decision rule for sample size 10

If $2 \leq Y \leq 8$, say the results are consistent with the null hypothesis.
If $Y \leq 1$ or $Y \geq 9$, say the results are inconsistent with the null hypothesis.

Using Table A, we find the significance level α associated with this decision rule to be

$$\text{Significance level} = P(Y \leq 1 \text{ or } Y \geq 9 \text{ when } H_0 \text{ is true}) = .022$$

Now let's calculate the power of the test for three specific alternatives.

Suppose the probability p that a taster selects Coke is really .6. The power is the probability that Y is in the rejection region under this alternative:

$$\text{Power} = P(Y \leq 1 \text{ or } Y \geq 9 \text{ when } p = .6) = .048$$

If the probability p of a taster choosing Coke is really .6, the chance of seeing Y in the rejection region is .048.

If the probability p of a taster choosing Coke is really .8, then the power of the test is

$$\text{Power} = P(Y \leq 1 \text{ or } Y \geq 9 \text{ when } p = .8) = .376$$

With $p = .8$, there is about a 38% chance of saying the observations support the alternative. A similar calculation shows that when $p = .95$ and the sample size is 10, the power of the test is .914.

Let's consider next a sample of 20 tasters. Under the no-preference null hypothesis, the test statistic Y has the Binomial(20, $\frac{1}{2}$) distribution. We would like a decision rule with significance level close to the value ($\alpha = .022$) we used for sample size 10. We have two choices of acceptance and rejection regions with significance levels close to .022, shown in Table 9-6.

We say the first choice in Table 9-6 is more conservative than the second, because we are less likely to decide in favor of the alternative with the first. Let's use this more conservative choice of acceptance and rejection regions. The decision rule then is:

Decision rule for sample size 20

If $5 \leq Y \leq 15$, say the results are consistent with the null hypothesis.
If $Y \leq 4$ or $Y \geq 16$, say the results are inconsistent with the null hypothesis.

The significance level for this decision rule is .012. The probability we say the results are inconsistent with the null hypothesis, when the null hypothesis is really true, is .012.

TABLE 9-6 Two choices of acceptance and rejection regions for sample size 20 in the taste-testing example

Choice	Acceptance region	Rejection region	Significance level, α
1	{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}	{0, 1, 2, 3, 4, 16, 17, 18, 19, 20}	.012
2	{6, 7, 8, 9, 10, 11, 12, 13, 14}	{0, 1, 2, 3, 4, 5, 15, 16, 17, 18, 19, 20}	.041

TABLE 9-7 Power of the test for sample sizes 10 and 20, under three specific alternatives. When the null hypothesis is true, $p = .5$. Here, p is the probability that a taster will select Coke.

Sample size	Rejection region	Significance level	Alternative value of p	Power of the test
10	{0, 1, 9, 10}	.022	.6	.048
			.8	.376
			.95	.914
20	{0, 1, 2, 3, 4, 16, 17, 18, 19, 20}	.012	.6	.051
			.8	.630
			.95	.997

Suppose now the probability p that a taster selects Coke is .6. The power is the probability that Y is in the rejection region:

$$\text{Power} = P(Y \leq 4 \text{ or } Y \geq 16 \text{ when } p = .6) = .051$$

There is about a 5% chance of saying the experimental results are inconsistent with the null hypothesis, when $p = .6$. Similar calculations show that the power of the test is .630 when $p = .8$; the power is .997 when $p = .95$.

The results of our power calculations are summarized in Table 9-7, illustrating the point we want to make for the investigators. The farther the alternative is from the state of nature specified under the null hypothesis, the greater our likelihood of deciding the experimental results are inconsistent with the null hypothesis. Also, for a specific alternative, larger sample sizes increase the chance of seeing results that are inconsistent with the null hypothesis. This is true here, even though for sample size 20 we used a test with smaller significance level than we used for sample size 10. Calculations such as these help us decide on a sample size that balances two opposing considerations: a smaller, less expensive experiment versus a larger, more powerful experiment.

Table 9-7 illustrates another point we addressed earlier. With small sample sizes, we may have little chance of seeing results that are inconsistent with the null hypothesis, even for alternatives of practical interest. This is a serious problem with small experiments.

We have a related concern with larger sample sizes. The likelihood increases that results will appear inconsistent with the null hypothesis, even for alternatives that are only slightly different from the null hypothesis. Then we must make a judgment about whether statistically significant results are really of any practical importance.

9-5 Some Comments on Experimental Design

In formal statistical inference, we make assumptions about the sample observations. We then make probability statements about the population, based on what we observe in the sample.

We should design an investigation so that our assumptions are reasonable. Design of experiments is extremely important, and too often ignored. First, write down the goals of the investigation, stating them in terms of parameters to be estimated and hypotheses to be tested. Then design the investigation to meet these goals.

> **Experimental design** is the area of statistics concerned with designing an investigation to best meet the study goals, as well as the assumptions for statistical inference.

A good experimental design minimizes the effects of factors not of interest. Then the effects of factors that are of interest can be detected more readily. We call the factors not of interest *extraneous factors*.

> An **extraneous factor** is a variable that is not of interest in the experiment, but might affect the outcome.

We design experiments to try to reduce the effects of extraneous factors, so that the effects of factors that are of interest can be detected more readily.

The taste test in Example 9-1 was a double-blind experiment. Neither the experimenter nor the tasters knew the identity of the beverages being tasted. The preconceptions of the experimenter and participants could not influence, whether blatantly or subtly, the results of the study. These preconceptions are extraneous factors.

The student in Example 9-2 tried to reduce the effects of extraneous factors in her bicycle experiment. She wore similar clothing for each trial, went through the same warm-up prior to bicycling, and kept the room temperature constant. The trials were run within a single week, all on weekdays when the student's study and sleeping habits were similar. She covered the odometer so she would not know how far she had pedaled during any trial. She also used a random process to determine the ordering of the experimental conditions. All of these precautions were to reduce the effects of extraneous factors. Because her study was not blinded, she could not eliminate her own preconceptions as possible extraneous factors. How could she have avoided this problem when she designed her experiment?

The manager in Example 9-3 selected a random sample of components from the day's production lot. She knew that random selection of the sample would eliminate her own biases from the selection process. She also hoped that random selection would balance other extraneous factors that might influence the results of the study.

Many investigations provide no information of value because of flaws in the experimental design. A common flaw is that the sample observations are not representative of the population of interest. The results of a taste test, for instance, cannot be generalized to a larger population unless the sample of tasters is representative of that population. One way to get a representative sample is to select the sample at random from the population.

Sometimes a random sample is not possible. Physicians at a cancer research center cannot obtain a random sample of cancer patients for their stud-

ies of new treatments. Instead, they rely on volunteers from among the patients referred to the center. These volunteers may not be representative of the population of cancer patients in the country; they may differ with respect to place of residence, racial background, socioeconomic status, and stage of disease. Researchers try to conduct the best experiments possible with the available volunteers. However, their experimental results may not apply to the larger patient population; a treatment that looked promising at the research center may not be successful in the general patient population.

We will see many examples of experimental designs in the chapters that follow. None of these investigations is perfect; few real experiments are. You are encouraged to think of limitations and extraneous factors that could affect the outcome of each investigation.

We begin in Chapter 10 with inferences about a measure of central tendency. Inferences about two measures of central tendency are considered in Chapter 11. In Chapter 12, we discuss comparisons of several means, via single-factor experiments and randomized block experiments. Chapter 13 covers two-factor experimental designs (specifically: balanced, completely randomized, factorial experiments). In Chapter 14 we address the problem of making inferences about one or more population variances. We consider ways to study the relationship between two quantitative variables in Chapter 15, and introduce the linear correlation coefficient as a measure of linear association between two variables. In addition, we discuss the method of least squares as a way of modeling a quantitative variable as a function of other variables. Chapter 16 introduces some simple ways to make inferences about a single qualitative (categorical) variable and about the relationship between two qualitative variables.

The hypothesis testing procedures in Chapters 10–16 are outlined in steps, as the general approach to hypothesis testing is outlined in Section 9-3. The purpose of this approach is to emphasize that the *strategy* of hypothesis testing is *the same,* regardless of the particular technique being used. The procedures are presented this way to help you learn the mechanics of the statistical techniques, not to encourage a mechanical approach to statistical analysis. On the contrary, I *strongly discourage a cookbook approach to statistical analysis.* In any experimental situation, we should consider the experimental design of primary importance. Often, a well-designed experiment needs little formal statistical analysis. On the other hand, a poorly designed experiment often cannot be salvaged, no matter how sophisticated the analysis technique. We should always consider the experimental design when deciding whether a particular procedure is appropriate. In addition, a complete analysis of any experiment requires data analysis, critical thinking, and common sense. The examples in this chapter have illustrated this approach.

After each of Chapters 10–16, there are exercises illustrating use of techniques covered in the chapter. At the end of Part III, there are additional exercises. For all exercises, consider the experimental design, use tools of data analysis, apply techniques of formal statistical inference when appropriate, and use critical thinking and common sense to make a thorough analysis of the problem.

Summary of Chapter 9

We should take care when interpreting the significance level and p-value in hypothesis testing. Some users of statistics *incorrectly* refer to the significance level (or p-value) as the probability that the null hypothesis is true; this is an *incorrect* interpretation! The significance level, as well as the p-value, is calculated under the null hypothesis probability model: It represents the likelihood of certain events occurring *if the null hypothesis were true.* We use the strategy of hypothesis testing to decide whether the experimental results are consistent with the null hypothesis probability model. That is, we assume the null hypothesis is true and then decide whether the experimental results are consistent with this assumption. (We *do not* calculate the probability that the null hypothesis is true!)

We define the power of a test and discuss it in relation to sample size. Power (the likelihood of saying the experimental results are inconsistent with the null hypothesis when a specific alternative is true) depends on the particular alternative of interest and on sample size. When deciding on a sample size for an experiment, we must consider alternatives that are of interest and the power we want under these alternatives.

When we design an experiment, we try to meet the assumptions that will allow us to compare the hypotheses of interest to us. We try to minimize the effects of extraneous factors—variables that could influence our results, but are not the object of our investigation. With careful design and conduct of an experiment, we hope to better assess the effects of factors that we are studying.

A confidence interval is an interval estimate of a population parameter. We will see several uses of confidence intervals in the chapters that follow.

Exercises for Chapter 9

EXERCISE 9-1 When we calculate the p-value based on experimental results, we are finding (select the correct answer):

a. The probability that the null hypothesis is true.

b. The probability that the alternative hypothesis is true.

c. The chance of seeing results at least as extreme in the direction of the alternative hypothesis as the results actually observed, if the null hypothesis were really true.

d. The likelihood of seeing results at least as extreme in the direction of the alternative hypothesis as the results actually observed, if the alternative hypothesis were really true.

EXERCISE 9-2 The significance level associated with a test of hypotheses is (select the correct answer):

a. The probability that the null hypothesis is true.

b. The probability that the alternative hypothesis is true.

c. The probability of seeing experimental results in the acceptance region (consistent with the null hypothesis) if the alternative hypothesis were really true.

d. The probability of seeing experimental results in the rejection region (inconsistent with the null hypothesis) if the alternative hypothesis were really true.

e. The probability of seeing experimental results in the acceptance region (consistent with the null hypothesis) if the null hypothesis were really true.

f. The probability of seeing experimental results in the rejection region (inconsistent with the null hypothesis) if the null hypothesis were really true.

EXERCISE 9-3

The power of the test is (select the correct answer):

a. The probability that the alternative hypothesis is true.

b. The probability that the null hypothesis is true.

c. The probability of seeing experimental results in the acceptance region (consistent with the null hypothesis) if a specific alternative were really true.

d. The probability of seeing experimental results in the rejection region (inconsistent with the null hypothesis) if a specific alternative were really true.

e. The probability of seeing experimental results in the acceptance region (consistent with the null hypothesis) if the null hypothesis were really true.

f. The probability of seeing experimental results in the rejection region (inconsistent with the null hypothesis) if the null hypothesis were really true.

EXERCISE 9-4

We make a Type I error if (select the correct answer):

a. We say the experimental results are consistent with the null hypothesis, when the null hypothesis is really true.

b. We say the experimental results are consistent with the null hypothesis, when the alternative hypothesis is really true.

c. We say the experimental results are inconsistent with the null hypothesis, when the null hypothesis is really true.

d. We say the experimental results are inconsistent with the null hypothesis, when the alternative hypothesis is really true.

EXERCISE 9-5

We make a Type II error if (select the correct answer):

a. We say the experimental results are consistent with the null hypothesis, when the null hypothesis is really true.

b. We say the experimental results are consistent with the null hypothesis, when the alternative hypothesis is really true.

c. We say the experimental results are inconsistent with the null hypothesis, when the null hypothesis is really true.

d. We say the experimental results are inconsistent with the null hypothesis, when the alternative hypothesis is really true.

EXERCISE 9-6

Consider the taste test in Example 9-1. Recall that the decision rule selected by the students was:

If $3 \leq Y \leq 9$, say the results are consistent with the no-preference null hypothesis.

If $Y \leq 2$ or $Y \geq 10$, say the results are inconsistent with the no-preference null hypothesis, suggesting there is a preference.

a. With 12 tasters, the significance level associated with this decision rule is .038, or about .04. What is the correct interpretation of this significance level?

b. Find the power of the test associated with the given decision rule if the probability p that a student selects Coke is: .1, .2, .4, .6, .8, .9.

c. What is the correct interpretation of each power value you found in part (b)? Compare the values of power you calculated under the six separate alternatives in part (b) and discuss your findings.

EXERCISE 9-7

Consider the taste test in Example 9-1, with 12 tasters.

a. Find the significance level associated with this decision rule:

If $2 \leq Y \leq 10$, say the results are consistent with the no-preference null hypothesis.

If $Y \leq 1$ or $Y \geq 11$, say the results are inconsistent with the no-preference null hypothesis, suggesting there is a preference.

What is the correct interpretation of this significance level?

b. Find the power of the test associated with the decision rule in part (a) if the probability p that a student selects Coke is: .1, .2, .4, .6, .8, .9.

c. What is the correct interpretation of each power value you found in part (b)? Compare the values of power that you calculated under the six separate alternatives in part (b) and discuss your findings.

d. Compare your calculations in this exercise with your calculations for the decision rule in Exercise 9-6. Discuss how power depends on the decision rule you select.

EXERCISE 9-8

Consider the taste test in Example 9-1, with 12 tasters.

a. Find the significance level associated with this decision rule:

If $1 \leq Y \leq 11$, say the results are consistent with the no-preference null hypothesis.

If $Y = 0$ or $Y = 12$, say the results are inconsistent with the no-preference null hypothesis, suggesting there is a preference.

What is the correct interpretation of this significance level?

b. Find the power of the test associated with the decision rule in part (a) if the probability p that a student selects Coke is: .1, .2, .4, .6, .8, .9.

c. What is the correct interpretation of each power value you found in part (b)? Compare the values of power that you calculated under the six separate alternatives in part (b) and discuss your findings.

d. Compare your calculations in this exercise with your calculations for the decision rules in Exercises 9-6 and 9-7. Discuss how power depends on the decision rule you select.

EXERCISE 9-9

Consider a taste test such as the one in Example 9-1, only now there are 6 tasters rather than 12.

a. Find the significance level associated with this decision rule:

If $2 \leq Y \leq 4$, say the results are consistent with the no-preference null hypothesis.

If $Y \leq 1$ or $Y \geq 5$, say the results are inconsistent with the no-preference null hypothesis, suggesting there is a preference.

What is the correct interpretation of this significance level?

b. Find the power of the test associated with the decision rule in part (a) if the probability p that a student selects Coke is: .1, .2, .4, .6, .8, .9.

c. What is the correct interpretation of each power value you found in part (b)? Compare the values of power that you calculated under the six separate alternatives in part (b) and discuss your findings.

d. Compare your results in this exercise with what you found in Exercises 9-6, 9-7, and 9-8. Discuss how power depends on sample size.

EXERCISE 9-10

Consider a taste test such as the one in Example 9-1, only now there are 6 tasters rather than 12.

a. Find the significance level associated with this decision rule:

If $1 \leq Y \leq 5$, say the results are consistent with the no-preference null hypothesis.

If $Y = 0$ or $Y = 6$, say the results are inconsistent with the no-preference null hypothesis, suggesting there is a preference.

What is the correct interpretation of this significance level?

b. Find the power of the test associated with the decision rule in part (a) if the probability p that a student selects Coke is: .1, .2, .4, .6, .8, .9.

c. What is the correct interpretation of each power value you found in part (b)? Compare the values of power that you calculated under the six separate alternatives in part (b) and discuss your findings.

d. Compare your calculations in this exercise with your calculations for the decision rule in Exercise 9-9. Discuss how power depends on the decision rule you select.

e. Comparing your results in Exercises 9-9 and 9-10 with your results in Exercises 9-6, 9-7, and 9-8, discuss how power depends on sample size.

EXERCISE 9-11 Consider a stationary bicycle experiment such as the one in Example 9-2.

a. Suppose the student makes four runs with the Walkman and four runs without the Walkman. For each run with the Walkman, she pedals farther than for any run without the Walkman. Using a probability model similar to the one in Example 9-2, calculate the p-value associated with these results. What is the correct interpretation of this p-value?

b. Repeat part (a) if she makes eight runs with the Walkman and eight runs without the Walkman, and she pedals farther in all runs with the Walkman than in any run without the Walkman.

c. Compare the results in parts (a) and (b) with those we found in Example 9-2. Interpret your results. Discuss the sample size considerations.