CHAPTER

18

The Binomial Test

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Binomial distribution (Chapter 6)
- *z*-score hypothesis tests (Chapter 8)
- Chi-square test for goodness of fit (Chapter 17)

Preview

In 1960, Gibson and Walk designed a classic piece of apparatus to test depth perception. Their device, called a *visual cliff*, consisted of a wide board with a deep drop (the cliff) to one side and a shallow drop on the other side. An infant was placed on the board and then observed to see whether he or she crawled off the shallow side or crawled off the cliff. Infants who moved to the deep side actually crawled onto a sheet of heavy glass, which prevented them from falling. Thus, the deep side only appeared to be a cliff—hence the name *visual cliff*.

Gibson and Walk reasoned that if infants are born with the ability to perceive depth, they would recognize the deep side and not crawl off the cliff. On the other hand, if depth perception is a skill that develops over time through learning and experience, then infants should not be able to perceive any difference between the shallow and the deep sides.

Out of 27 infants who moved off the board, only 3 ventured onto the deep side at any time during the experiment. The other 24 infants stayed exclusively on the shallow side. Gibson and Walk interpreted these data as convincing evidence that depth perception is innate. The infants showed a systematic preference for the shallow side.

The Problem: You should notice immediately that the data from this experiment are different from the data that we usually encounter. There are no scores. Gibson and Walk simply counted the number of infants who went off the deep side and the number who went to the shallow side. Still, we would like to use these data to make statistical decisions. Do these sample data provide sufficient evidence to make a confident conclusion about depth perception in the population? Suppose that 8 of the 27 infants had crawled to the deep side. Would you still be convinced that there is a significant preference for the shallow side? What about 12 out of 27?

The Solution: We are asking a question about statistical significance and need a hypothesis test to obtain an answer. The null hypothesis for the Gibson and Walk study would state that infants have no depth perception and cannot perceive a difference between the shallow and deep sides. In this case, their movement should be random with half going to either side. Notice that the data and the hypothesis both concern frequencies or proportions. This situation is perfect for the chi-square test introduced in Chapter 17, and a chi-square test can be used to evaluate the data. However, when individuals are classified into exactly two categories (for example, shallow and deep), a special statistical procedure exists. In this chapter, we introduce the *binomial test*, which is used to evaluate and interpret frequency data involving exactly two categories of classification.

18.1 OVERVIEW

Data with exactly two categories are also known as dichotomous data.

In Chapter 6, we introduced the concept of *binomial data*. You should recall that binomial data exist whenever a measurement procedure classifies individuals into exactly two distinct categories. For example, the outcomes from tossing a coin can be classified as heads and tails; people can be classified as male or female; plastic products can be classified as recyclable or non-recyclable. In general, binomial data exist when

1. The measurement scale consists of exactly two categories.
2. Each individual observation in a sample is classified in only one of the two categories.
3. The sample data consist of the frequency of, or number of individuals in, each category.

The traditional notation system for binomial data identifies the two categories as A and B and identifies the probability (or proportion) associated with each category as p and q, respectively. For example, a coin toss results in either heads (A) or tails (B), with probabilities $p = \frac{1}{2}$ and $q = \frac{1}{2}$.

In this chapter, we examine the statistical process of using binomial data for testing hypotheses about the values of p and q for the population. This type of hypothesis test is called a *binomial test*.

DEFINITION

A **binomial test** uses sample data to evaluate hypotheses about the values of p and q for a population consisting of binomial data.

Consider the following two situations:

1. In a sample of $n = 34$ color-blind students, 30 are male, and only 4 are female. Does this sample indicate that color blindness is significantly more common for males in the general population?
2. In 2005, only 10% of American families had incomes below the poverty level. This year, in a sample of 100 families, 19 were below the poverty level. Does this sample indicate that there has been a significant change in the population proportions?

Notice that both of these examples have binomial data (exactly two categories). Although the data are relatively simple, we are asking a statistical question about significance that is appropriate for a hypothesis test: Do the sample data provide sufficient evidence to make a conclusion about the population?

HYPOTHESES FOR THE BINOMIAL TEST

In the binomial test, the null hypothesis specifies exact values for the population proportions p and q. Theoretically, you could choose any proportions for H_0, but usually there is a clear reason for the values that are selected. The null hypothesis typically falls into one of the following two categories:

1. **Just Chance.** Often the null hypothesis states that the two outcomes, A and B, occur in the population with the proportions that would be predicted simply by chance. If you were tossing a coin, for example, the null hypothesis might specify $p(\text{heads}) = \frac{1}{2}$ and $p(\text{tails}) = \frac{1}{2}$. Notice that this hypothesis states the usual, chance proportions for a balanced coin. Also notice that it is not necessary to specify both proportions. Once the value of p is identified, the value of q is determined by $1 - p$. For the coin toss example, the null hypothesis would simply state

   H_0: $p = p(\text{heads}) = \dfrac{1}{2}$ (The coin is balanced.)

   Similarly, if you were selecting cards from a deck and trying to predict the suit on each draw, the probability of predicting correctly would be $p = \frac{1}{4}$ for any given trial. (With four suits, you have a 1-out-of-4 chance of guessing correctly.) In this case, the null hypothesis would state

   H_0: $p = p(\text{guessing correctly}) = \dfrac{1}{4}$ (The outcome is simply the result of chance.)

   In each case, the null hypothesis states that there is nothing unusual about the proportions in the population; that is, the outcomes are occurring by chance.
2. **No Change or No Difference.** Often you may know the proportions for one population and want to determine whether the same proportions apply to a different population. In this case, the null hypothesis would simply specify that there is no difference between the two populations. Suppose that national

statistics indicate that 1 out of 12 drivers will be involved in a traffic accident during the next year. Does this same proportion apply to 16-year-olds who are driving for the first time? According to the null hypothesis,

H_0: For 16-year-olds, $p = p(\text{accident}) = \frac{1}{12}$ (Not different from the general population)

Similarly, suppose that last year, 30% of the freshman class failed the college writing test. This year, the college is requiring all freshmen to take a writing course. Will the course have any effect on the number who fail the test? According to the null hypothesis,

H_0: For this year, $p = p(\text{fail}) = 30\%$ (Not different from last year's class)

THE DATA FOR THE BINOMIAL TEST

For the binomial test, a sample of n individuals is obtained and you simply count how many are classified in category A and how many are classified in category B. We focus attention on category A and use the symbol X to stand for the number of individuals classified in category A. Recall from Chapter 6 that X can have any value from 0 to n and that each value of X has a specific probability. The distribution of probabilities for each value of X is called the *binomial distribution*. Figure 18.1 shows an example of a binomial distribution for which X is the number of heads obtained in four tosses of a balanced coin.

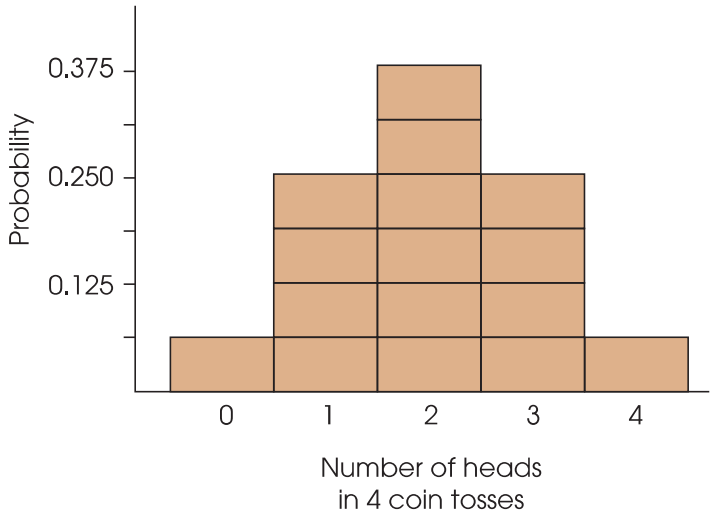
THE TEST STATISTIC FOR THE BINOMIAL TEST

As we noted in Chapter 6, when the values pn and qn are both equal to or greater than 10, the binomial distribution approximates a normal distribution. This fact is important because it allows us to compute z-scores and use the unit normal table to answer probability questions about binomial events. In particular, when pn and qn are both at least 10, the binomial distribution has the following properties:

1. The shape of the distribution is approximately normal.
2. The mean of the distribution is $\mu = pn$.

FIGURE 18.1

A binomial distribution for the number of heads obtained in four tosses of a balanced coin.



3. The standard deviation of the distribution is

$$\sigma = \sqrt{npq}$$

With these parameters in mind, it is possible to compute a z-score corresponding to each value of X in the binomial distribution.

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}} \qquad \text{(See Equation 6.3.)} \tag{18.1}$$

This is the basic z-score formula that is used for the binomial test. However, the formula can be modified slightly to make it more compatible with the logic of the binomial hypothesis test. The modification consists of dividing both the numerator and the denominator of the z-score by n. (You should realize that dividing both the numerator and the denominator by the same value does not change the value of the z-score.) The resulting equation is

$$z = \frac{X/n - p}{\sqrt{pq/n}} \tag{18.2}$$

For the binomial test, the values in this formula are defined as follows:

1. X/n is the proportion of individuals in the sample who are classified in category A.
2. p is the hypothesized value (from H_0) for the proportion of individuals in the population who are classified in category A.
3. $\sqrt{pq/n}$ is the standard error for the sampling distribution of X/n and provides a measure of the standard distance between the sample statistic (X/n) and the population parameter (p).

Thus, the structure of the binomial z-score (Equation 18.2) can be expressed as

$$z = \frac{X/n - p}{\sqrt{pq/n}} = \frac{\begin{matrix}\text{sample}\\ \text{proportion}\\ \text{(data)}\end{matrix} - \begin{matrix}\text{hypothesized}\\ \text{population}\\ \text{proportion}\end{matrix}}{\text{standard error}}$$

The logic underlying the binomial test is exactly the same as we encountered with the original z-score hypothesis test in Chapter 8. The hypothesis test involves comparing the sample data with the hypothesis. If the data are consistent with the hypothesis, then we conclude that the hypothesis is reasonable. But if there is a big discrepancy between the data and the hypothesis, then we reject the hypothesis. The value of the standard error provides a benchmark for determining whether the discrepancy between the data and the hypothesis is more than would be expected by chance. The alpha level for the test provides a criterion for deciding whether the discrepancy is significant. The hypothesis-testing procedure is demonstrated in the following section.

LEARNING CHECK

1. In the Preview, we described a research study using a visual cliff. State the null hypothesis for this study in words and as a probability value (p) that an infant will crawl off the deep side.
2. If the visual cliff study had used a sample of $n = 15$ infants, would it be appropriate to use the normal approximation to the binomial distribution? Explain why or why not.

3. If the results from the visual cliff study showed that 9 out of 36 infants crawled off the deep side, what z-score value would be obtained using Equation 18.1?

ANSWERS

1. The null hypothesis states that the probability of choosing between the deep side and the shallow side is just chance: $p(\text{deep side}) = \frac{1}{2}$.

2. The normal approximation to the binomial distribution requires that both pn and qn are at least 10. With $n = 15$, $pn = qn = 7.5$. The normal approximation should not be used.

3. With $n = 36$ and $p = \frac{1}{2}$, the binomial distribution has $\mu = \frac{1}{2}(36) = 18$, and $\sigma = \sqrt{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)(36)} = \sqrt{9} = 3$. $X = 9$ corresponds $z = -9/3 = -3.00$

18.2 THE BINOMIAL TEST

The binomial test follows the same four-step procedure presented earlier with other examples for hypothesis testing. The four steps are summarized as follows.

STEP 1 *State the hypotheses*. In the binomial test, the null hypothesis specifies values for the population proportions p and q. Typically, H_0 specifies a value only for p, the proportion associated with category A. The value of q is directly determined from p by the relationship $q = 1 - p$. Finally, you should realize that the hypothesis, as always, addresses the probabilities or proportions for the *population*. Although we use a sample to test the hypothesis, the hypothesis itself always concerns a population.

STEP 2 *Locate the critical region*. When both values for pn and qn are greater than or equal to 10, then the z-scores defined by Equation 18.1 or 18.2 form an approximately normal distribution. Thus, the unit normal table can be used to find the boundaries for the critical region. With $\alpha = .05$, for example, you may recall that the critical region is defined as z-score values greater than $+1.96$ or less than -1.96.

STEP 3 *Compute the test statistic (z-score)*. At this time, you obtain a sample of n individuals (or events) and count the number of times category A occurs in the sample. The number of occurrences of A in the sample is the X value for Equation 18.1 or 18.2. Because the two z-score equations are equivalent, you may use either one for the hypothesis test. Usually Equation 18.1 is easier to use because it involves larger numbers (fewer decimals) and it is less likely to be affected by rounding error.

STEP 4 *Make a decision*. If the z-score for the sample data is in the critical region, then you reject H_0 and conclude that the discrepancy between the sample proportions and the hypothesized population proportions is significantly greater than chance. That is, the data are not consistent with the null hypothesis, so H_0 must be wrong. On the other hand, if the z-score is not in the critical region, then you fail to reject H_0.

The following example demonstrates a complete binomial test.

EXAMPLE 18.1

In the Preview section, we described the *visual cliff* experiment designed to examine depth perception in infants. To summarize briefly, an infant is placed on a wide board that appears to have a deep drop on one side and a relatively shallow drop on the

other. An infant who is able to perceive depth should avoid the deep side and move toward the shallow side. Without depth perception, the infant should show no preference between the two sides. Of the 27 infants in the experiment, 24 stayed exclusively on the shallow side and only 3 moved onto the deep side. The purpose of the hypothesis test is to determine whether these data demonstrate that infants have a significant preference for the shallow side.

This is a binomial hypothesis-testing situation. The two categories are

A = move onto the deep side

B = move onto the shallow side

STEP 1 The null hypothesis states that, for the general population of infants, there is no preference between the deep and the shallow sides; the direction of movement is determined by chance. In symbols,

$$H_0: \quad p = p(\text{deep side}) = \frac{1}{2} \quad \left(\text{and } q = \frac{1}{2}\right)$$

$$H_1: \quad p \neq \frac{1}{2} \quad (\text{There is a preference.})$$

We use $\alpha = .05$.

STEP 2 With a sample of $n = 27$, $pn = 13.5$ and $qn = 13.5$. Both values are greater than 10, so the distribution of z-scores is approximately normal. With $\alpha = .05$, the critical region is determined by boundaries of $z = \pm 1.96$.

STEP 3 For this experiment, the data consist of $X = 3$ out of $n = 27$. Using Equation 18.1, these data produce a z-score value of

$$z = \frac{X - pn}{\sqrt{npq}} = \frac{3 - 13.5}{\sqrt{27\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{-10.5}{2.60} = -4.04$$

To use Equation 18.2, you first compute the sample proportion, $X/n = 3/27 = 0.111$. The z-score is then

$$z = \frac{X/n - p}{\sqrt{pq/n}} = \frac{0.111 - 0.5}{\sqrt{\frac{1}{2}\left(\frac{1}{2}\right)/27}} = \frac{-0.389}{0.096} = -4.05$$

Within rounding error, the two equations produce the same result.

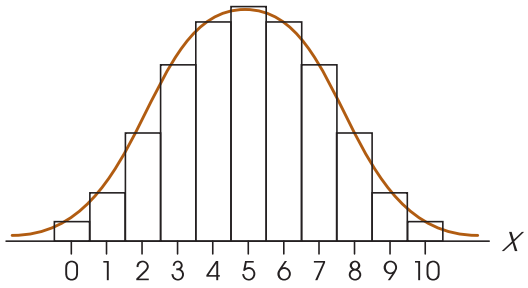
STEP 4 Because the data are in the critical region, our decision is to reject H_0. These data do provide sufficient evidence to conclude that there is a significant preference for the shallow side. Gibson and Walk (1960) interpreted these data as convincing evidence that depth perception is innate.

REAL LIMITS AND THE BINOMIAL TEST

In Chapter 6, we noted that a binomial distribution forms a discrete histogram (see Figure 18.1), whereas the normal distribution is a continuous curve. The difference between the two distributions was illustrated in Figure 6.18,

FIGURE 18.2

The relationship between the binomial distribution and the normal distribution. The binomial distribution is always a discrete histogram, and the normal distribution is a continuous, smooth curve. Each X value is represented by a bar in the histogram or a section of the normal distribution.



which is repeated here as Figure 18.2. In the figure, note that each score in the binomial distribution is represented by a bar in the histogram. For example, a score of $X = 6$ actually corresponds to a bar that reaches from a lower real limit of $X = 5.5$ to an upper real limit of 6.5.

When conducting a hypothesis test with the binomial distribution, the basic question is whether a specific score is located in the critical region. However, because each score actually corresponds to an interval, it is possible that part of the score is in the critical region and part is not. Fortunately, this is usually not an issue. When pn and qn are both equal to or greater than 10 (the criteria for using the normal approximation), each interval in the binomial distribution is extremely small and it is very unlikely that the interval overlaps the critical boundary. For example, the experiment in Example 18.1 produced a score of $X = 3$, and we computed a z-score of $z = -4.04$. Because this value is in the critical region, beyond $z = -1.96$, we rejected H_0. If we had used the real limit boundaries of $X = 2.5$ and $X = 3.5$, instead of $X = 3$, we would have obtained z-scores of

$$z = \frac{2.5 - 13.5}{2.60} \qquad \text{and} \qquad z = \frac{3.5 - 13.5}{2.60}$$

$$= -4.23 \qquad = -3.85$$

Thus, a score of $X = 3$ actually corresponds to an interval of z-scores ranging from $z = -3.85$ to $z = -4.23$. However, this entire interval is in the critical region beyond $z = -1.96$, so the decision is still to reject H_0.

In most situations, if the whole number X value (in this case, $X = 3$) is in the critical region, then the entire interval is in the critical region and the correct decision is to reject H_0. The only exception to this general rule occurs when an X value produces a z-score that is barely past the boundary into the critical region. In this situation, you should compute the z-scores corresponding to both real limits to determine whether any part of the z-score interval is not located in the critical region. Suppose, for example, that the researchers in Example 18.1 found that 8 out of 27 infants in the visual cliff experiment moved onto the deep side. A score of $X = 8$ corresponds to

$$z = \frac{8 - 13.5}{2.60}$$

$$z = \frac{-5.5}{2.60}$$

$$= -2.12$$

Because this value is beyond the − 1.96 boundary, it appears that we should reject H_0. However, this z-score is only slightly beyond the critical boundary, so it would be wise to check both ends of the interval. For $X = 8$, the real-limit boundaries are 7.5 and 8.5, which correspond to z-scores of

$$z = \frac{7.5 - 13.5}{2.60} = -2.31 \qquad \text{and} \qquad z = \frac{8.5 - 13.5}{2.60} = -1.92$$

Thus, a score of $X = 8$ corresponds to an interval extending from $z = -1.92$ to $z = -2.31$. However, the critical boundary is $z = -1.96$, which means that part of the interval (and part of the score) is not in the critical region for $\alpha = .05$. Because $X = 8$ is not completely beyond the critical boundary, the probability of obtaining $X = 8$ is greater than $\alpha = .05$. Therefore, the correct decision is to fail to reject H_0.

In general, it is safe to conduct a binomial test using the whole-number value for X. However, if you obtain a z-score that is only slightly beyond the critical boundary, you also should compute the z-scores for both real limits. If any part of the z-score interval is not in the critical region, the correct decision is to fail to reject H_0.

IN THE LITERATURE

REPORTING THE RESULTS OF A BINOMIAL TEST

Reporting the results of the binomial test typically consists of describing the data and reporting the z-score value and the probability that the results are caused by chance. It is also helpful to note that a binomial test was used because z-scores are used in other hypothesis-testing situations (see, for example, Chapter 8). For Example 18.1, the report might state:

> Three out of 27 infants moved to the deep side of the visual cliff. A binomial test revealed that there is a significant preference for the shallow side of the cliff, $z = -4.04$, $p < .05$.

Once again, p is *less than* .05. We have rejected the null hypothesis because it is very unlikely, probability less than 5%, that these results are simply caused by chance.

ASSUMPTIONS FOR THE BINOMIAL TEST

The binomial test requires two very simple assumptions:

1. The sample must consist of *independent* observations (see Chapter 8, page 254).
2. The values for pn and qn must both be greater than or equal to 10 to justify using the unit normal table for determining the critical region.

LEARNING CHECK

1. For a binomial test, the null hypothesis always states that $p = 1/2$. (True or false?)
2. The makers of brand X beer claim that people like their beer more than the leading brand. The basis for this claim is an experiment in which 64 beer drinkers compared the two brands in a side-by-side taste test. In this sample, 40 preferred brand X, and 24 preferred the leading brand.
   a. If you compute the z-score for $X = 40$, do these data support the claim that there is a significant preference? Test at the .05 level.
   b. If you compute z-scores for the real limits for $X = 40$, do the data support the claim that there is a significant preference? Test at the .05 level.

ANSWERS

1. False.

2. a. H_0: $p = \frac{1}{2} = q$, $X = 38$, $\mu = 32$, $\sigma = 4$, $z = 2.00$, reject H_0. Conclude that there is a significant preference.

b. The real limits of 39.5 and 40.5 correspond to z-scores of 1.88 and 2.13. The entire interval is not in the critical region so fail to reject H_0 and conclude that there is not a significant preference.

18.3 THE RELATIONSHIP BETWEEN CHI-SQUARE AND THE BINOMIAL TEST

You may have noticed that the binomial test evaluates the same basic hypotheses as the chi-square test for goodness of fit; that is, both tests evaluate how well the sample proportions fit a hypothesis about the population proportions. When an experiment produces binomial data, these two tests are equivalent, and either may be used. The relationship between the two tests can be expressed by the equation

$$\chi^2 = z^2$$

where χ^2 is the statistic from the chi-square test for goodness of fit and z is the z-score from the binomial test.

To demonstrate the relationship between the goodness-of-fit test and the binomial test, we reexamine the data from Example 18.1.

STEP 1 *Hypotheses.* In the visual cliff experiment from Example 18.1, the null hypothesis states that there is no preference between the shallow side and the deep side. For the binomial test, the null hypothesis states

$$H_0: \ p = p(\text{deep side}) = q = p(\text{shallow side}) = \frac{1}{2}$$

The chi-square test for goodness of fit would state the same hypothesis, specifying the population proportions as

	Shallow Side	Deep Side
H_0:	$\frac{1}{2}$	$\frac{1}{2}$

STEP 2 *Critical region.* For the binomial test, the critical region is located by using the unit normal table. With $\alpha = .05$, the critical region consists of any z-score value beyond ± 1.96. The chi-square test would have $df = 1$, and with $\alpha = .05$, the critical region consists of chi-square values greater than 3.84. Notice that the basic relationship, $\chi^2 = z^2$, holds:

$$3.84 = (1.96)^2$$

STEP 3 *Test statistic.* For the binomial test (Example 18.1), we obtained a z-score of $z = -4.04$. For the chi-square test, the expected frequencies are

	Shallow Side	Deep Side
f_e	13.5	13.5

With observed frequencies of 24 and 3, respectively, the chi-square statistic is

$$\chi^2 = \frac{(24 - 13.5)^2}{13.5} + \frac{(3 - 13.5)^2}{13.5}$$

$$= \frac{(10.5)^2}{13.5} + \frac{(-10.5)^2}{13.5}$$

$$= 8.167 + 8.167$$

$$= 16.33$$

With a little rounding error, the values obtained for the z-score and chi-square are related by the equation

$$\chi^2 = z^2$$

$$16.33 = (-4.04)^2$$

Caution: The χ^2 value is already squared. Do not square it again.

STEP 4 *Decision.* Because the critical values for both tests are related by the equation $\chi^2 = z^2$ and the test statistics are related in the same way, these two tests *always* result in the same statistical conclusion.

18.4 THE SIGN TEST

Although the binomial test can be used in many different situations, there is one specific application that merits special attention. For a repeated-measures study that compares two conditions, it is often possible to use a binomial test to evaluate the results. You should recall that a repeated-measures study involves measuring each individual in two different treatment conditions or at two different points in time. When the measurements produce numerical scores, the researcher can simply subtract to determine the difference between the two scores and then evaluate the data using a repeated-measures t test (see Chapter 11). Occasionally, however, a researcher may record only the *direction* of the difference between the two observations. For example, a clinician may observe patients before therapy and after therapy and simply note whether each patient got better or worse. Note that there is no measurement of how much change occurred; the clinician is simply recording the direction of change. Also note that the direction of change is a binomial variable; that is, there are only two values. In this situation it is possible to use a binomial test to evaluate the data. Traditionally, the two possible directions of change are coded by signs, with a positive sign indicating an increase and a negative sign indicating a decrease. When the binomial test is applied to signed data, it is called a *sign test*.

An example of signed data is shown in Table 18.1. Notice that the data can be summarized by saying that seven out of eight patients showed a decrease in symptoms after therapy.

The null hypothesis for the sign test states that there is no difference between the two treatments. Therefore, any change in a participant's score is the result of chance. In terms of probabilities, this means that increases and decreases are equally likely, so

$$p = p(\text{increase}) = \frac{1}{2}$$

$$q = p(\text{decrease}) = \frac{1}{2}$$

A complete example of a sign test follows.

TABLE 18.1

Hypothetical data from a research study evaluating the effectiveness of a clinical therapy. For each patient, symptoms are assessed before and after treatment and the data record whether there is an increase or a decrease in symptoms following therapy.

Patient	Direction of Change After Treatment
A	− (decrease)
B	− (decrease)
C	− (decrease)
D	+ (increase)
E	− (decrease)
F	− (decrease)
G	− (decrease)
H	− (decrease)

EXAMPLE 18.2

A researcher testing the effectiveness of acupuncture for treating the symptoms of arthritis obtains a sample of 36 people who have been diagnosed with arthritis. Each person's pain level is measured before treatment starts, and measured again after 4 months of acupuncture treatment. For this sample, 25 people experienced a reduction in pain, and 11 people had more pain after treatment. Do these data indicate a significant treatment effect?

STEP 1 State the hypothesis. The null hypothesis states that acupuncture has no effect. Any change in the level of pain is caused by chance, so increases and decreases are equally likely. Expressed as probabilities, the hypotheses are

$$H_0:\ p = p(\text{increased pain}) = \frac{1}{2} \text{ and } q = p(\text{decreased pain}) = \frac{1}{2}$$

H_1: $p \neq q$ (Changes tend to be consistently in one direction.)

Set $\alpha = .05$.

STEP 2 Locate the critical region. With $n = 36$ people, both pn and qn are greater than 10, so the normal approximation to the binomial distribution is appropriate. With $\alpha = .05$, the critical region consists of z-scores greater than +1.96 at one extreme and z-scores less than −1.96 at the other.

STEP 3 Compute the test statistic. For this sample we have $X = 25$ people with decreased pain. This score corresponds to a z-score of

$$z = \frac{X - pn}{\sqrt{npq}} = \frac{25 - 18}{\sqrt{36\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{7}{3} = 2.33$$

Because the z-score is only slightly beyond the 1.96 critical boundary, we consider the real limits for $X = 25$ to be certain that the entire interval is beyond the boundary. For $X = 25$, the real limits are 24.5 and 25.5, which correspond to z-scores of

$$z = \frac{24.5 - 18}{3} = 2.17 \quad \text{and} \quad z = \frac{25.5 - 18}{3} = 2.50$$

Thus, a score of $X = 25$ corresponds to an interval of z-scores ranging from $z = 2.17$ to $z = 2.50$. Note that this entire interval is beyond the 1.96 critical boundary.

STEP 4 Make a decision. Because the data are in the critical region, we reject H_0 and conclude that acupuncture treatment has a significant effect on arthritis pain, $z = 2.33$, $p < .05$.

ZERO DIFFERENCES IN THE SIGN TEST

You should notice that the null hypothesis in the sign test refers only to those individuals who show some difference between treatment 1 versus treatment 2. The null hypothesis states that if there is any change in an individual's score, then the probability of an increase is equal to the probability of a decrease. Stated in this form, the null hypothesis does not consider individuals who show zero difference between the two treatments. As a result, the usual recommendation is that these individuals be discarded from the data and the value of n be reduced accordingly. However, if the null hypothesis is interpreted more generally, it states that there is no difference between the two treatments. Phrased this way, it should be clear that individuals who show no difference actually are supporting the null hypothesis and should not be discarded. Therefore, an alternative approach to the sign test is to divide individuals who show zero differences equally between the positive and negative categories. (With an odd number of zero differences, discard one, and divide the rest evenly.) This alternative results in a more conservative test; that is, the test is more likely to fail to reject the null hypothesis.

EXAMPLE 18.3

It has been demonstrated that stress or exercise causes an increase in the concentration of certain chemicals in the brain called endorphins. Endorphins are similar to morphine and produce a generally relaxed feeling and a sense of well-being. The endorphins may explain the "high" experienced by long-distance runners. To demonstrate this phenomenon, a researcher tested pain tolerance for 40 athletes before and after they completed a mile run. Immediately after running, the ability to tolerate pain increased for 21 of the athletes, decreased for 12, and showed no change for the remaining 7.

Following the standard recommendation for handling zero differences, you would discard the 7 participants who showed no change and conduct a sign test with the remaining $n = 33$ athletes. With the more conservative approach, only 1 of the 7 who showed no difference would be discarded and the other 6 would be divided equally between the two categories. This would result in a total sample of $n = 39$ athletes with $21 + 3 = 24$ in the increased-tolerance category and $12 + 3 = 15$ in the decreased-tolerance category.

WHEN TO USE THE SIGN TEST

In many cases, data from a repeated-measures experiment can be evaluated using either a sign test or a repeated-measures t test. In general, you should use the t test whenever possible. Because the t test uses the actual difference scores (not just the signs), it makes maximum use of the available information and results in a more powerful test. However, there are some cases in which a t test cannot or should not be used, and in these situations, the sign test can be valuable. Four specific cases in which a t test is inappropriate or inconvenient are described below.

Before	After	Difference
20	23	+3
14	39	+25
27	Failed	+??
.	.	.
.	.	.
.	.	.

1. When you have infinite or undetermined scores, a *t* test is impossible, and the sign test is appropriate. Suppose, for example, that you are evaluating the effects of a sedative drug on problem-solving ability. A sample of rats is obtained, and each animal's performance is measured before and after receiving the drug. Hypothetical data are shown in the margin. Note that the third rat in this sample failed to solve the problem after receiving the drug. Because there is no score for this animal, it is impossible to compute a sample mean, an *SS*, or a *t* statistic. However, you could do a sign test because you know that the animal made more errors (an increase) after receiving the drug.
2. Often it is possible to describe the difference between two treatment conditions without precisely measuring a score in either condition. In a clinical setting, for example, a doctor can say whether a patient is improving, growing worse, or showing no change even though the patient's condition is not precisely measured by a score. In this situation, the data are sufficient for a sign test, but you could not compute a *t* statistic without individual scores.
3. Often a sign test is done as a preliminary check on an experiment before serious statistical analysis begins. For example, a researcher may predict that scores in treatment 2 should be consistently greater than scores in treatment 1. However, examination of the data after 1 week indicates that only 8 of 15 subjects showed the predicted increase. On the basis of these preliminary results, the researcher may choose to reevaluate the experiment before investing additional time.
4. Occasionally, the difference between treatments is not consistent across participants. This can create a very large variance for the difference scores. As we have noted in the past, large variance decreases the likelihood that a *t* test will produce a significant result. However, the sign test only considers the direction of each difference score and is not influenced by the variance of the scores.

LEARNING CHECK

1. A researcher used a chi-square test for goodness of fit to determine whether people had any preferences among three leading brands of potato chips. Could the researcher have used a binomial test instead of the chi-square test? Explain why or why not.
2. A researcher used a chi-square test to evaluate preferences between two logo designs for a minor-league hockey team. With a sample of $n = 100$ people, the researcher obtained a chi-square of 9.00. If a binomial test had been used instead of chi-square, what value would have been obtained for the *z*-score?
3. A developmental psychologist is using a behavior-modification program to help control the disruptive behavior of 40 children in a local school. After 1 month, 26 of the children have improved, 10 are worse, and 4 show no change in behavior. On the basis of these data, can the psychologist conclude that the program is working? Test at the .05 level.

ANSWERS

1. No, the binomial test cannot be used when there are three categories.
2. The *z*-score would be $\sqrt{9} = 3.00$.
3. Discarding the four participants who showed zero difference, $X = 26$ increases out of $n = 36$; $z = 2.67$; reject H_0; the program is working. If the four participants showing no change are divided between the two groups, then $X = 28$ out of $n = 40$; $z = 2.53$ and H_0 is still rejected.

SUMMARY

1. The binomial test is used with dichotomous data—that is, when each individual in the sample is classified in one of two categories. The two categories are identified as A and B, with probabilities of

$$p(A) = p \text{ and } p(B) = q$$

2. The binomial distribution gives the probability for each value of X, where X equals the number of occurrences of category A in a sample of n events. For example, X equals the number of heads in $n = 10$ tosses of a coin.

3. When pn and qn are both at least 10, then the binomial distribution is closely approximated by a normal distribution with

$$\mu = pn \qquad \sigma = \sqrt{npq}$$

   By using this normal approximation, each value of X has a corresponding z-score:

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}} \quad \text{or} \quad z = \frac{X/n - p}{\sqrt{pq/n}}$$

4. The binomial test uses sample data to test hypotheses about the binomial proportions, p and q, for a population. The null hypothesis specifies p and q, and the binomial distribution (or the normal approximation) is used to determine the critical region.

5. Usually the z-score in a binomial test is computed using the whole-number X value from the sample. However, if the z-score is only marginally in the critical region, you should compute the z-scores corresponding to both real limits of the score. If either one of these real-limit z-scores is not in the critical region, then the correct decision is to fail to reject the null hypothesis.

6. One common use of the binomial distribution is for the sign test. This test evaluates the difference between two treatments using the data from a repeated measures design. The difference scores are coded as being either increases (+) or decreases (−). Without a consistent treatment effect, the increases and decreases should be mixed randomly, so the null hypothesis states that

$$p(\text{increase}) = \frac{1}{2} = p(\text{decrease})$$

   With dichotomous data and hypothesized values for p and q, this is a binomial test.

KEY TERMS

binomial data (634)
binomial test (635)
binomial distribution (636)
sign test (643)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 18 on the book companion website.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGEbrain.com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are found in Appendix D. If you are testing a null hypothesis specifying that $p = q = \frac{1}{2}$, then you can use SPSS to perform The Binomial Test presented in this chapter. Following are detailed instructions for the test. For other versions of the null hypothesis, use the equivalent chi-square test for goodness of fit presented in Chapter 17 (p. 594). The chi-square test allows you to specify expected frequencies, which is equivalent to specifying values for p and q.

Data Entry

1. Enter the category labels A and B in the first column of the SPSS data editor.
2. In the second column, enter the frequencies obtained for the two binomial categories. For example, if 21 out of 25 people were classified in category *A* (and only 4 people in category *B*), you would enter the values 21 and 4 in the second column.

Data Analysis

1. Click **Data** on the tool bar at the top of the page and select **weight cases** at the bottom of the list.
2. Click the **Weight cases by** circle, then highlight the label for the column containing the frequencies for the two categories and move it into the **Frequency Variable** box by clicking on the arrow.
3. Click **OK.**
4. Click **Analyze** on the tool bar, select **Nonparametric Tests,** and click on **One Sample.**
5. Select **Automatically compare observed data to hypothesis**.
6. Click **RUN.**

SPSS Output

We used SPSS to analyze the data from Example 18.1 and the output is shown in Figure 18.3. The output reports the null hypothesis for the test and the level of significance, which is rounded to .000 for this example.

FIGURE 18.3
The SPSS output for the binomial test in Example 18.1.

Hypothesis Test Summary

Null Hypothesis	Test	Sig.	Decision
The categories defined by VAR00001 = A and B occur with probabilities 0.5 and 0.5.	One-Sample Binomial Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

FOCUS ON PROBLEM SOLVING

1. For all binomial tests, the values of p and q must add up to 1.00 (or 100%).
2. Remember that both pn and qn must be at least 10 before you can use the normal distribution to determine critical values for a binomial test.

DEMONSTRATION 18.1

THE BINOMIAL TEST

The population of students in the psychology department at State College consists of 60% females and 40% males. Last semester, the Psychology of Gender course had a total of 36 students, of whom 26 were female and only 10 were male. Are the proportions of females and males in this class significantly different from what would be expected by chance from a population with 60% females and 40% males? Test at the .05 level of significance.

STEP 1 ***State the hypotheses, and specify alpha.*** The null hypothesis states that the male/female proportions for the class are not different from what is expected for a population with these proportions. In symbols,

$$H_0: \quad p = p(\text{female}) = 0.60 \text{ and } q = p(\text{male}) = 0.40$$

The alternative hypothesis is that the proportions for this class are different from what is expected for these population proportions.

$$H_1: \quad p \neq 0.60 \text{ (and } q \neq 0.40)$$

We set alpha at $\alpha = .05$.

STEP 2 ***Locate the critical region.*** Because pn and qn are both greater than 10, we can use the normal approximation to the binomial distribution. With $\alpha = .05$, the critical region is defined as a z-score value greater than -1.96 or less than -1.96.

STEP 3 ***Calculate the test statistic.*** The sample has 26 females out of 36 students, so the sample proportion is

$$\frac{X}{n} = \frac{26}{36} = 0.72$$

The corresponding z-score (using Equation 18.2) is

$$z = \frac{X/n - p}{\sqrt{pq/n}} = \frac{0.72 - 0.60}{\sqrt{\dfrac{0.60(0.40)}{36}}} = \frac{0.12}{0.0816} = 1.47$$

STEP 4 ***Make a decision about H_0, and state a conclusion.*** The obtained z-score is not in the critical region. Therefore, we fail to reject the null hypothesis. On the basis of these data, you conclude that the male/female proportions in the gender class are not significantly different from the proportions in the psychology department as a whole.

PROBLEMS

1. To investigate the phenomenon of "home team advantage," a researcher recorded the outcomes from 64 college football games on one Saturday in October. Of the 64 games, 42 were won by home teams. Does this result provide enough evidence to conclude that home teams win significantly more than would be expected by chance? Use a two-tailed test with $\alpha = .05$.

2. Insurance companies charge young drivers more for automobile insurance because they tend to have more accidents than older drivers. To make this point, an insurance representative first determines that only 16% of licensed drivers are age 20 or younger. Because this age group makes up only 16% of the drivers, it is reasonable to predict that they should be involved in only 16% of the accidents. In a random sample of 100 accident reports, however, the representative finds 31 accidents that involved drivers who were 20 or younger. Is this sample sufficient to show that younger drivers have significantly more accidents than would be expected from the percentage of young drivers? Use a two-tailed test with $\alpha = .05$.

3. Güven, Elaimis, Binokay, and Tan (2003) studied the distribution of paw preferences in rats using a computerized food-reaching test. For a sample of $n = 144$ rats, they found 104 right-handed animals. Is this significantly more than would be expected if right- and left-handed rats are equally common in the population? Use a two-tailed test with $\alpha = .01$.

4. During the 2004 Olympic Games, Hill and Barton (2005) monitored contests in four combat sports: Greco-Roman wrestling, freestyle wrestling, boxing, and tae kwon do. Half of the competitors were assigned red outfits and half were assigned blue. The results indicate that participants wearing red won significantly more contests. Suppose that a sample of $n = 100$ contests produced 60 winners wearing red and only 40 wearing blue.
   a. Is this enough evidence to justify a conclusion that wearing red produces significantly more wins than would be expected by chance? Use a two-tailed test with $\alpha = .05$.
   b. Because the outcome of the binomial test is a borderline z-score, use the real limits for $X = 60$ to determine if the entire z-score interval is located in the critical region. If any part of the interval is not in the critical region, the correct decision is to fail to reject the null hypothesis.

5. Problem 6 in Chapter 17 cited a study showing that people tend to choose partners who are similar to themselves. Jones, Pelham, Carvallo, & Mirenberg, (2004) demonstrated that people have a tendency to select marriage partners with surnames that begin with the same last letter as their own. The probability of randomly matching two last names beginning with the same letter is only $p = 0.065$ (6.5%). The researchers looked at marriage records and found that 38 out of 400 brides and grooms had surnames beginning with the same last letter. Is this significantly more than would be expected by chance? Use a two-tailed test with $\alpha = .05$.

6. A researcher would like to determine whether people really can tell the difference between bottled water and tap water. Participants are asked to taste two unlabeled glasses of water, one bottled and one tap, and identify the one they thought tasted better. Out of 40 people in the sample, 28 picked the bottled water. Was the bottled water selected significantly more often than would be expected by chance? Use a two-tailed test with $\alpha = .05$.

7. In 1985, only 8% of the students in the city school district were classified as being learning disabled. A school psychologist suspects that the proportion of learning-disabled children has changed dramatically over the years. To demonstrate this point, a random sample of $n = 300$ students is selected. In this sample there are 42 students who have been identified as learning-disabled. Is this sample sufficient to indicate that there has been a significant change in the proportion of learning-disabled students since 1985? Use the .05 level of significance.

8. In the Preview section for Chapter 17, we discussed a study by Loftus and Palmer (1974) examining how different phrasing of questions can influence eyewitness testimony. In the study, students watched a video of an automobile accident and then were questioned about what they had seen. One group of participants was asked to estimate the speed of the cars when they "smashed into" each other. Another group of was asked to estimate the speed of the cars when they "hit" each other. Suppose that the actual speed of the cars was 22 miles per hour.
   a. For the 50 people in the "smashed-into" group, assume that 32 overestimated the actual speed, 17 underestimated the speed, and 1 was exactly right. Is this result significantly different from what would be expected by chance? Use a two-tailed test with $\alpha = .05$.
   b. For the 50 people in the "hit" group, assume that 27 overestimated the actual speed, 22 underestimated the speed, 1 was exactly right. Again, use a two-tailed test with $\alpha = .05$ to determine whether this result significantly different from what would be expected by chance.

9. A recent survey of practicing psychotherapists revealed that 25% of the individuals responding agreed with the

statement, "Hypnosis can be used to recover accurate memories of past lives" (Yapko, 1994). A researcher would like to determine whether this same level of belief exists in the general population. A sample of 192 adults is surveyed and 65 believe that hypnosis can be used to recover accurate memories of past lives. Based on these data, can you conclude that beliefs held by the general population are significantly different from beliefs held by psychotherapists? Test with $\alpha = .05$.

10. In 2005, Fung et al. published a study reporting that patients prefer technical quality versus interpersonal skills when selecting a primary care physician. Participants were presented with report cards describing pairs of hypothetical physicians and were asked to select the one that they preferred. Suppose that this study is repeated with a sample of $n = 150$ participants, and the results show that physicians with greater technical skill are preferred by 92 participants and physicians with greater interpersonal skills are selected by 58. Are these results sufficient to conclude that there is a significant preference for technical skill?

11. Danner and Phillips (2008) report the results from a county-wide study showing that delaying high school start times by one hour significantly reduced the motor vehicle crash rate for teen drivers in the study. Suppose that the researchers monitored 500 student drivers for 1 year after the start time was delayed and found that 44 were involved in automobile accidents. Before delaying the start time, the accident rate was 12%. Use a binomial test to determine whether these results indicate a significant change in the accident rate following the change in school start time. Use a two-tailed test with $\alpha = .05$

12. For each of the following, assume that a two-tailed test using the normal approximation to the binomial distribution with $\alpha = .05$ is being used to evaluate the significance of the result.

a. For a true-false test with 20 questions, how many would you have to get right to do significantly better than chance? That is, what X value is needed to produce a z-score greater than 1.96?

b. How many would you need to get right on a 40-question true-false test?

c. How many would you need to get right on a 100-question true-false test?

Remember that each X value corresponds to an interval with real limits. Be sure that the entire interval is in the critical region.

13. On a multiple-choice exam with 100 questions and 4 possible answers for each question, you get a score of $X = 32$. Is your score significantly better than would be expected by chance (by simply guessing for each question)? Use a two tailed test with $\alpha = .05$.

14. For each of the following, assume that a two-tailed test using the normal approximation to the binomial distribution with $\alpha = .05$ is being used to evaluate the significance of the result.

a. For a multiple-choice test with 48 questions, each with 4 possible answers, how many would you have to get right to do significantly better than chance? That is, what X value is needed to produce a z-score greater than 1.96?

b. How many would you need to get right on a multiple-choice test with 192 questions to be significantly better than chance?

Remember that each X value corresponds to an interval with real limits. Be sure that the entire interval is in the critical region.

15. Reed, Vernon, and Johnson (2004) examined the relationship between brain nerve conduction velocity and intelligence in normal adults. Brain nerve conduction velocity was measured three separate ways and nine different measures were used for intelligence. The researchers then correlated each of the three nerve velocity measures with each of the nine intelligence measures for a total of 27 separate correlations. Unfortunately, none of the correlations were significant.

a. For the 186 males in the study, however, 25 of the 27 correlations were positive. Is this significantly more than would be expected if positive and negative correlations were equally likely? Use a two-tailed test with $\alpha = .05$.

b. For the 201 females in the study, 20 of the 27 correlations were positive. Is this significantly more than would be expected if positive and negative correlations were equally likely? Use a two-tailed test with $\alpha = .05$.

16. In the Preview section for Chapter 11, we presented a study showing that swearing can help relieve pain (Stephens, Atkins, & Kingston, 2009). In the study, participants placed one hand in freezing cold water for as long as they could bear the pain. In one condition, they shouted a swear word over and over while the hand was in the water. In the other condition, they shouted a neutral word. Suppose that 18 of the 25 participants tolerated the pain longer while swearing than while shouting neutral words. Is this result significantly different from chance? Use a two-tailed test with $\alpha = .01$.

17. Thirty percent of the students in the local elementary school are classified as only children (no siblings). However, in the special program for talented and gifted children, 43 out of 90 students are only

children. Is the proportion of only children in the special program significantly different from the proportion for the school? Test at the .05 level of significance.

18. Stressful or traumatic experiences can often worsen other health-related problems such as asthma or rheumatoid arthritis. However, if patients are instructed to write about their stressful experiences, it can often lead to improvement in health (Smyth, Stone, Hurewitz, & Kaell, 1999). In a typical study, patients with asthma or arthritis are asked to write about the "most stressful event of your life." In a sample of $n = 112$ patients, suppose that 64 showed improvement in their symptoms, 12 showed no change, and 36 showed worsening symptoms.

a. If the 12 patients showing no change are discarded, are these results sufficient to conclude that the writing had a significant effect? Use a two-tailed test with $\alpha = .05$.

b. If the 12 patients who showed no change are split between the two groups, are the results sufficient to demonstrate a significant change? Use a two-tailed test with $\alpha = .05$.

19. Langewitz, Izakovic, and Wyler (2005) reported that self-hypnosis can significantly reduce hay-fever symptoms. Patients with moderate to severe allergic reactions were trained to focus their minds on specific locations where their allergies did not bother them, such as a beach or a ski resort. In a sample of 64 patients who received this training, suppose that 47 showed reduced allergic reactions and 17 showed an increase in allergic reactions. Are these results sufficient to conclude that the self-hypnosis has a significant effect? Use a two-tailed test with $\alpha = .05$.

20. Group-housed laying hens appear to prefer having more floor space than height in their cages. Albentosa and Cooper (2005) tested hens in groups of 10. The birds in each group were given free choice between a cage with a height of 38 cm (low) and a cage with a height of 45 cm (high). The results showed a tendency for the hens in each group to distribute themselves evenly between the two cages, showing no preference for either height. Suppose that a similar study tested a sample of $n = 80$ hens and found that 47 preferred the taller cage. Does this result indicate a significant preference? Use a two-tailed test with $\alpha = .05$.

21. In Problem 21 in Chapter 11, we described a study showing that students are likely to improve their test scores if they go back and change answers after reconsidering some of the questions on the exam (Johnston, 1975). In the study, one group of students was encouraged to reconsider each question and to change answers whenever they felt it was appropriate. The students were asked to record their original answers as well as the changes. For each student, the exam was graded based on the original answers and on the changed answers. For a group of $n = 40$ students, suppose that 29 had higher scores for the changed-answer version and only 11 had higher scores for the original-answer version. Is this result significantly different from chance? Use a two-tailed test with $\alpha = .01$.

22. The habituation technique is one method that is used to examine memory for infants. The procedure involves presenting a stimulus to an infant (usually projected on the ceiling above the crib) for a fixed time period and recording how long the infant spends looking at the stimulus. After a brief delay, the stimulus is presented again. If the infant spends less time looking at the stimulus during the second presentation, it is interpreted as indicating that the stimulus is remembered and, therefore, is less novel and less interesting than it was on the first presentation. This procedure is used with a sample of $n = 30$ 2-week-old infants. For this sample, 22 infants spent less time looking at the stimulus during the second presentation than during the first. Do these data indicate a significant difference? Test at the .01 level of significance.

23. Most children and adults are able to learn the meaning of new words by listening to sentences in which the words appear. Shulman and Guberman (2007) tested the ability of children to learn word meaning from syntactical cues for three groups: children with autism, children with specific language impairment (SLI), and children with typical language development (TLD). Although the researchers used relatively small samples, their results indicate that the children with TLD and those with autism were able to learn novel words using the syntactical cues in sentences. The children with SLI, on the other hand, experienced significantly more difficulty. Suppose that a similar study is conducted in which each child listens to a set of sentences containing a novel word and then is given a choice of three definitions for the word.

a. If 25 out of 36 autistic children select the correct definition, is this significantly more than would be expected if they were simply guessing? Use a two-tailed test with $\alpha = .05$.

b. If only 16 out of 36 children with SLI select the correct definition, is this significantly more than would be expected if they were simply guessing? Use a two-tailed test with $\alpha = .05$.

24. A researcher is testing the effectiveness of a skills-mastery imagery program for soccer players. A sample of $n = 25$ college varsity players is selected and each player is tested on a ball-handling obstacle course before beginning the imagery program and again after

completing the 5-week program. Of the 25 players, 18 showed improved performance on the obstacle course after the imagery program and 7 were worse.

a. Is this result sufficient to conclude that there is a significant change in performance following the imagery program? Use a two-tailed test with $\alpha = .05$.

b. Because the outcome of the binomial test is a borderline z-score, use the real limits for $X = 18$ and verify that the entire z-score interval is located in the critical region.

25. Last year the college counseling center offered a workshop for students who claimed to suffer from extreme exam anxiety. Of the 45 students who attended the workshop, 31 had higher grade-point averages this semester than they did last year. Do these data indicate a significant difference from what would be expected by chance? Test at the .01 level of significance.

26. Trying to fight a drug-resistant bacteria, a researcher tries an experimental drug on infected subjects. Out of 70 monkeys, 42 showed improvement, 22 got worse, and 6 showed no change. Is this researcher working in the right direction? Is there a significant effect of the drug on the infection? Use a two-tailed test at the .05 level of significance.

27. Biofeedback training is often used to help people who suffer migraine headaches. A recent study found that 29 out of 50 participants reported a decrease in the frequency and severity of their headaches after receiving biofeedback training. Of the remaining participants in this study, 10 reported that their headaches were worse, and 11 reported no change.

a. Discard the zero-difference participants, and use a sign test with $\alpha = .05$ to determine whether the biofeedback produced a significant difference.

b. Divide the zero-difference participants between the two groups, and use a sign test to evaluate the effect of biofeedback training.

PART V

REVIEW

After completing this part, you should be able to calculate and interpret correlations, find linear regression equations, conduct the chi-square tests for goodness of fit and for independence, and do a binomial test.

The most commonly used correlation is the Pearson correlation, which measures the direction and degree of linear relationship between two variables (X and Y) that have been measured on interval or ratio scales (numerical scores). The regression equation determines the best fitting line to describe the relationship between X and Y, and to compute predicted Y values for each value of X. A partial correlation can be used to reveal the underlying relationship between X and Y when the influence of a third variable is eliminated.

The Pearson formula is also used in a variety of other situations to compute special correlations. The Spearman correlation uses the Pearson formula when X and Y are both measured on ordinal scales (ranks). The Spearman correlation measures the direction and the degree to which the relationship is consistently one directional. When one of the variables consists of numerical scores and the other has only two values, the two values of the dichotomous variable can be coded as 0 and 1, and the Pearson formula can be used to find the point-biserial correlation. The point-biserial correlation measures the strength of the relationship between X and Y, and can be squared to produce the same r^2 value that is used to measure effect size for the independent-measures t test. When both variables are dichotomous, they can both be coded as 0 and 1, and the Pearson formula can be used to find the phi-coefficient. As a correlation, the phi-coefficient measures the strength of the relationship and is often used as a measure of effect size to accompany a chi-square test for independence for a 2×2 data matrix.

The chi-square test for goodness of fit uses the frequency distribution from a sample to evaluate a hypothesis about the corresponding population distribution. The null hypothesis for the goodness-of-fit test typically falls into one of two categories:

1. Equal proportions: The null hypothesis states that the population is equally distributed across the set of categories.
2. No difference: The null hypothesis states that the distribution for one population is not different from the known distribution for another population.

The chi-square test for independence uses frequency data from a sample to evaluate a hypothesis about the relationship between two variables in the population. The null hypothesis for this test can be phrased two different ways:

1. No relationship: The null hypothesis states that there is no relationship between the two variables in the population.
2. No difference: One variable is viewed as defining a set of different populations. The null hypothesis states that the frequency distribution for the second variable has the same shape (same proportions) for all the different populations.

The binomial test uses the frequencies or proportions from a sample to test a hypothesis about the corresponding population proportions for a binomial variable in the population. Because the binomial distribution approximates the normal distribution when pn and qn are both at least 10, it uses z-scores and proportions from the unit normal table for the test.

REVIEW PROBLEMS

1. The following scores are related by the equation $Y = X^2$. Note that this is not a linear relationship, but every time X increases, Y also increases.

X	Y
2	4
4	16
6	36
8	64
10	100

   a. Compute the Pearson correlation between X and Y. You should find a positive, but not perfect, correlation.

   b. Convert the scores to ranks and compute the Spearman correlation. You should find a perfect, positive correlation.

2. It is well known that similarity in attitudes, beliefs, and interests plays an important role in interpersonal attraction (see Byrne, 1971, for example). Thus, correlations for attitudes between married couples should be strong. Suppose that a researcher developed a questionnaire that measures how liberal or conservative one's attitudes are. Low scores indicate that the person has liberal attitudes, whereas high scores indicate conservatism. The following hypothetical data are scores for married couples.