Lesson

11-8 The Chi-Square Statistic

Vocabulary

expected number

deviation

chi-square statistic

BIG IDEA The chi-square statistic is found by adding squares of binomials and provides evidence for whether data found in certain tables represent events that are occurring randomly.

The *chi-square* statistic is different from any statistic you have yet seen. This statistic compares actual frequencies with the frequencies that would be expected by calculating probabilities, as shown below.

The following table shows the average daily numbers of live births in California between 1995 and 1997. Are birthdays randomly distributed among the days of the week?

Day of the Week	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.
Actual Numbers of Births	1,473	1,629	1,602	1,588	1,593	1,272	1,159

Source: *Journal of the American Medical Association*

The **expected number** of births is the mean number of births for a given day that is predicted by a probability. If the births occurred randomly, then the expected number for each day would be the same. There were 10,316 births each week on average. So the expected number of births for each day is $\frac{10{,}316}{7}$ which, rounded to the nearest integer, is 1,474.

Day of the Week	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.
Expected Numbers of Births	1,474	1,474	1,474	1,474	1,474	1,474	1,474

As you know, even if events occur randomly, it is not common for all events to occur with the same frequency. When you toss a coin ten times, you would not usually get 5 heads even if the coin were fair. Similarly, if there were 1,460 births on 4 days and 1,425 on the other three days, that would not seem to be much of a difference from the expected numbers. So the question is: Do the actual numbers deviate enough from the expected numbers that we should think that the births happen more often on certain days of the week?

If we let an expected number be e and an actual observed number be a, then $|e - a|$. The absolute value of the difference between these numbers is called the **deviation** of a from e. For example, the deviation on Saturday is $|1{,}474 - 1{,}272|$, or 202.

Mental Math

Solve.

a. $|n| = 14$

b. $|n| = -14$

c. $|-n| = 14$

d. $-|n| = 14$

In 1900, the English statistician Karl Pearson introduced the **chi-square statistic** as a way of determining whether the difference in two frequency distributions is greater than that expected by chance. ("Chi" is pronounced *ky* as in *sky*.) The algorithm for calculating this statistic uses the squares of deviations, which is why we study it in this lesson.

Calculating the Chi-Square Statistic

Step 1 Count the number of events. Call this number n. In the above situation, there are 7 events, one each for Mon., Tues., Wed., Thurs., Fri., Sat., and Sun.

Step 2 Let a_1, a_2, a_3, a_4, a_5, a_6, and a_7 be the actual frequencies.

In this example, $a_1 = 1{,}473$, $a_2 = 1{,}629$, $a_3 = 1{,}602$, $a_4 = 1{,}588$, $a_5 = 1{,}593$, $a_6 = 1{,}272$, and $a_7 = 1{,}159$.

Step 3 Let e_1, e_2, e_3, e_4, e_5, e_6, and e_7 be the expected frequencies.

In this example, $e_1 = e_2 = e_3 = e_4 = e_5 = e_6 = e_7 = 1{,}474$.

Step 4 Calculate $\frac{(a_1 - e_1)^2}{e_1}, \frac{(a_2 - e_2)^2}{e_2}, \ldots, \frac{(a_n - e_n)^2}{e_n}$. Each number is the square of the deviation, divided by the expected frequency.

$\frac{(a_1 - e_1)^2}{e_1} = \frac{1}{1{,}474}, \frac{(a_2 - e_2)^2}{e_2} = \frac{24{,}025}{1{,}474}, \frac{(a_3 - e_3)^2}{e_3} = \frac{16{,}384}{1{,}474}, \frac{(a_4 - e_4)^2}{e_4} = \frac{12{,}996}{1{,}474}, \frac{(a_5 - e_5)^2}{e_5} = \frac{14{,}161}{1{,}474}, \frac{(a_6 - e_6)^2}{e_6} = \frac{40{,}804}{1{,}474}, \frac{(a_7 - e_7)^2}{e_7} = \frac{99{,}225}{1{,}474}$

Step 5 Add the n numbers found in Step 4. This sum is the chi-square statistic.

$\frac{1}{1{,}474} + \frac{24{,}025}{1{,}474} + \frac{16{,}384}{1{,}474} + \frac{12{,}996}{1{,}474} + \frac{14{,}161}{1{,}474} + \frac{40{,}804}{1{,}474} + \frac{99{,}225}{1{,}474} = \frac{207{,}596}{1{,}474} \approx 140.8$

The chi-square statistic measures how different a set of actual observed numbers is from a set of expected numbers. The larger the chi-square statistic is, the greater the difference. But is 140.8 unusually large? You can find that out by looking in chi-square tables. These tables give the values for certain values of n and certain probabilities. On the next page is such a table. In this table, n is the number of events. The other columns of the table correspond to probabilities of 0.10 (an event expected to happen $\frac{1}{10}$ of the time), 0.05 (or $\frac{1}{20}$ of the time), 0.01 (or $\frac{1}{100}$ of the time), and 0.001 (or $\frac{1}{1{,}000}$ of the time). You are not expected to know how the values in the table were calculated. The mathematics needed to calculate them is normally studied in college.

Critical Chi-Square Values				
$n-1$	0.10	0.05	0.01	0.001
1	2.71	3.84	6.63	10.8
2	4.61	5.99	9.21	13.8
3	6.25	7.81	11.34	16.3
4	7.78	9.49	13.28	18.5
5	9.24	11.07	15.09	20.5
6	10.6	12.6	16.8	22.5
7	12.0	14.1	18.5	24.3
8	13.4	15.5	20.1	26.1
9	14.7	16.9	21.7	27.9
10	16.0	18.3	23.2	29.6
15	22.3	25.0	30.6	37.7
20	28.4	31.4	37.6	45.3
25	34.4	37.7	44.3	52.6
30	40.3	43.8	50.9	59.7
50	63.2	67.5	76.2	86.7

How to Read a Chi-Square Table

Examine the number 14.1, which appears in column 0.05, row 7. This means that, with 8 events, a chi-square value greater than 14.1 occurs with probability 0.05 or less.

On page 698, we obtained a chi-square value of 140.8 with $n = 7$ events. So we look in row $n - 1$, which is row 6. A value as large as 140.8 would occur with probability less than 0.001, that is, less than 1 in 1,000 times. So we have evidence that the births in California are not evenly distributed among the days of the week. The data should be examined to determine why Saturdays and Sundays have fewer births.

Suppose the frequencies of the births had led to a chi-square value of 10.9. Then, looking across row 6, we would see that this value is between the listed values 10.6 and 12.6. So 10.9 has a probability between 0.10 and 0.05. That means that a chi-square value as high as 10.9 would occur between $\frac{1}{10}$ and $\frac{1}{20}$ of the time just by chance. Statisticians normally do not consider this probability to be low enough to think there is reason to question the expected values.

When a chi-square value is found that occurs with probability less than 0.05, statisticians question whether the assumptions that led to the expected values are correct. With this criterion, the above distribution of births is highly unusual. So we would question whether the births are occurring randomly.

STOP QY

▶ QY

a. Why do you think there are fewer births on Saturday than on Monday through Friday?

b. Why do you think there are even fewer births on Sunday?

Example

Suppose 90 students were asked to name the United States President in 1950 from the names listed below. Suppose: 24 picked Dwight Eisenhower, 31 picked John Kennedy, and 35 picked Harry Truman (the correct answer). Is there evidence to believe the people were just guessing?

Harry S. Truman was the 33rd President of the United States.

Solution Calculate the chi-square statistic following the steps given above.

Step 1 Find the number of events. $n = 3$.

Step 2 Identify the actual observed values. $a_1 = 24$; $a_2 = 31$; $a_3 = 35$.

Step 3 Calculate the expected values. If people were just guessing, we would expect each of the three names to be picked by the same number of people. Since there were 90 people in all, each name would be picked by 30. So, $e_1 = 30$; $e_2 = 30$; $e_3 = 30$.

Step 4 Calculate $\frac{(a_1 - e_1)^2}{e_1}$, $\frac{(a_2 - e_2)^2}{e_2}$, and $\frac{(a_3 - e_3)^2}{e_3}$.

$$\frac{(a_1 - e_1)^2}{e_1} = \frac{(24 - 30)^2}{30} = \frac{36}{30}; \frac{(a_2 - e_2)^2}{e_2} = \frac{(31 - 30)^2}{30} = \frac{1}{30};$$

$$\frac{(a_3 - e_3)^2}{e_3} = \frac{(35 - 30)^2}{30} = \frac{25}{30}$$

Step 5 The sum of the numbers in Step 4 is $\frac{36 + 1 + 25}{30} = \frac{62}{30} \approx 2.07$.

Now examine the table. When $n = 3$, $n - 1 = 2$. So, look at the second row. The number 2.07 is less than the value 4.61 that would occur with probability 0.10. The numbers 24, 31, and 35 are like those that could randomly appear more than 10% of the time. It is quite possible that the people were guessing.

The chi-square statistic can be used whenever there are actual frequencies and you have some way of calculating expected frequencies. However, the chi-square value is not a good measure of the deviation from the expected frequencies when there is an expected frequency that is less than 5.

Questions

COVERING THE IDEAS

1. What does the chi-square statistic measure?
2. When was the chi-square statistic developed, and by whom?
3. For what expected frequencies should the chi-square statistic not be used?

In 4 and 5, average number of traffic deaths per day of the week in the United States are given for a particular year.

a. Calculate the chi-square statistic assuming that traffic deaths occur randomly on days of the week.

b. Is there evidence to believe that the deaths are not occurring randomly on the days of the week?

		Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.
4.	Year 1985	100	105	105	110	145	170	140
5.	Year 1995	100	100	100	105	140	150	130

6. Suppose in the Example of this lesson that 40 students had picked Harry Truman, 30 had picked Dwight Eisenhower, and 20 had picked John Kennedy. Would there still be evidence that students were guessing randomly?

APPLYING THE MATHEMATICS

7. You build a spinner as shown at the right and spin it 50 times with the following outcomes. Use the chi-square statistic to determine whether or not the spinner seems to be fair.

Outcome	1	2	3	4	5
Frequency	13	13	9	8	7

8. A coin is tossed 1,000 times and lands heads up 537 times. Compare the numbers of heads and tails with what would be expected if the coin were fair. Use the chi-square statistic to test whether the coin is fair.

9. *The World Almanac and Book of Facts 2006* lists 64 notable tornadoes in the United States since 1925. The table at the right shows their frequencies by season of the year. Use the chi-square statistic to determine whether these figures support a view that more tornadoes occur at certain times of the year than at other times of the year.

Season	Number of Tornadoes
Autumn	8
Winter	15
Spring	38
Summer	3

10. Here are the total points scored in each quarter from the 16 National Football League games played December 17–19, 2005.

Quarter	1	2	3	4	Total
Points	122	196	133	150	601

Source: National Football League

Use the chi-square statistic to answer this question. Do football teams tend to score more points in one quarter than in any other?

REVIEW

11. a. How many different permutations can be made using the letters of HORSE?
    b. How many different permutations can be made using the letters of MONKEY? **(Lesson 11-7)**

12. a. Which holds more, a cube with edges of length 6, or a rectangular box with dimensions 5 by 6 by 7? **(Lessons 11-6, 11-5)**
    b. Which holds more, a cube with edges of length x, or a box with dimensions $x - 1$ by x by $x + 1$?; Justify your answer.

In 13–15, expand and simplify the expression. (Lessons 11-6, 11-5)

13. a. $(2x + y)^2$ b. $(2x - y)^2$ c. $(2x - y)(2x + y)$

14. $(8 - a)(a - 8)$

15. $(3k^2 - 6km + 3m^2)^2$

16. Draw a picture of the following multiplication using rectangles. $4x(x + 6) = 4x^2 + 24x$ **(Lesson 11-3)**

17. On the fifth day after planting, a Moso bamboo tree was 38.5 cm. On the 14th day, the tree was 70.9 cm. **(Lessons 6-1, 5-4)**
    a. What was the average rate of change in height per day between the 5th and 14th days?
    b. Express the rate in Part a in cm/week.

EXPLORATION

18. A new high school for mathematics and science was opened in Cityville. Four hundred girls and 600 boys applied for 150 slots. A committee considered the applications and accepted 65 girls and 85 boys. Some people complained that there was discrimination.
    a. One complaint was that too few girls were accepted. This person's opinion was that there should have been equal numbers of boys and girls accepted and that the numbers accepted deviated too much from equality. Use the chi-square statistic to test whether this deviation could have occurred easily by chance.
    b. A second complaint was that too many girls were accepted. This person's position was that the numbers of boys and girls accepted should have been proportional to the number of applicants who were boys and girls. Use the chi-square statistic to test whether this deviation could have occurred easily by chance.
    c. If you were on the school board, would you agree with either complaint? Explain your answer.

Moso bamboo was introduced into the United States in about 1890.

Source: BAMBOO The Magazine of The American Bamboo Society

QY ANSWERS

a. Answers vary. Sample answer: With the ability to schedule deliveries, many physicians opt to schedule them during the week.

b. Answers vary. Sample answer: Many physicians choose Sunday as their day off and do not schedule deliveries for that day.