Math Objectives

- Students will be able to interpret whether values represented in a two-way table represent independent events by using conditional probabilities.
- Students will be able to derive a chi-square statistic for two categorical variables from a single population.
- Students will be able to determine if the value of a chi-square statistic indicates an association between two categorical variables.
- Students will make sense of problems and persevere in solving them (CCSS Mathematical Practices).
- Students will reason abstractly and quantitatively (CCSS Mathematical Practices).

Vocabulary

- association
- chi-square
- conditional probability
- independent

About the Lesson

- This lesson involves deriving and interpreting the chi-square statistic as an indication of whether two variables in a population are independent or associated.
- As a result, students will:
  - Derive the chi-square statistic based on observed values from a two-way table.
  - Analyze chi-square values calculated from samples drawn from a population in which the categorical variables are independent, and observe the distribution of such chi-square values.
  - Use a simulated chi-square distribution to decide whether the original value of the chi-square statistic indicates an association between the variables.

TI-Nspire™ Navigator™ System

- Use Class Capture to monitor students' progress.

Contingency Tables and Chi-Square

To begin Problem 1, move to page 1.2.
Answer the questions on the worksheet to investigate a possible association between gender and having a curfew.
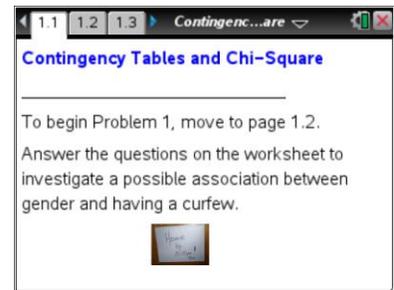
TI-Nspire™ Technology Skills:

- Download a TI-Nspire document
- Open a document
- Move between pages

Tech Tips:

- Make sure the font size on your TI-Nspire handhelds is set to Medium.

Lesson Files:
Student Activity

- Contingency_Tables_and_ Chi_Square_Student.pdf
- Contingency_Tables_and_ Chi_Square_Student.doc

TI-Nspire document

- Contingency_Tables_and_ Chi_Square.tns

Visit www.mathnspired.com for lesson updates and tech tip videos.

Discussion Points and Possible Answers

Move to pages 1.2 and 1.3.

This activity involves generating a number of random samples from a population. In order to avoid having your results be identical to those for another student in the room, it is necessary to "seed" the random number generator. Read the instructions on Page 1.2 for seeding your random number generator, then carry out that seeding on Page 1.3.
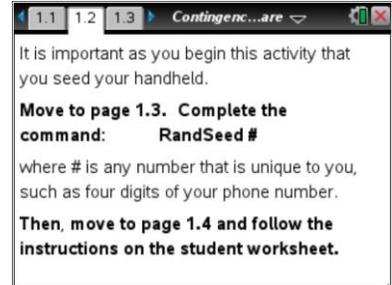
In the activity entitled **Two-way Tables and Association**, a randomly selected sample of 100 students from Rufus King High School was asked the question, "Do you have a curfew?" The results from the 100 students are summarized in the following table:

	Had a curfew	**Did not have curfew**	**Total**
Females	33	35	68
Males	20	12	32
Total	53	47	100

Teacher Tip: You might want to have students discuss the graph on Page 1.4, raising questions such as: Looking at the graph, do you think gender is associated with curfew? Explain your reasoning. (Note that hovering over a bar will display the percentage represented by that bar.) Some students might say more females have a curfew because the bar is much higher than for males with a curfew. Others might note that the difference in those that have and do not have a curfew for males is greater than it is for females.
This lesson sets the stage for applying the chi-square as a *test of independence*.

Move to page 1.4.

1. Based on the conditional probabilities displayed in this table, determine the following:
    a. What is the probability that if a male is selected from the sample, he has a curfew?

        Sample Answers: If a male is selected from this sample, the probability he has a curfew is 0.625, or 62.5%.

    b. What is the probability that if a female is selected from the sample, she has a curfew?

        Sample Answers: If a female is selected from this sample, the probability she has a curfew is 0.485, or 48.5%.

    c. Do you think this sample provides us with enough evidence to conclude that males at King are more likely to have a curfew? Explain your reasoning.

        Sample Answers: At this point, it is not clear what this sample indicates about any association between gender and curfew.
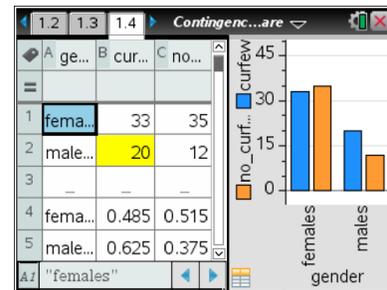
The sample of King students represents the **observed values of one sample**. If gender and having a curfew were **independent variables**, then the probability of males having a curfew would be the same as the probability of females having a curfew. If, however, the probability of males with a curfew was not the same as the probability of females with a curfew, then the variables would be **associated** and not independent.

2. If a student's gender and having a curfew were independent, what percent of males and what percent of females would have a curfew? Explain your reasoning.

    Sample Answers: The sample indicates 53 of the 100 students indicated they had a curfew. If having a curfew were independent of gender, then 53% of the 32 males and 53% of the 68 females have a curfew.

3. The highlighted cell in the table on page 1.4 represents the number of males with a curfew. Enter values in the highlighted cell to estimate the number of males and the number of females that result in approximately the same proportion of students with a curfew.

    a. How many of the males would have a curfew? How many of the females?
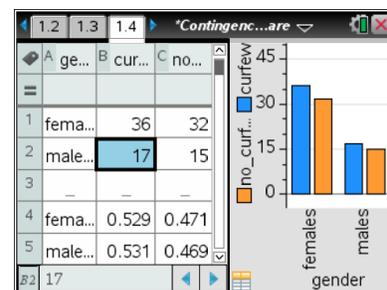
Sample Answers: The closest estimate would be 17
males and 36 females. If those values were represented in
the table, then 0.531 of the males have a curfew and 0.529
of the females have a curfew.

> **Teacher Tip:** The following problem asks students to determine the values
> that would result in exactly the same proportion of males and females with
> a curfew. The values could be derived by reviewing with students the
> properties of independent events and the multiplication rule of independent
> events that is often summarized as P(A and B) = P(A) x P(B). Similarly,
> there is a procedure to find the missing values by using the formula:
> expected value = (row total)(column total) / (overall total). If students use
> the formula, be sure they can explain why it makes sense.

b. Fill in the following table with the exact values that would indicate the proportion of males and
females with a curfew are the same. (Note that the number of males and females cannot be
represented by decimals values in reality.) These values, however, are the **expected values**
as they represent the values you would expect if curfew and gender were independent.

Sample Answers:

Expected values

	Had a curfew	Did not have curfew	Total
Females	**36.04**	**31.96**	68
Males	**16.96**	**15.04**	32
Total	53	47	100

A statistic, called a chi-square, is derived from the following formula:

$$\chi^2 = \sum \frac{(Expected - Observed)^2}{Expected}$$

4. a. Based on the formula, explain what you think a chi-square statistic represents.

Sample Answers: Essentially, the chi-square is a statistic that indicates how different the
observed sample is from a sample that would be obtained if the variables were independent.
It is a measure of how different the sample from King students is from a sample in which
gender and curfew at King were independent (question 3b).

b. Why do you think the square of the differences is divided by the expected value rather than summing the squared differences and calculating the average?

Sample Answers: It is important that the differences are relative to what was expected. A larger (or smaller) residual must be weighed against what was expected in order to determine if the residual value is truly different.

> **Teacher Tip:** The idea of "residual" runs through all of statistics and is clearly evident in the chi-square statistic. The difference, "Observed – Expected", provides residuals in the chi-square statistics. Squaring the difference will provide a positive measure for the differences. Dividing by the "Expected" value provides a way to judge the impact of the residual. For example, if a residual of 10 was obtained by a difference of 15 – 5 when 15 was observed and 5 was expected, then the square of the residual divided by the expected is 20. If a residual of 10 was obtained from 110 – 100, when 110 was observed and 100 was expected, and then the square of the residual divided by the expected is 1. The difference in the 20 and the 1 is important in evaluating the impact of the residual.

Move to page 1.5.

Assume the following key for the cells in the table displayed on Page 1.5:

- MC represents males with a curfew.
- FC represents females with a curfew.
- M_NC represents males with no curfew.
- F_NC represents females with no curfew.

5. Refer back to the original set of observations. The observed value for the number of males from King who had a curfew was 20. The table on Page 1.5 begins with a sample in which 20 males indicated they had a curfew and 33 females indicated they had a curfew. What is the value of the chi-square statistic for 20 males and 33 females with a curfew?

    Sample Answers: The chi-square value for 20 males is 1.705.

The question to think about is whether a chi-square value indicates that the differences between the observed and expected values are extreme or not. The following questions are designed to figure out the answer to that question by assuming it is possible to change the number of males with a curfew.

6. The cell highlighted on Page 1.5 in the row labeled as MC represents the number of males with a curfew.

    a. Change the number of observed males with a curfew to determine the following chi-square values.

Answers:

Number of males with a curfew	Chi-square
27 males	18.597
22 males	4.686
20 males	1.705
17 males	2.952 E-4 (.00030 or approximately 0)
12 males	4.539
5 males	26.389

b. What was the smallest chi-square value in your answer to part a? Explain why this number of males with a curfew results in a small chi-square value.

    Sample Answers: The smallest chi-square value was obtained from a sample with 17 males. If 17 of the 32 males had a curfew, then the percent of males with a curfew is close to the percent of the total number of students with a curfew, or 53%. The observed values are close to the expected values resulting in small residuals and a small chi-square value.

c. Find the largest chi-square value. Explain why this number of males with a curfew results in such a large chi-square value.

    Sample Answers: The largest chi-square value was obtained from a sample with 5 males. If 5 males have a curfew in a sample with 32 males and 53 students with a curfew, then the occurrence of curfew is very different than expected for males. The difference of the observed from the expected, or the residual, is larger than the previous differences and results in a larger chi-square value.

    Teacher Tip: As students continue to examine the formula for the chi-square statistic, they should realize it creates a statistic that indicates how different the values from an observed sample are from a sample in which the values are independent. The differences could be based on the number of males less than the expected value or greater than the expected value.

7. Use the table on Page 1.5 to determine the value of the chi-square in the following situations:

   a. Consider the case where none of the males had a curfew. What would be the value of the chi-square for this sample?

      Sample Answers: I entered 0 into the shaded cell. The chi-square value is 53.066.

   b. Consider the case where all of the males had a curfew. What would be the value of the chi-square for this sample?

      Sample Answers: I entered 32 into the shaded cell. The chi-square value is 41.731.

> **Teacher Tip:** The table is based on the number of males with a curfew. Questions involving the number of females with a curfew are also important and could have been used to determine the chi-square values. The table on Page 1.5, however, will not correctly derive the chi-square value for these questions as it is programmed to calculate the chi-square value involving males with a curfew and adjusting the other cells. Students could derive the chi-square value using their calculator or by hand. See the teacher notes at the end of this activity that directs you to a TI-Nspire activity involving how to create a chi-square model. This activity builds towards formal chi-squared hypothesis testing.

   c. What conjecture might you make about the association between gender and curfew if you observed either of the chi-square values in parts a and b?

      Sample Answer: The above chi-squares values are large. It is very unlikely we would obtain either the sample in 7a or the sample in 7b if there were little difference in the percent of males and females with a curfew at King. If I obtained one of these samples, I would conjecture that there is an association between curfew and gender at King.

Given your answer to part c, the question to be addressed is how large would a chi-square value have to be to conjecture that curfew and gender are associated.

Move to page 1.6.

Page 1.6 will generate simulated samples drawn from a population in which the variables gender and curfew are not associated. Each sample has 32 males and 68 females. 53 of the students are randomly selected from the 100 students and are considered the students who have a curfew (recall that 53% of the students have a curfew). The value of interest is the number of males selected from the 53 students with a curfew.
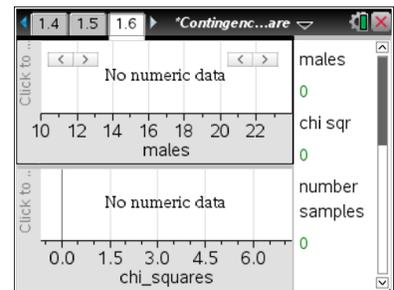
> **Teacher Tip:** Students should recognize that the simulated samples would provide a "snapshot" of the random variation in the chi-square values that are calculated from a population in which gender and curfew are independent (or nearly independent). Answers for the following questions will vary as each student will obtain different samples. The questions are designed to begin developing an understanding of two important observations: (1) the range of chi-square values that could be derived from a population in which the variables are independent, and (2) at what value is a chi-square considered large enough to indicate the sample probably did not come from a population in which the variables are independent.

8. Use the arrow on the left to select a sample.

    a. How many males have a curfew if 53 students were randomly selected from this sample of 32 males and 68 females?

        Sample Answers: My sample had 18 males with a curfew. (Answers will vary.)

    b. If this sample were compared to the expected values in which gender and curfew are independent, what is the value of the chi-square for this sample?

        Sample Answers: The chi-square value for 18 males is 0.20.

9. Generate at least five more samples. Note that clicking on a dot in either the top or bottom dot plot will highlight the corresponding dot in the other dot plot. Move the cursor over a dot to display its value. (Note: Click in an empty region in either plot to deselect the dots.)
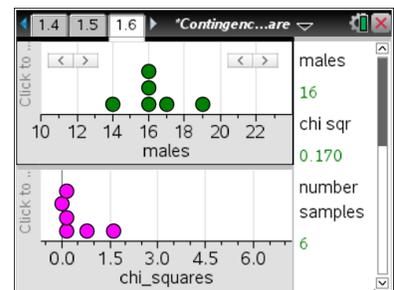

    Use the dot plots to find the observed number of males that resulted in the smallest chi-square value and the observed number of males that resulted in the largest chi-square value. Compare your answer to others in the class.

    Sample Answers: Most of us have the largest chi-square values for observed numbers of 12 males or less and 19 males or more. When the observed number of males was around 17, the chi-square values were close to 0. (Answers will vary based on students' samples.)

    TI-Nspire Navigator Opportunity: *Class Capture*
    See Note 1 at the end of this lesson.

---

10. Continue to use the arrow to obtain a total of 100 samples. As you generate each sample, note the relationship between the observed number of males and the location of the dot that represents the chi-square value.

    a. How does this relationship seem to fit with your answer to 7c?
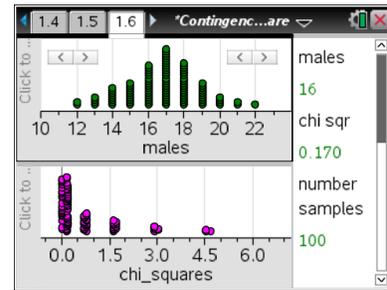
       Sample Answers: The conjecture I made still holds as I observe more samples. Samples in which most of the males had a curfew or samples in which very few of the males had a curfew result in chi-square values that are in the tail of the simulated chi-square distribution.

> **Teacher Tip:** The following problems ask students to describe the simulated distributions generated by the 100 samples. Discuss with the students the differences between the two distributions. For example, you might want to ask students to explain why one is skewed while the other is bell-shaped. This question provides an opportunity to review the chi-square formula and how the number of males with a curfew is connected to the chi-square values.

    b. Describe the simulated distribution for the number of males with a curfew.

       Sample Answers: The number of males is between 11 to 22 males. In general, the distribution is mound shaped around the middle value of 17 males.

    c. Describe the simulated distribution representing the chi-squares values for the 100 samples.

       Sample Answers: Most of the chi-squares values are close to zero. The distribution is skewed right; very few chi-squares values are large.

A sample that results in a large chi-square value is a noticeably different sample than the one identified with the expected values. A sample resulting in a large chi-square value indicates the variables are not independent. The question remains, however, what value of a chi-square is considered sufficiently large to suggest that gender and having a curfew are not likely to be independent? If we can identify a large chi-square value, then any sample resulting in this chi-square value or greater would be used to conjecture that there is an association between the variables in our population. Statisticians often identify "large" chi-squares as the largest chi-square values representing 5% or less of the chi-square distribution.

11. Consider the chi-square distribution for the 100 samples:

    a. Estimate the five largest chi-square values in the simulated sampling distribution. Click each point, and estimate the corresponding number of males with a curfew identified in the simulated distribution of males with a curfew. (There might be a cluster of points around some of the chi-square values. Estimate the largest chi-square values in the cluster and the connections to the number of males with a curfew.)

        Sample Answers: Chi-square values: 6.7 (23 males), 6.5 (11 males), 4.7 (22 males), 4.5 (12 males), and 3.01 (21 males). (Answers will vary based on students' samples; also, estimates might be very general.)

    b. Explain why some of the large chi-square values were a result of more males than expected having a curfew and other large chi-square values were a result of fewer males than expected having a curfew.

        Sample Answers: Large chi-square values are a result of observed values being very different from the values expected if the variables are independent. Some of the large differences resulted from the observed values being much greater than the expected values. In other cases, the observed values were much less than the expected values.

12. Recall the original sample obtained from King at the beginning of this activity.

    a. What was the number of males with a curfew in the original sample? Click on one of the points representing that number and estimate the corresponding chi-square value in the simulated chi-square distribution.

        Sample Answers: The number of males in the survey with a curfew was 20; the corresponding chi-square value was 1.71.

    b. Look at the distribution in the lower panel on Page 1.6. Approximately how many of the 100 chi-square values were greater than or equal to 1.7?

        Sample Answers: About 11 out of the 100 chi-square values, or 11%, were greater than or equal to 1.7. (Answers will vary.)

13. Based on the general standard for evaluating large chi-square values, does the observed number of 20 males with a curfew indicate there is an association between curfew and gender at King? Explain your answer.

   Sample Answers: The sample with 20 males with a curfew would not be unusual as 11% of 100 samples had a chi-square value equal to or greater than this chi-square value. This value is not equal to or greater than the values of 5% of the chi-squares generated in the distribution. Thus, the sample does not suggest that there is an association.

14. Henry was overheard complaining that it was unfair that he had curfew, while the females in his school were less likely to have a curfew. Based on the random sample that started this activity, is Henry's claim that females are less likely to have a curfew at King an accurate statement? Explain your answer.

   Sample Answers: Henry should not complain! Even though it looks like a greater percentage of males have curfews than females, the greater percentage could have happened by chance even if the variables of gender and curfew are independent. The chi-square value would suggest that the variables of gender and curfew are not associated.

   Teacher Tip: The chi-square statistic is also considered in the Math Nspired create activity entitled *Chi-Square Tests*. This activity explores three different chi-squared tests.

Wrap Up

Upon completion of the lesson, the teacher should ensure that students are able to:

- Use a statistic obtained from two categorical variables to determine whether the variables are independent.
- Interpret a chi-square value in terms of how "likely" the value is to have come from a population in which the variables are independent.
- Identify and interpret a "large" chi-square value as whether or not an observed sample came from a population in which the variables are associated.

Assessment

1. A random sample of 100 dogs (60 large and 40 small) was taken at a dog obedience school. Each of the dogs enrolled in the school was involved in a four-week obedience course. 80% of the dogs passed the course. If the size of a dog and whether or not the dog passed the course are independent, how many large dogs in the sample would you expect to have passed the course?

   Sample Answers: You would expect that 80% of the 60 dogs in the sample passed the course, or 48 large dogs.

2. What is the first step to determine if the size of a dog and whether or not the dog passed the course are associated?

   Sample Answers: The first step would be to determine what was expected if the size of the dogs and passing the course were independent. For a sample of 100 dogs, with 60 large dogs, 40 small dogs, and 80 dogs that passed the course, I would expect 48 large dogs passed the course, 32 small dogs passed the course, 12 large dogs did not pass the course, and 8 small dogs did not pass the course. These values represent what I would expect if the size of the dog and passing the course were independent.

3. Suppose 45 large dogs passed the course in our sample. How would a chi-square statistic indicate whether or not large dogs were more likely to pass the obedience course?

   Sample Answers: A chi-square statistic for the observed sample would need to be calculated. This value is then compared to a simulated distribution of chi-square values that occurs if the size of a dog and the result of passing the obedience course are independent. Finding the distribution would involve simulating a large number of samples in which the proportion of the size of the dogs and the results of the obedience course are the same as the observed sample. Examining a simulated distribution of chi-square values and identifying the values at which 5% or less of the chi-squares are equal to or greater than this value would provide an estimate of the large chi-square values.

TI-Nspire Navigator

Note 1

Question 10, *Class Capture*

As each student's dot plot will be slightly different, this question provides an opportunity to use TI-Nspire Navigator to capture screen shots of students' distributions. This will allow students to see how the simulated sampling distributions vary. Discuss with them how they are alike and how they are different. If using TI-Nspire Navigator is not possible, have several students display their simulated sampling distributions either with a document camera or on an overhead projector and discuss the features of the distributions.