What Did You Expect, Big Chi?

How did the U.S. Surgeon General determine that cigarette smoking is hazardous to your health? In this module, you investigate the statistical tests that medical researchers—and others—use to assess the information they collect.

Bill Chalgren • Robbie Korin • Pete Stabio

What Did You Expect, Big Chi?

Introduction

The Office of the U.S. Surgeon General requires cigarette packages to carry warning labels. Before concluding that cigarette smoking is a health hazard, researchers collected and analyzed thousands of pieces of information. For example, Table **1** lists the cause of death, along with the smoking habits, for 1000 randomly selected males, ages 45–64.

Table 1: Causes of death for 1000 males, ages 45–64

	Cancer	Heart Disease	Other
Nonsmokers	56	153	141
Smokers	136	308	206

Source: U.S. Department of Health, Education, and Welfare.

Using similar information, researchers attempted to determine whether smoking increased an individual's chances of dying from cancer or heart disease.

Discussion

a. Researchers compared the observed number of deaths for smokers with the number of deaths that would be expected if smoking did not increase the risk of contracting cancer or heart disease.

At the time the data in Table **1** was gathered, about 40% of the adult male population of the United States were smokers. Based on this statistic, how many smokers would you expect to find in a random sample of 1000 adult males?

b. After researchers identified differences between what they observed and what they expected, they had to decide if these differences were the result of increased risk, or were due to the chance variations in outcomes that occur in the sampling process.

1. Of the 1000 deceased adult males represented in Table **1**, how many had been smokers?

2. How does this value compare with your response to Part **a**?

3. Do you think the discrepancy between the expected number of smokers and the observed number of smokers is due to chance variation, or due to the possibility that smoking increases a male's chances of dying between the ages of 45 and 64?

c. Describe some real-world events that have an extremely small probability of occurring and yet do occur.

Activity 1

Before conducting an experiment, researchers typically state a hypothesis about the possible outcomes. Once the data has been collected, the actual results may differ from their expectations. The researchers then must decide if the differences between the observed frequencies and the expected frequencies are due to chance, or to an incorrect hypothesis.

Mathematics Note

The **expected frequency** of an outcome in an experiment is the number of times the outcome should theoretically occur. The **observed frequency** is the actual number of times (a non-negative integer) that the outcome occurs. When describing an experiment in this module, the following notation will be used:

- n represents the number of trials in an experiment
- k represents the number of different outcomes possible in each trial
- O_i represents the observed frequency of the ith outcome, where $i \in \{1,\ 2,\ 3,\ \ldots,\ k\}$
- E_i represents the expected frequency of the ith outcome
- p_i represents the theoretical probability of the ith outcome

For example, consider an experiment that involves rolling a fair die 30 times. Since there are 30 trials, each with 6 possible outcomes, $n = 30$ and $k = 6$. Since each outcome is equally likely, $p_i = 1/6$ for $i = 1,\ 2,\ \ldots,\ 6$. The expected frequency of each outcome is $(1/6) \bullet 30 = 5$, so $E_1 = 5,\ E_2 = 5,\ \ldots,\ E_6 = 5$. Table **2** shows the expected frequency of each outcome, along with some sample results for this experiment.

Table 2: Expected and observed frequencies for 30 rolls of a die

Outcome	1	2	3	4	5	6
Expected Frequency	5	5	5	5	5	5
Observed Frequency	7	3	5	8	2	5

Discussion 1

a. In an experiment that involves 200 flips of a fair coin, what is the expected frequency of heads? Explain your response.

b. The frequency of an outcome can be thought of as a random variable. Do you think that the observed frequency of heads in the experiment will always be the same as the expected frequency? Explain your response.

c. Is it possible to get 185 heads when tossing a coin 200 times?

d. When the observed number of heads differs from the expected number, at what point might you suspect that the coin is not fair?

e. Using the data in Table **2**, how would you analyze the differences between the observed frequencies and the expected frequencies to determine if these variations are due to chance?

Exploration 1

In most areas of research, repeating an experiment again and again would require too much time and cost too much money. To reduce the need for repeating experiments, statisticians have developed a method for determining whether observed differences are significant, or simply due to chance.

Mathematics Note

The **chi-square statistic**, denoted χ^2, is a measure of the difference between what actually occurred in an experiment and what was expected to occur. In an experiment with *k* possible outcomes, the chi-square statistic is the sum of the ratios of the squared differences of the observed and the expected frequencies $(O_i - E_i)^2$ to the expected frequency, where $i \in \{1,\ 2,\ 3,\ \ldots,\ k\}$. This can be denoted as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

For example, the chi-square statistic for the data in Table **2** can be found as follows:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_6 - E_6)^2}{E_6} \\ &= \frac{(7-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(5-5)^2}{5} = 5.2\end{aligned}$$

Once the chi-square value is calculated, it can be used to determine whether the differences between observed and expected frequencies are significant or the result of chance variation.

In the following exploration, you use histograms to investigate chi-square probability distributions. These distributions can be used in much the same way as the normal distribution described in the Level 6 module, "To Null or Not to Null."

a. Create a table to record the outcomes for an experiment that involves tossing a six-sided die. Your table should have rows and columns similar to those in Table **3** below.

Table 3: Calculation of chi-square

Outcome	1	2	3	4	5	6	Sum of Row
Expected Frequency (E_i)							
Observed Frequency (O_i)							
$O_i - E_i$							
$(O_i - E_i)^2$							
$(O_i - E_i)^2 / E_i$							$\chi^2 =$

b. Determine the expected frequency for each outcome if the die is tossed 30 times. Record these values in your table.

c. Perform the experiment by tossing a die 30 times. Record the observed frequency of each outcome in your table and calculate the chi-square statistic for the data.

d. Use technology to simulate the experiment in Part **c** 99 more times and calculate the corresponding chi-square values.

e. Create a histogram of the frequencies of the chi-square values from Part **d**. Sketch a smooth curve that closely fits the histogram and describe its shape and characteristics. **Note:** Save your work for use in Exploration **2**.

f. Consider an experiment that involves tossing five coins 160 times and counting the number of heads that appear each time. Table **4** shows the chi-square values for 99 such experiments. Create a histogram of the frequencies of these chi-square values. Sketch a smooth curve that closely fits the histogram and describe its shape and characteristics.

Table 4: 99 chi-square values for tossing 5 coins 160 times

3.16	2.36	5.08	3.98	10.70	1.26	6.12	0.86	4.36
10.20	6.62	10.38	10.72	7.58	7.42	1.66	3.58	8.44
1.72	10.04	10.94	3.72	3.18	4.06	8.44	5.24	2.88
4.92	3.82	3.02	6.36	0.72	6.52	1.82	5.82	8.78
16.70	5.42	0.56	8.86	2.48	1.42	3.52	2.40	15.90
0.94	2.22	4.76	3.10	2.94	19.26	0.80	3.12	3.96
11.66	5.68	3.88	3.66	4.14	4.54	3.92	2.86	2.52
5.64	5.04	7.00	2.86	3.64	6.78	4.06	3.82	1.92
4.06	2.52	1.82	3.9	2.12	3.04	4.14	1.24	0.80
7.58	6.14	5.04	5.64	6.08	2.04	3.34	6.78	6.86
8.64	1.04	7.18	17.36	3.64	3.50	2.16	6.86	3.32

Discussion 2

a. Will the sum of the entries in the $O_i - E_i$ row always equal 0? Explain your response.

b. **1.** How many different outcomes are possible when tossing an ordinary die?

2. What is the probability of each outcome?

c. **1.** How many different outcomes are possible when tossing five fair coins and counting the number of heads that appear?

2. What is the probability of each outcome?

d. Compare the shapes of the two curves you sketched in Exploration **1**.

e. Do you think that the difference in the probabilities of the outcomes when tossing a die and tossing five coins has any effect on the shape of the two curves? Justify your response.

Exploration 2

a. Consider an experiment that involves 20 tosses of a four-sided die with faces labeled 1, 2, 3, and 4. Determine the number of possible outcomes on each toss, their probabilities, and the expected frequency of each.

b. Use technology to simulate the experiment described in Part **a** 99 times and calculate the corresponding chi-square values.

c. Create a histogram of the frequencies of the chi-square values from Part **b**. Sketch a smooth curve that closely fits the histogram and describe its shape and characteristics.

d. Repeat Parts **a**–**c** for an experiment that involves tossing a 12-sided die 60 times.

e. Compare the smooth curves you sketched in Parts **c** and **d** with those you sketched in Exploration **1**.

Discussion 3

a. Compare the number of outcomes for each of the four experiments performed in Explorations **1** and **2**.

b. What characteristic of the experiment do you think affects the shape of the curves sketched in the explorations? Explain your response.

Mathematics Note

As shown in Figure **1**, a probability distribution of chi-square values, unlike a normal distribution, is not symmetric.

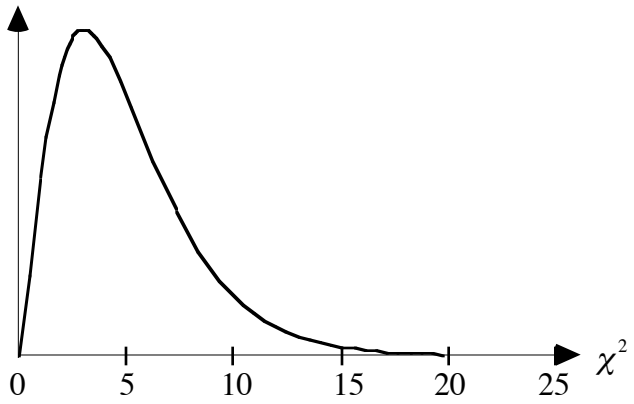


Figure 1: A chi-square distribution

As with a normal curve, however, the area of the region under the curve is 1. The shape of a chi-square distribution is determined by the **degrees of freedom** of the experiment. The degrees of freedom is based on the number of outcomes in an experiment.

For an experiment in which the events are independent and the theoretical probability for each outcome remains the same every time the experiment is performed, the degrees of freedom is equal to the number of possible outcomes minus 1, or $k - 1$.

For example, consider an experiment in which a single trial involves tossing three coins and counting the number of heads that appear. The four possible outcomes are 3 heads, 2 heads, 1 head, or 0 heads. Each trial is independent of other trials and the probability of obtaining each outcome remains the same. Since there are four possible outcomes, the degrees of freedom for the chi-square statistic in this experiment is $4 - 1 = 3$.

Note: In other types of experiments, the degrees of freedom must be determined using a different method and may not equal $k - 1$. You will investigate some examples in Activity **3**.

c. Explain how you know that the events in each experiment performed in the explorations are independent.

d. What are the degrees of freedom for each experiment in Explorations **1** and **2**?

e. As the degrees of freedom increase, what shape do you think the chi-square distribution will approach? Explain your response.

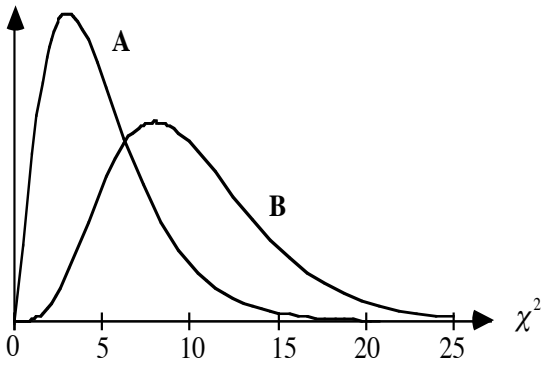
Assignment

1.1 **a.** Determine the degrees of freedom in an experiment that involves tossing one coin.

b. How many degrees of freedom are there in an experiment that involves drawing one card from an ordinary deck of playing cards and recording its suit? Explain your response.

c. How would the number of degrees of freedom in the playing-card experiment change if the face value of the card were recorded? Explain your response.

1.2 **a.** Based on your experience in the explorations, which of the chi-square probability distributions shown below involves more degrees of freedom? Justify your selection.



b. On each curve, estimate the chi-square value that divides the area under the curve in half. How do these values compare?

c. Consider the degrees of freedom and the chi-square values that divide the area under a curve in half. Based on your response to Part **a**, make a generalization about the relationship between the two.

d. Explain whether you think your generalization from Part **c** holds true for a value that divides the curve into any proportion.

1.3 **a.** Use the population in your mathematics class to complete the following table. Assume that males and females are equally likely to enroll in the class.
Note: Save your work for use in Problem **2.2**.

	Males	**Females**
Expected (*E*)		
Observed (*O*)		
$O - E$		
$(O - E)^2$		
$(O - E)^2/E$		

b. Determine the value of χ^2.

c. Determine the degrees of freedom for this experiment.

1.4 **a.** Develop a simulation for an experiment that involves flipping a fair coin 50 times.

b. Use the data from one simulation to complete the following table and calculate the corresponding chi-square value.

	Heads	**Tails**
Expected (*E*)		
Observed (*O*)		
$O - E$		
$(O - E)^2$		
$(O - E)^2 / E$		

c. Repeat the simulation 40 times to obtain 40 chi-square values.

d. Create a stem-and-leaf plot of the 40 values for χ^2. On your plot, locate the χ^2 that best estimates the value that separates the upper 10% of the data from the lower 90%.

e. Describe how you might modify the simulation to improve your estimate of the chi-square value that divides the upper 10% of the data from the lower 90%.

* * * * *

1.5 A chamber of commerce estimates that the numbers of small, medium, and large businesses in the city are approximately equal. To verify this estimate, the group conducts a survey. The results of a survey of 200 randomly selected businesses are shown in the following table.

Size of Business	**Small**	**Medium**	**Large**
Observed Frequency	42	67	91
Expected Frequency			

a. Determine the degrees of freedom for this situation.

b. Calculate the value of χ^2 for this data.

1.6 The student council estimates that 30% of the high school population hold part-time jobs. To support their estimate, the council selects a random sample of 32 students. The results of the survey are shown in the following table.

Part-time Job	**Yes**	**No**
Observed Frequency	14	18
Expected Frequency		

a. Determine the degrees of freedom for this situation.

b. Calculate the value of χ^2 for this data.

* * * * * * * * * *

Activity 2

At first glance, the information shown in Table **5** seems to indicate that smoking is related to cancer and heart disease. By itself, however, this intuitive observation does not provide sufficient reason to accept the connection.

Table 5: Causes of death for 1000 males, ages 45–64

	Cancer	**Heart Disease**	**Other**
Nonsmokers	56	153	141
Smokers	136	308	206

When a claim is tested using statistical methods, it is stated in the form of a hypothesis. Recall that the **null hypothesis** (H_0) is a statement about one or more parameters. The **alternative hypothesis** (H_a) is a hypothesis that is true if the null hypothesis is false.

In the Level 6 module "To Null or Not to Null," you learned to test hypotheses that compare a sample mean with a population mean. That test allowed you to compare a single observed value with the expected one. The chi-square statistic provides a tool for testing how well a set of observed data matches the expected frequencies for two or more categories.

Mathematics Note

The results of hypothesis testing are not expressed as certainties or absolutes. When testing a hypothesis, it is customary to limit the maximum probability of rejecting a true null hypothesis. This probability is the **level of significance** or **significance level**.

On a chi-square distribution with the appropriate degrees of freedom, the significance level is indicated by the area under the curve to the right of a given chi-square value. The significance level identifies the set of values that would lead to the rejection of the null hypothesis. The corresponding region under the curve is the **critical region**.

For example, Figure **2** shows a chi-square distribution with 5 degrees of freedom. The area under the curve to the right of 9.24 is approximately 10% of the total area. At a 0.10 level of significance, therefore, the shaded portion under the curve represents the critical region. For any chi-square value greater than 9.24, the null hypothesis should be rejected at the 0.10 significance level. This also means that the probability of incorrectly rejecting the null hypothesis is 10%. In other words, approximately 10% of true null hypotheses would be rejected using this significance level.

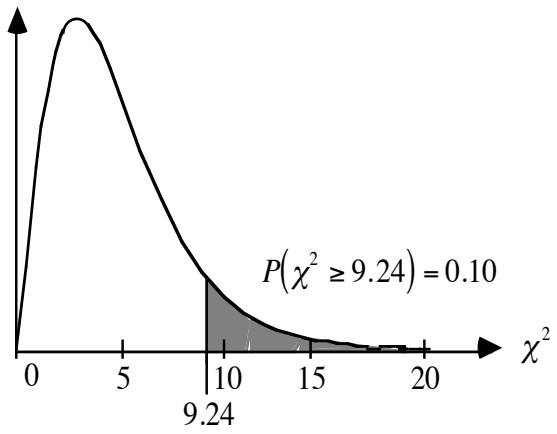


Figure 2: Chi-square distribution with 5 degrees of freedom

Chi-square values for various significance levels and degrees of freedom are often given in table form. Table **6** shows a portion of such a table. The value in the fifth row of the first column indicates that, in an experiment with 5 degrees of freedom, there is a 0.10 probability that a chi-square value greater than or equal to 9.24 will occur. This corresponds to the graph in Figure **2**. **Note:** A table with more values appears at the end of this module.

Table 6: Portion of a chi-square distribution table

	Significance Level				
Degrees of Freedom	**0.10**	**0.050**	**0.025**	**0.010**	**0.005**
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75

Discussion 1

a. **1.** Describe how to use a chi-square distribution table to determine if a chi-square value should result in a rejection of the null hypothesis at the 0.10 significance level with 5 degrees of freedom.

2. Explain how to use a graph to make the decision described above.

b. One traditional method of making decisions about statistical data is to formulate a hypothesis, then choose a significance level on which to base the decision to reject or fail to reject that hypothesis. At which significance level, 0.10 or 0.05, are you more likely to reject a true null hypothesis? Explain your response.

c. **1.** If you select a significance level of 0.05, approximately what percentage of the time would you fail to reject a valid null hypothesis? Explain your response.

2. If you select a significance level of 0.01, approximately what percentage of the time would you fail to reject a valid null hypothesis? Explain your response.

d. A null hypothesis that is rejected at one significance level might not be rejected at another significance level. Why does this occur?

e. In Figure **2**, why would a chi-square value greater than 12.83 indicate that the results of the experiment may not be due to chance?

f. Explain why a given significance level can also be described as the probability of incorrectly rejecting a true null hypothesis.

g. Table **7** contains some information on fatal automobile accidents in Montana in 1993, according to the day of the week on which they occurred.

Table 7: Fatal accidents, by day of the week

Day	Mon.	Tues.	Wed.	Thur.	Fri.	Sat.	Sun.
Fatal Accidents	15	16	15	23	34	38	25

Source: *1993 Montana Highway Patrol Annual Report*.

1. At first glance, the information in Table **7** seems to indicate that more fatal accidents occur on weekend days than on weekdays. What factors might contribute to higher numbers of fatal accidents on weekends?

2. State the hypothesis that you would test if you wished to determine whether this difference in number of fatal accidents is due to chance.

3. If you can reject your null hypothesis, would that imply anything about the potential causes for the observed data?

Exploration

In the following exploration, you use the chi-square statistic to decide whether or not an altered die is still fair. After collecting data, you test the null hypothesis H_0: The probability of each face occurring equals 1/6.

a. Obtain a fair die, an altered die, and a cup. To ensure unbiased outcomes, shake both dice in the cup, then release the dice so that they roll several times before coming to rest. Roll the dice 60 times and record the observed frequencies for each die separately.

b. To determine if the altered die is still fair, the frequency of its outcomes must be compared with the expected frequencies from a fair die.

1. Calculate the expected frequencies for each outcome for a fair die.

2. Then calculate the chi-square statistic for your rolls of the fair die.

c. **1.** Collect the class data for χ^2 from Part **b**.

2. Create a stem-and-leaf plot of the class data.

3. On your plot, locate the χ^2 value that best estimates the value that divides the upper 10% of the data from the lower 90%. (This approximates the 0.10 significance level.)

d. **1.** Calculate the chi-square statistic for your altered die, using expected frequencies for a fair die.

2. Plot the chi-square value for the altered die on the stem-and-leaf plot from Part **c**.

e. Based on the position of the chi-square value for the altered die, decide if you believe the null hypothesis should be rejected. Write a statement defending your decision.

Discussion 2

a. If the altered die is still fair, should the chi-square value found in the experiment be large or small? Explain your response.

b. Compare the chi-square value you obtained for the altered die with those of others in your class.

c. If the observed frequencies equaled the expected frequencies for all six faces of the die, what would be the value of χ^2 ?

Assignment

2.1 Using the chi-square distribution table at the end of this module, write a statement that describes the approximate probability of obtaining a chi-square value at least as great as the value you obtained for the altered die in Part **d** of the exploration.

2.2 **a.** In Problem **1.3**, you recorded the numbers of males and females in your mathematics class and determined the value of a chi-square statistic. Can this statistic be used to test the null hypothesis below? Explain your response.

H_0: Males and females are equally likely to enroll in this math class.

b. Using the chi-square distribution table, write a statement that describes the approximate probability of obtaining a chi-square value that is greater than the value in Problem **1.3**.

2.3 **a.** Describe the relationship between the chi-square values and the degrees of freedom in an experiment. Is this relationship true for all significance levels?

b. Explain why this relationship occurs.

2.4 The table below shows one year's data for the number of fatal accidents by day of the week.

Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Observed (O)	15	16	15	23	34	38	25
Expected (E)							

Complete Parts **a–d** below to test H_0: The probability of a fatal accident occurring is the same for each day of the week.

a. Calculate the expected number of fatal accidents for each day and record these values in the table.

b. Calculate χ^2 for this data.

c. Test H_0 at the 0.05 significance level. Summarize the results of your test, indicating whether you rejected or failed to reject H_0, and explaining whether or not the differences between the observed and expected frequencies are due to chance at this significance level.

2.5 A greeting card distributor recommends that stores carry the following percentages of cards: 25% love/friendship, 30% birthday, 20% wedding/anniversary, 10% sympathy/get well, and 15% other/special occasion.

The new manager of a gift shop wants to see if actual sales closely follow the distributor's recommended percentages. The data collected for cards sold during one week are shown in the following table.

Type	Love	Birthday	Wedding	Sympathy	Other
Sales	54	71	42	19	43

a. State H_0 and H_a for this situation.

b. Find the value of χ^2 that could be used to test the null hypothesis.

c. Based on the value of χ^2 from Part **b**, would you reject or fail to reject H_0 at the 0.05 significance level? Make a sketch of the approximate chi-square distribution to justify your response. Include the significance level, the area that indicates rejection of H_0, and the location of the chi-square value from Part **b**.

2.6 According to genetic theory, when red snapdragons and white snapdragons are crossed, the first generation of hybrids will have pink flowers. When these hybrids are crossed, the second generation will produce flowers in the following ratio of colors: 1 red to 2 pink to 1 white (1:2:1).

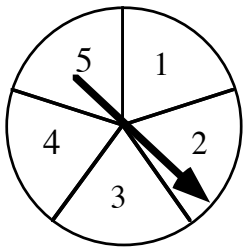
a. In a laboratory experiment involving snapdragons, the second generation consisted of 19 red, 59 pink, and 22 white flowers. Are these results consistent with genetic theory at the 0.05 significance level?

b. Can you be absolutely certain that your response to Part **a** is true? Explain your response.

* * * * *

2.7 A candy company claims that 30% of its popular candy mix is red, 30% is green, 20% is yellow, and 20% is brown. A random sample of 200 pieces of the mix contains 50 red, 54 green, 46 yellow, and 50 brown. Should the company adjust its current claims? Use a 0.05 significance level in your test.

2.8 A board game uses a spinner in order to determine the number of spaces a player may advance on each turn. The spinner is divided into five congruent sectors, as shown below.



One of the players thinks that the spinner is not fair. She tallies the results of 100 spins. The results of her experiment are shown in the following table.

Outcome	**1**	**2**	**3**	**4**	**5**
Observed Frequency	19	20	25	26	10

Determine an appropriate null hypothesis for this situation. Use a chi-square test with a 0.10 significance level to determine whether to reject or fail to reject your null hypothesis.

2.9 Toss a thumbtack into the air 100 times and record your observations. Is the probability that the tack lands point up the same as the probability that it lands point down? Use a chi-square statistic to test an appropriate hypothesis.

* * * * * * * * * *

Research Project

Formulate a null hypothesis regarding some topic of interest to you. Collect data and use the chi-square statistic to test your hypothesis. Write a report that includes the following:

- the null hypothesis
- a description of the method used to gather data
- the collected data
- the chi-square statistic generated
- your reasons for rejecting or failing to reject the null hypothesis
- a summary statement.

Activity 3

A clothing manufacturer plans to launch a new line of unisex apparel. To help maximize its appeal, the designers want to know if there is any relationship between gender and color preference. Do males tend to like blue or red, while females prefer purple or green?

In the previous activity, you used chi-square values to determine how well expected outcomes matched observed outcomes. In this activity, you use chi-square values to determine if two variables are dependent or independent.

Statistical dependence does not imply a cause-and-effect relationship between two variables. For example, consider an experiment that involves drawing 2 cards from a set of 4 red cards and 4 black cards. Although the probability of getting a red card on the second draw depends on the outcome of the first draw, the color of the first card does not *cause* the second draw to produce either a red or a black card.

Exploration

When the chi-square statistic is used to test whether or not two variables are dependent, the null hypothesis is stated assuming that the two variables are independent. If the null hypothesis is rejected, you treat the two variables as if they are dependent. In the following exploration, you will test the null hypothesis H_0: Color preference is independent of gender.

a. Do you think that color preference is independent of gender? For instance, given the colors blue, green, purple, and red, do you think that the distribution of color preferences in males will be about the same as in females? Record your prediction.

b. In order to test H_0, you must first design a survey and collect a sufficient amount of data. Record the gender and color preference for each person surveyed in a two-way table with headings like those in Table **8** below.

In order for the chi-square probability distribution to provide a reasonable model for an experiment, the expected frequency E_i for each possible outcome must be at least 5. If any cell in your expected frequency table contains a value less than 5, you must increase the sample size in the experiment. In other words, you must survey enough people to obtain expected frequencies of at least 5 for each outcome.

Table 8: Observed frequencies of color preferences

	Blue	**Green**	**Purple**	**Red**	**Total**
Female					
Male					
Total					

c. To record expected frequencies, create another two-way table with headings like those in Table **8**.

Since the totals in the right-hand column of the new table must be the same as those in Table **8**, enter these totals in the table now.

d. To find the expected frequency for females who prefer blue, complete the following steps.

1. Use the totals of the observed frequencies to determine the probability that a person in the sample is female.

2. Use the totals of the observed frequencies to determine the probability that a person in the sample prefers blue.

3. Your responses to Steps **1** and **2** are probabilities for events that are assumed to be independent. Use them to determine the probability that a person in the sample is female and prefers blue.

4. Use the sample size along with your response to Step **3** to determine the expected number of females who prefer blue. Record this number in the table from Part **c**.

e. Determine the expected frequencies of the other possible outcomes in the experiment and record them in the table.

f. In order to decide if color preference and gender are independent, you must next determine the degrees of freedom in the experiment. Table **9** shows some sample data for a survey of 70 people.

Table 9: Gender versus color preference

	Blue	Green	Purple	Red	Total
Female					34
Male					36
Total	22	15	16	17	70

Given these totals for the observed frequencies, determine the least number of values needed to allow you to find every other value in Table **9**. This value is the number of degrees of freedom for the experiment.

Mathematics Note

In an experiment to test the independence of two variables, the results can be displayed in a two-way table. In this case, the degrees of freedom can be calculated as follows, where r is the number of rows and c is the number of columns (not including the totals):

$$(r-1)\bullet(c-1)$$

For example, consider a survey in which a random sample of 100 students are asked to name their favorite academic subject. To test the independence of subject preference and gender, the researchers display the collected data in a two-way table, such as Table **10** below.

Table 10: Favorite subject of 100 students

	English	Math	History	Science	Total
Female	9	16	14	10	49
Male	17	10	11	13	51
Total	26	26	25	23	100

Considering only the cells that contain observed frequencies, this table has 2 rows and 4 columns. The degrees of freedom for this experiment can be calculated as follows:

$$(r-1)(c-1)=(2-1)(4-1)=1\bullet 3=3$$

g. Use the information given in the mathematics note to determine the degrees of freedom for your experiment involving gender and color preference.

h. Use the eight cells in your table of observed frequencies and the corresponding eight cells in the table of expected frequencies to calculate χ^2.

Discussion

a. When determining expected frequencies in Parts **d** and **e** of the exploration, why are gender and color preference assumed to be independent events?

b. Why would obtaining the same distribution of color preferences for males and females suggest that color preference is independent of gender?

c. In Parts **f** and **g** of the exploration, you determined the degrees of freedom for the experiment using two different methods. How did your answers compare?

d. If color preference and gender are found to be dependent, does this mean that being female causes a person to prefer a certain color? Explain your response.

Assignment

3.1 **a.** Test the hypothesis in the exploration at the 0.05 significance level and the appropriate degrees of freedom.

b. Summarize the results of your test, indicating whether you rejected or failed to reject H_0. Explain whether or not the differences between the observed and expected frequencies are due to chance at this significance level.

c. How would your decision regarding the null hypothesis change for a significance level of 0.025? of 0.005?

3.2 In the introduction to this module, you discussed the statistics shown in the table below, which shows cause of death for 1000 randomly selected males, ages 45–64, along with their smoking habits.

	Cancer	Heart Disease	Other	Total
Nonsmoker	56	153	141	350
Smoker	136	308	206	650
Total	192	461	347	1000

a. Use this data to test the null hypothesis, H_0: Cause of death is independent of smoking habits, at the 0.05 significance level.

b. Write a report on the results of your test, including your decision to reject or fail to reject the null hypothesis, the chi-square value obtained, the degrees of freedom in the experiment, the significance level used to test the null hypothesis, and the corresponding value of χ^2 in the chi-square distribution table.

3.3 A magazine subscription service wants to determine if readers' preferences for two national magazines are independent of geographical location. After selecting a sample from the national population, they classified readers by both magazine preference and geographical region. The table below shows the results of their survey.

	Magazine A	**Magazine B**	**No Preference**
Northeast	16	23	5
South	33	18	6
Midwest	15	20	7
West	20	32	5

a. Use this data to test the null hypothesis, H_0: Preference for magazine A or magazine B is independent of geographical region, at the 0.05 significance level.

b. Write a report on the results of your test, including your decision to reject or fail to reject the null hypothesis, the chi-square value obtained, the degrees of freedom in the experiment, the significance level used to test the null hypothesis, and the corresponding value of χ^2 from the chi-square distribution table.

3.4 The table below shows the results of another magazine preference survey, as described in Problem **3.3**.

	Magazine A	**Magazine B**	**No Preference**
Northeast	160	230	50
South	330	180	60
Midwest	150	200	70
West	200	320	50

a. In what ways does this data differ from the data in Problem **3.3**?

b. Does the probability that a person will prefer magazine A according to the table in Problem **3.3** differ from the probability according to the data in this table? Explain your response.

c. Do the probabilities of any preferences differ between the two tables? Explain your response.

d. Use the data in this table to test the null hypothesis, H_0: Preference for magazine A or magazine B is independent of geographical region, at the 0.05 significance level.

e. Use your results from Problems **3.3a** and **3.4d** to describe how the value of χ^2 was affected by the increase in sample size.

f. Do large sample sizes always imply large chi-square values? Explain your response.

* * * * *

3.5 Imagine that you are a quality control specialist at a manufacturing plant. As part of the quality control process, you select a sample from a week's production, then collect data on product performance by day manufactured. This information is shown in the table below.

	Mon.	**Tues.**	**Wed.**	**Thurs.**	**Fri.**
Acceptable	182	210	190	186	175
Not Acceptable	18	9	15	20	23

In your next report to the plant manager, you must analyze how product quality is related to the day manufactured. Write an appropriate statistical hypothesis, then use the data from the table to arrive at a decision about this hypothesis.

Your report should include an explanation of your decision and a discussion of the difference between statistical dependence (or independence) and a cause-and-effect relationship.

3.6 In the U.S. Armed Forces, the ready reserve includes those who are intended to assist active forces in a war. The table below shows some data collected in 1994 from a sample of 1000 ready-reserve personnel.

	Enlisted	**Officer**	**Total**
Female	101	21	122
Male	742	136	878
Total	843	157	1000

Consider the null hypothesis H_0: The gender of ready reserve personnel is independent of status. Use the data in the table to make a decision about this hypothesis.

* * * * * * * * * *

Summary Assessment

1. In the following problem, you will use an appropriate tool to generate 100 random digits (from 0 to 9, inclusive), then test to see if the digits have been generated in a truly random manner.

a. If the random number generator is truly random, what would you expect the probability of obtaining each digit to be?

b. Write a null hypothesis for this test.

c. Generate 100 random digits.

d. Record the frequency of each digit.

e. Calculate χ^2 and interpret its value at the 0.10 significance level.

f. Write a summary of your test.

2. In the exploration in Activity **3**, you used the chi-square statistic to test the hypothesis that color preference is independent of gender.

Choose another variable that might be related to gender. Design and conduct a survey to determine if this variable is independent of gender.

Write a report of your findings. Include a description of any factors that might create bias in your sample, an explanation of your decision to reject or not to reject the null hypothesis, the significance level used, the degrees of freedom in the experiment, the chi-square value obtained, and a discussion of the implications of your decision.

Module Summary

- The **expected frequency** of an outcome in an experiment is the number of times the outcome should theoretically occur. The **observed frequency** is the actual number of times the outcome occurs.
- The **chi-square statistic**, denoted χ^2, is a measure of the difference between what actually occurred in an experiment and what was expected to occur.

  In an experiment with k possible outcomes, the chi-square statistic is the sum of the ratios of the squared differences of the observed and the expected frequencies $(O_i - E_i)^2$ to the expected frequency, where $i \in \{1,\ 2,\ 3,\ \ldots,\ k\}$. This can be denoted as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)}{E_i}$$

- A probability distribution of chi-square values, unlike a normal distribution, is not symmetric. As with a normal curve, however, the area of the region under the curve is 1.
- The shape of a chi-square distribution is determined by the number of **degrees of freedom** of the experiment.

  For an experiment in which the events are independent and the theoretical probability for each outcome remains the same every time the experiment is performed, the degrees of freedom equal the number of possible outcomes minus 1, or $k - 1$.
- In an experiment to test independence of two variables, the results of the experiment can be displayed in a two-way table. In this case, the degrees of freedom can be calculated as follows, where r is the number of rows and c is the number of columns:

$$(r - 1) \bullet (c - 1)$$

- On a chi-square distribution with the appropriate degrees of freedom, the **significance level** is the area under the curve to the right of a given chi-square value. This represents the probability that a chi-square value greater than the given value can occur.

Selected References

Bailey, N. *Statistical Methods in Biology*. New York: Halsted Press, 1972.

Hamburg, M. *Statistical Analysis for Decision Making*. New York: Harcourt, Brace and World, 1970.

Huntsberger, D., and P. Billingsley. *Elements of Statistical Inference*. Boston: Allyn and Bacon, 1981.

Kvanli, A., C. Guynes, and R. Pavur. *Introduction to Business Statistics: A Computer Integrated Approach*. St. Paul, MN: West Publishing Co., 1986.

Montana Highway Patrol. *1993 Montana Highway Patrol Annual Report*. Helena, MT: Montana Highway Patrol, 1993.

National Council of Teachers of Mathematics (NCTM). *Curriculum and Evaluation Standards for School Mathematics Addenda Series. Grades 9–12: Data Analysis and Statistics*. Reston, VA: NCTM, 1992.

Reinhardt, H., and D. Loftsgaarden. *Elementary Probability and Statistical Reasoning*. Lexington, MA: D. C. Heath and Co., 1977.

Schefler, W. *Statistics: Concepts and Applications*. Menlo Park, CA: Benjamin/Cummings, 1988.

Triola, M. *Elementary Statistics*. New York: Addison-Wesley, 1992.

U.S. Bureau of the Census. *Statistical Abstracts of the United States: 1995*. Washington, DC: U.S. Government Printing Office, 1995.

U.S. Department of Health and Human Services. "A Physician Talks About Smoking: A Slide Presentation." (Document No. A14700.)

U.S. Department of Health, Education, and Welfare. "Chart Book on Smoking, Tobacco, and Health." (Document No. CDC75-7511.)

Witte, R. *Statistics*. New York: Holt, Rinehart, and Winston, 1985.

Chi-Square Distribution Table

	Significance Level				
Degrees of Freedom	**0.10**	**0.050**	**0.025**	**0.010**	**0.005**
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.65	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.72	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
⋮	⋮	⋮	⋮	⋮	⋮
20	28.41	31.41	34.17	37.57	40.00
30	40.26	43.77	46.98	50.89	53.67
40	51.80	55.76	59.34	63.69	66.77
50	63.17	67.50	71.42	76.15	79.49