Unit 29: Inference for Two-Way Tables

SUMMARY OF VIDEO

In this video, we visit the Broad Institute in Cambridge, Massachusetts, where our host, Dr. Pardis Sabeti, has a small research team investigating an ancient biological battle – the non-stop evolutionary arms race between our bodies and the infectious microorganisms that try to invade and inhabit them. The Broad Institute is home to new high tech tools such as the latest generation of genome sequencers. They allow us to sequence out the letters that code the genomes of both humans and our microbial enemies. In her research, Dr. Sabeti and her team use the data that these machines provide to find clues that might lead to new ways to battle some of our most dangerous diseases, diseases that we in the West rarely encounter.

One of the deadliest is Lassa fever, which, like the more notorious tropical disease Ebola, is caused by a virus and kills its victims with hemorrhagic fever. Throughout West Africa, thousands of people die of Lassa fever every year. But what is surprising is that many tens of thousands more throughout the region are exposed to the virus without getting sick. This suggests that these people have some sort of resistance to the virus. It is the source of this resistance that Dr. Sabeti wants to discover.

Dr. Sabeti's work on Lassa fever is still in its early stages, but one of the models for what she hopes to uncover can be found in the research on another tropical disease, malaria, which kills and sickens millions every year. With malaria we already know of one important source of resistance to the disease. It's a genetic mutation that is better known for the harm it does than for the good – sickle cell anemia.

(Continued on next page...)

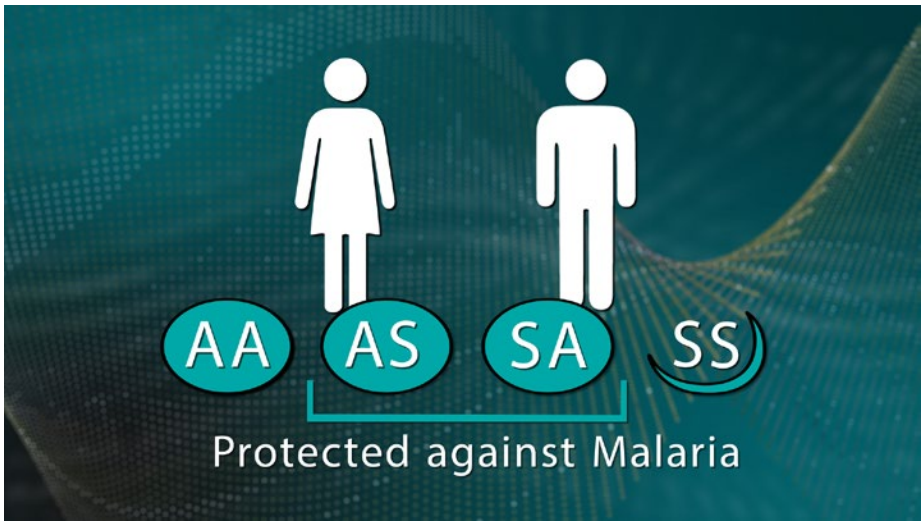

Figure 29.1. Inheriting the sickle cell mutation.

As we discovered in the module on binomial distributions (Unit 21), if a child inherits two copies of the sickle cell mutation (SS) from his or her parents, the child will have sickle cell anemia. If the child inherits only one copy of the gene, he or she is unaffected by the disease, but more importantly the child is protected against malaria. (See Figure 29.1.) It is this protective effect that is responsible for the sickle cell mutation becoming so prevalent and it is statistics that can reveal it.

To see how two-way tables can help reveal protective factors, Dr. Sabati has borrowed some data from Dr. Hans Ackerman. He and his colleagues looked at the genotypes of 315 children with severe malaria. Since each child inherits one hemoglobin gene from each parent, they examined 630 genes in total. The researchers wanted to quantify whether children who came down with malaria were less likely to have inherited the protective sickle cell version of the gene (HbS) rather than the normal version (HbA), as compared to the general population. Table 29.1 shows the breakdown of HbA and HbS in two groups of children. The top row of the table shows the genes they found in the children with severe malaria. The bottom row shows the genes they found in a control group of newborn babies.

	HbA Susceptible	HbS Protective	Total
Malaria	623	7	630
Control	1065	101	1166

Table 29.1. Table of hemoglobin gene in two groups of children.

Intuitively, we would expect to find the protective version of the gene less frequently in the children sick with malaria than in the control group. After all, if they were protected, they likely wouldn't have come down with the disease. Table 29.2 shows the conditional distribution of HbA and HbS for each group of children.

	HbA Susceptible	HbS Protective	Total
Malaria	98.89%	1.11%	100%
Control	91.34%	8.66%	100%

Table 29.2. Conditional distribution of HbA and HbS for each group.

Notice that HbS was inherited by the kids who caught malaria only 1.11% of the time compared to 8.66% of the time by the control group. Is that difference larger than would be expected just by chance? Is it statistically significant? We can conduct a test of hypotheses to find out whether there is sufficient evidence that the status of two variables – Malaria/General Population and HbS/HbA – are linked. Our null hypothesis is that there is no association between contracting malaria and having the HbS sickle cell gene. The alternative hypothesis is that there is an association between contracting malaria and having the protective HbS sickle cell gene.

H_0: No association between malaria and HbS

H_a: Association between malaria and HbS

We can compute what the expected counts in our two-way table would be if there really is no association between our variables as the null hypothesis states. Here's how to compute the expected counts:

$$\text{expected count} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

Table 29.3 shows the results of adding the expected counts to our two-way table.

		HbA Susceptible	HbS Protective	Total
Malaria	Observed	623	7	630
	Expected	592.1	37.9	630.0
Control	Observed	1065	101	1166
	Expected	1095.9	70.1	1166.0

Table 29.3. Adding the expected counts.

Now we can see that if there were no relationship between having the gene and coming down with the disease we would expect to find 37.9 HbS genes in the children with malaria. But in reality there are only 7 HbS genes in that group. Is that difference between 7 and 37.9 enough to tell us that there is an association between our two categorical variables? The next step in our analysis is to use the chi-square test statistic, given below, to figure out if that difference is significant.

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

The chi-square test statistic is a measure of how far the observed counts in the table are from the expected counts. Here are the calculations:

$$\chi^2 = \frac{(623 - 592.1)^2}{592.1} + \frac{(7 - 37.9)^2}{37.9} + \frac{(1065 - 1095.9)^2}{1095.9} + \frac{(101 - 70.1)^2}{70.1}$$

$$\chi^2 \approx 41.26$$

Using software, we find the *p*-value: $p \approx 0$. So, we have very strong evidence that there is an association between our variables and we can reject our null hypothesis. This result, together with the pattern of the data, gives support to the research hypothesis that the HbS sickle cell variant of the hemoglobin gene does protect against malaria.

STUDENT LEARNING OBJECTIVES

A. Understand the basic principles of the chi-square test.

B. Know how to calculate the expected cell counts in a two-way table.

C. Know the assumptions required for the chi-square test of independence.

D. Be able to conduct a chi-square test of independence.

CONTENT OVERVIEW

Each year, the study *Monitoring the Future: A Continuing Study of American Youth* (MTF) surveys 12th-grade students on a wide range of topics related to behaviors, attitudes, and values. It is a major source of information on smoking, drinking, and drug habits of American youth.

Suppose we want to investigate whether the environment in which students grow up is linked to the likelihood that they have consumed alcohol (more than just a few sips). We focus on three growing-up environments – a farm, the country, or a small-to-medium size city. Since we expect the growing-up environment may help us explain the likelihood of alcohol consumption, environment is the explanatory variable, and alcohol consumption is the response variable. We are interested in testing if there is an association between these two variables or if they are independent.

The two-way table in Table 29.4 shows the results on these questions from the 2011 MTF survey.

		Environment		
	Count	A Farm	Country	Small/Medium City
Alcohol	No	144	342	1366
	Yes	305	800	3049

Table 29.4. Results from questions on growing-up environment and drinking alcohol.

We begin analyzing these data using techniques covered in Unit 13, Two-Way Tables. Because we think that growing-up environment explains whether or not students might have consumed alcohol, we calculate the conditional percentages for the variable alcohol for each level of environment. In other words, we compute the column percentages, which appear in Table 29.5.

		Environment		
	Percent	A Farm	Country	Small/Medium City
Alcohol	No	30.94	29.95	32.07
	Yes	69.06	70.05	67.93
	Total	100.00	100.00	100.00

Table 29.5. Column percentages.

Based on Table 29.5, it looks as if students who grew up in the country were the most likely (70.05%) to have drunk alcoholic beverages and the students who grew up on a farm were

the least likely (67.93%). The question is whether these differences are due to an association between the two variables or could these differences be due simply to chance variation? In order to distinguish between these two cases, we set up hypotheses for a significance test:

H_0 : No association between drinking alcohol and growing-up environment.
H_a : Association between drinking alcohol and growing-up environment.

Remember, the data in Table 29.5 came from a sample of 12th-grade students. The meaning of the null hypothesis is that in the population of all 12th-grade students in America there is no difference among the distributions of alcohol consumption for the three growing-up environments. To test H_0 we compare the observed counts from Table 29.4 with the counts that we would expect to see if the two variables were independent (no association). If it turns out that the **observed counts** are far from the **expected counts**, then we would have evidence against the null hypothesis. Here's how to calculate the expected counts.

Calculating the Expected Counts

Assume that H_0 is true and that there is no association between two variables in a two-way table. Then the expected count in any cell of the table is computed as follows:

$$\text{expected count} = \frac{(\text{row total})(\text{column total})}{\text{grand total}},$$

where the grand total is the sum of the counts in all cells in the table.

Before calculating the expected counts, we add the row and column totals to our table of counts (See Table 29.6.).

		Environment			Total
		Farm	Country	City	
Alcohol	No	144	342	1366	1852
	Yes	305	800	3049	4154
Total		449	1142	4415	6006

Table 29.6. Addition of row and column totals to Table 29.4.

For example, the expected count for the cell in the first row, first column is:

$$\text{expected count} = \frac{(1852)(449)}{6006} = 138.45$$

Table 29.7 shows the expected counts added to the table. For each cell, the expected counts appear below the observed counts.

		Environment			Total
		Farm	Country	City	
Alcohol	No	144	342	1366	1852
		138.5	352.1	1361.4	
	Yes	305	800	3049	4154
		310.5	789.9	3053.6	
Total		449	1142	4415	6006

Table 29.7. Table 29.6 with expected counts added.

If there is no association between alcohol consumption and growing-up environment, the expected counts should be close to the observed counts. We compare the observed and expected counts by way of a **chi-square test statistic**, χ^2, which is given below.

Computing the Chi-Square Test Statistic

The chi-square test statistic measures how far the observed counts in a two-way table are from the expected counts.

The χ^2- test statistic is calculated by the following formula:

$$\chi^2 = \sum \frac{\left(\text{observed} - \text{expected}\right)^2}{\text{expected}}$$

Next, we calculate the value of the chi-square test statistic:

$$\chi^2 = \frac{\left(144 - 138.5\right)^2}{138.5} + \frac{\left(342 - 352.1\right)^2}{352.1} + \frac{\left(1366 - 1361.4\right)^2}{1361.40.76}$$

$$+ \frac{\left(305 - 310.5\right)^2}{310.5} + \frac{\left(800 - 789.9\right)^2}{789.5} + \frac{\left(3049 - 3053.6\right)^2}{3053.6}$$

$$\approx 0.76$$

If the null hypothesis is true and the cell counts are reasonably large, then the chi-square test statistic has an approximate chi-square distribution. Like *t*-distributions, chi-square distributions are specified by degrees of freedom. In this case, the degrees of freedom depend on the number of rows and columns of the table: $df = (r - 1)(c - 1)$, where r and c are the number of rows and columns, respectively. Table 29.4 has two rows and three columns.

So, we get $df = (2 - 1)(3 - 1) = 2$. To calculate a *p*-value, we find the probability of observing a value from the chi-square distribution with $df = 2$ that is at least as large as the one we observed, $\chi^2 = 0.76$. Using software, we determine that $p \approx 0.684$ as can be seen in Figure 29.2. Assuming the null hypothesis is true, we would expect to see χ^2-values at least as large as the one we observed around 68% of the time. That's pretty often. So, we have insufficient evidence to reject the null hypothesis. We conclude that there does not appear to be an association between students' drinking alcohol and whether students grew up on a farm, in the country, or in a city.

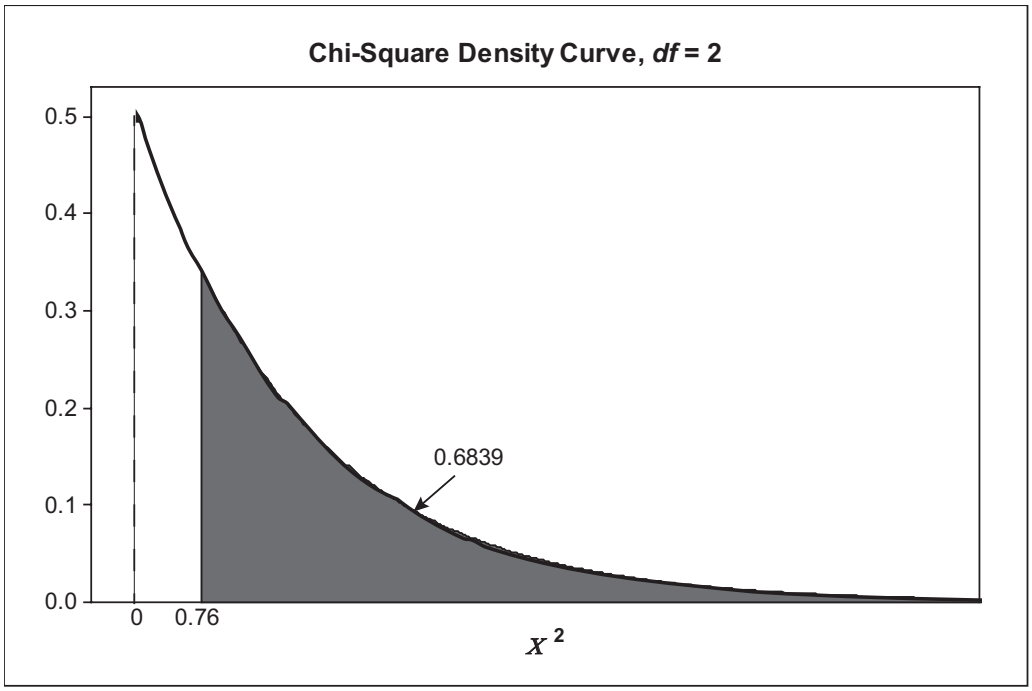

Figure 29.2. Calculating the p-value from a chi-square distribution.

Next, we ask whether the same results would be true for 12th-grade students' smoking habits. In other words, are the smoking habits of 12th-grade students independent of their growing-up environment? Table 29.8 gives results for these questions from the 2011 MTF survey. (More students answered the question on smoking than did on drinking alcohol.)

		Environment		
		Farm	Country	City
Smoking	Never	299	738	3218
	Occasionally	159	403	1521
	Regularly, now or past	103	236	596

Table 29.8. Table of smoking and growing-up environment.

Again we set up the null and alternative hypotheses:

H_0 : No association between smoking and growing-up environment.

H_a : Association between smoking and growing-up environment.

This time we leave the work of calculating the expected cell counts to the statistical software Minitab. Figure 29.3 shows the Minitab output.

```
Rows: SMOKING   Columns: Growing-Up Environment

                    FARM  COUNTRY    CITY   Total

Never                299      738    3218    4255
                   328.2    805.6  3121.2  4255.0

Occasionally         159      403    1521    2083
                   160.7    394.4  1528.0  2083.0

Regularly            103      236     596     935
 (now or past)      72.1    177.0   685.9   935.0

Total                561     1377    5335    7273
                   561.0   1377.0  5335.0  7273.0

Cell Contents:      Count
                    Expected count

Pearson Chi-Square = 56.154, DF = 4, P-Value = 0.000
```

Figure 29.3. Minitab chi-square analysis for smoking and growing-up environment.

Notice that the cell counts appear below the observed counts in the table. The value of the test statistic is $\chi^2 \approx 56.2$. Since this is a 3×3 table, *df* = (3 – 1)(3 – 1) = 4. Minitab reports the *p*-value as approximately 0. So, we conclude that the variables smoking and growing-up environment are not independent – there is an association. The results from the chi-square test do not tell us anything about the nature of the association, only that there is one. To learn

about the nature of that association, we look at the conditional distributions of smoking for each of the growing-up environments. Table 29.9 shows the column percentages.

		Environment		
	Percent	Farm	Country	City
Smoking	Never	53.3	53.6	60.3
	Occasionally	28.3	29.3	28.5
	Regularly, now or past	18.4	17.1	11.2
	Total	100.0	100.0	100.0

Table 29.9. Conditional distribution of smoking for each growing-up environment.

What we notice from Table 29.9 is that a higher percentage of students who grew up in a city never smoked (60.3%) compared to students who grew up on a farm (53.3%) or in the country (53.6%). The percentages for students who occasionally smoked (but not regular smokers) were about the same for all three growing-up environments. However, the percentage of regular smokers (either now or in the past) was higher for students who grew up on a farm (18.4%) or in the country (17.1%) compared to students who grew up in a city (11.2%).

The chi-square test, like the *z*-test for proportions, is an approximate method that becomes more accurate as the cell counts get larger. If the expected cell counts get too low, the test becomes untrustworthy. Here are some guidelines for when a chi-square test gives accurate results.

Guidelines for Using Chi-Square Test

The chi-square test gives trustworthy results provided the following are satisfied:

- All expected counts are greater than 1.
- No more than 20% of the expected counts are less than 5.

Statistical software will often give a warning if the guidelines have been violated. For example, energy drinks – non-alcoholic beverages that usually contain high amounts of caffeine (e.g., Red Bull, Full Throttle, and Monster) – have caused concern in the medical community. Suppose we wanted to know if the pattern of daily consumption of energy drinks was associated with students' growing-up environment.

The output from Minitab appears in Figure 29.4. Notice the software reports the value of the chi-square test statistic, but this time it does not provide a *p*-value. Instead it prints a warning, which we have highlighted.

```
Rows: Energy Drinks   Columns: Growing-Up Environment

                  FARM  COUNTRY    CITY    Total

None                57      144     598      799
                 52.55   150.44  596.01   799.00

One                 11       44     160      215
                 14.14    40.48  160.38   215.00

Two                  4       13      36       53
                  3.49     9.98   39.54    53.00

Three                1        3      20       24
                  1.58     4.52   17.90    24.00

Four                 0        2      10       12
                  0.79     2.26    8.95    12.00

Five or more         0        3       4        7
                  0.46     1.32    5.22     7.00

Total               73      209     828     1110
                 73.00   209.00  828.00  1110.00

Cell Contents:      Count
                    Expected count

Pearson Chi-Square = 7.773, DF = 10

* WARNING * 2 cells with expected counts less than 1
* WARNING * Chi-Square approximation probably invalid

* NOTE * 7 cells with expected counts less than 5
```

Figure 29.4. Minitab chi-square analysis for energy drinks and growing-up environment.

In this case, we could combine some of the categories for energy drinks. For example, we might combine categories Three, Four, and Five or more into a single category “Three or more.” You will get a chance to try this approach in the exercises.

Data for two-way tables can arise in different ways. In the case of the *Monitoring the Future* data, a single sample of high school students was chosen to take part in the survey. Their responses to two questions (two categorical variables) were organized into two-way tables. That was not the case for the data discussed in the video. Those data came from two different samples, a sample of children sick with malaria and a sample of newborns (control group),

which were then classified according to one categorical variable, HbS/HbA. In this case, the sample, malaria or control, was the second variable in the two-way table. There is no difference in the analysis used in these two situations. The expected counts and chi-square test statistics were computed using the same formulas in both cases.

KEY TERMS

The **observed cell counts** (or frequencies) are the actual number of observations that fall into each cell (class). The **expected counts** (or frequencies) are the number of observations that should fall into each class in a frequency distribution under the hypothesized probability distribution.

Chi-square statistic:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Degrees of freedom for chi-square test of independence: $(r-1)(c-1)$, where the number r and c are the number of rows and columns, respectively.

Expected count for chi-square test of independence:

$$\text{expected count} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}.$$

THE VIDEO

Take out a piece of paper and be ready to write down the answers to these questions as you watch the video.

1. What type of research is the host of this series, Dr. Pardis Sabeti, involved in?

2. Dr. Sabeti's work is modeled off of work done on malaria. What genetic mutation is an important source of resistance to malaria?

3. What were the null and alternative hypotheses for testing whether the sickle cell gene protects against malaria?

4. What is the rule for calculating the expected counts under the null hypothesis?

5. The *p*-value of the chi-square test statistic turned out to be approximately 0. What can you conclude based on this *p*-value?

UNIT ACTIVITY:
ASSOCIATIONS WITH COLOR

This activity is in three parts. In Part I, you will examine the reasoning behind the expected count formula. In Part II, you will need to collect data on eye color and gender from a sample of students. In Part III, there are different samples – different types of M&M candies. The candies are classified on one variable, color. In all three cases, you will conduct chi-square analyses.

Part I: Introduction – Assumption of Independence and Expected Count Formula

1. A survey given to 500 students asked: How would you describe your political preference? There were three response choices: GOP (Republican), DEM (Democrat), and IND (Independent). Keeping with the color theme of this activity, GOP is red (red states tend to vote Republican), DEM is the blue, and to make the color scheme patriotic, we'll let IND be represented by the color white. In addition to collecting information on political preference, the students indicated whether they were male or female. The results are given in Table 29.10.

	Count	Male	Female	Total
Political	DEM (Blue)	107	89	196
Preference	GOP (Red)	76	109	185
Color	IND (White)	63	56	119
	Total	246	254	500

Table 29.10. Distribution of political preference and gender.

We are interested in finding out whether there is an association between gender and political preference. We begin attacking this problem as a problem in probability. For example, to estimate the probability that a randomly selected student will be female and a Democrat (blue), we use the observed proportion 107/500. We can also calculate marginal probabilities using the row or column totals. For example, we estimate the probability that a student prefers the Democratic Party to be 196/500 and the probability that a randomly selected student is female as 246/500.

Using probability, we can examine what it would mean for the variables gender and political preference to be independent (or to have no association). If gender and political preference are independent, then we can use the Multiplication Rule to calculate this probability: *P*(political preference = DEM and gender = female).

a. Assume the variables gender and political preference are independent. Use the Multiplication Rule to estimate *P*(DEM and female) from the marginal probabilities. Show your calculations. (Give your answer to at least 4 decimals.)

b. Use your probability in (a) to determine the number of students out of the 500 observed that you would expect to fall into the category of being female and preferring the Democratic Party.

c. In a test of the null hypothesis H_0 : no association between the variables , the formula for calculating the expected count is

$$\text{expected count} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

For the cell corresponding to female and DEM, determine the expected count from the formula above. Compare your result with your answer to (b).

d. Repeat (a) - (c) for the cell corresponding to DEM and Male.

2. a. Assuming that the null hypothesis in 1(c) is true, calculate the expected counts for each cell in Table 29.10.

b. Calculate the value of the chi-square test statistic and the degrees of freedom. Then determine the *p*-value.

c. What can you conclude from your results in (b)?

Part II: Single Sample, Classified on Two Categorical Variables

One way to gather data that is appropriate for chi-square analysis is to select a single sample and then to classify the subjects in that sample by two categorical variables.

You will need a sample of students (your class, combined classes, friends). The two variables that you will use to classify the students in your sample are gender and eye color. The null hypothesis is:

H_0 : No association between gender and eye color.

or equivalently:

H_0 : The variables gender and eye color are independent.

3. a. Collect the data from your sample of students. Enter it into a copy of Table 29.11.

Count		Eye Color			Total
		Blue	Brown	Other	
Gender	Male				
	Female				
Total					

Table 29.11. Data on gender and eye color.

b. State the null and alternative hypotheses.

c. Calculate the expected cell counts and enter them into your table.

d. Perform a chi-square test. Report the value of the test statistic, the *p*-value, and your conclusion.

Part III: Multiple Samples, Classified on One Categorical Variable

Another data structure that is appropriate for chi-square analysis is when samples are drawn from different populations and classified on one categorical variable. In this case, we can think of "which sample" as the second variable. Next, your samples will be from different types of M&M candies. Given bags of at least two types of M&Ms, you will classify the M&Ms into colors, taking care to record which type of M&Ms candies you are classifying.

The null hypothesis is:

H_0 : No association between M&M type and color.

or equivalently:

H_0 : The color distributions are the same for the different M&M types.

4. a. Collect the color distribution from bags of up to four types of M&Ms. Then enter your data into a table similar to the one in Table 29.12. (Be sure to record the type.)

Count		Type 1 Regular	Type 2	Type 3	Type 4	Total
Color	Green					
	Blue					
	Yellow					
	Orange					
	Red					
	Brown					
Total						

Table 29.12. Data on M&Ms type and color.

b. State the null and alternative hypotheses.

c. Calculate the expected cell counts and enter them into your table.

d. Perform a chi-square test. Report the value of the test statistic, the *p*-value, and your conclusion.

EXERCISES

The questions in these exercises all relate to data collected from the study *Monitoring the Future: A Continuing Study of American Youth* (MTF).

1. One of the questions on the MTF survey asked the following: About how many (if any) energy drinks do you drink PER DAY, on average? Figure 29.4 (see Page 12) shows Minitab results from testing to see if there is an association between the number of energy drinks students consumed each day and their growing-up environment. As noted in the Content Overview, Minitab computed the value of the chi-square test statistic but did not compute a *p*-value.

a. Explain all ways in which this analysis failed to meet the guidelines for using a chi-square test.

b. In order to continue the investigation into an association between energy drink consumption and growing-up environment, we decided to combine the last three categories (Three, Four, and Five or more) into a single category Three+. Make a copy of Table 29.13. Use the data from Figure 29.4 to fill in the observed values in the third row of the table. Then find the row total and enter that into your table.

			Environment			Total
		Count	Farm	Country	City	
Energy Drinks	None	Observed	57	144	598	799
		Expected	52.55	150.44	596.01	
	One	Observed	11	44	160	215
		Expected	14.14	40.48	160.38	
	Two	Observed	4	13	36	53
		Expected	3.49	9.98	39.54	
	Three +	Observed				
		Expected				
Total			73	209	828	1110

Table 29.13. Two-way table of energy drinks and growing-up environment.

c. Use the row and column totals to calculate the expected counts for the third row. Enter the expected counts into your table. Do the expected counts in your completed table meet the guidelines for using the chi-square test?

d. Calculate the value of the chi-square test statistic. How many degrees of freedom are associated with this statistic?

e. Determine the *p*-value and state your conclusion.

2. Table 29.14 revisits data from Unit 12's exercises, which also dealt with responses to the MTF survey. Table 29.14 organizes data on gender and responses to the following question: How intelligent do you think you are compared with others your age?

		Intelligence			Total
Count		Below Average	Average	Above Average	
Gender	Female	437	2243	4072	
	Male	456	1643	4593	
Total					

Table 29.14. Results from questions on gender and intelligence rating.

a. We would like to test whether there is a statistical difference between how males and females rate their intelligence compared to their peers. In this context, which is the explanatory variable and which is the response variable? Explain.

b. State an appropriate null hypothesis and alternative hypothesis.

c. Make a copy of Table 29.14. Calculate the row totals and column totals and enter them into your table. Then calculate the expected counts for each cell and enter the expected counts into your table.

d. Calculate the chi-square test statistic. What are the degrees of freedom associated with the chi-square test statistic?

e. Calculate the *p*-value and state your conclusion.

3. We would expect that there is an association between how students rated their intelligence and their academic success. Table 29.15 organizes students responses rating their intelligence compared to their peers and their average grade in high school.

		Average Grade		
Count		A	B	C or Below
Intelligence	Above	2886	4044	1387
	Average	1335	1881	585
	Below	305	416	164

Table 29.15. Results from question on intelligence and average grade.

a. State the null and alternative hypotheses.

b. Calculate the expected counts and record them in a table.

c. Calculate the chi-square test statistic. State the degrees of freedom. Determine the *p*-value.

d. If the null hypothesis is true, how likely would it be to observe a value from the chi-square distribution in (c) at least as large as the value of the chi-square test statistic that you calculated in (c)? Does this provide strong evidence against the null hypothesis? Explain.

4. Another question on the MTF survey asked the following: On average over the school year, how many hours per week do you work in a paid or unpaid job? The survey results, classified into a two-way table, are shown in Figure 29.5. In addition, the Minitab output contains the conditional distributions of hours worked per week for each gender (row percentages). And finally, of particular interest is whether or not there is a statistical difference in work patterns between male and female 12th-grade students. The expected counts, under the hypothesis that there is no association between gender and work patterns, also appear in Figure 29.5. (See key at bottom of output for the order of appearance.)

```
Rows: Gender    Columns: Hours Worked/Week

                10 or            21 or
           None  fewer  11 to 20   more   Total

Female     2711   1439      1405   1124    6679
          40.59  21.55     21.04  16.83  100.00
           2808   1416      1350   1106    6679

Male       2846   1364      1266   1064    6540
          43.52  20.86     19.36  16.27  100.00
           2749   1387      1321   1082    6540

Total      5557   2803      2671   2188   13219
          42.04  21.20     20.21  16.55  100.00
           5557   2803      2671   2188   13219

Cell Contents:      Count
                    % of Row
                    Expected count

Pearson Chi-Square = 12.705, DF = 3, P-Value = 0.005
```

Figure 29.5. Minitab chi-square analysis for gender and weekly work hours.

a. State appropriate null and alternative hypotheses for this situation.

b. Report the outcome of the chi-square test and state your conclusion.

c. A chi-square test tells you whether or not there is an association between the two variables but it doesn't tell you anything about the nature of that association. Based on the row percentages, describe the nature of the association between gender and hours worked per week or describe evidence for the lack of such an association.

REVIEW QUESTIONS

1. The video for Unit 15, Designing Experiments, focused on an observational study of coral reefs. Moray eels are an important component of coral reef fish communities. Researchers Robert Young and Howard Winn conducted an observational study of moray eel behavior in the Belize Barrier Reef. They focused on two common species, the spotted moray and the purplemouth moray. For each eel they observed, they identified its species and classified the locations of the sightings into three categories, G for grass bed, S for sand or rubble, and B for within one meter of the border between grass and sand/rubble. The results are presented in Table 29.16.

	Count	Spotted	Purplemouth
Habitat Use	G	127	116
	S	99	67
	B	264	161

Table 29.16. Habitat types for two species of moral eels.
Source: Robert F. Young, Howard E. Winn, and W. L. Montgomery. Activity Patterns, Diet, and Shelter Site Use for Two Species of Moray Eels, Gymnothorax moringa and Gymnothorax vicinus, in Belize. Copeia: February 2003, Vol. 2003, No. 1, pp. 44-55.

a. Set up the hypotheses to test whether there is a relationship between eel species and habitat use.

b. Create a table of expected cell counts.

c. Calculate the chi-square test statistic. Show your calculations. Report the degrees of freedom, and the *p*-value. At the 0.05 level of significance, is the habitat use independent of the species of moray eel?

d. To examine the nature of any association between the two variables, habitat use and moray eel species, calculate either row or column percentages, whichever is more appropriate to the situation under study. Justify your choice of type of percentage. What do your percentages reveal about moray eels?

2. A random sample of registered voters was asked about their educational background and whether or not they voted in the November 2012 elections. Table 29.17 contains the results of the survey.

		Voted Nov. 2012	
		Yes	No
Highest Educational Attainment	Not HS Graduate	57	64
	HS Graduate/No College	227	163
	Some College or Associate's Degree	303	51
	Bachelor's Degree or Higher	303	51

Table 29.17. Survey results on voting and highest educational attainment.

a. In this situation, which is the explanatory variable and which is the response variable? Justify your answer.

b. Set up the hypotheses for testing whether educational attainment and voting in the 2012 presidential election are independent.

c. Calculate the expected counts for each cell.

d. Calculate the chi-square test statistic, state the degrees of freedom, and determine the *p*-value. Are the results significant?

e. Make a bar chart that displays how voting patterns are related to highest educational attainment. (Your choice of which variable is the explanatory variable should be evident in your display.) Label the bars with the corresponding percentages. Describe the nature of the relationship between the two variables.

3. Some tired, stressed-out students have turned to 2-ounce energy drink shots such as 5-Hour Energy to give them the energy boost they feel they need to make it through the day (or night). Compared to energy drinks that can run about 100 calories per 8-ounce serving, energy shots are sugar free and are around 4 calories per shot.

Because of the low calorie count, would female students be apt to drink more energy shots on a daily basis than male students? To find out, researchers asked a group of 12th-grade students the following question: How many (if any) energy drink shots do you drink PER DAY, on average? Table 29.18 gives the results from a survey given to a sample of 12th-grade students.

Count		Female	Male
Energy Shots Consumed Per Day	None	896	938
	Less than one	63	70
	One	16	19
	Two	5	16
	Three	7	5
	Four	1	0
	Five or Six	4	4
	Seven or more	4	7

Table 29.18. Student responses to question on energy drink shots.

a. The researchers wanted to see if there was an association between the daily number of energy drink shots consumed and gender. Calculate the expected cell counts for each cell.

b. Based on your answer to (a) do the expected counts satisfy the guidelines for using a chi-square test? Explain.

c. Combine some of the categories for the amount of energy shots consumed per day. Compute the expected counts and check to see if the guidelines for using the chi-square test are satisfied. If not, combine some additional categories until the guidelines are satisfied. (There are different choices for how the categories can be combined.)

d. Perform a chi-square test on your data from (c). What is the value of the chi-square test statistic? Report its *p*-value. What conclusions could the researchers draw from your results?