Unit 29: Inference for Two-Way Tables

PREREQUISITES

Unit 13, Two-Way Tables is a prerequisite for this unit. In addition, students need some background in significance tests, which was introduced in Unit 25.

ADDITIONAL TOPIC COVERAGE

Additional coverage of inference for two-way tables can be found in *The Basic Practice of Statistics*, Chapter 23, Two Categorical Variables: The Chi-Square Test.

ACTIVITY DESCRIPTION

Students should work in small groups on this activity. The activity consists of three parts. The first part provides a justification for the formula for computing the expected cell counts for chi-square tables. Students can work on Part I on their own or it could be part of a lecture/class discussion. Parts II and III involve two different structures for datasets, both of which are appropriate for the chi-square analysis covered in this unit.

Here are the two data structures: (1) subjects from a single sample are classified according to two categorical variables and (2) subjects from multiple samples (drawn from different populations) are classified according to a single categorical variable. In the latter case, “which sample” can be thought of as the second categorical variable. In the first case, a chi-square test for independence is performed; in the second case, a chi-square test for homogeneity is performed. The chi-square test statistics and the analyses are the same for both situations. So, in this unit, we have put little emphasis on distinguishing between these two situations.

MATERIALS

For Part III, bags of at least two different types of M&Ms are needed. Large-sized bags were used for the sample data, with the exception of the M&Ms minis, for which a medium bag was purchased. In addition, students will need paper plates or bowls to contain the M&Ms while they are being counted.

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

Part I provides an explanation of the expected counts formula used in a chi-square test of independence. Students need to be familiar with the Multiplication Rule from Unit 19, Probability Models. This part could be approached either as an activity or as part of an informal lecture that introduces the topic of this activity. It could also be skipped and students could move directly to Part II.

Part II: Single Sample, Classified on Two Categorical Variables

For this part, students will need to collect data from people. The class could serve as the sample, or perhaps combine this class with another class, or have students add their friends to the sample. Students will need to classify each individual in the sample by gender and eye color. An easy way to collect the data is to draw a table on the board. Each student should come up to the board and put a tally line in the appropriate box for gender and eye color. After students have completed their entries, numbers can replace the tally marks. Students can then copy the table from the board and begin work on Part II.

Part III: Multiple Samples, Classified on One Categorical Variable

Students should work in groups to collect the data on the M&Ms colors. Again, you may want to put a chart on the board and have students enter their results for each color as they finish sorting their M&Ms into colors. Once the data are collected, groups will need a copy of the class data. Since the resulting two-way table is quite large, group members should be encouraged to divide up the work of computing the expected cell counts.

The color distribution of M&Ms differs by types and has changed over the years. You can write to Mars, the makers of M&Ms, for the latest color distribution in its candies.

THE VIDEO SOLUTIONS

1. Dr. Pardis Sabeti investigates the nonstop evolutionary arms race between our bodies and the infectious microorganisms that invade and inhabit them. In other words, she investigates connections between genotypes and protections from infectious diseases. Her work on Lassa fever is still in its early stages.

2. Sickle cell anemia hemoglobin mutation, HbS.

3. H_0 : No association betweeen malaria and HbS.
   H_a : Association between malaria and HbS.

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

5. We reject the null hypothesis and conclude that there is an association between the HbS gene and malaria.

UNIT ACTIVITY:
ASSOCIATIONS WITH COLOR SOLUTIONS

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

1. a. $P(\text{DEM and female}) = P(\text{DEM})\,P(\text{female}) = \left(\frac{196}{500}\right)\left(\frac{246}{500}\right) \approx 0.1929$

b. Expected number = $\left(\frac{196}{500}\right)\left(\frac{246}{500}\right)(500) = \frac{(196)(246)}{500} \approx 96.432$

c. Expected count = $\frac{(196)(246)}{500} \approx 96.432$

d. $P(\text{DEM and male}) = P(\text{DEM})P(\text{male}) = \left(\frac{196}{500}\right)\left(\frac{254}{500}\right) \approx 0.1991$

Expected number = $\left(\frac{196}{500}\right)\left(\frac{254}{500}\right)(500) = \frac{(196)(254)}{500} \approx 99.57$

Expected count = $\frac{(196)(254)}{500} \approx 99.57$

2. a.

	Expected	Male	Female	Total
Political	DEM (Blue)	96.43	99.57	196
Preference	GOP (Red)	91.02	93.98	185
Color	IND (White)	58.55	60.45	119
	Total	246	254	500

b. $\chi^2 = + \frac{(107 - 96.43)^2}{96.43} + \frac{(89 - 99.57)^2}{99.57} + \ . \ . \ . \ + \frac{(56 - 60.45)^2}{60.45} \approx 7.825$

$df = (3 - 1)(2 - 1) = 2$; $p \approx 0.02$

c. There is sufficient evidence to reject the null hypothesis. There is association between these two variables. In other words, they are dependent.

3. a. Sample data will be used to provide sample answers.

	Count	Eye Color			Total
		Blue	Brown	Other	
Gender	Male	8	20	6	34
	Female	4	16	12	32
Total		12	36	18	66

b. H_0 : No association between gender and eye color.
H_a : Association between gender and eye color.

c. Sample answer:

	Count	Eye Color			Total
		Blue	Brown	Other	
Gender	Male	8 6.18	20 18.55	6 9.27	34
	Female	4 5.82	16 17.45	12 8.73	32
Total		12	36	18	66

d. Sample answer:

$$\chi^2 = \frac{(8-6.18)^2}{6.18} + \frac{(20-18.55)^2}{18.55} + \ . \ . \ . \ + \frac{(12-8.73)^2}{8.73} \approx 3.72 \ \ ; \ df = 2$$

$p \approx 0.151$. There is insufficient evidence to reject the null hypothesis. In other words, there is no strong evidence to suggest that there is an association between eye color and gender.

4. a. Sample data (will be used for sample answers) *(See next page...)*:

	Count	Type 1 Dark	Type 2 Regular	Type 3 Peanut	Type 4 Mini	Total
Color	Green	112	109	41	228	490
	Blue	188	160	39	203	590
	Yellow	75	91	47	210	423
	Orange	141	123	36	187	487
	Red	81	62	20	221	384
	Brown	59	84	30	100	273
	Total	656	629	213	1149	2647

b. H_0 : No association between M&M type and color distribution.
H_a : Association between M&M type and color distribution.

c. Sample answer:

	Count	Type 1 Dark	Type 2 Regular	Type 3 Peanut	Type 4 Mini	Total
Color	Green	112 121.4	109 116.4	41 39.4	228 212.7	490
	Blue	188 146.2	160 140.2	39 47.5	203 256.1	590
	Yellow	75 104.8	91 100.5	47 34	210 183.6	423
	Orange	141 120.7	123 115.7	36 39.2	187 211.4	487
	Red	81 95.2	62 91.2	20 30.9	221 166.7	384
	Brown	59 67.7	84 64.9	30 22	100 118.5	273
	Total	656	629	213	1149	2647

d. $\chi^2 \approx 100.3$; df = (6 – 1)(4 – 1) = 15; $p \approx 0$

There is an association between M&Ms type and color distribution. In other words, Different types of M&Ms have different color distributions.

EXERCISE SOLUTIONS

1. a. There were two cells with expected counts less than 1. The guidelines call for all expected counts to be greater than 1. Also, there were 7 cells with expected counts below 5. That means that around 39% of the cells have expected counts under 5. The guidelines state that no more than 20% of the cells should have expected counts less than 5.

b. See solution to (c).

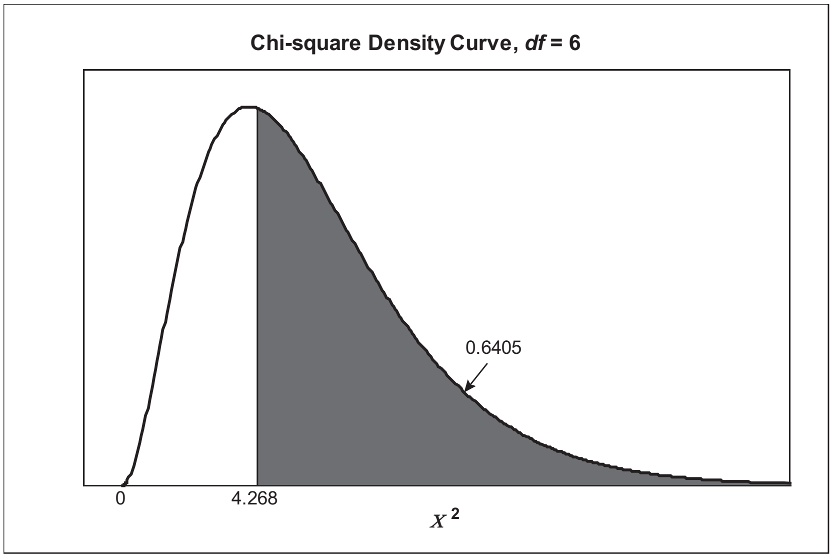
c. Based on the completed table below, all expected counts were greater than 1. Two expected counts were below 5, which is just under 17% of the cells. So, the expected counts in the table below meet the guidelines.

			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	1	8	34	43
		Expected	2.83	8.10	32.08	
Total			73	209	828	1110

d. This is a 4×3 table; *df* = (4 – 1)(3 – 1) = 6. The chi-square test statistic is calculated below:

$$\begin{aligned}\chi^2 &= \frac{(57-52.55)^2}{52.55} + \frac{(144-150.44)^2}{150.44} + \frac{(598-596.01)^2}{596.01} \\ &+ \frac{(11-14.14)^2}{14.14} + \frac{(44-40.48)^2}{40.48} + \frac{(160-160.38)^2}{160.38} \\ &+ \frac{(4-3.49)^2}{3.49} + \frac{(13-9.98)^2}{9.98} + \frac{(36-39.54)^2}{39.54} \\ &+ \frac{(1-2.83)^2}{2.83} + \frac{(8-8.10)^2}{8.10} + \frac{(34-32.08)^2}{32.08} \\ &\approx 4.268\end{aligned}$$

e. $p \approx 0.64$. (See area under density curve below.) There is insufficient evidence to reject the null hypothesis. We found no clear evidence of an association between 12th-grade students' consumption of energy drinks and their growing-up environment.



2. a. Gender is the explanatory variable. We would like to use gender to explain how students' rate their intelligence compared to their peers.

b. H_0 : No association between gender and intelligence rating.
H_a : Association between gender and intelligence rating.

c.

		Intelligence			Total
	Count	Below Average	Average	Above Average	
Gender	Female	437 448.5	2243 1951.7	4072 4351.8	6752
	Male	456 444.5	1643 1934.3	4593 4343.2	6692
Total		893	3886	8665	13444

d. *df* = (2 – 1)(3 – 1) = 2

$$\chi^2 = \frac{(437-448.5)^2}{448.5} + \frac{(2243-1951.7)^2}{1951.7} + \frac{(4072-4351.8)^2}{4351.8} + \frac{(456-444.5)^2}{444.5} + \frac{(1643-1934.3)^2}{1934.3} + \frac{(4593-4313.2)^2}{4313.2} \approx 124.1$$

(Answers may vary somewhat depending on the number of decimals used in the expected cell count.)

e. $p \approx 0$. Reject the null hypothesis. There is a statistically significant difference between how males and females rate their intelligence compared to their peers. (In other words, there is an association between gender and intelligence rating.)

3. a. H_0 : No association between intelligence rating and average grades.

H_a : Association between intelligence rating and average grades.

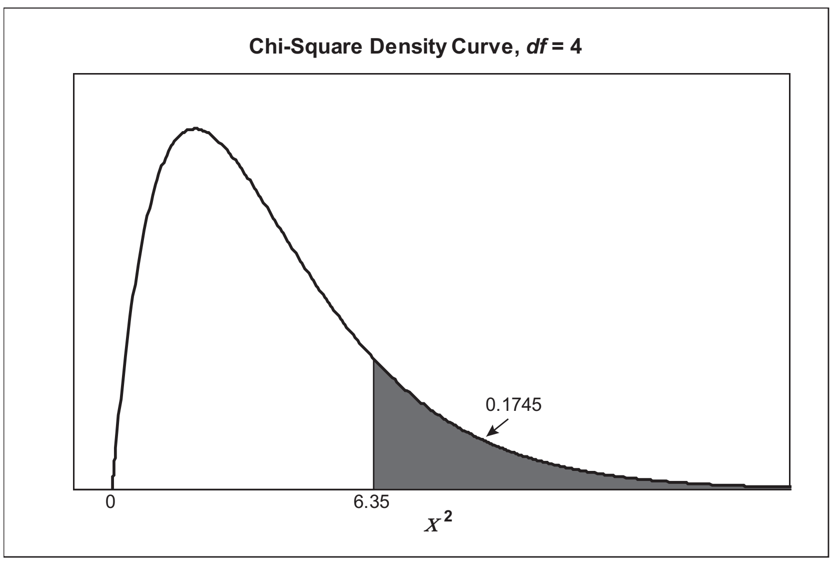
b.

	Count	A	B	C or Below	Total
		Average Grade			
Intelligence	Above	2886 2894.9	4044 4055.8	1387 1366.2	8317
	Average	1335 1323	1881 1853.6	585 624.4	3801
	Below	305 308	416 431.6	164 145.4	885
Total		4526	6341	2136	13003

c. $df = (3 - 1)(3 - 1) = 4$

$$\chi^2 = \frac{(2886 - 2894.9)^2}{2894.9} + \ . \ . \ . \ + \frac{(164 - 145.4)^2}{145.4} \approx 6.35$$

As shown below, $p \approx 0.174$



d. We would expect to see a value from a chi-square distribution with $df = 4$ as or more extreme than 6.35 roughly 17.4% of the time. So, this is a somewhat common occurrence. It does not provide strong evidence against the null hypothesis. Generally strong evidence means that the percentage should be below 5%.

4. a. H_0 : No association between gender and hours worked/week.

H_a : Association between gender and hours worked/week.

b. $\chi^2 = 12.705$; $p = 0.005 < 0.05$. Therefore, the results are significant. There is an association between gender and hours worked per week. (Note: The practical significance is another matter and cannot be determined by a *p*-value.)

c. The biggest discrepancy in work patterns is that a higher percentage of males did not work (43.52%) compared to females (40.59%). Furthermore, in every category of hours worked/ week, there is a higher percentage of females than males.

REVIEW QUESTIONS SOLUTIONS

1. a. H_0 : No association between habitat use and eel species.
H_a : Association between habitat use and eel species.

b.

	Count	Spotted	Purplemouth	Total
Habitat Use	G	127 142.8	116 100.2	243
	S	99 97.5	67 68.5	166
	B	264 249.7	161 175.3	425
Total		490	344	834

c. Here are the calculations for the chi-square test statistic:

$$\chi^2 = \frac{(127-142.8)^2}{142.8} + \frac{(116-100.2)^2}{100.2} + \frac{(99-97.5)^2}{97.5} + \frac{(67-68.5)^2}{68.5} + \frac{(264-249.7)^2}{249.7} + \frac{(161-175.3)^2}{175.3} \approx 6.28$$

The degrees of freedom are: $df = (3 – 1)(2 – 1) = 2$.

Using software, $p \approx 0.043$.

Since $p < 0.05$, we reject the null hypothesis and conclude that there is an association between habitat use and moray eel species.

d. Column percentages are more appropriate. The explanatory variable is the eel species. So, we should compare the conditional distributions of habitat use for each species of moray eel.

		Spotted	Purplemouth
Habitat Use	G	25.9%	33.7%
	S	20.2%	19.5%
	B	53.9%	40.8%
Total		100%	100%

We learn that a majority (53.9%) of the spotted moray eels were found in border habitats compared to only 46.8% of the purplemouth moray eels.

2. a. Educational attainment is the explanatory variable and voting is the response variable. We expect that a person's highest educational attainment will shed light on whether or not they voted in the 2012 elections.

b. H_0 : No association between education and voting.
H_a : Association between education and voting.

c.

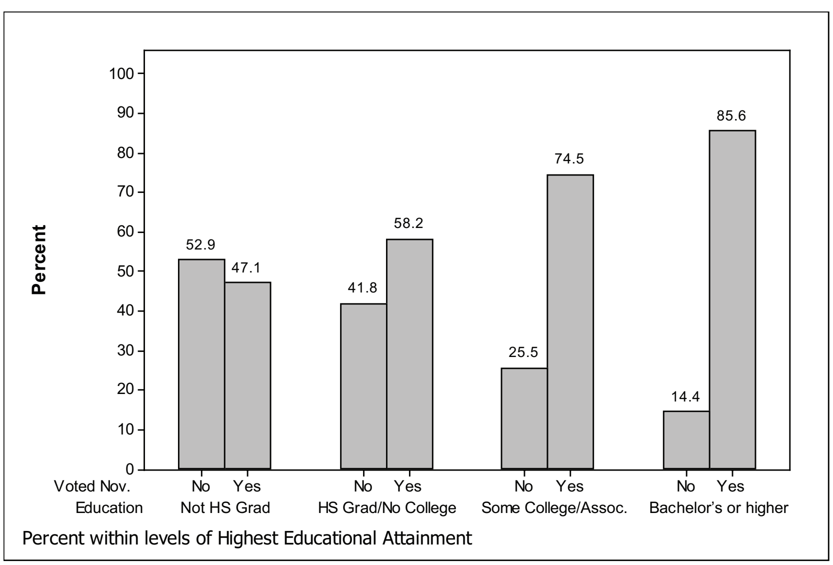
			Voted Nov. 2012		Total
		Count	Yes	No	
Highest Educational Attainment	Not HS Grad		57	64	121
		Expected	84.5	36.5	
	HS Grad/No College		227	163	390
		Expected	272.3	117.7	
	Some College/Associate's		303	51	364
		Expected	254.1	109.9	
	Bachelor's or Higher		303	51	354
		Expected	247.1	106.9	
	Total		858	371	1229

d. $$\chi^2 = \frac{(57-84.5)^2}{84.5} + \frac{(64-36.5)^2}{36.5} + \ . \ . \ . \ + \frac{(51-106.9)^2}{106.9} \approx 100.1$$

$df = (4 - 1)(2 - 1) = 3$; $p \approx 0.000$

Since $p < 0.5$, the results are significant. There is a relationship between these two variables.

e. Since the explanatory variable is highest educational attainment, the chart below represents graphically the conditional distributions of voting for each level of highest educational attainment.



As the level of highest educational attainment increases, so does the participation in voting. More educated people are more likely to vote than those who are not educated.

3. a.

		Female	Male	Total
	Count			
Energy Shots Consumed Per Day	None	896 888.89	938 945.11	1834
	Less than one	63 64.46	70 68.54	133
	One	16 16.96	19 18.04	35
	Two	5 10.18	16 10.82	21
	Three	7 5.82	5 6.18	12
	Four	1 0.48	0 0.52	1
	Five or Six	4 3.88	4 4.12	8
	Seven or more	4 5.33	7 5.67	11
Total		996	1059	2055

b. No, the guidelines are not satisfied. There are two cells that have counts below 1 (0.48 and 0.52). In addition, there are 4 cells with counts less than 5, which is 25% of the cells.

c. Sample answer (students may decide to combine different categories):

		Female	Male	Total
	Count			
Energy Shots Consumed Per Day	None	896 888.9	938 945.1	1834
	One or Less	79 81.4	89 86.6	168
	Two or Three	12 16	21 17	33
	Four or more	9 9.7	11 10.3	20
Total		996	1059	2055

d. Sample answer is based on sample answer to (c): $\chi^2 = 2.282$; $p \approx 0.52$.

There is insufficient evidence to reject the null hypothesis. There is insufficient evidence to indicate that there is a linkage between amounts of energy drink shots consumed and gender.