Unit 13: Two-Way Tables

SUMMARY OF VIDEO

This video deals with analysis of categorical variables (for example, gender, race, occupation) and relationships between categorical variables. The context is a Happiness Survey that was part of Somerville, Massachusetts' 2011 annual census. The video focuses on two of the survey questions, one that asks respondents to rate their current level of happiness and the other that asks them to rate the beauty of Somerville. Happiness ratings are boiled down into three categories: Unhappy, So-So, and Happy. Ratings of Somerville's physical beauty are categorized as Bad, OK, and Good. Results from these two questions are organized into a two-way table with Happiness as the row variable and Physical Beauty as the column variable (see Table 13.1). The marginal totals (bottom row and right-most column) have been added to the two-way table.

		Physical Beauty			
		Bad	OK	Good	Total
Happiness	Unhappy	90	123	62	275
	So-so	555	972	610	2137
	Happy	541	1426	1406	3373
Total		1186	2521	2078	5785

Table 13.1. Results from rating happiness and Somerville's physical beauty.

Notice that 5785 Somerville residents answered both of these questions. (The table only accounts for respondents who have answered both questions.) First, look at the distribution of each variable separately – this is called a marginal distribution. Computations of the marginal distributions of the two variables appear in Tables 13.2 and 13.3. From the marginal distributions we find that slightly more than 58% of respondents reported they were Happy and around 36% of the respondents rated Somerville's physical beauty as Good.

See tables on next page...

		Marginal Distribution
Happiness	Unhappy	275/5785 × 100% ≈ 4.75%
	So-so	2137/5785 × 100% ≈ 36.94%
	Happy	3373/5785 × 100% ≈ 58.31%

Table 13.2. Marginal distribution of Happiness.

	Physical Beauty		
	Bad	OK	Good
Marginal Distribution	1186/5785 × 100% ≈ 20.50%	2521/5785 × 100% ≈ 43.58%	2078/5785 × 100% ≈ 35.92%

Table 13.3. Marginal distribution of Physical Beauty.

Next, we dig even deeper into the two-way table's data by computing conditional distributions, distributions of one variable restricted to a single outcome of another variable. For example, we can investigate how just the Unhappy people rated Somerville's beauty. In this case, we are looking at the distribution of beauty ratings just within the Unhappy group (275 respondents). Here are the calculations:

Bad: 90/275 × 100% ≈ 32.73%
OK: 123/275 × 100% ≈ 44.73%
Good: 62/275 × 100% ≈ 22.55%

Table 13.4 shows the conditional distribution of Physical Beauty for each category of Happiness.

		Physical Beauty			Total
		Bad	OK	Good	
Happiness	Unhappy	32.73%	44.73%	22.55%	100%
	So-so	25.97%	45.48%	28.54%	100%
	Happy	16.04%	42.28%	41.68%	100%

Table 13.4. Conditional distribution of Physical Beauty for each Happiness category.

Notice that only 22.55% of Unhappy people rated Somerville's beauty as Good compared to 41.68% of the Happy people – clearly there is a connection between the Happiness and Physical Beauty variables. The graphic display in Figure 13.1 can help us visualize this linkage.

Figure 13.1. Conditional distribution of Physical Beauty for each level of Happiness.

The bar graph in Figure 13.1 shows that as the level of Happiness goes up, the percentage of Bad ratings for Physical Beauty goes down. In addition, as the level of Happiness goes up, the level of Good beauty ratings also goes up. As we know, correlation isn't necessarily causation. However, now that Somerville has identified a link between residents' happiness levels and their thoughts on the city's physical beauty, officials want to dig deeper on the next survey in an effort to improve residents' satisfaction with Somerville.

STUDENT LEARNING OBJECTIVES

A. Organize a small data set on two categorical variables into a two-way table by hand. Use software to classify data from large data sets into two-way tables.

B. Calculate the marginal distributions for each of the variables in a two-way table of counts.

C. Given a two-way table of counts, calculate the joint distribution of the two variables.

D. Given a two-way table of counts, calculate the conditional distribution of one variable for each level of the other variable.

E. Draw a bar graph that represents the conditional distribution of one variable at each level of another variable.

F. Understand the difference between (1) the conditional distribution of X for each level of Y and (2) the conditional distribution of Y for each level of X.

G. Recognize which type of percentage -- marginal, joint, or conditional -- is appropriate to answer a particular question.

CONTENT OVERVIEW

This unit discusses methods for studying relationships between two categorical variables. Some categorical variables – such as gender, eye color, occupation – are inherently categorical. Others – such as age in the following categories: under 30, between 30 and 60, and over 60 – are created by grouping values of a quantitative variable into categories. **Nominal** categorical variables have values with no inherent order; **ordinal** categorical variables have values with an inherent order. One example of an ordinal variable would be college class: freshman, sophomore, junior, and senior. Any table or graphic display involving an ordinal variable should preserve the inherent order of values for that variable.

A relationship between two categorical variables requires that both variables must be responses from the same individuals or cases. The first step in extracting information about a relationship between the two variables is to organize the raw data into a two-way table. Table 13.5 shows data from the first 10 respondents to Somerville's Happiness Survey.

Survey ID	Happiness	Physical Beauty
1	Happy	Good
2	Happy	Good
3	So-so	OK
4	Happy	Bad
5	So-so	Good
6	Happy	Good
7	Unhappy	Bad
8	So-so	Good
9	So-so	Bad
10	So-so	OK

Table 13.5. Data on first 10 respondents to Happiness Survey.

For the two-way table, we'll use Happiness as the row variable and Physical Beauty as the column variable (just as was done in the video). Respondents #1 and #2 replied Happy and Good to the questions on rating personal happiness and Somerville's physical beauty, respectively. Hence, we have entered two tally marks into the corresponding cell of Table 13.6. Respondent #3 replied So-so and OK and we have entered a single tally mark into the corresponding cell of Table 13.6. Table 13.7 shows the results from the completed tally converted to numbers.

		Physical Beauty		
		Bad	OK	Good
Happiness	Unhappy		|	
	So-so			
	Happy			||

Table 13.6. Making a two-way table from the data in Table 13.5.

		Physical Beauty		
		Bad	OK	Good
Happiness	Unhappy	1	0	0
	So-so	1	2	2
	Happy	1	0	3

Table 13.7. Two-way table for data in Table 13.5.

Although it's good to practice making a two-way table by hand on a small data set, there were 5785 respondents to these two questions in the Somerville survey. Organizing large data sets into two-way tables is tedious to do by hand and best left to technology.

Once we have organized the data into a two-way table, we can compare different types of percentages. Next, we look at responses to a survey of 12th grade students. Table 13.8 organizes their responses to questions on gender and how many hours per week they work at either a paid or unpaid job. The row variable is Hours and the column variable is Gender. The row and column totals have been added to the table.

Count		Gender		
		Female	Male	Total
Hours	None	10	3	13
	10 or fewer hours	7	4	11
	11 to 20 hours	2	7	9
	21 to 30 hours	8	3	11
	More than 30 hours	2	4	6
Total		29	21	50

Table 13.8. Two-way table for Hours and Gender.

From the marginal totals, Table 13.8 shows 13 respondents who did not work and 21 respondents who were male. From the joint distribution, there were three respondents who fell into both of these categories, males who did not work.

Computing Distributions

Joint distribution percentages of the two variables: (cell entry)/(grand total) × 100%

Marginal distribution percentages for one variable: (Total entry)/(grand total) × 100%

Table 13.9 shows the joint distribution percentages of Hours and Gender (white cells inside table) along with the marginal distributions for Hours and Gender (right-most column and bottom row, respectively).

	Percent	Gender		
		Female	Male	Total
Hours	None	20	6	26
	10 or fewer hours	14	8	22
	11 to 20 hours	4	14	18
	21 to 30 hours	16	6	22
	More than 30 hours	4	8	12
	Total	58	42	100

Table 13.9. Joint and marginal distributions as percentages.

From the marginal distributions in Table 13.9, we observe that 26% of the students did not work and 42% of the respondents were male. Now, remember those three respondents who were males and did not work? From the joint distribution we find that they make up 6% of the respondents.

Conditional distributions provide the most insight into relationships between the two variables. For the 12th grade survey, we are interested in comparing the work patterns of males to females. So, we need to calculate the conditional distribution of Hours for each level of Gender. To do that we calculate column percentages as described in the box below.

Computing Column Percentages

(cell entry)/(column total) × 100%

Column percentages are conditional distributions of the row variable for each level of the column variable.

The column percentages for the Hours-Gender data appear in Table 13.10.

	Percent	Gender	
		Female	Male
Hours	None	34.48	14.29
	10 or fewer hours	24.14	19.05
	11 to 20 hours	6.9	33.33
	21 to 30 hours	27.59	14.29
	More than 30 hours	6.9	19.05
	Total	100	100

Table 13.10. Conditional distribution of Hours for each level of Gender.

Sometimes it is easier to take in information if it is presented graphically. The bar chart in Figure 13.2 is a graphical representation of the numbers in 13.10. The conditional distribution of Hours for females is represented by the first 5 bars on the left and the conditional distribution of Hours for males is represented by the last 5 bars on the right. One result that jumps out from looking at the bar chart is that the highest bar for females, associated with the response None (34.5%), is higher than the highest bar for males, associated with the response of working 11 to 20 hours per week (33.3%).

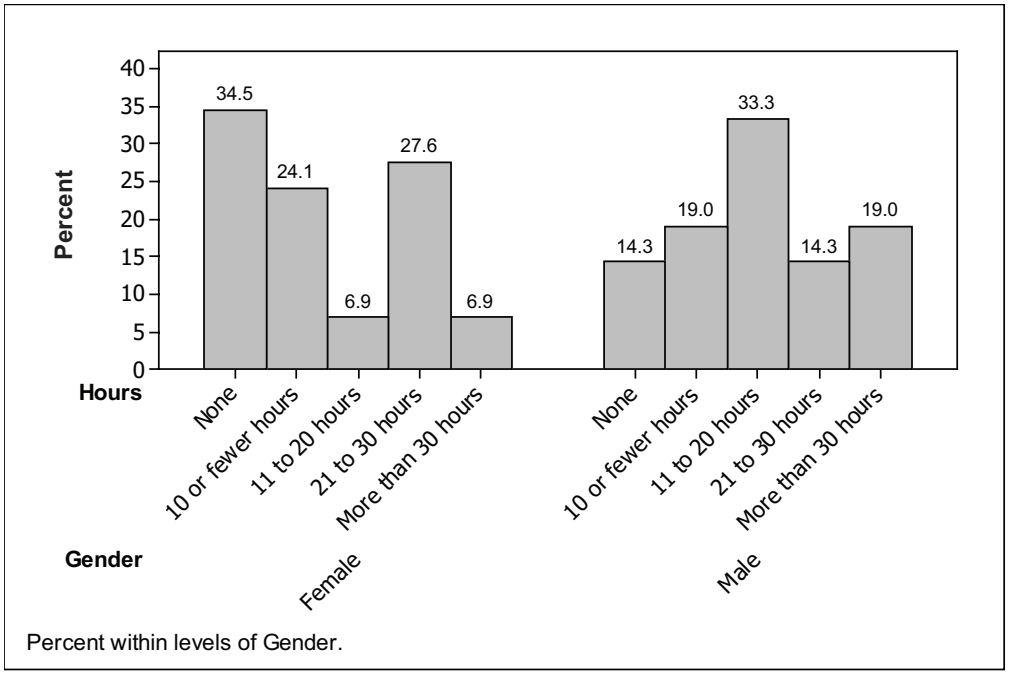

Figure 13.2. Bar chart of conditional distributions of Hours for each level of Gender.

Similarly, we can compute the conditional distribution of Gender for each level of Hours. Since there are five values for the variable Hours, there will be five conditional distributions, one for each row of the table. We calculate these percentages as follows.

Computing Row Percentages

(cell entry)/(row total) × 100%

Row percentages are conditional distributions of the column variable for each level of the row variable.

The results appear in Table 13.11.

		Gender		
	Percent	Female	Male	Total
Hours	None	76.92	23.08	100%
	10 or fewer hours	63.64	36.36	100%
	11 to 20 hours	22.22	77.78	100%
	21 to 30 hours	72.73	27.27	100%
	More than 30 hours	33.33	66.67	100%

Table 13.11. Conditional distribution of Gender for each level of Hours

From Table 13.11, we learn that nearly 77% of the student respondents who did not work were female and that nearly 67% of the students who worked more than 30 hours per week were male.

KEY TERMS

Categorical variables can be either **nominal**, values that have no inherent order, or **ordinal**, values that have an inherent order. The inherent order of the values of an ordinal categorical variable should be preserved in tables and charts involving that variable.

A **two-way table** of counts (or frequencies) organizes data about two categorical variables taken from the same individuals or subjects. Values of the **row variable** label the rows of the table; values of the **column variable** label the columns of the table. A two-way table in which the row variable has n values and the column variable has m values is called an $n \times m$ table.

The sum of the row entries or the sum of the column entries are called the **marginal totals**. **Marginal distributions** are computed by dividing the row or column totals by the overall total. Marginal distributions provide information about the individual variables but do not provide any information about the relationship between the two variables.

A two-way table of counts can be converted into a **joint distribution** by dividing each cell count by the grand total and multiplying by 100%.

There are two sets of **conditional distributions** for a two-way table:

- distributions of the row variable for each fixed level of the column variable
- distributions of the column variable for each fixed level of the row variable

Conditional distributions provide one way to explore the relationship between the row and column variables.

THE VIDEO

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

1. Give two (or more) examples of categorical variables.

2. What did Somerville include in its 2011 census that was unconventional?

3. In the two-way table used to organize the responses to rating personal happiness and Somerville's physical beauty, which variable was the row variable and which was the column variable? Explain.

4. As the level of happiness went up (from Unhappy to So-so to Happy), what happened to the percent of respondents who rated Somerville's physical beauty as Bad?

UNIT ACTIVITY:
HAPPINESS SURVEY

Complete the survey at the end of this activity (or a modified version that your instructor provides). After you have responded to the survey, your instructor will distribute the class data. Answer the following questions based on the class data.

1. Organize the data on rating physical beauty and happiness into a two-way table. Use Happiness for the row variable and Beauty for the column variable. Add the marginal totals to your table.

In the rest of this activity, round percentages to one decimal.

2. What percentage of your class responded Happy? Show the appropriate calculation.

3. What percentage of your class rated the Physical Beauty of your campus or school as Good? Show the appropriate calculation.

4. a. Create a table showing the conditional distribution of Physical Beauty for each level of Happiness.

b. Were Happy students or Unhappy students more likely to respond that the Physical Beauty of campus was good? Support your answer with appropriate percentages. Show how these percentages were calculated.

5. a. Make a bar chart that represents the conditional distributions of Happiness for each level of Physical Beauty. Use a percent scale for the vertical axis and label each bar with its corresponding percent.

b. Write a few sentences describing what can be learned from your bar chart in (a).

6. Write a brief report analyzing the data from the remaining survey question(s). Include analysis of relationships between responses to the remaining survey question(s) and the survey questions involving Physical Beauty and Happiness. Include two-way tables and at least one graphic display in the report.

HAPPINESS SURVEY

Circle your answers to the following questions:

What is your class year?

Fr So Jr Sr

Rate the physical beauty of your campus (or school):

Bad OK Good

Rate your level of happiness today:

Unhappy So-so Happy

EXERCISES

Each year the study *Monitoring the Future: A Continuing Study of American Youth* surveys students on a wide range of topics related to behaviors, attitudes, and values. These exercises are based on data collected from the 2011 survey of 12th grade students.

Table 13.12 organizes data on gender and responses to the following question:

> How intelligent do you think you are compared with others your age?

Responses to this question have been boiled down into three categories: Below Average, Average, and Above Average.

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

Table 13.12. Results from questions on gender and intelligence.

Refer to Table 13.12 for questions 1 and 2.

1. a. Copy Table 13.12. Add a row to the bottom and a column to the right-end of your table. Compute the marginal totals and enter them into your table.

b. What percentage of the students who answered both questions were male? Female? Show your calculations. (Round percentages to one decimal.)

c. What percentage of the students rated their intelligence as above average? What does this tell you about 12th grade students' assessment of their intelligence?

2. a. Compute conditional distributions of Intelligence for males and females. Record your results in a table. Show calculations. (Round percentages to one decimal.)

b. Represent the distributions in your table from (a) in a bar chart.

c. Write a brief description of how the male respondents rated their intelligence compared to female respondents.

Table 13.13 organizes data on gender and responses to the following question:

How would you describe your political preference?

Responses to this question have been categorized as Rep (Republican), Ind (Independent), Dem (Democrat), Oth (Other), and No Pref/Hvnt Decid (No preference or haven't decided).

		Political Preference				
		Rep	Ind	Dem	Oth	No Pref/ Hvnt Decid
Gender	Female	1275	723	1633	89	2917
	Male	1620	871	1332	183	2577

Table 13.13. Results from questions on gender and political preference.

Questions 3 and 4 refer to Table 13.13.

3. a. Create a table showing the joint distribution (percentage) of gender and political preference. Add a row to the bottom and a column to the right end of your table. Enter the marginal distributions for gender and political preference into the added row and column. (Round percentages to one decimal.)

b. Create a table showing the conditional percentages for Political Preference for each gender. (Round percentages to one decimal.)

c. Create a table showing the conditional percentages of Gender for each category of Political Preference. (Round percentages to one decimal.)

4. Use the tables you created in question 3 to answer (a) – (d).

a. What percent of the respondents were females and Democrats? What percent of the respondents were males who were Independents?

b. Were male students or female students more likely to respond they were Republicans? Include relevant percentages in your answer.

c. Were Republicans more likely to be male or female? Be sure to include relevant percentages in your answer. Explain how this question differs from (b).

d. Make a graphic display that represents the distribution of Political Preference for each gender. Compare the political preferences of the 12th grade male students to the 12th grade female students.

REVIEW QUESTIONS

The *Monitoring the Future Study* is a major source of information on smoking, drinking and drug habits of American youth. Based on data collected from the 2011 survey, you will decide whether or not smoking is linked to gender or if there is a linkage between high school grades and alcohol consumption. The review questions will focus on data collected from the following three survey questions:

I. On how many occasions (if any) have you had alcoholic beverages to drink – more than just a few sips during the last 30 days?

II. Have you ever smoked cigarettes?

III. Which of the following best describes your average grade so far in high school?

Question 1 explores the relationship between Smoking (Question II) and Gender. Results from these questions from students who answered both questions appear in Table 13.14. (You may notice that the overall total in this table differs from the overall total in Table 13.12. Some students chose not to respond to certain questions.)

		Smoking				
		Never	Once or twice	Occasionally/ not regularly	Regularly in past	Regularly now
Gender	Female	4244	1228	649	253	428
	Male	3957	1182	793	337	543

Table 13.14. Results from questions on gender and smoking.

1. a. Copy Table 13.14 and add a row to the bottom and a column to the right. Label the added row and column “Total” and enter the marginal totals.

b. Are male or female 12th grade students more likely to never have smoked? Support your answer with appropriate percentages. Show the calculations.

c. Create a bar chart representing the conditional distribution of Smoking for each gender. Label each bar with its corresponding percentage. (Round percentages to one decimal.)

Questions 2 – 4 explore relationships between grades in high school and smoking. Responses to survey questions II and III have been organized into the Table 13.15.

		Smoking				
		Never	Once or twice	Occasionally/ not regularly	Regularly in past	Regularly now
Grades	D	46	22	23	12	25
	C-, C, or C+	926	436	323	123	285
	B-, B, or B+	3792	1275	769	330	489
	A- or A	3465	665	327	126	168

Table 13.15. Responses to questions on grades and smoking.

2. a. How many students answered both the question on grades and the question on smoking?

b. What percentage of students answering both questions had never smoked? Had smoked at least once?

c. What percentage of students had average grades of A- or A and never smoked? Show the calculations. (Round answer to one decimal.)

3. a. What percentage of A- or A students have never smoked? Show the calculations.

b. What percentage of students' whose averages are C+ or below never smoked? Show the calculations.

4. Next, you will examine the relationship between Smoking status and having a B- average or better.

a. What percentage of students who are regular smokers had averages B- or better?

b. What percentage of students who were regular smokers in the past had averages B- or better?

c. What percentage of students who never smoked had averages B- or better?

5. The graphic display in Figure 13.3 represents the conditional distribution of alcohol usage for each level of Grade. The conditional percentages (rounded to the nearest percent) appear above each bar.

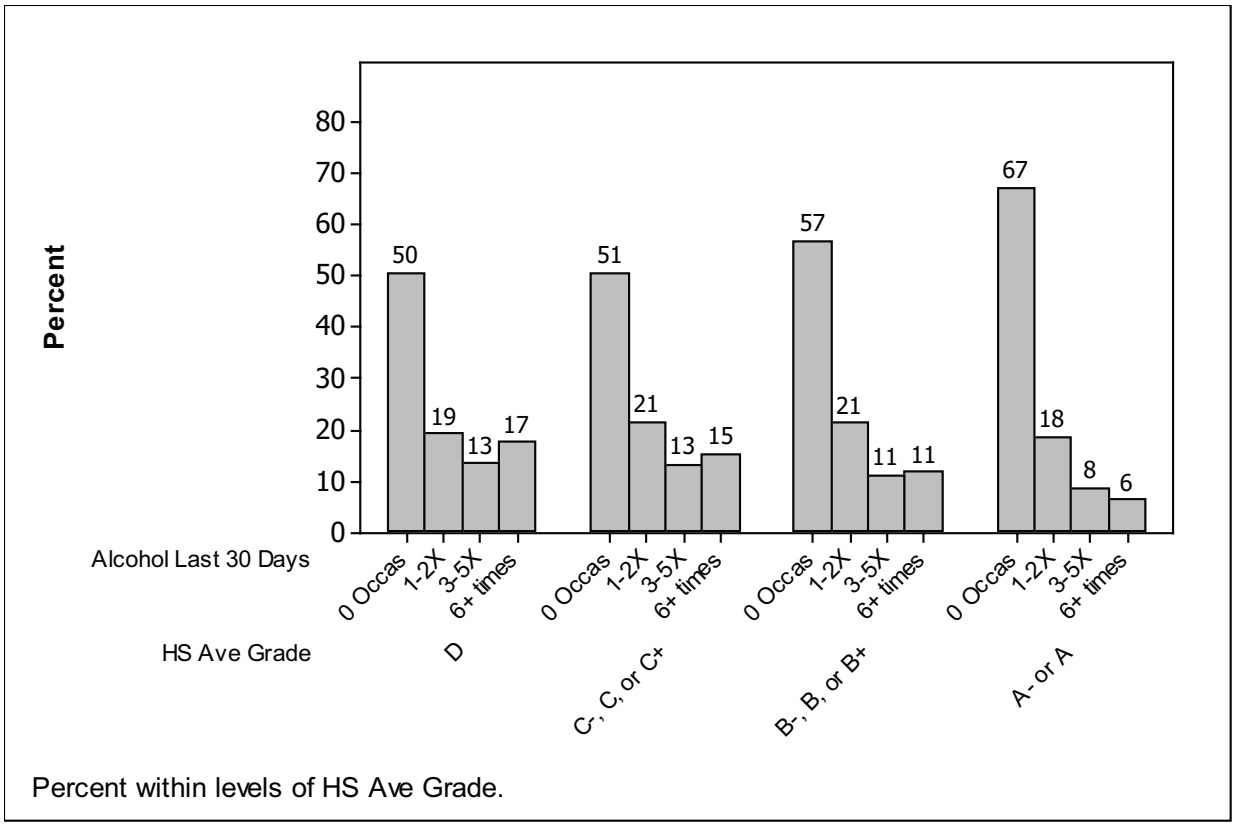

Figure 13.3. Conditional distributions of Alcohol for each level of GPA.

a. Do the conditional percentages for each level of Grades sum to 100%? If not, explain why they might not sum to 100%.

b. Write a brief description of alcohol usage among the different levels of Grades. What relationship, if any, can you find between alcohol usage and grades?