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%).

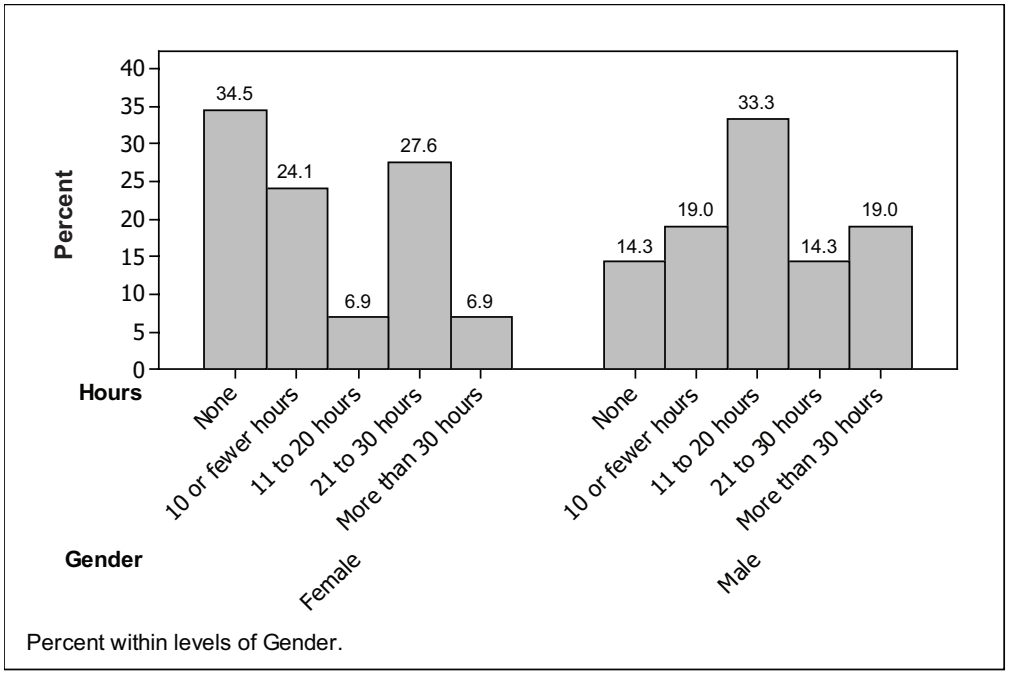

Figure 13.2. Bar chart of conditional distributions of Hours for each level of Gender.

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.