Chapter 6. Two-Way Tables

Note. This chapter deals with comparing two categorical variables.

Definition. A **two-way table** of counts organizes data about two categorical variables. Values of the **row variable** label the rows that run across the table, and values of the **column variable** label the columns that run down the table. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.

Marginal Distributions

Example S.6.1. Two-Way Stooges.
We now consider all 190 Three Stooges films and two categories. One category is "the role of third stooge" (Curly/Shemp/Joe) and the other is "number of slaps in the film" (which we break into intervals as $[0, 5]$, $[6, 10]$, $[11, 15]$, $[16, 20]$, $[21, 25]$, $[26, 30]$, $[31, 35]$, $[36, 40]$, and $[41, \infty)$). Notice that both of these are in fact categorical variables, even though

"number of slaps in the film" could be dealt with as a quantitative variable. The data can be put in a two-way table as follows.

	Curly	Shemp	Joe	TOTAL
0 to 5 slaps	20	9	5	34
6 to 10 slaps	29	25	5	59
11 to 15 slaps	20	16	2	38
16 to 20 slaps	16	5	3	24
21 to 25 slaps	3	7	1	11
26 to 30 slaps	4	7	0	11
31 to 35 slaps	1	2	0	3
36 to 40 slaps	2	0	0	2
more than 40 slaps	2	6	0	8
TOTAL	97	77	16	190

Note. In the table above, the distributions of "third stooge" and of "number of slaps" alone are called *marginal distributions* because they appear at the right and bottom margins of the two-way table. "Third Stooge" is the column variable and "number of slaps" is the row variable.

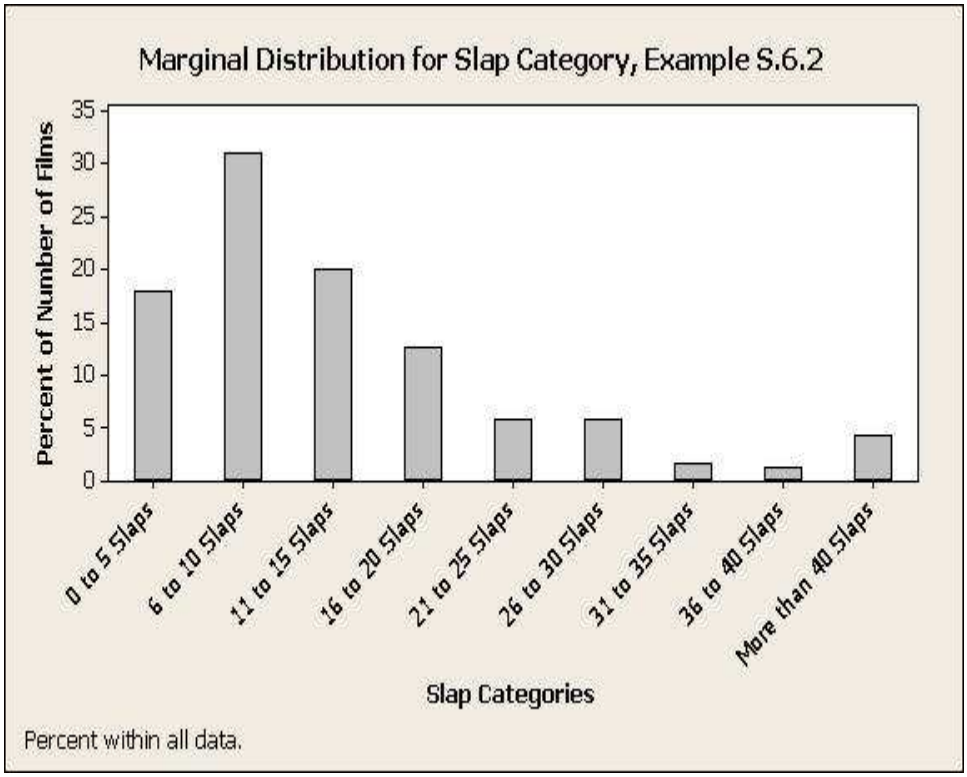
Conditional Distributions

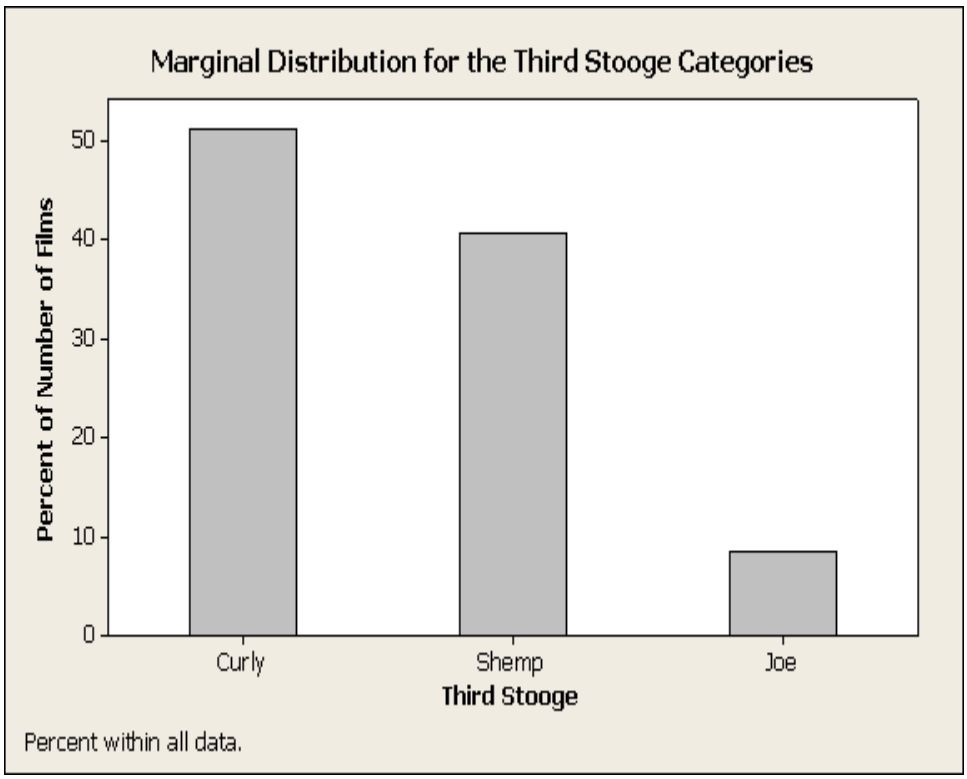
Note. We can describe relationships among categorical variables by calculating appropriate percents from the counts given.

Definition. The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table. A **conditional distribution** of a variable is the distribution of values of that variable among only individuals who have a given value of the other variable. There is a separate conditional distribution for each value of the other variable.

Example S.6.2. Marginal Stooges.
Give the marginal distribution (in terms of percentages) of each of the categorical variables in Example S.6.1. Give a conditional distribution for the Curly category (in terms of percentages).

Partial Solution. You will create the conditional distribution on Worksheet #3. Minitab gives the following bar charts for the marginal distributions:


Marginal Distribution for Slap Category, Example S.6.2
Percent of Number of Films
35
30
25
20
15
10
5
0
0 to 5 Slaps
6 to 10 Slaps
11 to 15 Slaps
16 to 20 Slaps
21 to 25 Slaps
26 to 30 Slaps
31 to 35 Slaps
36 to 40 Slaps
More than 40 Slaps
Slap Categories
Percent within all data.


Marginal Distribution for the Third Stooge Categories
Percent of Number of Films
50
40
30
20
10
0
Curly
Shemp
Joe
Third Stooge
Percent within all data.

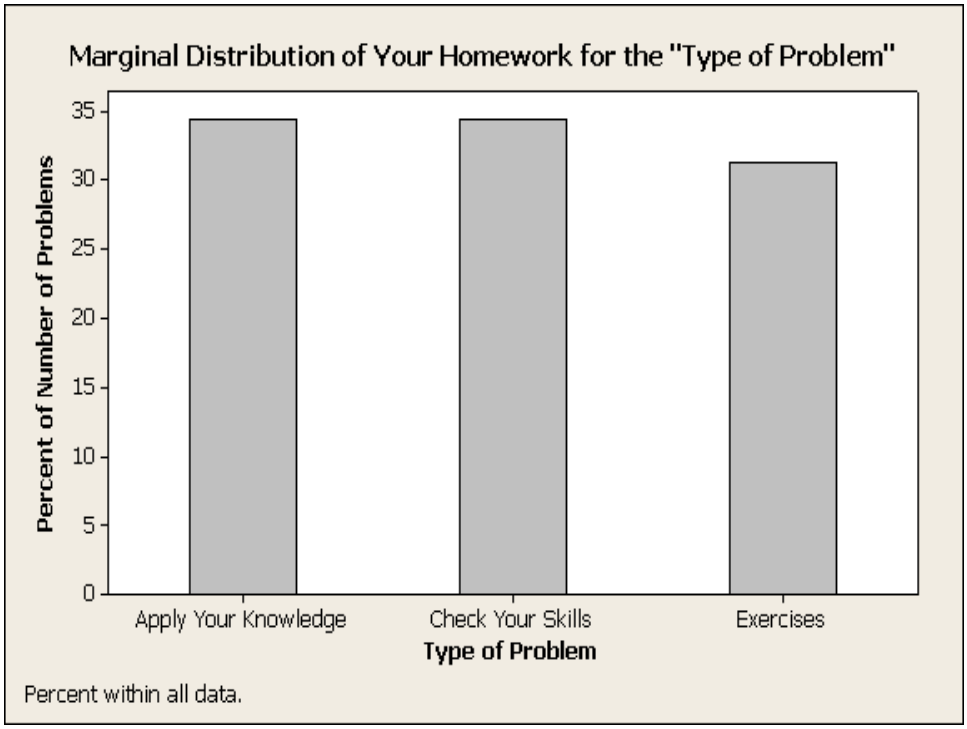
Note. Part of our plan in this class is to convince you that *there is data everywhere* (even in the Three Stooges)!!! In the next example, we turn the textbook on itself and create a two-way table using your homework assignments.

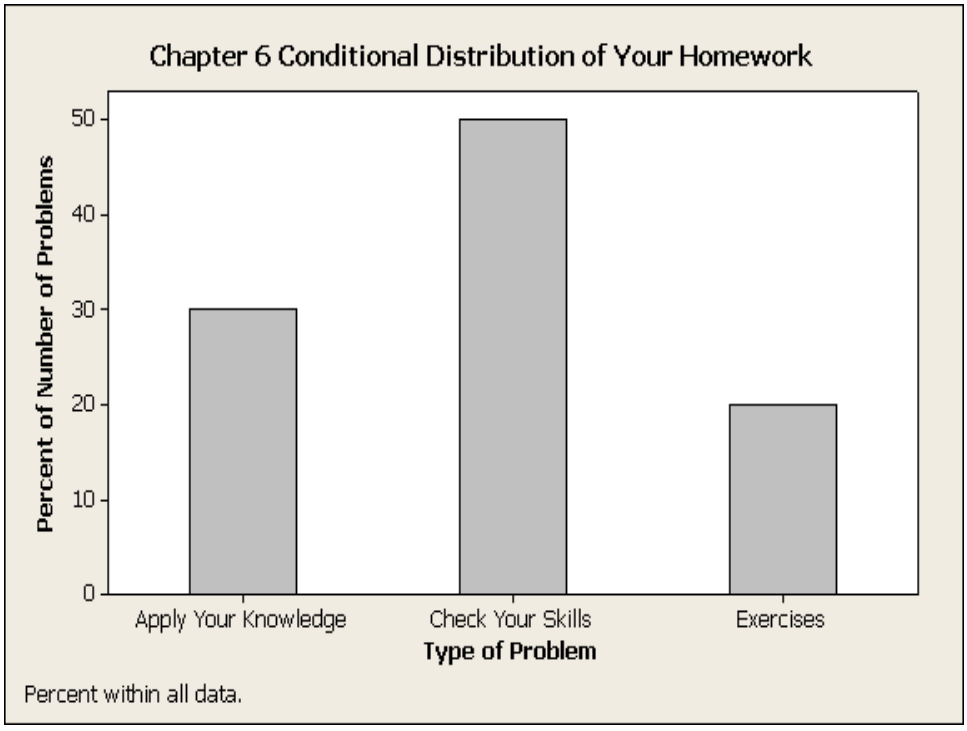
Example. As you know, the text breaks problems into three categories of "Type of Problem": "Apply Your Knowledge," "Check Your Skills," and "Exercises." If we use this as a column variable and we use as the other categorical variable "Chapter Number," then we find that the assigned homework problems for this class break down as follows.

	Apply Your Knowledge	Check Your Skills	Exercises	TOTAL
Chapter 1	10	10	7	27
Chapter 2	9	10	11	30
Chapter 3	9	10	18	37
Chapter 4	10	10	10	30
Chapter 5	7	10	14	31
Chapter 6	6	10	4	20
Chapter 8	10	10	6	26
Chapter 9	8	9	7	24
Chapter 10	11	10	9	30
Chapter 11	11	8	9	28
Chapter 12	4	4	7	15
Chapter 13	10	9	7	26
Chapter 14	8	10	8	26
Chapter 15	19	10	10	39
Chapter 16	10	9	9	28
Chapter 18	8	10	10	28
Chapter 20	13	10	5	28
Chapter 23	6	10	3	19
TOTAL	169	169	154	492

Use Minitab to create a bar chart for the marginal distribution of the "type of problem" variable. Use Minitab to create a bar chart for the conditional distribution of the "type of problem" variable for Chapter 6.

Partial Solution. In Minitab, enter in `C1` the data 169, 169, and 154. In `C2` enter the text "Apply Your Knowledge," "Check Your Skills," and "Exercises" (so that `C1` and `C2` are in correspondence). Click `Graph` and `Bar Chart`. Then in the "Bar Charts" box under "Bars Represent," select "Values from a table." Select `C1` as the "Graph variables" and `C2` as the "Categorical variable." You can click `Labels` to enter a title. To display the data as percentages, click `Bar Chart Options` and select `Show Y as Percent`. Click `OK` to get the following chart.





Example. Exercise 6.23, page 163. Exercise 6.24, page 163

Simpson's Paradox

Note. We now consider an implication of lurking variables. We start with an illustration.

Example. Exercise 6.30, page 165.

Here is the data on the survival of patients at two different hospitals where the patients have been categorized according to their condition:

Good Condition

	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	600

Poor Condition

	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1500	200

(a) Compare percents to show that Hospital A has a higher survival rate for both groups of patients.

(b) Combine the data into a single two-way table of outcome ("survival" or "died") by hospital (A or B). In this table,

which hospital has the higher rate?

Solution. For (a), we see that of the "good condition" patients, at Hospital A $6/600 = 1\%$ died and at Hospital B $8/600 = 1.3\%$ died (so Hospital A looks better). For the "poor condition" patients: $57/1500 = 3.8\%$ died at Hospital A and $8/200 = 4\%$ died at Hospital B (again, Hospital A looks safer).

For (b), the combined two-way table is:

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

We see that Hospital A loses $63/2100 = 3\%$ of its surgery patients and Hospital B loses only $16/800 = 2\%$ (so Hospital B appears better). This is an illustration of Simpson's Paradox. *The patient's condition is a lurking variable when we compare the death rates at the two hospitals.* Many more patients in poor condition go to Hospital A.

Definition. An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox** after British statistician Edward Hugh Simpson in "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society, Ser. B* **13**(2), (1951), 238-241. Simpson's original paper is posted online at:

http://www-stat.wharton.upenn.edu/ hwainer/Readings/

Simpson_The%20Interpretation%20of%20Interaction%20in%20Contingency%20Tables.pdf

Note. Simpson's paradox (as the text states on page 159) is just an extreme form of the fact that observed associations can be misleading when there are lurking variables.

rbg-2-25-2009