

UNIT 8

Describing Relationships with Contingency Tables

Objectives:

- To construct and interpret a contingency table for two qualitative variables
- To recognize independence or describe an association between two qualitative variables

We have already seen that data involving two qualitative variables can be displayed in a contingency table. Table 7-1 (Table A.1-8) is a contingency table displaying data from the variables “Residence” and “Political Party” in the SURVEY DATA, displayed as Data Set A.1 in the appendix. The stacked bar charts displayed in Figures 7-1a and 7-1b (Figures A.1-8(a) and A.1-8(b)) provide us with a visual display of the relationship between the two qualitative variables.

Recall that to describe the relationship between two qualitative variables, we look at how the distribution for one qualitative variable compares across the categories of the other qualitative variable. A stacked bar chart can be very a very helpful visual display in describing the relationship between two qualitative variables.

However, a stacked bar chart is generally easier to interpret when all the bars are the same height. In particular, Figure 7-1b (Figure A.1-8(b)) is easier to work with than Figure 7-1a (Figure A.1-8(a)).

Simply put, if there is no relationship between two qualitative variables, then we would expect that the distribution for one qualitative variable looks the same for each of the categories of the other qualitative variable. For instance, if our two qualitative variables were

residence (rural, suburban, urban) and political party (Republican, Democrat, Other, Independent), and if there were no relationship between these two variables, then we would expect the rural, suburban, and urban areas to all have about the same proportion of Republicans, about the same proportion of Democrats, about the same proportion of Others, and about the same proportion of Independents. Consequently, we would expect that each bar in a stacked bar chart to look like each of the other bars; that is, the division of stacks for one bar would be identical to the division of stacks for each of the other bars. When two variables show no relationship, we say that the variables are *independent*.

To illustrate, we shall imagine that a contingency table displaying the variables residence and political party is to be constructed from 1000 voters. Tables 8-1a and 8-2a represent two possible resulting contingency tables.

We shall first consider the contingency table displayed as Table 8-1a. From Table 8-1a, find the percentage of Republicans in the rural area, the percentage of Democrats in the rural area, the percentage of Others in the rural area, and the percentage of Independents in the rural area. You should find these percentages respectively to be $76/200 = 38\%$, $50/200 = 25\%$, $40/200 = 20\%$, and $34/200 = 17\%$. These percentages have been displayed in the row labeled “Rural” of Table 8-1b. Now, find the percentage of Republicans in the suburban area, the percentage of Democrats in the suburban area, the percentage of Others

Table 8-1a
Contingency Table for "Residence" and "Political Party"

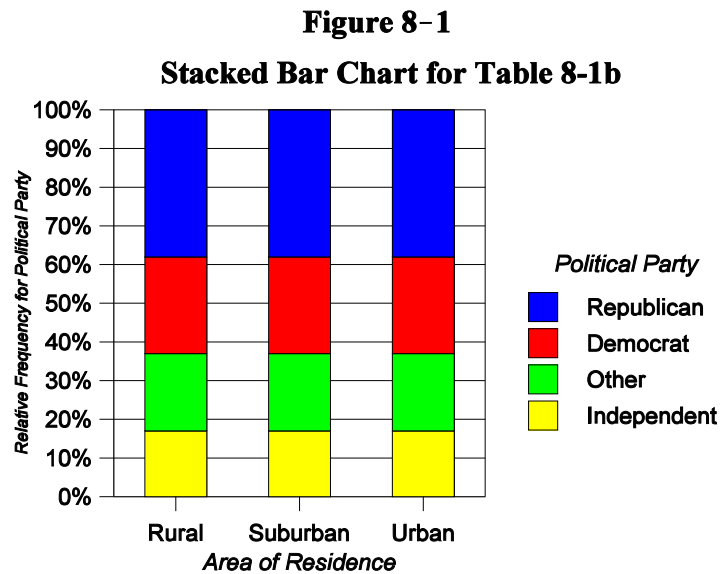
		Political Party				
		Republican	Democrat	Other	Independent	
Residence	Rural	76	50	40	34	200
	Suburban	114	75	60	51	300
	Urban	190	125	100	85	500
		380	250	200	170	1000

Table 8-1b
Contingency Table for "Residence" and "Political Party"

		Political Party				
		Republican	Democrat	Other	Independent	
Residence	Rural	38%	25%	20%	17%	100%
	Suburban	38%	25%	20%	17%	100%
	Urban	38%	25%	20%	17%	100%

in the suburban area, and the percentage of Independents in the suburban area. You should find these percentages respectively to be $114/300 = 38\%$, $75/300 = 25\%$, $60/300 = 20\%$, and $51/300 = 17\%$. Notice that these are exactly the same as the percentages we found for the rural area! These percentages have been displayed in the row labeled “Suburban” of Table 8-1b. We now leave it to you to find the percentages of each political party for the urban area and verify that these are the same as those in the rural and suburban areas, and these percentages have been displayed in the row labeled “Urban” of Table 8-1b.

In order to emphasize the fact that Table 8-1b displays the distribution of political party affiliations for each of the three areas of residence, each row total is shown to be 100%. Since it is difficult to make comparisons when the heights of the bars in a stacked bar chart are not equal, we prefer to construct a stacked bar chart from Table 8-1b rather than from Table 8-1a. Figure 8-1 displays Table 8-1b as a stacked bar chart with the bars representing the different areas of residence and the stacks representing the different political party affiliations. Since each bar in Figure 8-1 has been scaled to the same length representing 100%, a visual comparison of the distribution of political party affiliations for each area of residence is easy.



The fact that each bar in Figure 8-1 looks like each of the other bars is what tells us that residence and political party affiliation are independent, or in other words that the distribution of political party affiliations is the same for each area of residence. The illustration provided by Table 8-1a, Table 8-1b, and Figure 8-1 is somewhat unrealistic, because the percentage corresponding to any one of the four political party affiliations is exactly the same for all three areas of residence. When two qualitative variables are independent, very rarely (if ever) do we find in our data that the distribution for one qualitative variable is exactly identical for each of the categories of the other qualitative variable. Our data would show some small differences between the distributions resulting from random variation. The situation is analogous to flipping a perfectly balanced fair coin 100 times; we expect to see close to 50 heads and 50 tails but are not surprised if we don't see exactly 50 heads and 50 tails, since we expect there to be some random variation.

With real data then, we are looking to see if the distribution for one qualitative variable is similar for each of the categories of the other qualitative variable. Of course, this leaves us wondering just how much of a difference we have to see in the distributions in order for us to conclude that there is a relationship between the two qualitative variables. We are not yet prepared to discuss exactly how to make this decision, but we shall be at a later time. For now, we shall simply use our best judgement in looking at a stacked bar chart to decide whether or not two qualitative variables are independent.

When we constructed Table 8-1b and the corresponding stacked bar chart of Figure 8-1, we chose to compare the distribution of political party affiliations for each area of residence. However, we could have chosen instead to compare the distribution of areas of residence for each political party affiliation. Had we done this, we would have found that the distribution of areas of residence is the same for each political party affiliation, and that each bar in the corresponding stacked bar chart would look identical to each of the other bars, since the two variables are independent. To see that this is true, we invite you to find the percentage of rural voters among Republicans, the percentage of rural voters among Democrats, the percentage of rural voters among Others, and the percentage of rural voters among Independents. You should find these percentages respectively to be $76/380 = 20\%$, $50/250 = 20\%$, $40/200 = 20\%$, and $34/170 = 20\%$, demonstrating that the percentage of rural voters is the same for all political party affiliations. We leave it to you to verify that the percentage of suburban voters is the same for all political party affiliations and that the percentage of urban

voters is the same for all political party affiliations. (If we were to construct a table to compare the distribution of areas of residence for each political party affiliation with row and column headings the same as those in Tables 8-1a and 8-1b, then each column total would be shown to be 100%.)

Let us now consider the contingency table displayed as Table 8-2a. From Table 8-2a, find the percentage of Republicans in the rural area, the percentage of Democrats in the rural area, the percentage of Others in the rural area, and the percentage of Independents in the rural area. You should find these percentages

respectively to be $121/200 = 60.5\%$, $25/200 = 12.5\%$, $30/200 = 15\%$, and $24/200 = 12\%$. These percentages have been displayed in the row labeled "Rural" of Table 8-2b. Now, find the percentage of Republicans in the suburban area, the percentage of Democrats in the suburban area, the percentage of Others in the suburban area, and the percentage of Independents in the suburban area. You should find these percentages respectively to be $149/300 = 49.7\%$, $50/300 = 16.7\%$, $55/300 = 18.3\%$, and $46/300 = 15.3\%$. Notice that these are quite different from the percentages we found for the rural area! These percentages have been displayed in the row labeled "Suburban" of Table 8-2b. We now leave it to you to find the percentages of each political party for the urban area and verify that these are the percentages displayed in the row labeled "Urban" of Table 8-2b, which are all different from the corresponding percentages in the rural and suburban areas.

As in Table 8-1b, we emphasize the fact that Table 8-2b displays the distribution of political party affiliations for each of the three areas of residence by showing each row total to be 100%. Once again, since it is difficult to make comparisons when the heights of the bars in a

stacked bar chart are not equal, we prefer to construct a stacked bar chart from Table 8-2b rather than from Table 8-2a. Figure 8-2 displays Table 8-2b as a stacked bar chart with the bars representing the different areas of residence and the stacks representing the different political party affiliations. Once again since each bar in Figure 8-1 has been scaled to the same length representing 100%, a visual comparison of the distribution of political party affiliations for each area of residence is easy.

The fact that each bar in Figure 8-2 looks different from other bars is what tells us that residence and political party affiliation are not independent, or in other words that the distribution of political party affiliations is not the same for each area of residence. When two variables show a relationship, we say that the variables are *associated* or *dependent*.

To describe the relationship (or association or dependency), we can describe the differences in the distribution of political party affiliations among the areas of residence. Figure 8-2 makes this easy by providing a picture of how the distributions differ. To describe the relationship, we can say that the proportion of Republicans appears to be highest in the rural area and lowest in the urban area. We could also say that the proportion of Democrats, the proportion of Others, and the proportion of Independents each appear to be highest in the urban area.

As before, we may wonder just how much of a difference we have to see in the distributions in order for us to conclude that there is a relationship between the two qualitative variables. When we look at Figure 8-2, some of the differences in proportions might look rather large, but how can we decide when a difference

Table 8-2a
Contingency Table for "Residence" and "Political Party"

Residence	Political Party				
	Republican	Democrat	Other	Independent	
Rural	121	25	30	24	200
Suburban	149	50	55	46	300
Urban	110	175	115	100	500
	380	250	200	170	1000

Table 8-2b
Contingency Table for "Residence" and "Political Party"

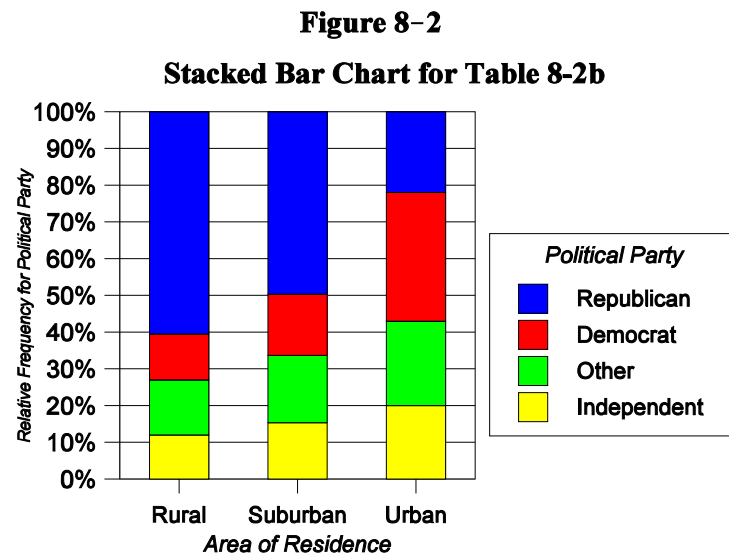
Residence	Political Party				
	Republican	Democrat	Other	Independent	
Rural	60.5%	12.5%	15.0%	12.0%	100%
Suburban	49.7%	16.7%	18.3%	15.3%	100%
Urban	22.0%	35.0%	23.0%	20.0%	100%

is large enough to conclude that a relationship exists. As we said earlier, we are not yet prepared to discuss exactly how to make this decision, but we shall be later, and for now, we simply use our best judgement in deciding whether or not a relationship appears to exist.

When we constructed Table 8-2b and the corresponding stacked bar chart of Figure 8-2 from Table 8-1b, we chose to compare the distribution of political party affiliations for each area of residence. However, we could have chosen instead to compare the distribution of areas of residence for each political party affiliation. Had we done this, we would have found that each bar looks different from each of the other bars, since the two variables are associated. Our description of the relationship in this case would be in terms of the proportion rural voters, the proportion of suburban voters, and the proportion of urban voters among each political party affiliation instead of the proportion of Republicans, the proportion of

Democrats, the proportion of Others, and the proportion of Independents among each area of residence. For instance, we invite you to find the percentage of rural voters among Republicans, the percentage of rural voters among Democrats, the percentage of rural voters among Others, and the percentage of rural voters among Independents. You should find these percentages respectively to be $121/380 = 31.8\%$, $25/250 = 10\%$, $30/200 = 15\%$, and $24/170 = 14.1\%$. We see that the percentage of rural voters looks to be quite different among the political party affiliations. One comment we might make in describing the relationship between residence and political party is that the proportion of rural voters appears to be highest among Republicans and lowest among Democrats. We leave it to you to compare the percentage of suburban voters among the political party affiliations and to compare the percentage of urban voters among the political party affiliations. (If we were to construct a table to compare the distribution of areas of residence for each political party affiliation with row and column headings the same as those in Tables 8-1a and 8-1b, then each column total would be shown to be 100%.)

When constructing a stacked bar chart, it is a matter of personal choice as to which qualitative variable is represented by the bars and which categorical variable is represented by the stacks. This may be dictated by the way in which the data is collected or the way in which it seems more “natural” to look at the data.



Self Test Problem 8-1. A satellite television company gathers data from customers concerning area of residence and choice of favorite entertainment channel; results are displayed in the contingency table on the right.

(a) Construct a contingency table which contains the row totals, the column totals, and the grand total.

(b) What proportion of customers selected the Science Fiction Channel as their favorite?

(c) What proportion of customers did not select the Comedy Channel as their favorite?

(d) What proportion of rural customers selected the Western Channel as their favorite?

(e) What proportion of customers are from the suburban area among those who selected the Cartoon Channel as their favorite?

(f) What proportion of urban customers chose either the Cartoon or Comedy Channel as their favorite?

(g) What proportion of customers are from either the suburban or urban area among those who selected Science Fiction Channel as their favorite?

(h) Identify which of the following are not stated properly, and indicate why. The data are to be used to study the

- (i) relationship between channel preference among rural, suburban, and urban residents,
- (ii) difference in channel preference among rural, suburban, and urban residents,
- (iii) difference between area of residence and channel preference,
- (iv) relationship between area of residence and channel preference.

(i) Choose the appropriate graphical display for the data with the goal in part (h): multiple boxplots, a scatterplot, a stacked bar chart, a histogram.

(j) Construct two tables containing relative frequencies: one based on relative frequencies for channel preference among each of the three areas of residence, and one based on relative frequencies for the three areas of residence among each of the channel preferences.

(k) For each of the tables in part (j), construct a stacked bar chart (with each bar scaled to a height representing 100%) to display the relative frequencies.

(l) For each of the stacked bar charts constructed in part (k), discuss whether or not a relationship between area of residence and favorite channel appears to exist and how this possible relationship might be described.

(m) Could either of the two variables in this data be treated as qualitative-ordinal?

(n) Construct a pie chart displaying the relative frequencies for choice of favorite channel. Does the pie chart provide the same information as a stacked bar chart?

		Favorite Channel			
		Cartoon	Comedy	Sci-Fi	Western
Residence	Rural	28	7	17	53
	Suburban	22	34	36	46
	Urban	19	51	60	27

Answers to Self Test Problems

- 8-1** (a) See Data Set A.2 in the appendix, the TVSAT DATA. (b) $113/400 = 28.25\%$ (c) $308/400 = 77\%$ (d) $53/105 = 50.5\%$ (e) $22/69 = 31.9\%$ (f) $70/157 = 44.6\%$ (g) $96/113 = 85.0\%$ (h) The goal is stated properly in (ii) and (iv); the goal is not stated properly in (i), since rural, suburban, and urban are categories, not variables; the goal is not stated properly in (iii), since area of residence and channel preference are not commensurate variables. (i) stacked bar chart (j) See Tables A.2-1 and A.2-2. (k) See Figures A.2-1 and A.2-2. (l) Since the bars do not appear to be similar to one another in each stacked bar chart, it is reasonable to think that a relationship might exist. From Figure A.2-1, the rural area appears to have the highest percentage of customers who favor the Cartoon Channel and the highest percentage of customers who favor the Western Channel; the urban and suburban areas each appear to have a higher percentage of customers who favor the Science Fiction Channel and a higher percentage of customers who favor the Comedy Channel. From Figure A.2-2, the percentage of rural customers appears to be highest among customers who favor the Cartoon Channel and among customers who favor the Western Channel; the percentage of urban customers appears to be highest among customers who favor the Comedy Channel and among customers who favor the Science Fiction Channel. (m) Favorite channel is qualitative-nominal; but area of residence could be treated as qualitative-ordinal, since the categories have a natural ordering. (n) See Figure A.2-3; a pie chart only provides information about one qualitative variable, while a stacked bar chart provides information about two qualitative variables.

Summary

To describe the relationship between two qualitative variables, we look at how the distribution for one qualitative variable compares across the categories of the other qualitative variable. A stacked bar chart can be a very helpful visual display in describing the relationship between two qualitative variables; however, a stacked bar chart is generally easier to interpret when all the bars have been scaled to a height representing 100%. When constructing a stacked bar chart, it is a matter of personal choice as to which qualitative variable is represented by the bars and which categorical variable is represented by the stacks. This may be dictated by the way in which the data is collected or the way in which it seems more “natural” to look at the data.

When two variables show no relationship, we say that the variables are *independent*. When two qualitative variables are independent, the distribution for one qualitative variable is the same for each of the categories of the other qualitative variable; each bar in a stacked bar chart will look like each of the other bars. When two variables show a relationship, we say that the variables are *associated* or *dependent*. When two qualitative variables are associated, the distribution for one qualitative variable is not the same for each of the categories of the other qualitative variable; at least one bar in a stacked bar chart will look different from other bars.