SPSS PC Version 10: Frequency Distributions and Crosstabs[1]

The following uses a set of variables from the “1995 National Survey of Family Growth” to demonstrate how to use some procedures available in SPSS PC Version 10.

Producing frequency distributions with SPSS:

Following is an example of getting a frequency distribution on a data set. First, bring the file into SPSS either using point a click, or using syntax.

- When you bring in data, either by pointing and clicking, or by using syntax, the 'SPSS Data Editor' screen will pop up. Click on the *Analyze* menu item, then *Descriptive Statistics*, and then *Frequencies*. This will bring up a dialog box containing a list of all the variables in the data set you currently have open.

- Scroll down the list until you find the variable(s) you want to include in your frequency table(s). In this example we are looking for the variable *region*. Click on the variable name/label to highlight the variable, and then click on the right-pointing arrow to move the variable to the list of variables you want frequency distributions for. If you accidentally move the wrong variable to the "Variable(s)" list, simply highlight it and click on the left-pointing arrow to remove it and then add the correct variable to the "Variable(s) list." When you have the correct variable in your list, click on the "*OK*" button and SPSS will create your frequency table. SPSS will automatically open a separate "Output viewer" window containing this frequency table. For the variable *region* you get the following output:

Statistics

REGION Region Where R Lives

N	Valid	10847
	Missing	0

REGION Region Where R Lives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 Northeast	2029	18.7	18.7	18.7
	2.00 Midwest	2593	23.9	23.9	42.6
	3.00 South	3751	34.6	34.6	77.2
	4.00 West	2474	22.8	22.8	100.0
	Total	10847	100.0	100.0	

The syntax that would be created from this pointing and clicking sequence is:

```
GET
  FILE='D:\help\nsfg95.sav'.
FREQUENCIES
  VARIABLES=region
  /ORDER=  ANALYSIS .
```

- The first table shows the number of valid cases (remember that we use the letter N to indicate this valid number of cases) and the number of missing cases for the variable we are interested in. You can see that there are no individuals in our data set that are missing information on their region of residence; all 10,847 cases have valid information on this variable.

- The second table is the actual frequency table, containing five columns: The first column lists the value categories of the variable, including any designated missing categories, and the second shows the number of cases – that is, the frequency – of cases falling in each category. Here we can see that 2,029 of the respondents live in the Northeast

[1] Prepared by Kyle Crowder of the Sociology Department of Western Washington University, and by Patty Glynn, University of Washington. November 2, 2000

region, 2,593 live in the Midwest, etc. The third column is labeled "Percent" and the fourth is labeled "Valid Percent." If there were missing values for this variable, the columns labeled “Percent” and “Valid Percent” would have different values.

Q: What is the difference between the "Percent’ column and "Valid Percent" column is this distinction important?
A: The "Percent" column represents the percentage of all cases, including the missing cases, constituted by each category, while the "Valid Percent" category presents the percentage of only the non-missing cases falling into each category. In some cases we will be interested in treating the missing values on a variable as just another category of that variable. In many cases, however, this missing information is to be ignored. When we are not interested in the missing values, the "Valid percent" column provides an accurate picture of the distribution of the valid cases since these "valid" percentages are not deflated by the inclusion of the missing cases in the denominator. *You should always use the "Valid Percent" column unless the missing values are of some substantive interest to you (and they usually are not).*

In our example, the "Percent" and "Valid percent" columns are identical because there are no missing values for the variable *region*. We can see, for example, that 34.6% of our NSFG respondents live in the South.

The final column of the table presents the cumulative percent at each of the non-missing categories.

Q: When is the "cumulative frequency" column meaningful or useful?
A: Although SPSS always reports them in all frequency tables, the cumulative percentages are meaningful only for variables measured at the ordinal and interval-ratio levels. The values of nominal variables, by definition, cannot be ordered. Without ordering it is meaningless to talk about the percentage of cases falling "at or below" a particular value on the variable -- that is, having that value or a "lower" value -- since no values are higher or lower than the others.

In our example, the "Cumulative percent" column is meaningless since we have a variable measured at the nominal level.

Producing cross-tabulations with SPSS:

- Now let's pretend that I have a burning desire to see how the region of residence affects the likelihood that someone has ever cohabited with a partner outside of marriage as indicated by the variable *cohever* ('Whether R has ever cohabited'). To start to explore this relationship in the sample, we can have SPSS create a cross-tabulation table, or "crosstab."

- At the top of the 'SPSS Data Editor' screen choose *Analyze*, then *Descriptive Statistics…*, and *Crosstabs*. This will open a dialog box in which you can choose the two variables to be included in the bivariate table and chi-square test. Remember that the values of the independent variable should occupy the columns and the values on the dependent variable should occupy the rows of your table. Choose these by highlighting variables from your list and clicking the appropriate right-pointing arrows next to the "Column(s)" and "Row(s)" boxes. In our example we are interested in the impact or effect of region on cohabitation, so our independent variable is region of residence (*region*) and the dependent variable is our indicator of whether the respondent has ever cohabited (*cohever*). Following convention, *region* will be our column variable and *cohever* will be our row variable.

- Now you should request column percentages so you can get an idea of the magnitude of the differences between the conditional distributions. Remember that the differences in the distribution of the dependent variable across different values of the independent variable (i.e., the conditional distributions) give us some basic idea of whether a relationship between the two variables exists and how strong that relationship might be. You can request the column percentages needed for this comparison by clicking on "Cells" and choosing "Column" under the "Percentages" heading in the dialog box that pops up. After you've made your selection return to the main dialog box by hitting the "Continue" button. Once you have the correct variables selected into the appropriate boxes and have requested column percentages, hit "OK."

- SPSS will spit out the following output:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
COHEVER Whether R has Ever Cohabited * REGION Region Where R Lives	10847	100.0%	0	.0%	10847	100.0%

COHEVER Whether R has Ever Cohabited * REGION Region Where R Lives Crosstabulation

			REGION Region Where R Lives				
			1.00 Northeast	2.00 Midwest	3.00 South	4.00 West	Total
COHEVER Whether R has Ever Cohabited	1.00 Yes, has cohabited	Count	853	1090	1496	1183	4622
		% within REGION Region Where R Lives	42.0%	42.0%	39.9%	47.8%	42.6%
	2.00 No, has never cohabited	Count	1176	1503	2255	1291	6225
		% within REGION Region Where R Lives	58.0%	58.0%	60.1%	52.2%	57.4%
Total		Count	2029	2593	3751	2474	10847
		% within REGION Region Where R Lives	100.0%	100.0%	100.0%	100.0%	100.0%

The syntax that would be created from this pointing and clicking sequence is:

```
CROSSTABS
  /TABLES=cohever  BY region
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT ROW COLUMN .
```

- As usual, the first box, entitled "Case Processing Summary" summarizes the number of valid cases for the two variables used in the analysis. Each of the variables of interest has valid information for all 10,847 respondents. The second box contains the actual bivariate cross-tabulation including the column percentages you requested. The cells of this table each tell us the number of cases falling into a particular combination of values on the two variables, *region* and *cohever*. For example, there are 853 people who live in the Northeast and have cohabited while 1,496 of the people from the South region have cohabited.

- The column percentages provide a better comparison of the regions in terms of the likelihood of cohabiting because they, by definition, take into consideration the number of people living into each of the regions (the number of people in the various categories of the independent variable). For example, the 853 northeasterners who cohabited represent 42.0% of all the northeasterners (total=2,029) in the sample while the 1,496 southern cohabitors represent only 39.9% of the larger group of southerners (total=3,751). Again, these differences in column percentages give us a clue about the nature of the relationship between region and cohabitation. We can see that the conditional distributions of the dependent variable (conditional on the independent variable) do vary: a 47.0% of the westerners have cohabited while only 39.9% of the southerners have cohabited, with the northeasterners and midwesterners falling between these extremes at 42.0%.

  Q: Why focus on one of the categories of the dependent variable (value of 1 = 'yes, have cohabited) and ignored the other value of the dependent variable?
  A: Because the dependent variable is dichotomous (has only two values) the information in the cells for the second value of *cohever* (2 = 'no, has never cohabited') is redundant. For example, if 42.0% of the northeasterners have cohabited, I know that the other 58.0% of the people must fall in the other category and have not cohabited.