DLI Atlantic Training 2006

Session 1.2

A Taste of SPSS (v.13)

“I need some variables…”

Standard Disclaimer

This is an “EH never learned the 8 times tables” document.

This is not a sophisticated, nor inclusive, document and it does gloss over important details. All errors are mine, of course.

But
this can serve as a reminder of what many students are doing with data sets and what the students need
in terms of their own preparation
to do many of the research assignments given to introduce them to research methods.

For further reading, see some of the books recommended last year on basic statistics and SPSS.

Preliminaries

"Which variables do I need?"

What do you want to find out?

"I want to explore the factors that contribute to a sense of health."

Chuck's easy-to-use form (center insert) is a useful way of putting the task back into the user's hands, rather than yours. Ask the student to fill out the form using the documentation/user guide before taking them to the SPSS workstation.

The student should have some knowledge of the problem (background reading) so that they can come up with a reasonable hypothesis. The hypothesis often involves the relationship between variables, such as the extent to which sex or education affects happiness.

Students usually have to find out whether there is a relationship (or an association) between the variables and, if there is, how strong it is, in what direction is it, and is there any statistical significance to that numerical relationship.

Some considerations in variable selection?

- Decide upon the dependent variable (that is, what is being explored) and list it first.
- It is all much easier if you have variables that are nicely distributed (that is, eschew the variable that has 5,000,014 responses (or 96%) for a value of "excellent", and 14 responses (less than 1%) each for "very good", "fair", and "poor" and "missing". Something exists to explain if respondents are spread fairly evenly across the variable's categories, which is the ideal situation for research.
- Continuous non-nominal variables (see glossary on the back page) have the most straight forward statistical procedures for assignments testing the strength of a relationship between variables. This type of measurement is essential for the dependent variable in class assignments for econometrics where students are to conduct a regression analysis.
- It may be desirable to recode a variable for analysis—but keep track of your work. The time that you don't document the recode instructions will be the time that the student returns needing to regenerate the recoded variable. When recoding a variable, the best practice is to create a new variable to contain the recodes and not to change the values in the original variable.
- Be alert for skip patterns in the selection of variables (read the documentation!) Such patterns can result in large numbers of missing responses in variables for those respondents for whom the variables were not applicable.

There is no right or wrong choice of variables. If the student asks, remind him/her that it is a research decision. It is up to the student to make the decision as to what to investigate.

IMPORTANT NOTE

In this booklet, the right hand page gives basic instructions on how to do a frequency table, how to create a cross-tab, how to recode a variable (and make it beautiful), how to do a statistical test and finally, how to cite the work.

The left-hand page provides a view of the results, and a suggestion of how you might interpret the results or read the table.

WHAT DOES IT MEAN?

What does this table tell us?

Self-assessed health rating

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Excellent?	6255	25.1	25.1	25.1
	Very good?	8669	34.7	34.8	59.9
	Good?	6700	26.9	26.9	86.8
	Fair?	2461	9.9	9.9	96.7
	Poor?	826	3.3	3.3	100.0
	Total	24911	99.8	100.0	
Missing	Not stated	11	.0		
	Don't know	29	.1		
	Total	40	.2		
Total		24951	100.0		

Observations:

a. There is a fairly good distribution in this table. There are some people who regard their health as being poor, and a much larger group saying that their health is excellent.
b. There were 24,951 people in this survey and 24,911 responded with an evaluation of their own health
c. 29 individuals said that they did not know how they would assess their health while 11 individuals did not supply any response to this question
d. 86% of those who contributed a valid response indicated responses of good, very good, or excellent health.

For the report

In Cycle 17 of the General Social Survey, 4 out of 5 Canadians sampled for this survey regard their health as good to excellent. Less than 15% reported their health as fair or poor.

Table 2: Self-assessed health rating (weighted)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Excellent?	6787670	26.6	26.6	26.6
	Very good?	8947154	35.0	35.1	61.7
	Good?	6827255	26.7	26.8	88.4
	Fair?	2240472	8.8	8.8	97.2
	Poor?	706061	2.8	2.8	100.0
	Total	25508612	99.9	100.0	
Missing	Not stated	10004	.0		
	Don't know	25807	.1		
	Total	35811	.1		
Total		25544423	100.0		

Note the changes that occur when the weight variable is added. The total number of responses goes up to 25,544,423 and the percentage of those with self-reported excellent health rises from 25.1% to become 26.6%.

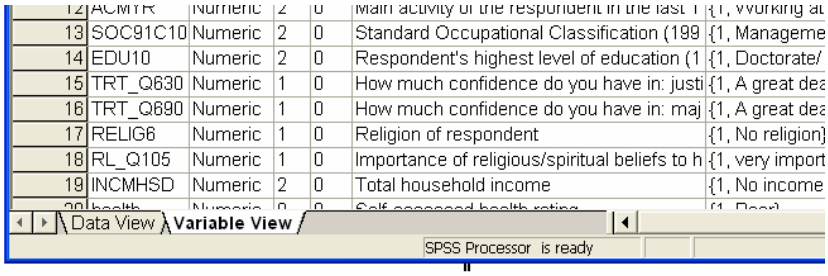
With the addition of the weight variable, we can now generalize to Canadians rather than only those who were sampled for the survey.

For the report

In Cycle 17 of the General Social Survey, 4 out of 5 Canadians regard their health as good to excellent. Less than 15% reported their health as fair or poor.

HOW DO I GENERATE FREQUENCIES?

When you open an SPSS save file, the **Data Editor** will appear on the screen. The ***Data View*** screen has the numerical responses by each respondent to the questionnaire; the ***Variable View*** shown below provides handy (and valuable) attribute information for the variables. The **Output Viewer** displays the results of your analysis.



Get to know your variables by running frequencies on them. From the SPSS menu at the top of the **Data Editor** screen, click on **Analyze** to get the pull-down menu and proceed as follows:

Analyze → Descriptive Statistics → Frequencies

In the dialogue box that appears, scroll down and highlight **HAL_Q110** (*Self-assessed health rating*). Double-click to move **HAL_Q110** into the variable list column. Click the **OK** button. The output screen pops up with the table seen on the left hand page.

It is a thing of beauty. Now try doing a frequency for the variable **PE_Q330**

~ ~ ~

HOW DO I APPLY THE WEIGHT VARIABLE?

In the user guide for GSS17, users are cautioned not to release unweighted estimates. How do we ensure that the results we produce will be properly weighted to represent the entire Canadian population?

Refer the student to the GSS 17 User Guide. The guide gives information on which variable to use (depending on what the user wants to do) and it also gives guidelines on when and how to produce weighted results.

Under the section “Release Guidelines and Data Reliability”, for example, Statistics Canada cautions against using weights if any cell value in your work is less that 15. That is one reason why we usually apply the weight as a final step in our analysis.

Applying the weight variable is simple enough in SPSS, but be aware that the user should be reading the section on Release Guidelines very carefully. Begin at the Data Editor screen and select DATA from the toolbar.

Data→ Weight Cases→

Tick the button “Weight Cases,” highlight the weight variable (in this case, WGHT_PER) and use the arrow key to select WGHT_PER. Press OK. You will notice that the screen doesn’t change—except that the words “Weight On” appear at the very bottom of the Data Editor screen.

Now, go back and repeat the process but select “Do not weight cases” so that we can continue working only with the survey sample, as we investigate our variables a bit more..

WHAT DOES IT MEAN?

What does this table tell us?

Table 2: Self-assessed health rating * Hours spent watching TV during typical week Crosstabulation

			Hours spent watching TV during typical week				
			less than 5 hours?	5 to 14 hours?	15 to 29 hours?	30 hours or more?	Total
Self-assessed health rating	Excellent?	Count	1921	3067	914	269	6171
		% within Hours spent watching TV during typical week	29.8%	26.5%	19.6%	14.6%	25.2%
	Very good?	Count	2291	4249	1587	457	8584
		% within Hours spent watching TV during typical week	35.5%	36.8%	34.0%	24.8%	35.0%
	Good?	Count	1580	2991	1430	578	6579
		% within Hours spent watching TV during typical week	24.5%	25.9%	30.6%	31.4%	26.8%
	Fair?	Count	500	975	546	365	2386
		% within Hours spent watching TV during typical week	7.7%	8.4%	11.7%	19.8%	9.7%
	Poor?	Count	161	270	193	173	797
		% within Hours spent watching TV during typical week	2.5%	2.3%	4.1%	9.4%	3.3%
Total		Count	6453	11552	4670	1842	24517
		% within Hours spent watching TV during typical week	100.0%	100.0%	100.0%	100.0%	100.0%

First, find the cells indicating 100%. Note that all the columns arrive at a total of 100%, which means that your primary focus will be in reading up and down the table.

Look for the large percentages and for patterns.

In the table above, of those who watched TV for less than 5 hours, 29.8% of respondents indicated that they felt their health was excellent. In looking at the column marked 30 hours or more, only 14.6% of that group felt that their health was excellent.

For the report

In Cycle 17 of the General Social Survey, a third of the respondents who watched television less than 5 hours a week reported their health as being excellent. Of those who watched more than 30 hours of television a week, 30% described their health as fair or poor.

HOW DO I DO A CROSS-TAB?

Do the number of hours watching television every week affect the assessment by Canadians of their own health? We can explore this using variables HS_Q110 and PE_Q330 and creating a cross-tabulation.

One way to run crosstabs is to click on **Analyze** to get the pull-down menu and proceed as follows:

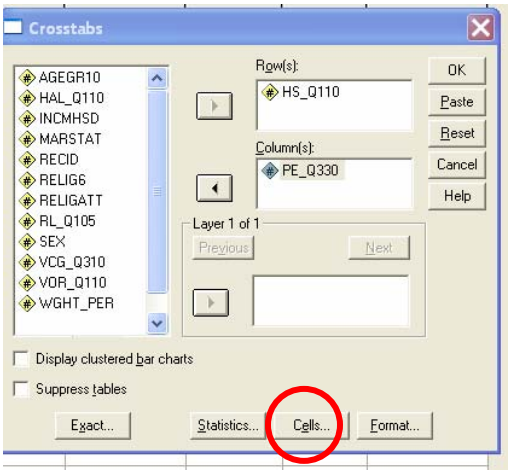
Analyze
→ Descriptive Statistics
→ Crosstabs

Oh, no—more choices! What goes in which box, and which percentages should I use?

One convention is to keep your **Dependent** Variable in the Table as **Rows,** with the **Independent** Variable in **columns** and to request ***column percentages***. If you do this consistently, you will become more adept at quickly finding any interesting patterns in the data.

	Independent variable Factors that might impact DV
Dependent variable **To be explored**	

Move down the list of variables using arrows to highlight the variable you wish to explore (in this case, the Dependent variable will be **HS_Q110).** Click on the right arrow key to move it into the Row Box. Move the Independent variable (**PE_Q330**) into the Column Box.



HOW DO I GET PERCENTAGES?

You could simply click on OK at this point but, if you are not terribly numerate, it is easier to see patterns if you get the percentages at the same time.

To do this, click on the **CELLS** button, and then click on "Percentages" box beside "Column." Press the **CONTINUE** button; then press **OK** to create a cross tabulation of the two variables.

WHAT DOES IT MEAN?

Frequencies before the Recode:

Total household income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No income or loss	45140	.2	.2	.2
	Less than $5,000	96109	.4	.5	.7
	$5,000 to $9,999	326542	1.3	1.7	2.4
	$10,000 to $14,999	747006	2.9	3.9	6.3
	$15,000 to $19,999	747571	2.9	3.9	10.2
	$20,000 to $29,999	1817365	7.1	9.5	19.7
	$30,000 to $39,999	2238274	8.8	11.6	31.3
	$40,000 to $49,999	2058797	8.1	10.7	42.0
	$50,000 to $59,999	2211403	8.7	11.5	53.5
	$60,000 to $79,999	3149874	12.3	16.4	69.9
	$80,000 to $99,999	2187424	8.6	11.4	81.3
	$100,000 or more	3589676	14.1	18.7	100.0
	Total	19215180	75.2	100.0	
Missing	Not stated	2499806	9.8		
	Don't know	3829437	15.0		
	Total	6329243	24.8		
Total		25544423	100.0		

Statistical or Methodological Decision?

Cut-point statistics: 5 equal groups

Total household income

N	Valid	19215180
	Missing	6329243
Percentiles	20	7.00
	40	8.00
	60	10.00
	80	11.00

Research decision: Recode to show poverty line

1	No income or loss	7	$30,000 to $39,999
2	Less than $5,000	8	$40,000 to $49,999
3	$5,000 to $9,999	9	$50,000 to $59,999
4	$10,000 to $14,999	10	$60,000 to $79,999
5	$15,000 to $19,999	11	$80,000 to $99,999
6	$20,000 to $29,999	12	$100,000 or more

In this case, we will recode values of 1 to 5 into a value of 1 to reflect household income at the poverty line or below, rather than 1 to 6, which would reflect a statistically-based decision.

What is the right way to recode a variable? Is there a right way?

It is a matter of choice for the researcher: what is important to him/her? What knowledge of the topic do they have that will influence how they want to explore a topic?

In analysis, however, it is very often useful to have distributions in your recoded values that are approximately equal or that reflect the distribution of the original variable.

In the report…

The presence of low income is associated in the literature with negative health measures. Using the 2002/2003 low income cut-off range for a family of two of $18,560-$19,044, we can see the distribution of according to household income status. Less than 15% of those below the poverty line report that they feel their health is excellent.

HOW DO I DO A RECODE?

First of all, when would you want to modify a variable by recoding it?

Recoding is a useful way to group continuous variables into digestible chunks that can be fit into a table, and it is also helpful in producing tables that are easier to read and that more readily let you see patterns. This time, let us select another variable that might have an impact on self-reported health status: household income.

Step 1: Decide how to handle the variable.

The choice is a research decision. It can be done based on the distribution of the data for the variable, or you can model the groups according to your research question (as we will do).

Total household income

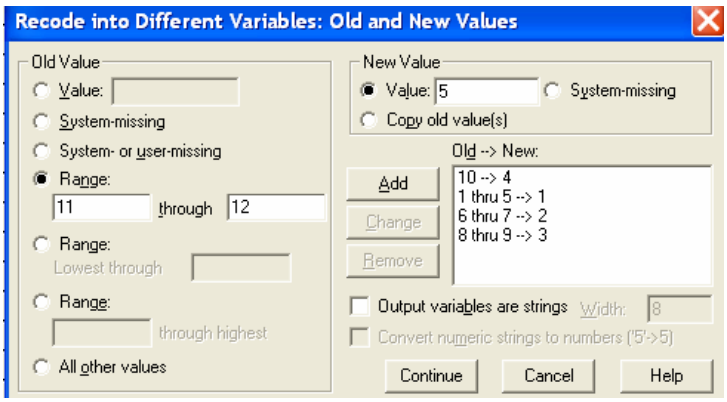
Code	Content	Freq.	Valid %	Recode
1	No income or loss	63	.3	
2	Less than $5,000	125	.7	Codes **1-5** to be the new value of **1** (poverty line)
3	$5,000 to $9,999	504	2.7	
4	$10,000 to $14,999	1140	6.0	Codes **6-7** to be the new value of **2** (lower middle class)
5	$15,000 to $19,999	969	5.1	
6	$20,000 to $29,999	2139	11.3	Codes **8-9** to be the new value of **3** (middle class)
7	$30,000 to $39,999	2435	12.8	
8	$40,000 to $49,999	2102	11.1	Code **10**-to be the new value of **4** (upper middle class)
9	$50,000 to $59,999	2168	11.4	
10	$60,000 to $79,999	2765	14.5	Code **11-12** to be the new value of **5** (rich)
11	$80,000 to $99,999	1803	9.5	
12	$100,000 or more	2795	14.7	

Step 2: Recode the variable

Transform → Recode→ Into a New Variable

A. Begin with the old variable (INCMHSD) and double-click to make it the input variable. Add a new variable name and a description for the variable (in this case, the name of the new variable will be ***Income*** and the Label will be ***Household Income***.). Click on the **CHANGE** button.

B. Now that you have created a new variable ready for definition, use the values of the old variable and indicate how you want them treated in the new variable. Click on the button marked **OLD AND NEW VALUES**. Fill out the dialogue box using the decisions you have made (above). Click **CONTINUE** when done.



WHAT DOES IT MEAN?

Self-assessed health rating * Household income Crosstabulation

			Household income					
			1.00	2.00	3.00	4.00	5.00	Total
Self-assessed health rating	Excellent?	Count	292074	892603	1183871	926977	2040397	5335922
		% within Household income	14.9%	22.0%	27.7%	29.5%	35.3%	27.8%
	Very good?	Count	509612	1345741	1543128	1242135	2313295	6953911
		% within Household income	26.0%	33.2%	36.1%	39.5%	40.1%	36.2%
	Good?	Count	634825	1224288	1139042	774221	1142283	4914659
		% within Household income	32.4%	30.2%	26.7%	24.6%	19.8%	25.6%
	Fair?	Count	362433	465001	316725	163856	228655	1536670
		% within Household income	18.5%	11.5%	7.4%	5.2%	4.0%	8.0%
	Poor?	Count	158057	125827	87435	37681	50416	459416
		% within Household income	8.1%	3.1%	2.0%	1.2%	.9%	2.4%
Total		Count	1957001	4053460	4270201	3144870	5775046	19200578
		% within Household income	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Observations:

Well, you hypothesized that income has an affect on a person's assessment of their own health and it certainly seems that this is true for those in excellent health.

However, the wording for this finding is also bit unwieldy for a newcomer to SPSS and statistics. That is because your values are increasing for one variable and decreasing in the other variable. Try another recode to reorder the values of Self-assessed health rating.

Reordered recode:

5	→	1	Poor
4	→	2	Fair
3	→	3	Good
4	→	4	Very Good
5	→	5	Excellent

Self-assessed health rating

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Poor	826	3.3	3.3	3.3
	Fair	2461	9.9	9.9	13.2
	Good	6700	26.9	26.9	40.1
	Very good	8669	34.7	34.8	74.9
	Excellent	6255	25.1	25.1	100.0
	Total	24911	99.8	100.0	
Missing	System	40	.2		
Total		24951	100.0		

HOW DO I MAKE IT BEAUTIFUL?

If you now do a frequency on the new variable you have created **(Analyze→ Descriptive→ Frequencies→ *income*)**, you will get a rather homely sight (and are missing some important information to help interpret the display). It is time to make the new variable more meaningful by providing value labels and refining the variable type.

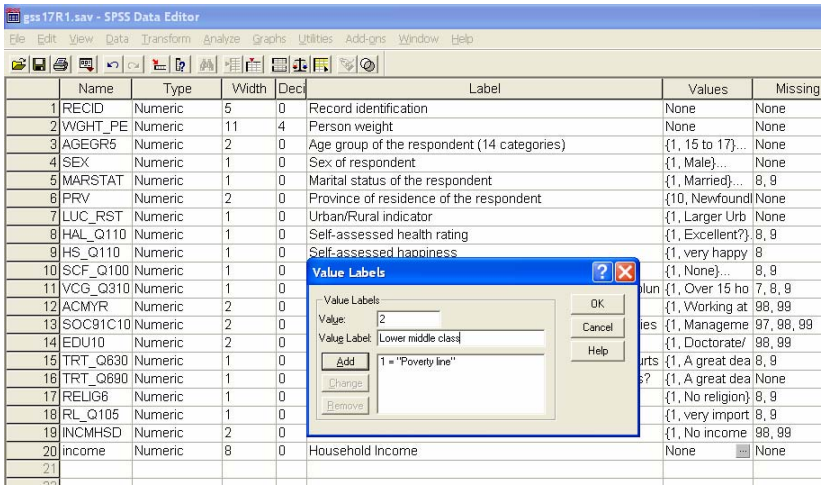
Household Income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	2801	11.2	14.7	14.7
	2.00	4574	18.3	24.1	38.8
	3.00	4270	17.1	22.5	61.3
	4.00	2765	11.1	14.5	75.8
	5.00	4598	18.4	24.2	100.0
	Total	19008	76.2	100.0	
Missing	System	5943	23.8		
Total		24951	100.0		

How do we give readable, easily communicated labels to these values?

First, make sure that you are in **Data Editor** and select the **Variable View** screen. At the bottom of the list of variables, you will see your new variable, *income*.

Because this is a category rather than a decimal number, remove the two decimal points by clicking on the up and down symbol. Reset it from 2 to 0. Now, move over to the variable characteristic of Labels and click on the three dots. This will bring up a dialogue box into which you can put the value and the label you want assigned to it.



When you have finished, click **OK** and run a frequency on the new value of ***incomes*** again. Is it not much more beautiful?

Now, run the cross-tab with your brand new, fully dressed variable, ***incomes***.

Three Item Glossary

Self-assessed health rating * Household income Crosstabulation

			Household income					
			Poverty line	Lower class	Middle Class	Upper middle class	Rich	Total
Self-assessed health rating	Poor	Count	240	148	86	37	37	548
		% within Household income	8.6%	3.2%	2.0%	1.3%	.8%	2.9%
	Fair	Count	541	534	301	145	188	1709
		% within Household income	19.4%	11.7%	7.0%	5.3%	4.1%	9.0%
	Good	Count	879	1343	1125	649	897	4893
		% within Household income	31.4%	29.4%	26.3%	23.5%	19.5%	25.8%
	Very good	Count	739	1557	1591	1122	1840	6849
		% within Household income	26.4%	34.1%	37.3%	40.7%	40.0%	36.1%
	Excellent	Count	396	988	1167	807	1634	4992
		% within Household income	14.2%	21.6%	27.3%	29.2%	35.6%	26.3%
Total		Count	2795	4570	4270	2760	4596	18991
		% within Household income	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.274	.008	35.706	.000
N of Valid Cases		18991			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Ahh. That looks better. It is easier to see the high values and low values. Only .8% of rich people responded that their health was poor, whereas 8.6% of the poverty line (or below) group indicated a self-assessed status of health as poor. At the other end of the scale, 35.6% of the rich group indicated that their health was excellent, but only 14.2 % of the low-income group reported their health to be excellent.

But how strong is the relationship between income and self-reported health rating? We used the Gamma test because it is one of the tests that are appropriate for two ordinal variables.

What does the approximate significance of .000 mean? That there is nothing there? And what does the value of .274 mean? It looks like a batting score.

If the Approx. Sig. is low (typically with a number of less than .05), it indicates that there is a relationship between the variables.

If there is a positive number, it suggests that as the independent value rises, so does the dependent value. In other words, when income goes up, the self-rated health status also rises. How strong is the link?

The gamma value of +.274 indicates that our ability to accurately which of two people has the better self-assessed health status is improved by 25% if we know their respective income levels. Students should have some sort of guideline from their professors on interpreting test values (see previous page for samples of appropriate language).

In the report…

Does money guarantee a feeling of good health? The relationship between income and status of health is a positive, albeit somewhat weak, association. It would be useful to look at other factors, such as age, to see how other variables affect self-assessed health status.

When in doubt, highlight the test result on the output screen, right click, and select **SPSS coach**.

HOW DO I DO A STATISTICAL TEST?*

Students are often asked to include a "statistical test"—to determine how strong the level of association between their variables is and what sort of association it is. Which statistical test to select is up to them!!!!

SPSS makes it look easy—as long as the student knows the type of variables they are using and how they are suppose to interpret the results. In the case of income and the apparent effect on a person's assessment of their own health, both variables in the public use file for GSS 17.are ordinal (see glossary). This determines the type of statistical test that is appropriate in this case.

Begin the cross-tab **Analyze → Descriptive Statistics → Crosstabs**

Before clicking on the OK button, click on STATISTICS. You will see a number of choices. Because we have two ordinal variables, we can select Gamma (or one of the other statistical tests under the Ordinal category) and press OK to perform the operation. The results appear in a small box under the cross-tab (see table 3).

To interpret the test, students usually follow guidelines provided by their professor. For example:

Measure of Association	Gamma
Types of Variables	Ordinal x ordinal
Values (strengths)	-1.00 to +1.00 0=no association -1.00 = perfect (negative) association +1.00 = perfect (positive) association So 0.10 would indicate a very weak negative association while 0.70 would indicate a very positive association
Direction	+ indicates positive association (variables change or move in same direction (i.e., as one increases in value, the other increases or as one decreases, the other decreases) - indicates negative association (variables change or move in opposite directions (i.e. as one goes up, the other goes down--or vice versa)

Suggestions for describing values of Percentage Reduction in Error Measures of Association

Appropriate phrase for neg value	**Value**	**Appropriate phrase for pos. value**
Perfect negative association	1.00	Perfect positive association
Virtually perfect positive association	0.90 to 0.99	Virtually perfect positive association
Very strong negative association	0.70 to 0.89	Very strong positive association
Substantial negative association	0.50 to 0.69	Substantial positive association
Moderate negative association	0.30 to 0.49	Moderate positive association
Weak negative association	0.10 to 0.29	Weak positive association
Negligible negative associate	0.01 to 0.09	Negligible positive associate
No association	0.0	No association

QUICK REFERENCE

1. Frequencies

Analyze → Descriptive Statistics → Frequencies

2. Cross-tabulations

Analyze → Descriptive Statistics → Crosstabs

Move dependent variable into row box;
move independent variable(s) into column box

- **Calculate percentages in crosstabs**

  After adding dependent and independent variables, click on Cells, and tick Column percentages (you can vary this choice, according to your preference).

3. Recode a Variable

Transform → Recode → Into a New Variable

Define new variable by selecting the variable you wish to recode, give the new variable a name of 8 characters or less, and provide a "simple English/French" label.

Indicate how you want the old values to be transformed in the new variable. Click OK when done. Run frequencies to check your work.

4. Beautify (that is, document) New Variables

Using the Variable View screen in the Data Editor, click on the three dots in the Values column beside your new variable (which will be at the end of the list). This will bring up a dialogue box where you can indicate what each value means. (Value *1* will get the Value Label *Rich*, for example). Click OK when done.

Since you may often be working with categorical data, click on the Decimal Column and switch the default of 2 to a zero.

5. Add a statistical test

Analyze → Descriptive Statistics → Crosstabs

After adding dependant and independent variables, click on Statistics and tick the test that is appropriate to the type of variables involved in the cross-tab. Click OK.

HOW DO I CITE IT?

Use the bibliographic guide prepared by Gaetan Drolet (University of Laval/Statistics Canada). It is not yet publicly released but can be accessed in draft format:

<http://www.statcan.ca/english/Dli/gaetan/citationstyle.htm>

The username and password are *gaetan*

For the subset of the public use microdata file:

Statistics Canada. 2004. "General Social Survey 2003, Cycle 17" [subset compiled from public-use microdata file]. Statistics Canada (producer), using IDLS (distributor); (raw data, 6.0 MB, 24,951 cases). Accompanying documentation: user guide and questionnaire (PDF). Released July 6, 2004.
http://janus.ssc.uwo.ca/idls/
(accessed April 6, 2006).

For the User Guide to the General Social Survey, cycle 17:

Statistics Canada. 2004. *2003 General Social Survey, Cycle 17: Social Engagement Public Use Microdata File Documentation and User's Guides*. Statistics Canada Catalogue no. 12M0017GPE.
http://www.statcan.ca/english/Dli/Metadata/gss/cycle17-2003/gssc17gid.pdf
(accessed April 6, 2006).

Three Item Glossary

Variable types (levels of measurement)

There is hierarchy in the statistical world.

Nominal: responses describe qualities or attributes that do not imply magnitude or value.
Example? Variable "Coffee shops": Values: Tim Hortons/Starbucks/Dunkin' Donuts/Irving Gas Stop/Other
The categories differ in name only and have the least amount of information of the four levels of measurement in the group.

Ordinal responses order data from lowest to highest (or vice versa) so they rank data—but there is no idea of the magnitude or difference between values.
Example: Variable: Coffee size: Small, medium, large, extra large We know the greater than/less than relationship between the values—but still no discernible elementst of measurement

The above variables are called discrete or categorical values. The following two types are sometimes referred to as scale or continuous variables—they have indicators of measurement: Think POWER!

Interval responses can be put in ascending or descending order and have measurable distances between them. Example: IQ is an example of an interval variable. If you register 0 on the test, it mean that you have a total absence of intelligence test, but that you really didn't get enough sleep—enough to put your name on the form, but that's about it.

Ratio responses can not only be put in ascending or descending order and have measurable distances between them, but there is a starting point of zero—an absence of the value. Age is a good example of a ratio varialbe. Do you know anyone who is minus 35 years old? Number of children is another example.

Variables and their Relationships

Dependent Variable: the variable to be explained or predicted. It is the result or outcome variable in an equation.

Independent variable(s): the variable(s) that predicts a change to the Dependent Variable, those elements that have an effect on the subject of analysis.

For example, you could decide (for it is <u>your</u> decision) that you want to explore the impact of various sociological factors on income. Income will be your dependent variable. You might select a few independent variables to see which ones have a greater (or less) impact on income. You might select sex as an independent variable, and then you might select years of education. Or age. Or how often the person buys lottery tickets.

Dummy variables.

Well, you know that nominal and ordinal variables (categorical variables) do not contain measurable numbers. The problem is that some students have to do regression analyses and they cannot use categorical data—unless they transform the values associated with the variable into simple indicators.

That is, they substitute a 1 for the presence of a category and zeros for the absence of that category. The key to the power of a dummy variable is that its average or mean is the proportion of one's that it embodies. This proportion multiplied by 100 is equal to the percentage of this category in its original nominal variable. Ask Chuck what a dummy variable looks like in action.