UBC POLI380 SPSS WORKBOOK

cutler@politics.ubc.ca

Table of Contents

Introduction

This workbook consists of exercises and explanations that demonstrate the use of statistical software. We use this software to answer social scientific questions using large datasets. There are many such programs, but the one we will use in this course is SPSS (Statistical Package for the Social Sciences). Doing these exercises is important not only for learning how to use the software, but also for mastering the conceptual content of the course and completing the assignments successfully. It is impossible to succeed in this course without doing the lessons in this workbook.

The workbook uses three different datasets so that you get some practice with different kinds of political science data. One is the Canadian Census of Population and Housing (commonly known as "the census"), where the units of analysis are people. Another is the Canadian Addiction Survey, where the units of analysis are youth in Canada. Another is a survey of Americans during the Presidential election campaign of 2000 (the really close one where George W. Bush beat Al Gore and **changed the course of history…**). The last is a dataset of countries of the world with indicators of each country's economy, society, demographics, and politics.

The workbook contains numerous screen shots of SPSS so that you can check to see that you are doing the analysis correctly.[1] There are a number of conventions used here:

- SPSS commands and variable names appear in ARIAL font
- SPSS menu items are referred to with a dash separating the menu levels. For example to open a file, you will be told to Click File-Open
- Each week's material is divided into numbered sections

[1] This workbook was originally prepared using SPSS v. 12. Some of the screen pictures are from v. 12. Users of SPSS v. 13 through v.18 will find only minor cosmetic differences between what they see on their screen and the v. 12 screenshots in this workbook. For instance, in some dialog boxes one set of horizontal buttons has become vertical, and vice versa.

EXERCISE 1 for Lab 1

Do Exercise 1 on p.20 of K & W. But make sure you pose a research question explicitly (write it down) before you write the paragraph.

Then do exercise 5 but applied to the research question you just wrote about (not a question from Exercise 4, as it says there). If you think there might be a nonlinear relationship, feel free to use low, moderate, and high as points for the independent variable on the x-axis.

Bring this stuff to your tutorial/lab next week and be prepared to spend no more than 2 minutes explaining what you've done. (You're allowed to read it and show the graph.) In the tutorial you'll then discuss as a group how you would apply the material in K & W section 3.2 to your research question and variables. This will be about 5 minutes per student.

LESSON 1 (for lab 2, week of Sept 20) – Data

Objectives: First look at SPSS. SPSS windows. Output file. Relationship between menus and syntax. Data browser. Variable names, labels, value labels.

Commands: RECODE. COMPUTE. DESCRIPTIVES.

1.1 OPEN and LOOK at DATA

The first dataset we'll use is personalized. Each student has a different random sample of a random number of cases (between 50 and 1050) from the Canadian Census Microdata file. That file is itself a sample of 2.8% of all Canadians. Each row in your dataset is a real, randomly selected Canadian. Your file has 27 census variables measuring characteristics of Canadians (the other 2 are just your student number and the number of cases in your dataset) .

You get the data, and all further data in this workbook, from: http://faculty.arts.ubc.ca/fcutler/teaching/POLI380/data

Find the dataset named with your student number. They're in numerical order. (If yours is not there, please email me).
Open it in one of two ways:

1) Just double-click and then choose Open. SPSS should open up with the data visible to you.

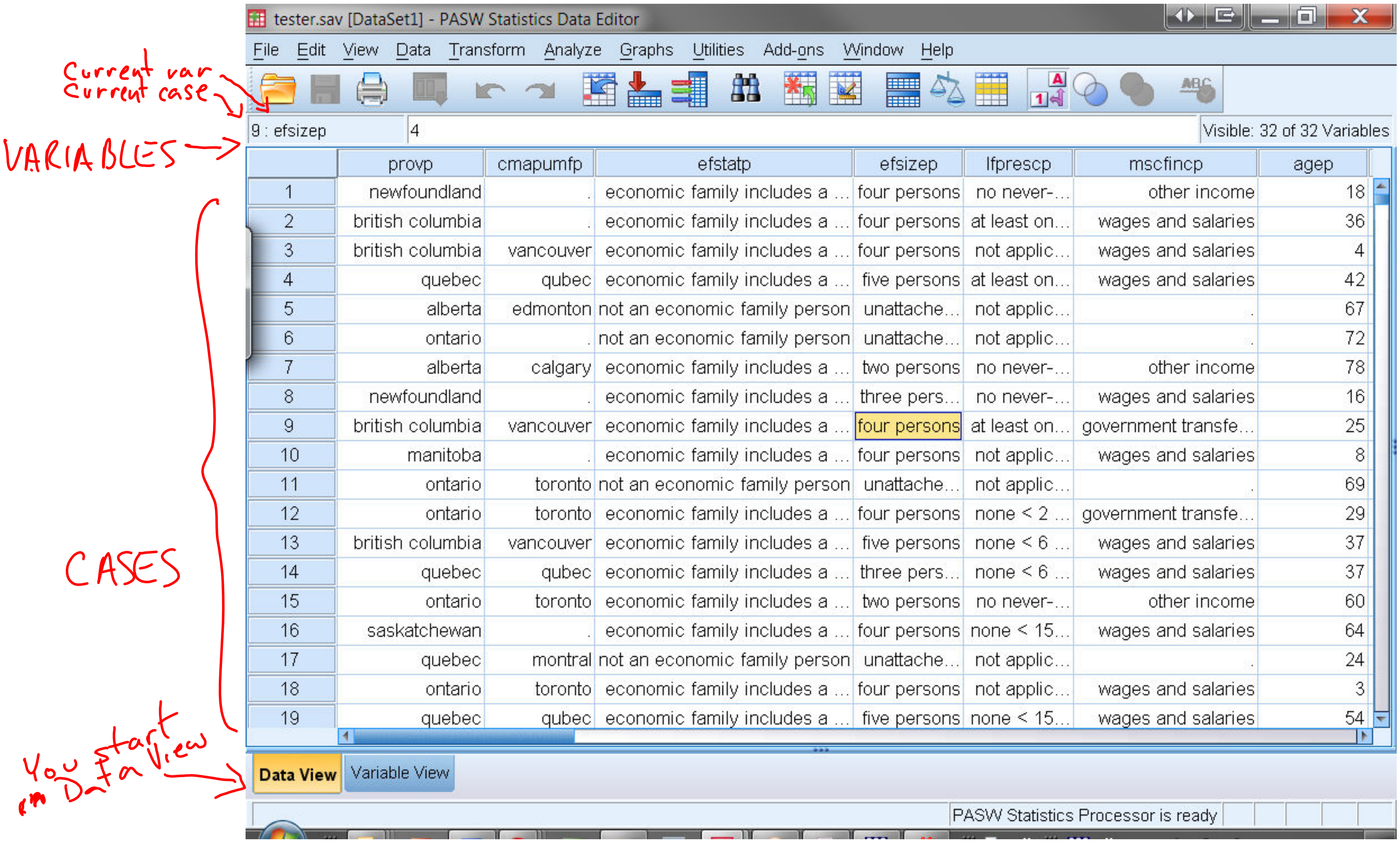
or

2) Right-Click on it. Use Save As to save it somewhere on the machine you're working on (or a USB key or something).
Then, in SPSS, Click File-Open (That means Open in the File menu), find the file, then open it.

Whichever you use, I want you to work with this file more than once, so find some convenient way to store it where you can get at it easily. (A hotmail or gmail file space is probably the best thing). To save the file at the end of the lab, just use Save As from the File menu and then put it in your file space or email it to yourself or something.

First, let's make sure we see variables the same way: Go to Edit-Options. On the General tab (it should come up with that tab visible first), in the top left box called Variable Lists, you should select Display names and Alphabetical.That will mean the program will show you a list of variables alphabetically and using the variable name rather than the label.

SPSS has three important windows: the Data Editor, the Syntax Window, and the Output window. You start by looking at your data in the Data Editor. When you start up you'll see numbers in the data cells, not labels as in this picture. For variables that don't have a natural numerical meaning, these numbers are given labels.



tester.sav [DataSet1] - PASW Statistics Data Editor

9 : efsizep 4

Visible: 32 of 32 Variables

	provp	cmapumfp	efstatp	efsizep	lfprescp	mscfincp	agep
1	newfoundland	.	economic family includes a ...	four persons	no never-...	other income	18
2	british columbia	.	economic family includes a ...	four persons	at least on...	wages and salaries	36
3	british columbia	vancouver	economic family includes a ...	four persons	not applic...	wages and salaries	4
4	quebec	qubec	economic family includes a ...	five persons	at least on...	wages and salaries	42
5	alberta	edmonton	not an economic family person	unattache...	not applic...	.	67
6	ontario	.	not an economic family person	unattache...	not applic...	.	72
7	alberta	calgary	economic family includes a ...	two persons	no never-...	other income	78
8	newfoundland	.	economic family includes a ...	three pers...	no never-...	wages and salaries	16
9	british columbia	vancouver	economic family includes a ...	four persons	at least on...	government transfe...	25
10	manitoba	.	economic family includes a ...	four persons	not applic...	wages and salaries	8
11	ontario	toronto	not an economic family person	unattache...	not applic...	.	69
12	ontario	toronto	economic family includes a ...	four persons	none < 2 ...	government transfe...	29
13	british columbia	vancouver	economic family includes a ...	five persons	none < 6 ...	wages and salaries	37
14	quebec	qubec	economic family includes a ...	three pers...	none < 6 ...	wages and salaries	37
15	ontario	toronto	economic family includes a ...	two persons	no never-...	other income	60
16	saskatchewan	.	economic family includes a ...	four persons	none < 15...	wages and salaries	64
17	quebec	montral	not an economic family person	unattache...	not applic...	.	24
18	ontario	toronto	economic family includes a ...	four persons	not applic...	wages and salaries	3
19	quebec	qubec	economic family includes a ...	five persons	none < 15...	wages and salaries	54

PASW Statistics Processor is ready

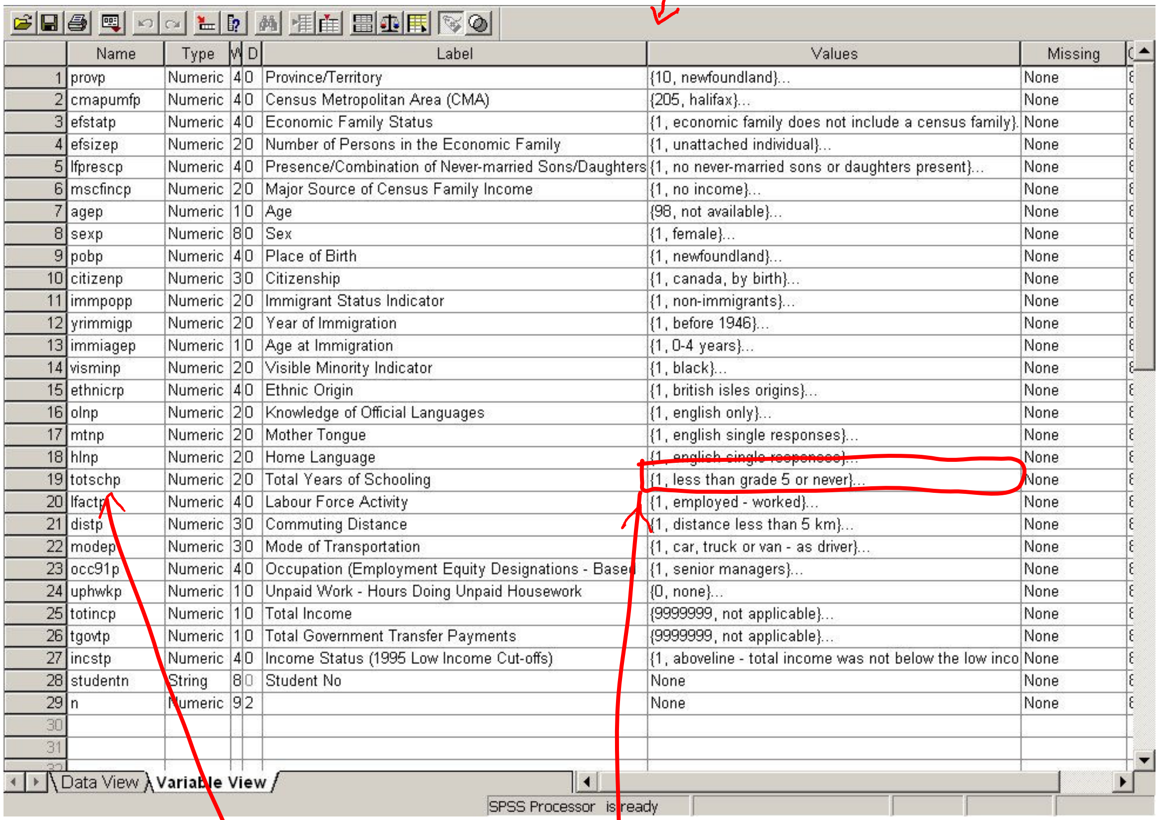
For example, the number for the Vancouver Census Metropolitan Area is 933, and that number is assigned the label "Vancouver" by the folks who made the dataset. To switch back and forth between numbers and labels, go to the View menu and check Value Labels or just click on the toolbar. Once you do, your data should look something like this (but you all have different actual data, so the data in the main part of the window will not be the same as in this example). Look at this example and spend a couple of minutes getting acquainted with looking at your data this way.

The key thing to notice is that the cases or "units of analysis" appear in the rows of the data matrix. Here, the cases are people – REAL CANADIANS! The variables that describe them are the columns.

So, for example, on the screen in this example, case 13 is from Vancouver, is 37 years old, female, born in BC.

Many of the variable names are weird abbreviations because it used to be that variable names could only be 8 characters long. So to find out what each one measures, just hover your mouse pointer over the variable name at the top of a column.

When you're done looking at this, click on the Variable View tab at the bottom. You get a screen like this (though I've narrowed the unimportant columns here and you should widen your columns so you can see the Label and Values like this. Just hold down the mouse on the bars separating the column headings and drag them wider):



	Name	Type	W	D	Label	Values	Missing	C
1	provp	Numeric	4	0	Province/Territory	{10, newfoundland}...	None	8
2	cmapumfp	Numeric	4	0	Census Metropolitan Area (CMA)	{205, halifax}...	None	8
3	efstatp	Numeric	4	0	Economic Family Status	{1, economic family does not include a census family}.	None	8
4	efsizep	Numeric	2	0	Number of Persons in the Economic Family	{1, unattached individual}...	None	8
5	lfprescp	Numeric	4	0	Presence/Combination of Never-married Sons/Daughters	{1, no never-married sons or daughters present}...	None	8
6	mscfincp	Numeric	2	0	Major Source of Census Family Income	{1, no income}...	None	8
7	agep	Numeric	1	0	Age	{98, not available}...	None	8
8	sexp	Numeric	8	0	Sex	{1, female}...	None	8
9	pobp	Numeric	4	0	Place of Birth	{1, newfoundland}...	None	8
10	citizenp	Numeric	3	0	Citizenship	{1, canada, by birth}...	None	8
11	immpopp	Numeric	2	0	Immigrant Status Indicator	{1, non-immigrants}...	None	8
12	yrimmigp	Numeric	2	0	Year of Immigration	{1, before 1946}...	None	8
13	immiagep	Numeric	1	0	Age at Immigration	{1, 0-4 years}...	None	8
14	visminp	Numeric	2	0	Visible Minority Indicator	{1, black}...	None	8
15	ethnicrp	Numeric	4	0	Ethnic Origin	{1, british isles origins}...	None	8
16	olnp	Numeric	2	0	Knowledge of Official Languages	{1, english only}...	None	8
17	mtnp	Numeric	2	0	Mother Tongue	{1, english single responses}...	None	8
18	hlnp	Numeric	2	0	Home Language	{1, english single responses}...	None	8
19	totschp	Numeric	2	0	Total Years of Schooling	{1, less than grade 5 or never}...	None	8
20	lfactp	Numeric	4	0	Labour Force Activity	{1, employed - worked}...	None	8
21	distp	Numeric	3	0	Commuting Distance	{1, distance less than 5 km}...	None	8
22	modep	Numeric	3	0	Mode of Transportation	{1, car, truck or van - as driver}...	None	8
23	occ91p	Numeric	4	0	Occupation (Employment Equity Designations - Based	{1, senior managers}...	None	8
24	uphwkp	Numeric	1	0	Unpaid Work - Hours Doing Unpaid Housework	{0, none}...	None	8
25	totincp	Numeric	1	0	Total Income	{9999999, not applicable}...	None	8
26	tgovtp	Numeric	1	0	Total Government Transfer Payments	{9999999, not applicable}...	None	8
27	incstp	Numeric	4	0	Income Status (1995 Low Income Cut-offs)	{1, aboveline - total income was not below the low inco	None	8
28	studentn	String	8	0	Student No	None	None	8
29	n	Numeric	9	2		None	None	8
30								
31								

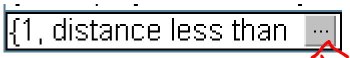
This screen (after I've narrowed the unimportant columns) lists:

the variables' Names, their storage Type, the variable Label to provide more description of what it is, the Values that the variable takes on with the labels for those values,

and a list of which values indicate that the measurement is Missing.

You can change the names and labels of variables just by clicking and editing.

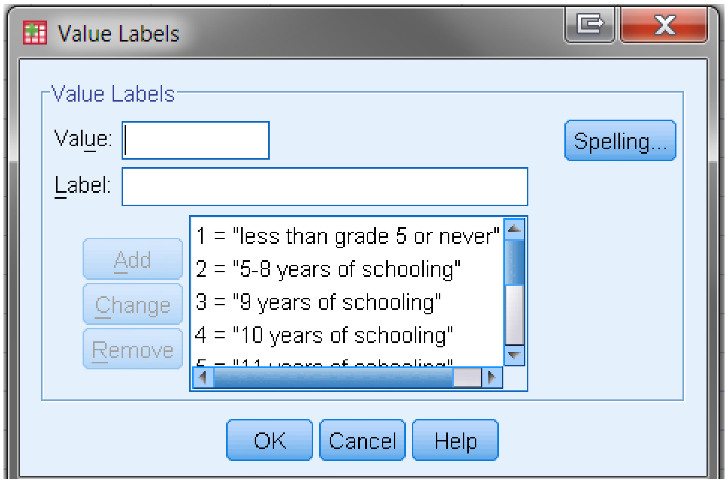
Do that for totschp. Double-Click on totschp and rename it Education.



{1, distance less than ...

When you click on a cell in the Values column and then click on the little three dots thing that appears in the box,

you get another box with the values for the variable and their labels. Click on the Values cell for the variable now labelled Education (Total Years of Schooling) and you get a box like this:



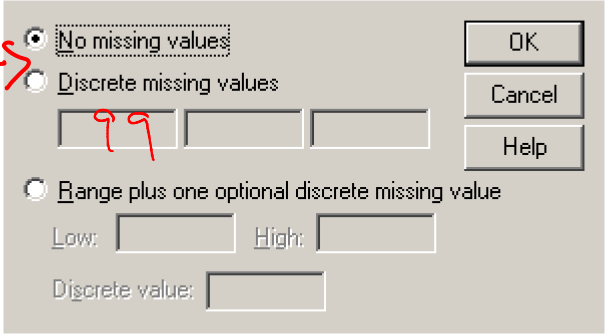
1 = "less than grade 5 or never"
2 = "5-8 years of schooling"
3 = "9 years of schooling"
4 = "10 years of schooling"
5 = "11 years of schooling"

Notice something REALLY IMPORTANT: Even though this is a natural numeric variable (years of schooling), it isn't *coded* that way. Coding is the process of assigning numeric values to the things you have measured. This variable (totschp that you renamed education) takes on the **values** 1 through 9, and these denote *levels* of education.

Scroll down in the little box, find 14-17 years of schooling, and click it. The **Value** is 8. Change the Value Label to "Post-Secondary" (type it in) and then click the Change button.

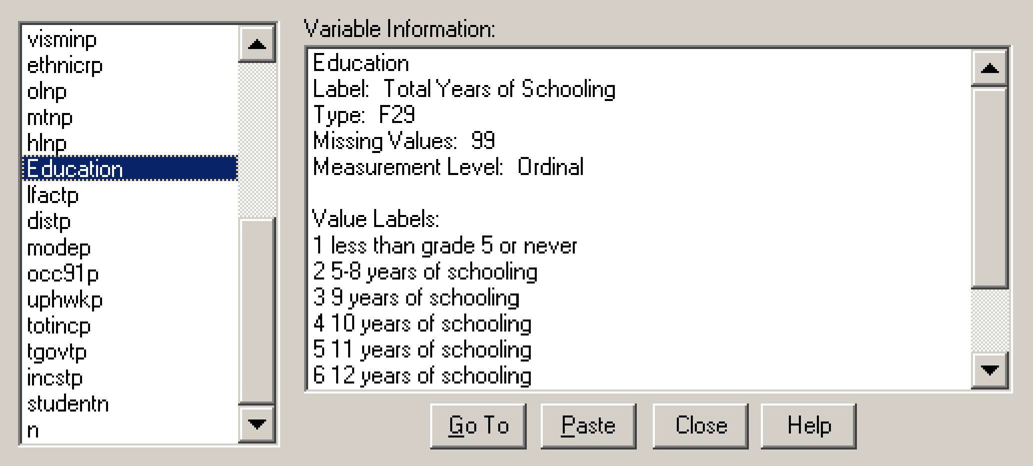
Notice one more thing: the last value is 99! What the heck is that? If a measurement was not obtained for a particular case on a particular variable, it still has to get a value. We usually give missing data a value way outside the range of the data so that it's pretty obvious it is not a real value. But we still have to tell SPSS that it's missing, otherwise the value will be used in the calculation of statistics.
(Obviously, including some 99s would drastically inflate the average education level).

So hit OK to get out of the box and get back to the Variable View. Go over to the Missing column for your Education variable. Click in the box and then click the **...** (three dots) that you see there. You'll get the Missing Values box. Now tell SPSS that 99 is missing for the education variable. Click on the second button (Discrete missing values) and then fill one of the blanks with 99. Then hit OK and make sure that the Variable View shows you that 99 is the missing value for education.



● No missing values
○ Discrete missing values
99
○ Range plus one optional discrete missing value
Low: High:
Discrete value:

Finally, another way to get information on the variables in your dataset is to use the Utilities menu and choose Variables. This brings up a box that gives you all the info about a variable in one handy display. Use it to check what you've done.



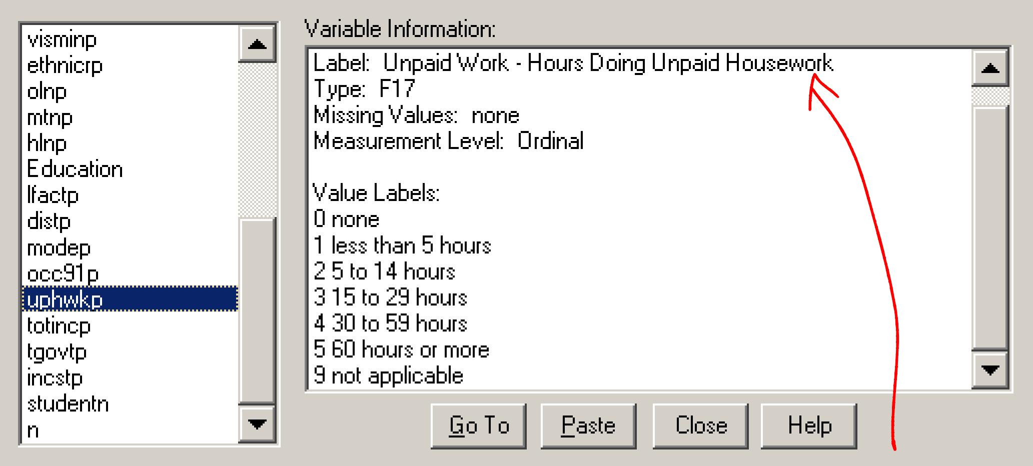
Variable Information:

Education
Label: Total Years of Schooling
Type: F29
Missing Values: 99
Measurement Level: Ordinal

Value Labels:
1 less than grade 5 or never
2 5-8 years of schooling
3 9 years of schooling
4 10 years of schooling
5 11 years of schooling
6 12 years of schooling

1.2 The RECODE Command

Sometimes the values aren't quite what you want. Consider the variable **uphwkp**: Use the Utilities-Variables menu selection to have a look at the variable's label and the range of the values for the variable.



Variable Information:

Label: Unpaid Work - Hours Doing Unpaid Housework
Type: F17
Missing Values: none
Measurement Level: Ordinal

Value Labels:
0 none
1 less than 5 hours
2 5 to 14 hours
3 15 to 29 hours
4 30 to 59 hours
5 60 hours or more
9 not applicable

It's supposed to be the **Hours doing unpaid housework** (see the label), but again the values denote *ranges* of hours. What if you want to get an average number of hours quickly? If you do it with the current StatCan coding, you'll get an average that you have to translate back to real hours. One way to make this variable easier to use is to RECODE the values to correspond to real hours, which is what they're supposed to mean in the first place.

Remember, you can do this for any variable.
You will RECODE stuff whenever you use a program like this to analyze data.

One way to RECODE it into hours would be to simply give each value the midpoint of the range it indicates. To do that, we'd better write down how we want the old values to correspond to new ones (let's round up, except for 60 which seems pretty unlikely anyway):

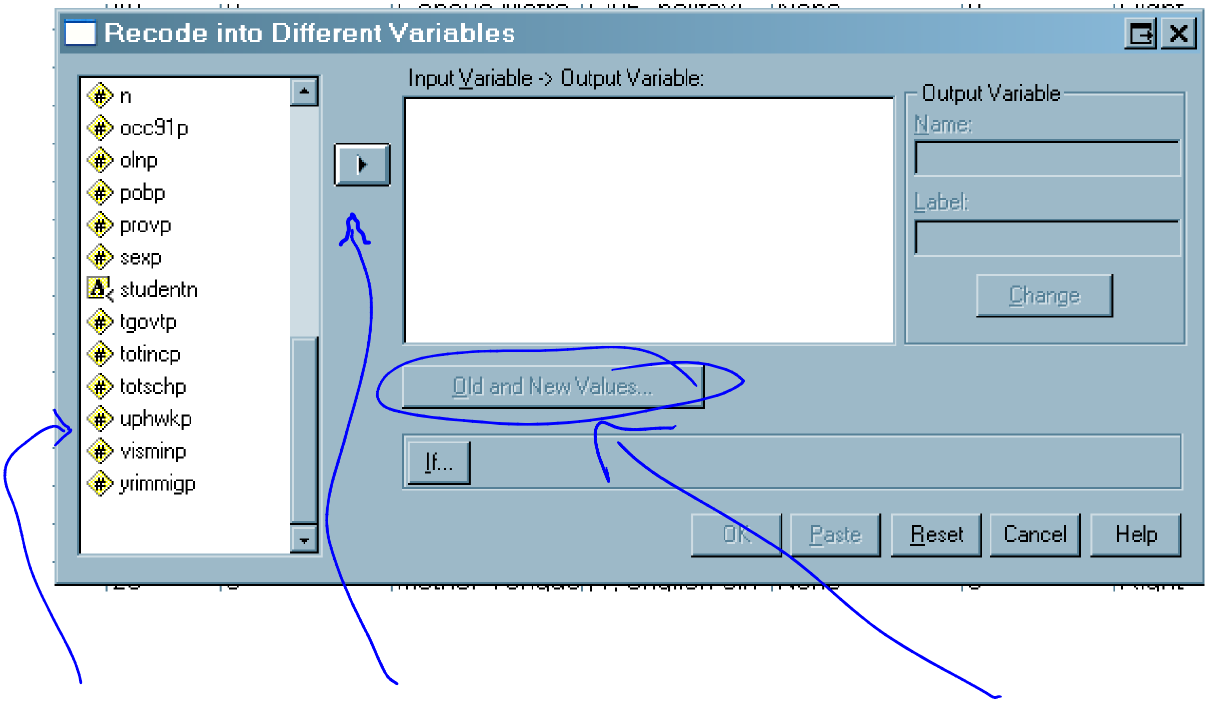
Here are the set of recode parameters in the form: oldcode=newcode

0=0 1=3 2=10 3=22 4=45 5=65 9=missing

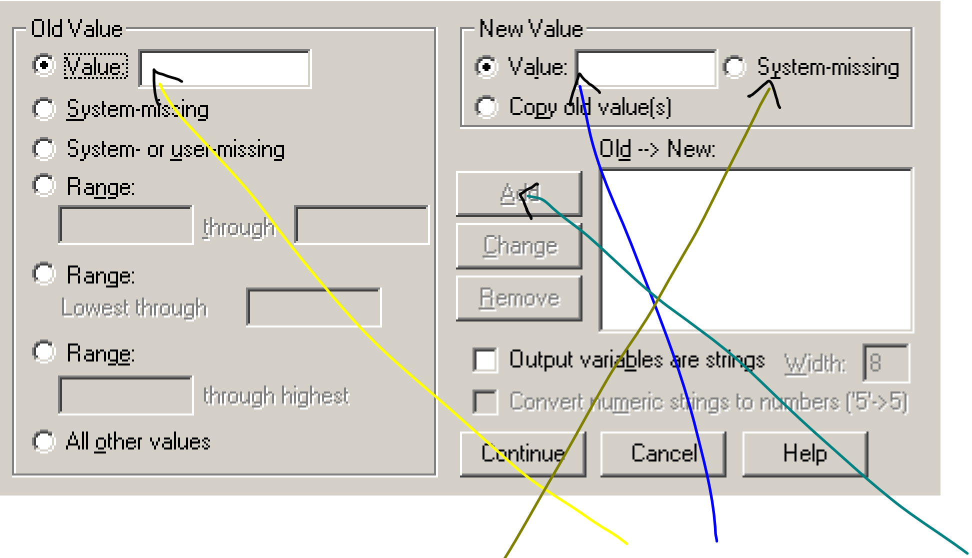
[Notice that we are arbitrarily assigning values within these ranges. But we're not really changing the information contained in the data. We still know that 22 will refer to someone estimating that they do 15-29 hours of unpaid housework in an average week.]

That's how recoding works. You tell the program to change zero to zero, 1 to 3, 2 to 10, and so on. Now, we could change the current variable, but it's probably safer in most cases to generate a new variable with the new codes.

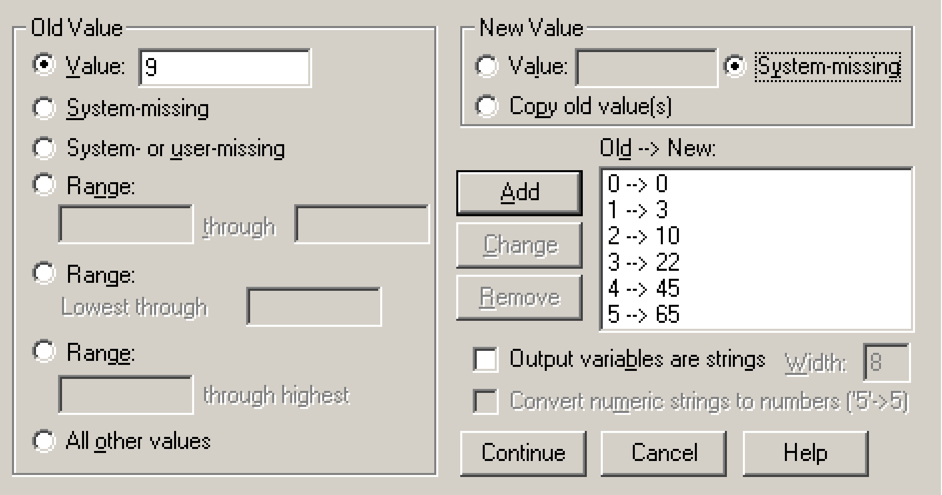
Go to the Transform menu, choose Recode, and then Into Different Variables. Scroll down the list of variables until you find uphwkp.



Select the variable uphwkp and then press the little arrow to put it into the white box. Or just drag it – wow; it took til 2008 for them to let you drag it!
Then click the Old and New Values button. Here's what you get:



Refer to your list of old and new values. Type in 0 in the top-left Value box and 0 in the right-hand value box. Then click Add. You have to do this for each one of your recodes. For 9=missing, type 9 in the first box and then click the System-missing button. Remember to click Add for the last one you do.
Then when your box looks like this:



● Value: 9

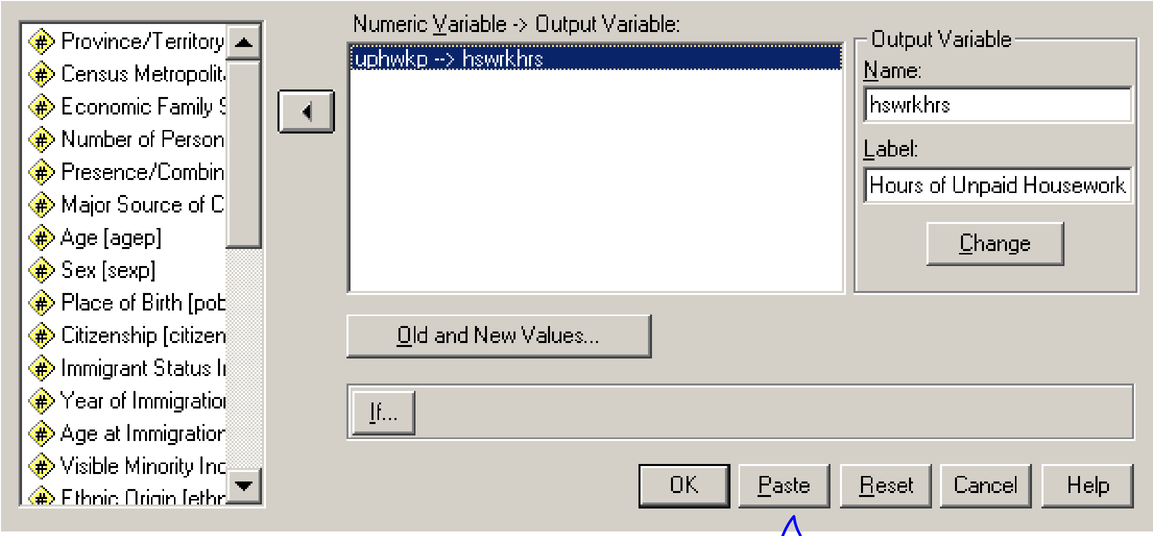
● System-missing

Old --> New:

0 --> 0
1 --> 3
2 --> 10
3 --> 22
4 --> 45
5 --> 65

Click Continue.

Then you're back in the Recode into Different Variables box. Enter the Output Variable as hswrkhrs and label it "Hours of Unpaid Housework". Then click the Change button on the right. (picture on next page).



Numeric Variable -> Output Variable: uphwkp --> hswrkhrs

Output Variable

Name: hswrkhrs

Label: Hours of Unpaid Housework

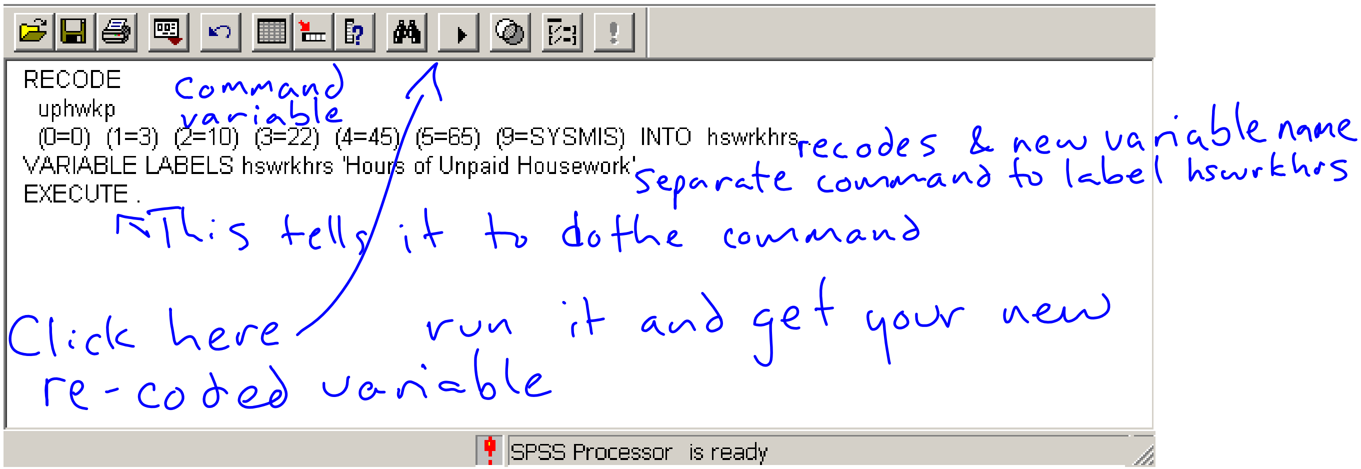
Finally, don't hit OK!!!!!!!!

Instead, hit Paste.

What's this?????? What happened??????

You got the Syntax Window, that's what.

Syntax Window



```
RECODE
  uphwkp
  (0=0)  (1=3)  (2=10)  (3=22)  (4=45)  (5=65)  (9=SYSMIS)  INTO  hswrkhrs .
VARIABLE LABELS hswrkhrs 'Hours of Unpaid Housework'.
EXECUTE .
```

command

variable

recodes & new variable name

Separate command to label hswrkhrs

This tells it to do the command

Click here run it and get your new re-coded variable

SPSS Processor is ready

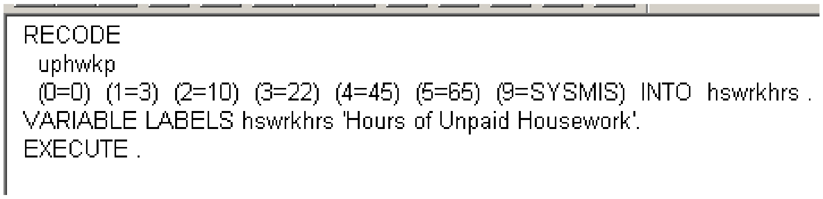
SPSS, like every other program, used to be entirely command-driven. None of this fancy windows and menu stuff back in the day. In fact, a bunch of things are still easier when done as commands rather than through the menus. I think this is one of them.

When you hit Paste, it brought up a separate window called the Syntax Window. It is just a place for commands to be written and then run. The commands you could have issued by hitting OK instead of PASTE appear there. Have a look at my annotations on the Syntax Window above. Notice the periods at the end of each command to tell SPSS that it is the end of the command. (Hey, just like a sentence!)

RECODE is the command, uphwkp is the variable, the parenthetical stuff is the old and new values, and INTO hswrkhrs tells it to put all this in a new variable.

Then the VARIABLE LABELS command gives the new variable a descriptive label.

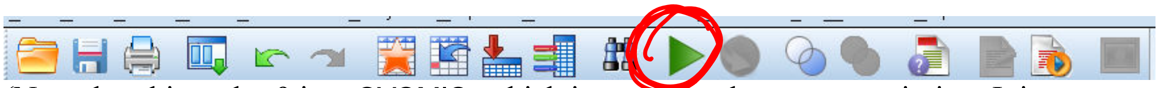
And EXECUTE runs this little program.



```
RECODE
  uphwkp
  (0=0)  (1=3)  (2=10)  (3=22)  (4=45)  (5=65)  (9=SYSMIS)  INTO  hswrkhrs .
VARIABLE LABELS hswrkhrs 'Hours of Unpaid Housework'.
EXECUTE .
```

SO Select All of this text in your syntax window (Ctrl-A is the shortcut to select all text).

And then Click on the green arrow button on the toolbar to run it.

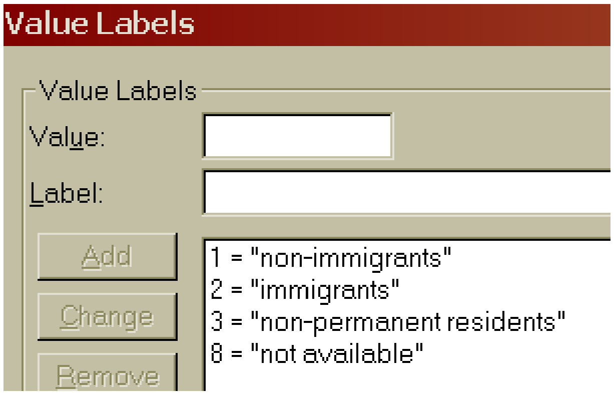


(Note that this codes 9 into SYSMIS, which is not quite the same as missing. It is stronger, but works similarly. It just won't show up as missing in the variable view).

We need to check that this works. To do so, we need SPSS to show us what the variable consists of now. We'll use the Frequencies command. Select ANALYZE-DESCRIPTIVE STATISTICS-FREQUENCIES from the menu. Put hswrkhrs into the blank box in the middle and then hit OK at the bottom. You'll end up in yet another window: the Output window. It's where you'll get the results of anything you ask SPSS to do. (Even errors will show up there.) Now, I hope you'll see two little tables. The first one just tells you how many of the cases (people) in your dataset are "valid" cases for this analysis. For this variable, it's the people for whom the question about housework makes sense.

The second table shows you how many of your cases fall into each of the new categories. For now, all we care about is that those new categories (0, 3, 10, 22, 45, 65) are there. Most of you won't have all of them, but you should have some of them.

Let's do one more quick example with the syntax window. The variable immpopp tells you whether someone is an immigrant. You would think it would be a binary variable (just two values): either immigrant or non-immigrant. Unfortunately, if we were to check its Value Labels (like I did here→), we'd see that it has a third category: "non-permanent residents" (some of you may have none of these folks in your personalized census dataset, however).



1 = "non-immigrants"
2 = "immigrants"
3 = "non-permanent residents"
8 = "not available"

Let's just categorize these people as immigrants, since even though they're not citizens yet, all of them must have been born outside Canada. So let's make a new variable called immigrant which we code 1 for all immigrants and 0 for non-immigrants.

So, to do this, put the following into the syntax window (TAs will explain each line):

```
RECODE  immpopp   (1=0)  (2=1)  (3=1)  (8=SYSMIS)  INTO  immigrant .
VARIABLE LABELS immigrant 'immigrant'.
VALUE LABELS  immigrant  0  'nonimmigrant'   1  'immigrant'  .
FREQUENCIES VARIABLES =immigrant.
EXECUTE .
```

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	nonimmigrant	100	78.7	78.7	78.7
	immigrant	27	21.3	21.3	100.0
	Total	127	100.0	100.0	

You should get something like this (with different numbers, of course!). The FREQUENCIES command is what you just did from the menus. It counted the nonimmigrants and immigrants according to the new classification you made. We'll do more with FREQUENCIES and explain the Valid Percent and Cumulative Percent Columns later on.

In this example, you can see that the 22 immigrants and 5 non-permanent residents in the example above were combined into one value, so we now have 27 classified as immigrants on this variable.

1.3 COMPUTE New Variables

Now we're going to make a new variable that we might need to examine a hypothesis. The new variable is the percentage of the person's income that comes from government (EI, Pension, Welfare, Disability pension, etc). The command we use is COMPUTE. We compute it from variables already in the dataset to make a new one. Notice that the census includes the following two variables:

totincp - Total Income

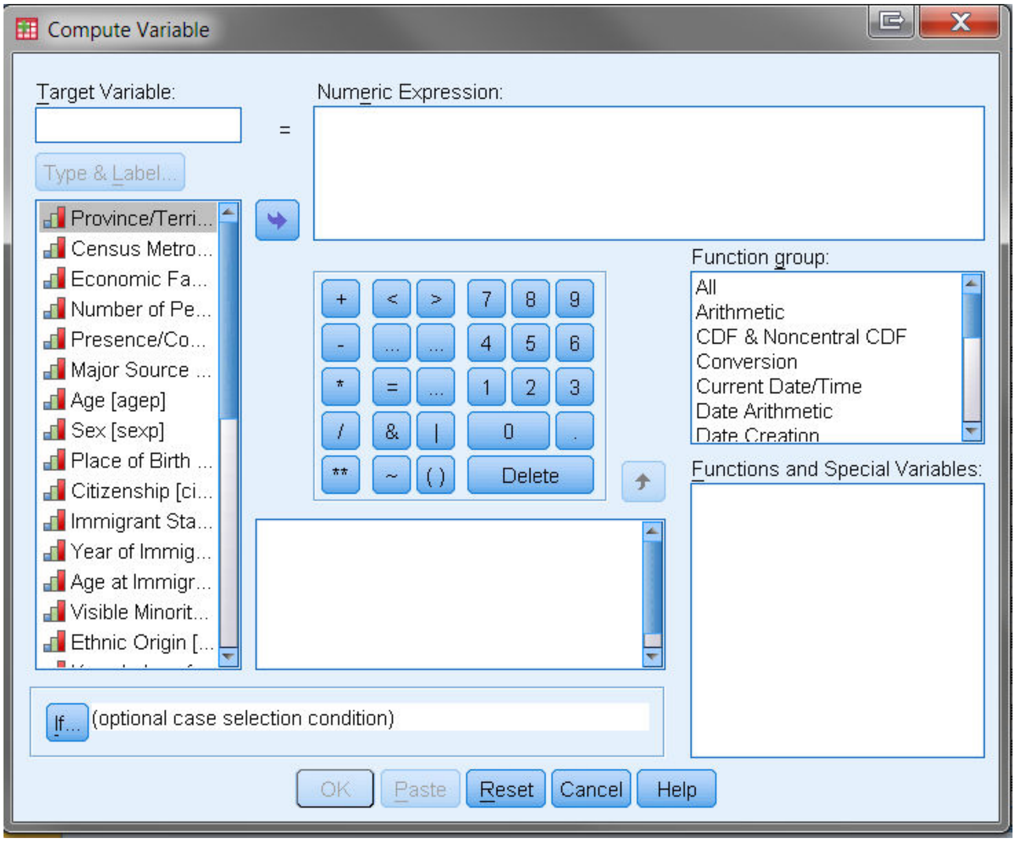
tgovtp – Total Income from Government Tranfers

These aren't very useful on their own if we care about the impact of government. If we want to measure a concept we might call *dependence on government*, we need a variable measuring the proportion of a person's income that comes from government.

What we need is simply tgovtp divided by totincp. $\frac{tgovtp}{totincp}$

Before doing so, you need to click on each variable's values entry in the Variable View screen. As you did earlier, tell SPSS that 9999999 is missing for both variables. Otherwise people with 9999999 on either variable will seem pretty rich!

Then go to Transform – Compute Variable and you get the **Compute Variable** box:



Type a new variable name in the Target Variable box, perhaps govpcinc (my shorthand for government percent of income). Then in the Numeric Expression box just type tgovtp/totincp. (income from government *divided by* total income)

(Or you can click on the variables in the list and then put a division sign in between).

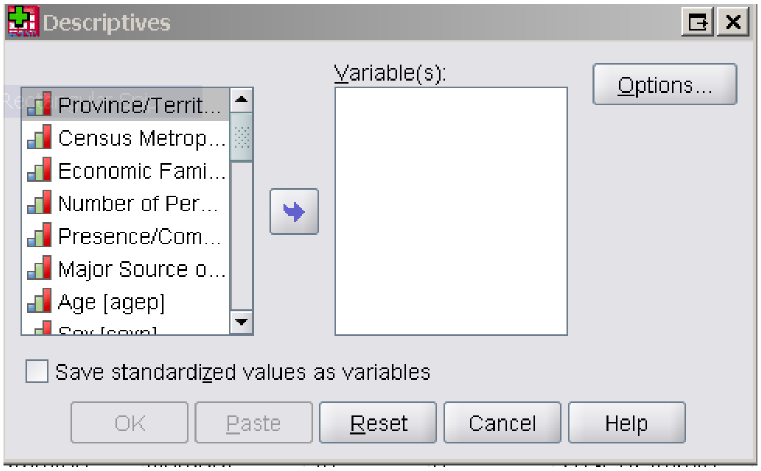
Then hit **Paste** and the **COMPUTE** command will go into the syntax window. Simple as this:

```
COMPUTE govpcinc = tgovtp/totincp .
EXECUTE .
```

Select these two lines and then click the green run button on the toolbar

Then you get the Output Viewer warning you – just making sure you're aware – that for a bunch of cases, the denominator (total income) was zero, so that for those cases the new **govpcinc** variable is set to system missing. This means that you have this govpcinc variable only for those people with income, which is probably what we would want to examine a hypothesis that involved government transfers as a source of income. For instance, it'll now exclude children.

1.4 DESCRIPTIVE Statistics



Now you can do your first simple analysis. We'll just ask for the most basic *summary statistics* from our new variable: govpcinc .

Go to Analyze – Descriptive Statistics – Descriptives

Put your new govpcinc variable into the empty Variables box by selecting it and then clicking the little arrow.

Hit OK.

You will get something like the following in the Output Viewer window:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
govpcinc	554	-.03	1.00	.2922	.39540

Of course, your numbers won't be the same because you have your own personalized sample of Canadians! Everyone has a *different* random sample, of *different sizes*, so you will all have *different* answers to all questions you ask of this data. These are different *sample estimates*, but you're all trying to estimate the same population value (what I'll call the 'true value' or the 'right answer').

N indicates the number of cases that have valid (non-missing) data on this variable.

The **Mean** is just the average proportion (i.e. the 0-to-1 version of a percentage) of total personal income that comes from government tranfers. So in this particular sample, **the average person with income gets 29% of their income from the government**. Think about that phrase. It does not indicate that:

- 29% of people get income from government; or
- 29% of personal income comes from government

Please think about why it means what I said it means:
the average person with income gets 29% of their income from the government

(That seems HIGH!
We will see later on that the mean is a poor summary of the data for this variable. There aren't many average people on this variable.)

1.5 LAST STEPS

Keep the data so you can use it for your next assignment. To do this…

Before you Exit SPSS save your data. File-Save the data to a disk, USB key, or just save it somewhere easy to find and then email it to yourself through my.ubc email or hotmail or gmail or dropbox or whatever. Please try not to lose it, you'll be making changes to the data and you don't want to have to do them over again.

And just save the Syntax so you can look at it if you need it. Call it Lab2 or something

You don't need to save the Viewer Output.

LESSON 2: Summarizing Data (Lab 4, week of Oct 4)

Objectives: Use FREQUENCIES & MEANS to describe your data.

Commands: FREQUENCIES. HISTOGRAM. MEANS.

2.1. FREQUENCIES

If a variable has few enough categories, you can see the whole *distribution* (how many cases fall into each of these categories). This concept of a *distribution* is absolutely crucial to the rest of the course, so make sure you understand it (see lecture notes and textbook).

First, set SPSS so that you can see the values and their labels in your output. Pull down EDIT-OPTIONS and then go to the Output Labels tab. Change the two boxes in each of the boxes to be Names and Labels & Values and Labels respectively (that is, you change all four things on this screen). Then hit OK.

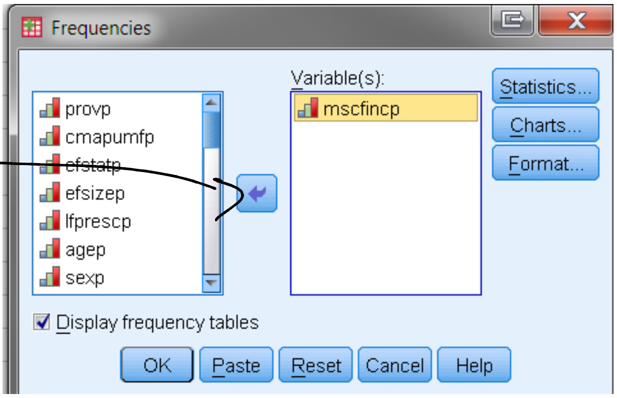
Here's a tip: in any box with variables, you can **right-click** on the variable and choose Variable Information. That'll give you the coding for the variable – what each number means.

Select ANALYZE-DESCRIPTIVE STATISTICS-FREQUENCIES from the menu.

Find the variable mscfincp, labelled "Major Source of Census Family Income". Move it to the Variables box with that silly little arrow. Or, double click.

(hint: when the window comes up and the first variable is highlighted, you can type the first letter of the variable you're looking for to go there quickly).

Run this by clicking OK.



Variable(s): mscfincp

☑ Display frequency tables

Your Viewer window should come up with the **frequency distribution** for this variable something like this (with different numbers, of course).

mscfincp Major Source of Census Family Income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 no income	1	.1	.2	.2
	2 wages and salaries	452	58.6	68.9	69.1
	3 self-employment income	34	4.4	5.2	74.2
	4 government transfer payments	129	16.7	19.7	93.9
	5 investment income	11	1.4	1.7	95.6
	6 other income	29	3.8	4.4	100.0
	Total	656	85.1	100.0	
Missing	System	115	14.9		
Total		771	100.0		

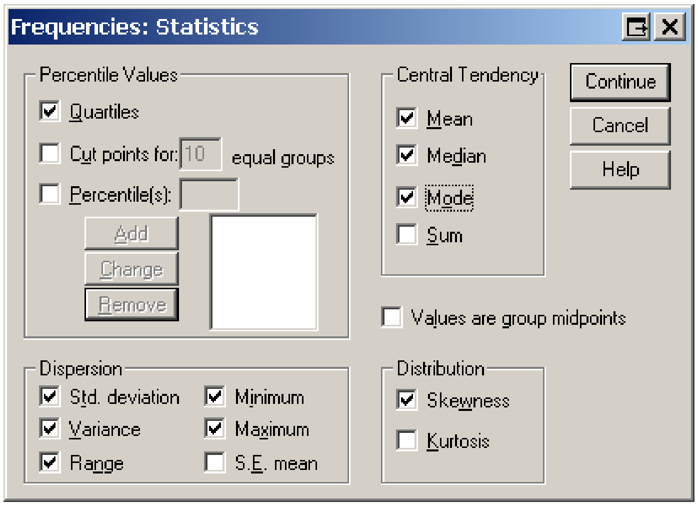
Notice how the variable is *distributed*. Look at the Valid Percent column, which tells you the percentages in each category, ignoring the missing values. For instance, 68.9% of these Canadians have wages and salaries as their principal source of income. In my dataset here, only one person has 'no income'; many of you will have no-one in this category. This is because the value 9 ('not applicable') was set to be SYSTEM MISSING so SPSS would ignore it. This is because the idea of the major source of income doesn't make sense for people with no income. I don't know why there is even this one person here, except that data almost always have some errors and this might be one.

Now let's look at the *distribution* of a continuous variable. Same concept, but we need to think of it differently because a continuous variable has too many different values for us to learn anything by listing them, like Canadians' ages.

Go back to the data editor window by choosing it from the Window menu or getting it from the Windows bar at the bottom of your screen. Then pull up the FREQUENCIES box again. Notice that your mscfincp variable is still there. You have to select it and move it out of the Variables box because you don't want to get its frequencies again.

Find the variable age (agep) and move it to the variable box.

Before you run this, click on the Statistics button at the top right. Then click the boxes for Quartiles, Std. deviation, Variance, Range, Minimum, Maximum, Mean, Median, Mode, and Skewness!


Frequencies: Statistics

Percentile Values
☑ Quartiles
☐ Cut points for: 10 equal groups
☐ Percentile(s):

Central Tendency
☑ Mean
☑ Median
☑ Mode
☐ Sum

☐ Values are group midpoints

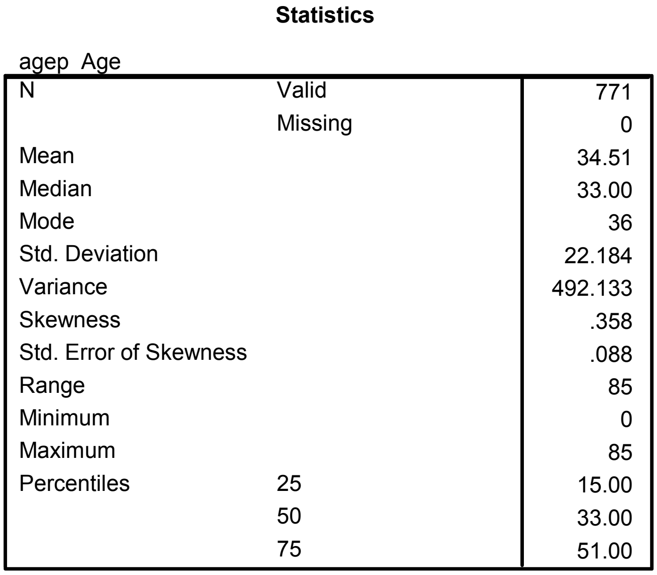
Dispersion
☑ Std. deviation
☑ Variance
☑ Range
☑ Minimum
☑ Maximum
☐ S.E. mean

Distribution
☑ Skewness
☐ Kurtosis

Then hit continue and then OK in the Frequencies box.

Here you get a whole lot of *summary* information about the distribution of age and then the frequencies report for all of the ages in your sample.

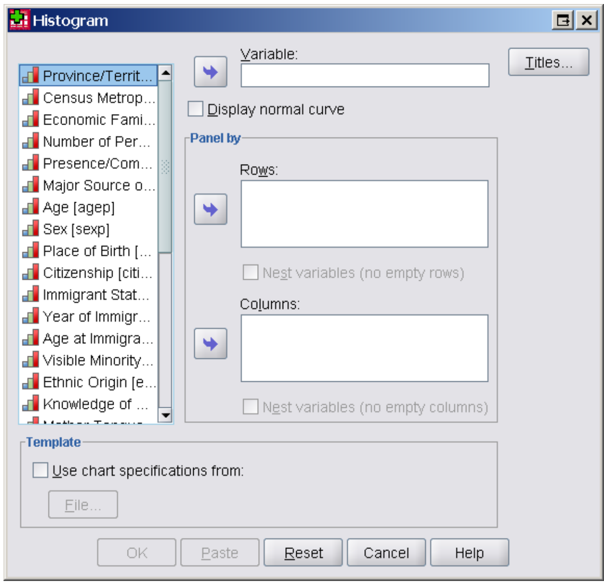
You don't need much more than this to tell you all you need to know about the distribution. You see the mean, minimum, maximum, and the ages of the people who are at the 25th, 50th, and 75th percentile of the ages in your sample. Note, of course, that the 50th percentile and the median are the same thing. Think about that if you're unfamiliar with medians and percentiles.


Statistics

agep Age

N	Valid	771
	Missing	0
Mean		34.51
Median		33.00
Mode		36
Std. Deviation		22.184
Variance		492.133
Skewness		.358
Std. Error of Skewness		.088
Range		85
Minimum		0
Maximum		85
Percentiles	25	15.00
	50	33.00
	75	51.00

2.2 HISTOGRAM

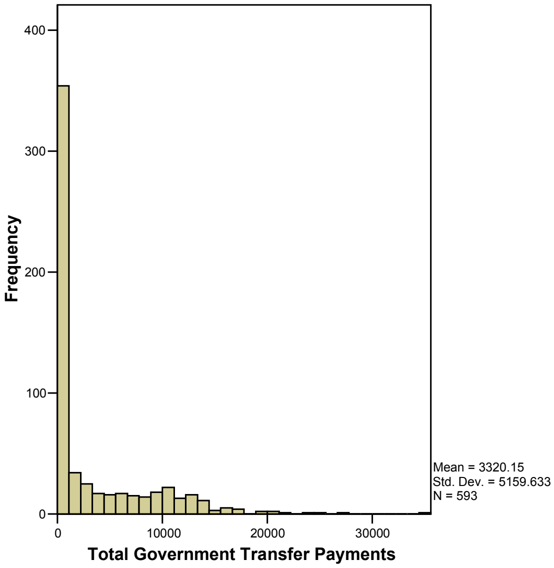


I was going to wait to present graphics, but you can learn a lot about distributions by looking at them. Let's use the most common graph to describe one variable. That is, to describe the ***distribution*** of the variable.

The key graph for this purpose is the HISTOGRAM. Histograms are pictures of distributions, with the values from lowest to highest on the x-axis, and the prevalence of those values (as raw numbers or as percentages) on the y-axis.

It creates 'bins' for certain values of the variable and then the bar shows how many cases fall into those 'bins'. I'm sure you know that – my kids were doing this in Grade 2. (You'd better, it's a big part of the course. We'll use histograms over and over).

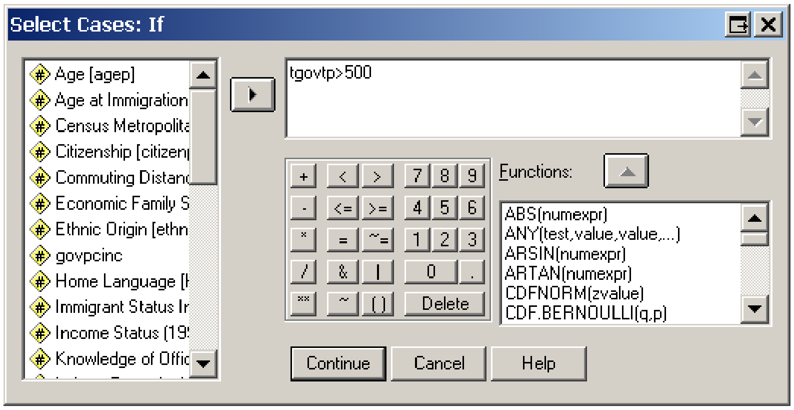
Go to the menu item: GRAPHS-LEGACY DIALOGS-HISTOGRAM.
Put the Total Government Transfer Payments variable (tgovtp) into the Variable box and then click OK.



You'll get something like this. It's that easy to see the whole ***distribution*** of the variable. But perhaps you don't think you're seeing enough. The ZERO category is dominating the graph, so you're not seeing much of the distribution for the people who **do** get government transfer income. What can we do to get just those folks?

We want to select only the cases with some income from government. To do that, we use the SELECT CASES command. It's on the Data menu at the bottom.
Choose DATA-SELECT CASES.
Then in the box that comes up select the second radio button down ("If condition is satisfied") and then click the If... button that shows up.

You get this box. In the white space you put an expression that defines the cases you want to select for analysis. Here, I put in tgovtp>500. That selects those people who got more than $500

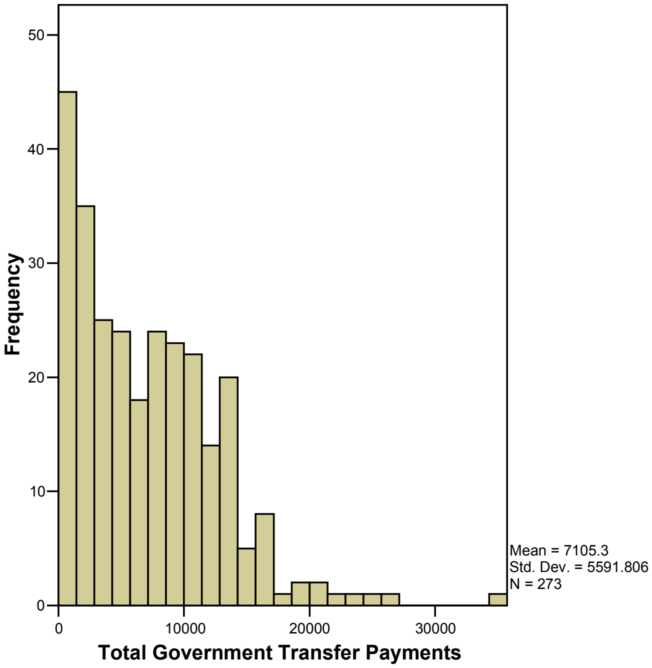


tgovtp>500

in government transfers.

Hit continue. Then hit OK.

If you look in the data editor, the cases that you're not using have a line through the case number.

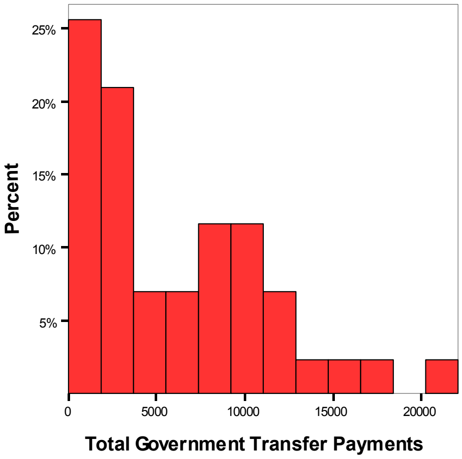


Finally, run your histogram again. GRAPHS-LEGACY DIALOGS-HISTOGRAM. The variable you last used is still in the box, so you just hit OK.

You get something like this. Now you get a much better idea of the ***distribution*** of transfer payments for those with more than $500 of income from government.

This shows you how many people in your sample fall in each of those bins.

IMPORTANT: Histograms can show frequencies (number of cases) OR percent of cases on the y-axis. Below is a graph of the same information with percent. (You'll see how to do this in another lesson).



Finally, make sure you take off the select cases filter.
Go back to DATA-SELECT CASES. Check the All Cases circle and then hit OK.

2.3 MEANS (AVERAGES in different groups)

Finally, we'll switch gears and start to move towards multi-variable analysis, just to get you thinking along these lines.

The MEANS procedure separates the sample automatically (without using the SELECT CASES thing) so you can get averages for different groups. For example, average income for men and for women, or average age for immigrants and non-immigrants. This is the foundation for quantitative social science, so make sure you understand *why* we would do this. Say, you might want to compare the number of persons per capita held without trial by the government in countries receiving development assistance versus countries not receiving assistance.

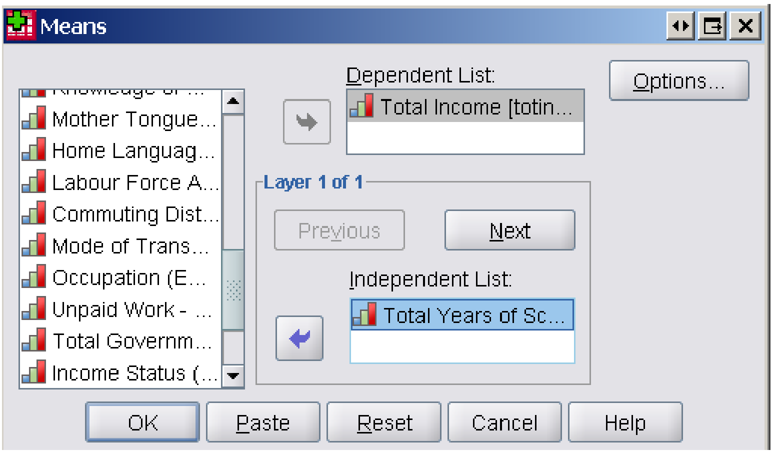
Let's just get the mean and median income for each of the education categories. *Why?* Well, to see whether people with more education have higher incomes. The answer may be obviously "yes", but we would like a more precise answer of how much higher.

Go to ANALYZE-COMPARE MEANS-MEANS.

Notice that you have to fill in two boxes. This is *bivariate* analysis (two variables!).
SPSS calls one the Dependent and the other the Independent.

Put the variable you want averages for in the Dependent List box.

Put the variable you want to split the sample by in the Independent List box.



Dependent List: Total Income [totin...

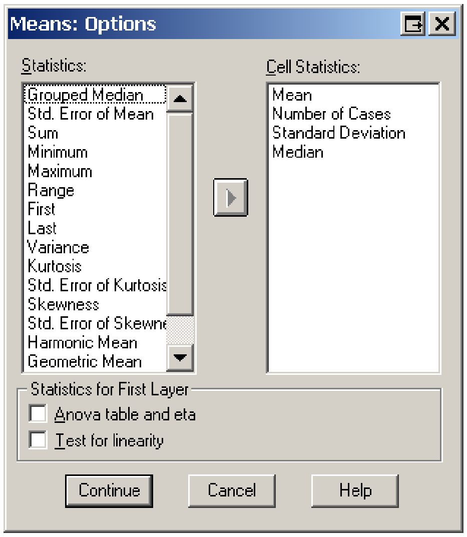
Independent List: Total Years of Sc...

Let's do income by education, as I've done in the picture above. Except your variable name for education is now just "Education".

Before you run it, go to the Options button.

Select Median from the left column and use the little arrow button to put it in the Cell Statistics Box. Notice that you can get a lot of summary statistics from the MEANS procedure.

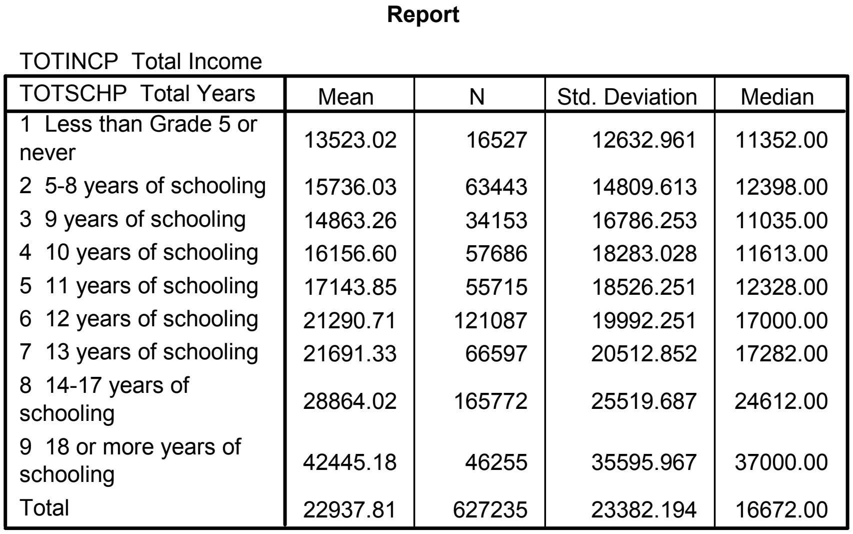
Then Continue. And then OK from the previous box.



Cell Statistics:
Mean
Number of Cases
Standard Deviation
Median

☐ Anova table and eta
☐ Test for linearity

You'll get a report that looks like this:



Report

TOTINCP Total Income

TOTSCHP Total Years	Mean	N	Std. Deviation	Median
1 Less than Grade 5 or never	13523.02	16527	12632.961	11352.00
2 5-8 years of schooling	15736.03	63443	14809.613	12398.00
3 9 years of schooling	14863.26	34153	16786.253	11035.00
4 10 years of schooling	16156.60	57686	18283.028	11613.00
5 11 years of schooling	17143.85	55715	18526.251	12328.00
6 12 years of schooling	21290.71	121087	19992.251	17000.00
7 13 years of schooling	21691.33	66597	20512.852	17282.00
8 14-17 years of schooling	28864.02	165772	25519.687	24612.00
9 18 or more years of schooling	42445.18	46255	35595.967	37000.00
Total	22937.81	627235	23382.194	16672.00

There you go. Means, Standard Deviations, and Medians of income for each educational category. The education categories are in the rows. So: Does more education lead to higher incomes? By how much? Does every year of schooling help?

(Hmmm. Should we limit this to the people who are actually earning employment or self-employment income?? How would we do that? Would it make the gradient steeper? Hmmm.)

LESSON 3 - Getting the Data in Shape for Analysis

Objectives: Practice coding, recoding, computing, use arithmetic functions.

Commands: RECODE. COMPUTE.

In this lab we're going to use American Presidential Election survey data. Sorry, it's for the year 2000, not this week's latest poll. So you're encountering a different dataset, but the principles we've covered so far remain the same. In this data, each row (case) is still a person. But they are different people representing a different population.

This data is from the 2000 election that got George W. Bush into office.
It's a random sample survey of Americans interviewed AFTER the election.

It is called the National Election Study, 2000 and the file is Nes2000depr.sav (at the top or bottom of the list, depending on how it's sorted). The codebook is Nes2000deprCodebook.doc, but it doesn't give you any more than in in the SPSS variable labels. You get these files from http://faculty.arts.ubc.ca/fcutler/teaching/POLI380/data.

You'll probably have to right-click and save each of them to the computer (or your USB drive) and then open it in SPSS. Or, you can try to open it by double-clicking on it.

Right away, go to the file menu and Save As. Save it to a disk or a USB drive or to the My Documents directory on the computer you're using. IF YOU DO THE LATTER, make sure you email it to yourself at the end of the lab so you can use the changed dataset for your next assignment and for next week's lab.

First, take a quick look at the Variable View (remember, bottom left of the SPSS screen). Go to the top of the Label column, put your mouse on the dividing line between Label and Values, and drag the handle to the left so you can see the whole label. Now do the opposite to narrow the Type, Width, and Decimals columns. Have a look at the variable names and their labels. They'll give you a good idea about some of the questions in the survey. Remember that these are all survey questions so each data point is a person's answer to an interview question. 158 Variables!

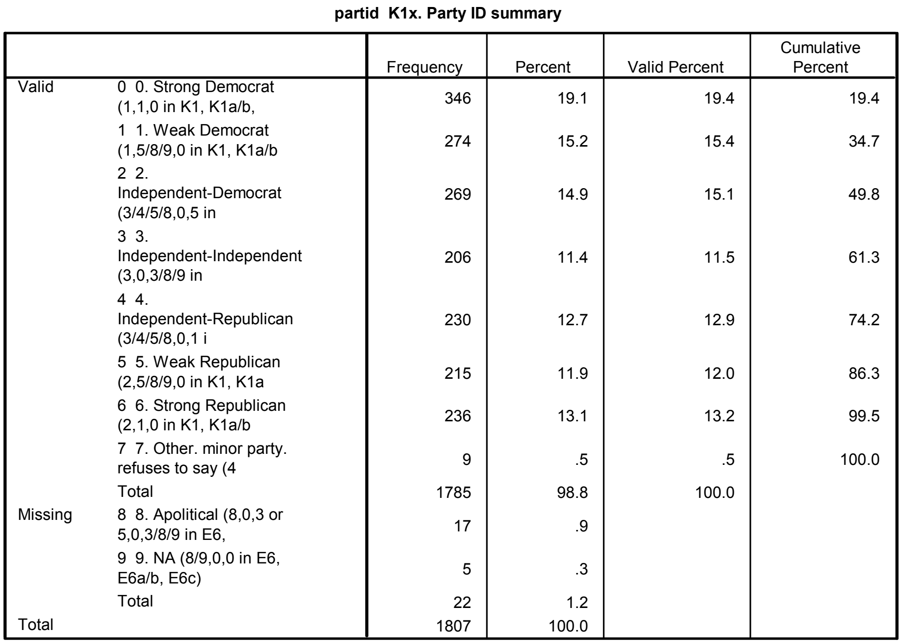
Notice that missing values are already programmed in here. Phew!

3.1 Create Dummy (Indicator) Variables

The first thing we'll do is create three *dummy variables* indicating people's 'party identification'. ("In politics in general, do you usually think of yourself as a Democrat, Republican, or Independent"). A dummy variable is one that has only two possible values; usually zero and one.

Now, you might want variables that indicate each party group so you can compare, for instance, Democrats and non-Democrats or Independents versus everyone else.

First, let's have a look at the party identification variable: partid. This question asks which party the person feels closest to, or a part of.



partid K1x. Party ID summary

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 0. Strong Democrat (1,1,0 in K1, K1a/b,	346	19.1	19.4	19.4
	1 1. Weak Democrat (1,5/8/9,0 in K1, K1a/b	274	15.2	15.4	34.7
	2 2. Independent-Democrat (3/4/5/8,0,5 in	269	14.9	15.1	49.8
	3 3. Independent-Independent (3,0,3/8/9 in	206	11.4	11.5	61.3
	4 4. Independent-Republican (3/4/5/8,0,1 i	230	12.7	12.9	74.2
	5 5. Weak Republican (2,5/8/9,0 in K1, K1a	215	11.9	12.0	86.3
	6 6. Strong Republican (2,1,0 in K1, K1a/b	236	13.1	13.2	99.5
	7 7. Other. minor party. refuses to say (4	9	.5	.5	100.0
	Total	1785	98.8	100.0	
Missing	8 8. Apolitical (8,0,3 or 5,0,3/8/9 in E6,	17	.9		
	9 9. NA (8/9,0,0 in E6, E6a/b, E6c)	5	.3		
	Total	22	1.2		
Total		1807	100.0		

Run a Frequencies (Analyze – Descriptive Statistics – Frequencies) on variable **partid** (in the variable list it appears under K1x). This table (above) is what it should look like.

Notice that partid runs from 0, meaning Strong Democrat to 6 meaning Strong Republican, and even 7 meaning Other, Minority Party, or respondent Refuses to Say.

Open a new Syntax window. Go to FILE-NEW-SYNTAX.

Type in the following code (in bold below). Each time, we generate a new variable (democrat, indep, and gop) equal to 0 for everyone. Then we use IF to change it to 1 if a case satisfies the IF condition in parentheses.

[gop stands for 'Grand Old Party' – The Republicans]

```
COMPUTE democrat=0.
IF (partid>=0 and partid<=2) democrat=1.
COMPUTE indep=0.
IF (partid=3) indep=1.
COMPUTE gop=0.
IF (partid>=4 and partid<=6) gop=1.
EXECUTE.
```

See, we made a new variable democrat that equals zero for EVERYONE in the dataset. Then, *if* partid is 0, 1, or 2, change democrat to equal 1.
And so on for the Republicans and Independents.

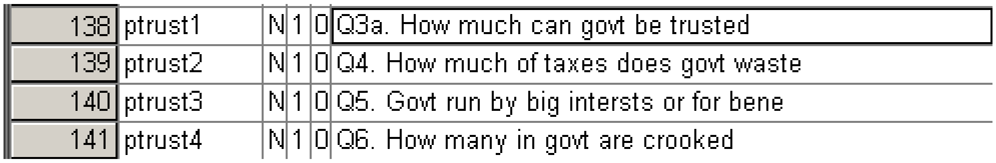
Now run a frequencies on all three of these variables and they should match the frequencies you saw in the partid variable. [hint: you can put all three of the variables in the frequencies window at once to get output for all three frequencies at once].

So you just created three new variables. Go and look at them in the variable view of the data editor; they appear at the bottom. NOTE that there are no value labels or a variable label or missing values. You can fill the labels in. There will not be missing values because the zero group for each variable includes all of the other party identifiers AND all the missing values. Think about that for a second or two, maybe three.

3.2 Create an Additive Index: add up a bunch of variables

Now, let's create a summary index of a person's trust in government built from four questions: ptrust1 through ptrust4. The concept we're measuring is a citizen's trust in government, but we're using four measurements (variables) to do so by combining them into one variable. This is tricky – you may have to spend more time thinking here.

An index is "a method of accumulating scores on individual items to form a composite measure of a complex phenomenon. An index is constructed by… combining the scores for each observation (person) across all the items (questions)".
Here are the questions we can use.



138	ptrust1	N	1	0	Q3a. How much can govt be trusted
139	ptrust2	N	1	0	Q4. How much of taxes does govt waste
140	ptrust3	N	1	0	Q5. Govt run by big intersts or for bene
141	ptrust4	N	1	0	Q6. How many in govt are crooked

All these things plausibly measure trust in government, so we're combining them to get an overall measure or "index".

First, do a Frequencies on all of them to get a sense of the distribution of each one. Stop and think. Have a look at each one of them; interpret the frequencies in your head. [TAs, please pause so students can do this].

Now, if you're building an index, and you're going to combine some variables by adding them up, you'd better make sure they all run in the same direction and they're coded somewhat similarly.

!@$)(%&)$! These ones aren't (and there's a good reason we'll talk about next week!). In the first one, higher values indicate less trust; in the other three questions higher values indicate more trust. Which do you think is more intuitive? Furthermore, the first one is coded 1 through 4, while the last three are coded 1 through 5. So let's go with the majority: positive means higher trust and we'll use the 1 to 5 scale. So we only need to change ptrust1 so that it matches the others. Recode ptrust into a new variable called ptrust1a.

```
RECODE ptrust1 (4=1) (3=2) (2=3) (1=4) into ptrust1a.

EXECUTE.
FREQUENCIES VARIABLES ptrust1a.
```

Now, 4 indicates that government can be trusted "Just about Always".

So now we can just add up these variables to make an index from 4 to 19. (19 is because the maximum values are 4,5,5,5.) Put the following commands in the syntax window, highlight them, and run this mini-program with the little arrow button. NOTICE that the first variable in the list is ptrust1***a*** !

```
COMPUTE trust=ptrust1a+ptrust2+ptrust3+ptrust4.
EXECUTE.
FREQUENCIES VARIABLES trust.
```

trust

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4.00	11	.6	.8	.8
	5.00	285	15.8	19.6	20.4
	6.00	71	3.9	4.9	25.3
	7.00	224	12.4	15.4	40.7
	8.00	107	5.9	7.4	48.0
	9.00	148	8.2	10.2	58.2
	10.00	103	5.7	7.1	65.3
	11.00	85	4.7	5.8	71.2
	12.00	96	5.3	6.6	77.8
	13.00	76	4.2	5.2	83.0
	14.00	138	7.6	9.5	92.5
	15.00	26	1.4	1.8	94.3
	16.00	56	3.1	3.9	98.1
	17.00	12	.7	.8	99.0
	18.00	13	.7	.9	99.9
	19.00	2	.1	.1	100.0
	Total	1453	80.4	100.0	
Missing	System	354	19.6		
Total		1807	100.0		

There's your distribution of the new trust index, scaled 4 to 19. Think about how you got this. Now this isn't a very nice measurement, 4 to 19, is it? You don't want to have to write up in a report something like: "Measured on a scale of 4 to 19, Americans' average trust in government is 8.64". So what to do?

Many variables with strange measurement scales get *rescaled* to run from 0 to 1. This does nothing to our conclusions. It doesn't change the *information* contained in the variable, it just means we can think of the variable more clearly, particularly if we are presenting averages of the variable.

We can do this for this new trust index, trust,very easily by subtracting 4 to get the lowest value equal to zero. Think about that. Then divide by the new maximum, which is 15, to squish it down to between zero and one.
So do it and then run another frequencies, all at once, like this:

```
COMPUTE trust=(trust-4)/15.
EXECUTE.

FREQUENCIES VARIABLES trust.
```

The frequencies will look the same, but the values now run from 0 to 1.

If you want a histogram of this variable to have a look, just put this in the syntax window and run it:

```
GRAPH BAR(SIMPLE)=PCT BY trust .
```

Or do this from GRAPHS-LEGACY DIALOGS-HISTOGRAM
You can use this command for any variable and just replace trust with the one you want. It might be quicker than using the pull-down menu.

3.3 More Recoding Into New Variables Practice

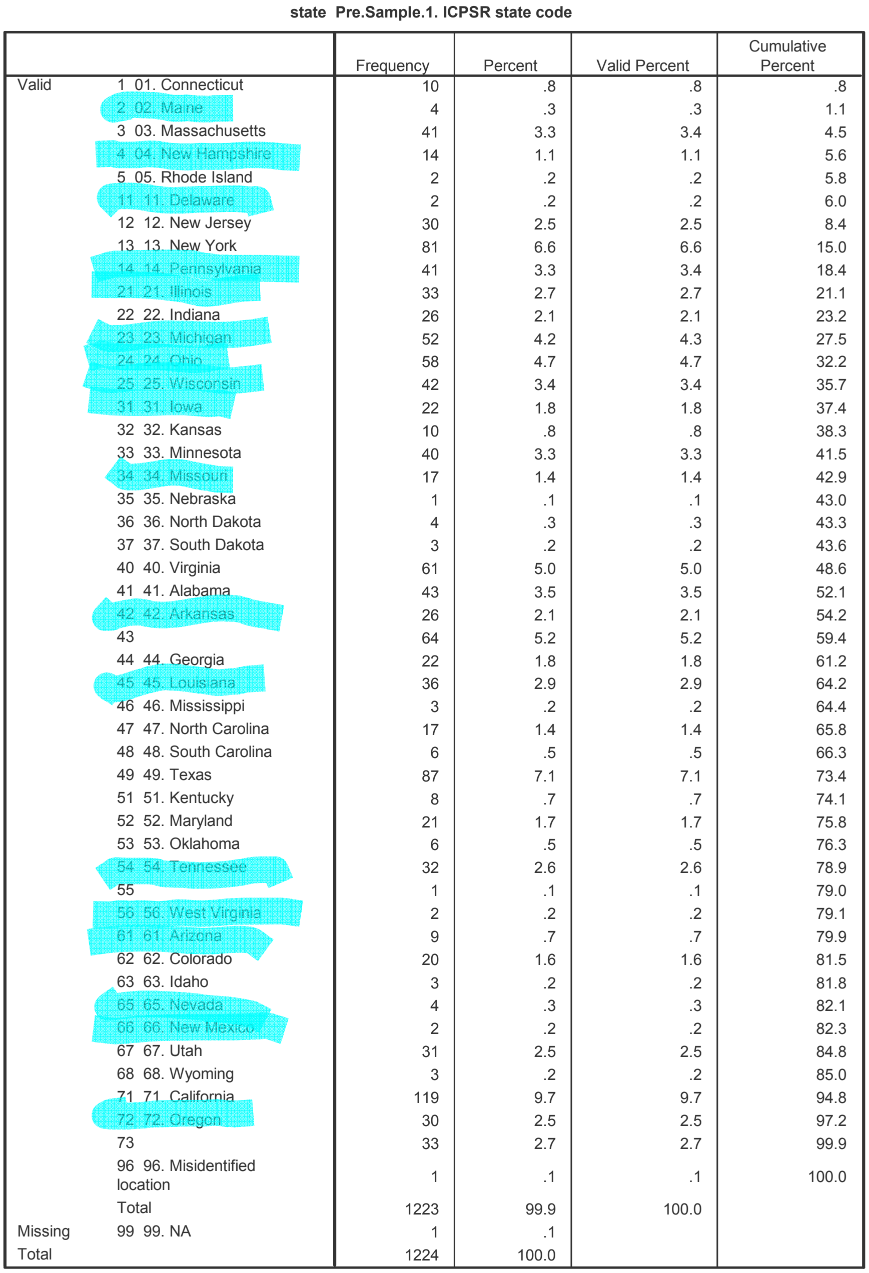
Let's create a variable indicating whether the respondent was in a 'battleground state' – a state that was considered close enough that the candidates poured campaign resources (adverstising, organization) into it. Just FYI: Some states had real campaign activity while others were abandoned by *both* sides because they concluded that the outcome in that state was a certain win for one candidate or the other.

This recoding is going to be *tedious* because of the range of values we need. The states were (according to CNN): WA, OR, NV, NM, AZ, WI, IL, IA, MO, AR, LA, TN, MI, OH, WV, PA, FL, DE, ME, NH.

The easiest way to start is to run a FREQUENCIES on State. This will give you the number codes for these states.

So then we need to do the recode. Try it through the menu.

Use TRANSFORM-RECODE-INTO DIFFERENT VARIABLES.

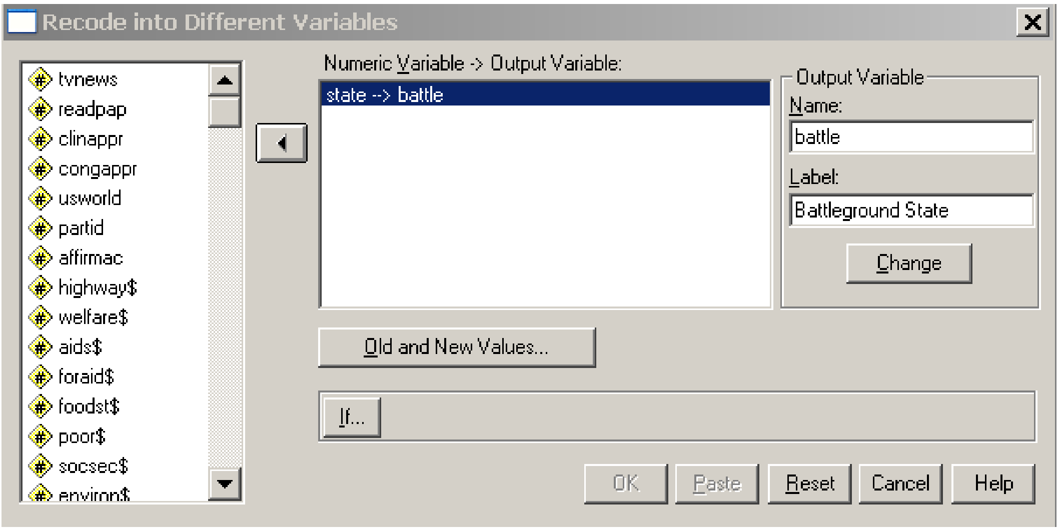


state Pre.Sample.1. ICPSR state code

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 01. Connecticut	10	.8	.8	.8
	2 02. Maine	4	.3	.3	1.1
	3 03. Massachusetts	41	3.3	3.4	4.5
	4 04. New Hampshire	14	1.1	1.1	5.6
	5 05. Rhode Island	2	.2	.2	5.8
	11 11. Delaware	2	.2	.2	6.0
	12 12. New Jersey	30	2.5	2.5	8.4
	13 13. New York	81	6.6	6.6	15.0
	14 14. Pennsylvania	41	3.3	3.4	18.4
	21 21. Illinois	33	2.7	2.7	21.1
	22 22. Indiana	26	2.1	2.1	23.2
	23 23. Michigan	52	4.2	4.3	27.5
	24 24. Ohio	58	4.7	4.7	32.2
	25 25. Wisconsin	42	3.4	3.4	35.7
	31 31. Iowa	22	1.8	1.8	37.4
	32 32. Kansas	10	.8	.8	38.3
	33 33. Minnesota	40	3.3	3.3	41.5
	34 34. Missouri	17	1.4	1.4	42.9
	35 35. Nebraska	1	.1	.1	43.0
	36 36. North Dakota	4	.3	.3	43.3
	37 37. South Dakota	3	.2	.2	43.6
	40 40. Virginia	61	5.0	5.0	48.6
	41 41. Alabama	43	3.5	3.5	52.1
	42 42. Arkansas	26	2.1	2.1	54.2
	43	64	5.2	5.2	59.4
	44 44. Georgia	22	1.8	1.8	61.2
	45 45. Louisiana	36	2.9	2.9	64.2
	46 46. Mississippi	3	.2	.2	64.4
	47 47. North Carolina	17	1.4	1.4	65.8
	48 48. South Carolina	6	.5	.5	66.3
	49 49. Texas	87	7.1	7.1	73.4
	51 51. Kentucky	8	.7	.7	74.1
	52 52. Maryland	21	1.7	1.7	75.8
	53 53. Oklahoma	6	.5	.5	76.3
	54 54. Tennessee	32	2.6	2.6	78.9
	55	1	.1	.1	79.0
	56 56. West Virginia	2	.2	.2	79.1
	61 61. Arizona	9	.7	.7	79.9
	62 62. Colorado	20	1.6	1.6	81.5
	63 63. Idaho	3	.2	.2	81.8
	65 65. Nevada	4	.3	.3	82.1
	66 66. New Mexico	2	.2	.2	82.3
	67 67. Utah	31	2.5	2.5	84.8
	68 68. Wyoming	3	.2	.2	85.0
	71 71. California	119	9.7	9.7	94.8
	72 72. Oregon	30	2.5	2.5	97.2
	73	33	2.7	2.7	99.9
	96 96. Misidentified location	1	.1	.1	100.0
	Total	1223	99.9	100.0	
Missing	99 99. NA	1	.1		
Total		1224	100.0		

First, you have to choose the state variable and then indicate the new variable name and label. Let's call the variable battle.

It should look like this.



Recode into Different Variables

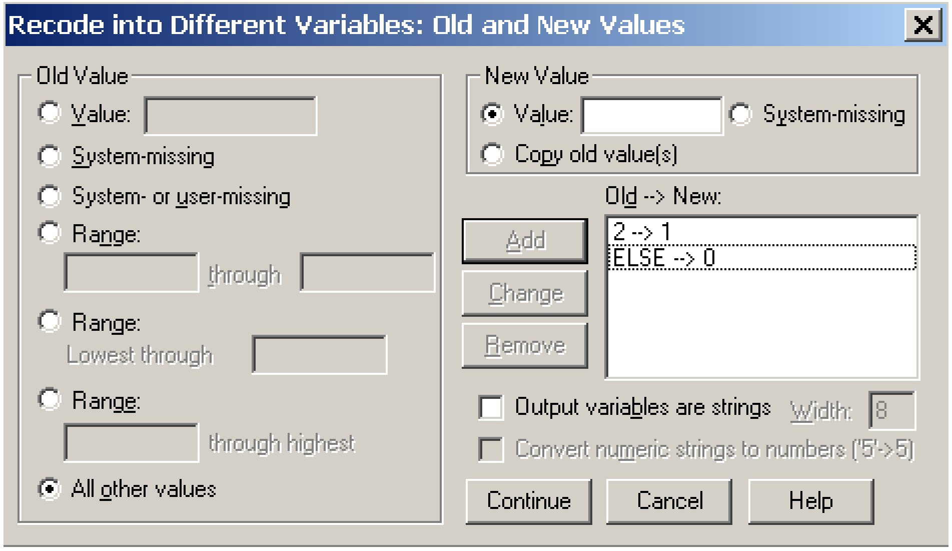
Numeric Variable -> Output Variable: state --> battle

Output Variable

Name: battle

Label: Battleground State

Then click on Old and New Values to type in your recodes. But just do a couple. Make 2 equal to 1 and All other values (bottom left of box) equal to 0. Each time, type the Old Value into the left column and the new value into the New Value box.



Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

○ System- or user-missing

○ Range: through

○ Range: Lowest through

○ Range: through highest

● All other values

New Value

● Value: ○ System-missing

○ Copy old value(s)

Old --> New:

2 --> 1
ELSE --> 0

□ Output variables are strings Width: 8

□ Convert numeric strings to numbers ('5'->5)

Then hit Continue and then Paste. Go to the Syntax Window (from the Window Menu) and your Syntax should look like this:

```
RECODE
  state
  (2=1)  (ELSE=0)  INTO  battle .
VARIABLE LABELS battle 'Battleground State'.
EXECUTE .
```

So we got the menu to tell us how to put the command in, but we haven't done all the values. You need to add a bunch of recodes like the (2=1) there, and it's probably easier to do that in the syntax. That only did Maine (i.e. state coded 2). Now let's do the rest. The values are 2,4,11,14,21,23, 24, 25, 31, 34, 42,45,54,56,61,65,66,72.

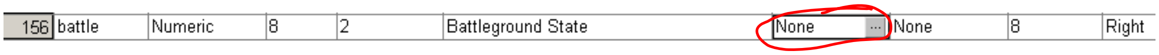
So you need to make the RECODE command look like this before you run it:

```
RECODE state
  (2,4,11,14,21,23, 24, 25, 31, 34, 42,45,54,56,61,65,66,72=1) (ELSE=0)  INTO
battle .
VARIABLE LABELS battle 'Battleground State'.
EXECUTE .
```

So all these values become 1. Notice what the (ELSE=0) statement does. It takes all other values and sets them to 0.

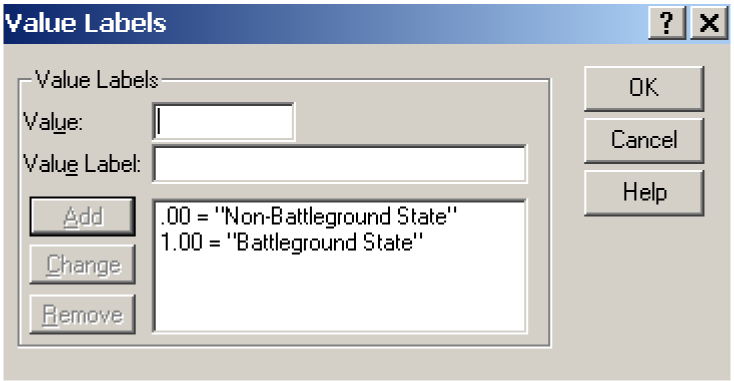
Quite often, recoding *is* as tedious as this!

OK, now you can highlight this command and run it with the green run button.



156	battle	Numeric	8	2	Battleground State	None ...	None	8	Right

Then go back to the Variable View of the Data Editor and enter value labels. This battle variable is now the last on the list. Go to it and click on the three dots at the right side of the None entry in the Values column for this variable.



.00 = "Non-Battleground State"
1.00 = "Battleground State"

Add two value labels to the variable like so…

And then hit OK.

Finally, check how it turned out by running a FREQUENCIES on the battle variable.

What you did is REALLY COMMON. I've done this literally thousands of times. You took a variable with a lot of different values and make it a binary or "dummy" variable.

Now, are there differences between people in battleground and non-battleground states, as we might expect?
Run a Means (Analyze-Compare Means-Means) comparing the variable you generated (trust) in battleground and non-battleground states.
IS there any DIFFERENCE? Figure out what you would say about this relationship.
Oh well, welcome to the sometimes-disappointing world of social science.

Y'know what else? You could have done this recoding with

```
COMPUTE battle=0.
IF (state=2 or state=4 or state=11 or state=…..) battle=1.
EXECUTE.
```

See why?

LESSON 4 – Sampling Demonstration

Objectives: See Sampling in Action.

Commands: Select IF. Frequencies.

First of all, there's no point in doing this lab unless you've read KW Chapter 7 carefully. Twice. Second time more carefully than the first! Read it. And while you're at it, the coincidentally numbered Chapter SEVEN in *Drunk*.

To recap the concept of sampling:

1. In an ideal world you'd be able to collect data on every case to which your question applies. If it's about the Canadian public, for example, you'd get information from 31 million Canadians. You'd get 31 million (the population size) measurements on the variable you're measuring. If you did, you could just calculate the mean or the variance or whatever and not worry about any of this stuff that we think about when we have to take a sample.

2. In a less ideal world, but still totally impossible, you wouldn't be able to interview every case but you would be able to take an infinite number of repeated samples of the same size from the population. Youd take one sample of, say, 1000, and then another and then another and so on. For each one, you'd calculate the average of whatever you're trying to measure. If we then graphed the results of the averages from these repeated samples, given what we know about the properties of the sampling distribution, the average of these averages would be exactly equal to the average from number 1 above. That is, the average for the whole population. And, we would know how spread out the averages would be, so we could say how often we'd get a sample with an average a certain distance from the true population average. For example, that we'd get a sample with an average more than two standard deviations away from the mean in only 5% of the samples of this size.

NOTE that ***This is a hypothetical, a thought experiment, a theoretical concept. No-one ever takes real repeated samples because in fact that would just be one big sample and it would end up not being a sample at all because you'd eventually get every member of the population***.

3. In a less ideal, and still nearly impossible world, you'd just have one sample of a given size and so you'd calculate an average. But the nearly-impossible thing is you would actually know the *population* variance – how spread out are the values in the population. So you'd be able to say how spread out your averages would be if you were to take all possible samples (an infinite number of samples) of that size.

4. In the **real world**, you have just the one sample and you don't really know how spread out are all the values of your variable in the population. You use the mean in your sample to estimate the mean in the population. But you admit it's just an estimate, that can be close to (more probable) or a long way (less probable) from that true mean in the population. But how far away, with what probability? You use the variation of the values of your variable in your one sample to estimate how spread out are all the values in the population. You plug this into the formulas below (and on p.126 onwards in KW), and you use what we know about the *normal distribution*, so that you can say how far away

your one-sample average would be, with a given probability. For example: Using a sample of 1000 Canadians, my estimate of the average age of Canadians is 47.3 years, and I expect that I would be within ± 2.1 years of the true average age in 95% of samples; or further than ± 2.1 years in 5% of samples.

This lab will just demonstrate the fact that averages (and other statistics) will be different in different samples.

In this lab each student will take a random sample from a large dataset. You'll get means and standard deviations for a variable and then report these to the TA. The TA will enter these in a spreadsheet and then graph them. If there aren't very many of you, you can do two or three each so you get 20-30 means.

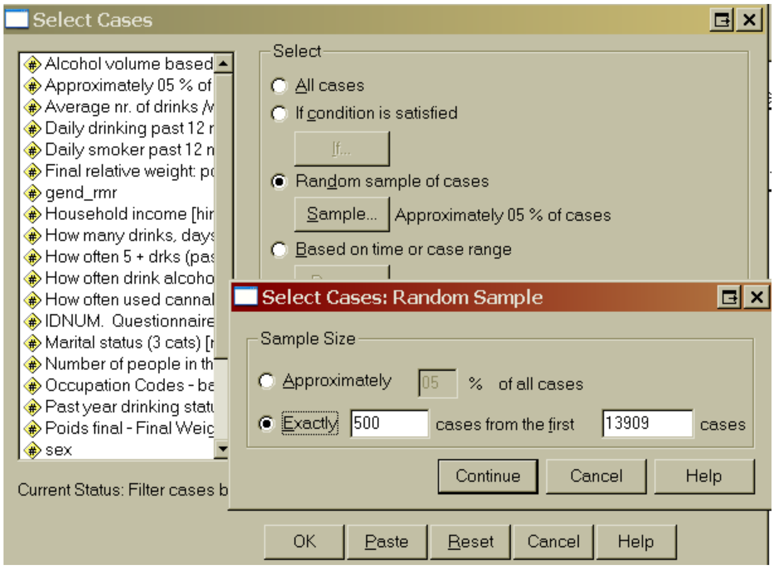
The data is the *Canadian Addiction Survey* (CAS). The whole study had 13,909 participants selected randomly from the Canadian population (age 15+), enabling very confident generalizations to the whole Canadian population. You're going to take a smaller sample of all these folks, so you won't be able to be as confident!

Get the data at http://faculty.arts.ubc.ca/fcutler/teaching/POLI380/data. It's called CAS2004.sav. Open it in SPSS just like you opened you own Census dataset.

4.1 Random Sample

Take a random sample. But first go to Transform-Random Number Generators. Check the Set Starting Point box in the middle. Then turn on Fixed Value and enter your student number. Hit OK. That way you'll all get different samples. This step is crucial

Now go to Data-Select Cases. Click the button for Random sample of cases. Then click the Sample button to tell it how much of a sample you want. Use the second line and put in 500 cases from the first 13909 cases. Then click Continue and then OK.



Select Cases

Select
○ All cases
○ If condition is satisfied
● Random sample of cases
Approximately 05 % of cases
○ Based on time or case range

Current Status: Filter cases b

Select Cases: Random Sample

Sample Size
○ Approximately 05 % of all cases
● Exactly 500 cases from the first 13909 cases

Now if you go to the Data Editor you should see that a lot of cases have diagonal lines through the case number on the left. That means they're not going to be included in any analysis you perform on the data. The ones without lines through them were the lucky 500 who get included in your personal random sample. Actually you don't have 500 because of the way weights have to be applied to the data to make it representative of the Canadian population. You'll all have a slightly different number around 500.

4.2. Number of Men & Women (Repeated Samples of a PROPORTION)

Now we'll just run a Frequencies on the sex variable. We might want to do this to estimate the proportions of men and women in Canada if we didn't have the Census and we only had enough money for 500 telephone interviews.

Remember a proportion is the decimal version of a percentage. And you must remember it's the mean of a binary variable, one that can only take on values of zero and one.

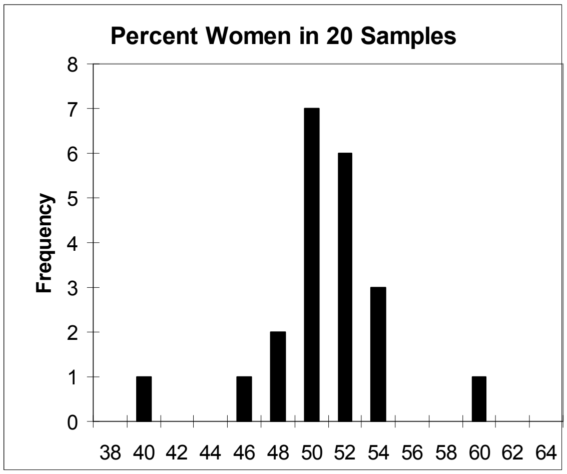
Now, go to ANALYZE-DESCRIPTIVE STATISTICS-FREQUENCIES and move the sex variable into the empty Variables box. Then just hit OK. You'll get the usual Frequencies output: numbers of cases in each category and their proportion of the total, given in percent.

Keep the output up in front of you while the TA asks all of you for the percent of women in your sample.

Now, you should know that the true proportion of women in the population 15 years and over is **.515** (or, as a percentage: 51.5%). So that's the population characteristic that you are trying to *estimate* with a sample of 500. The survey actually got about 58% women, but the data you are working with has *weights* applied to it so that the full sample of 13,909 shows 51.5% women, which is really close to the true proportion of women in the Canadian population.

Now, write down on a scrap paper how far off your estimate was from 51.5%. We'll use that later.

For those of you following along at home, I generated 20 of these samples (n=500) and got the following percent female: 50.2, 51.7, 48.2, 49.1, 52.7, 46.6, 51.5, 46.5, 50.9, 45.5, 51.9, 49.4, 49.5, 48.2, 52.7, 51.4, 53.7, 58.4, 49.8, 38.5, 50. The histogram of these values is there → Notice that I got a couple of wacky samples below 40% and above 60% female. That's just the luck of the draw.



Now, this is only 20 samples. Wouldn't it be nice if we could have some idea of how spread out our answers (like % female) would be in an infinite number repeated samples. Then we could say how likely it is that we would be 'off' from the true answer by a certain amount. For example, that we'd expect to get within 3% of the true answer in 95% of repeated samples (i.e. 19 times out of 20).

Turns out we do: the theoretical ***sampling distribution***! We've covered this in class and in the textbook (sec 7.2 & 7.3). We start with the standard error of a proportion; it's easy to calculate, especially for a binary variable which when added up gives a proportion.

From p.126 of K&W

$$\boxed{\sigma_{\bar{Y}} = \frac{S_Y}{\sqrt{N}}} = \sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{\text{proportion} * (1 - \text{proportion})}{\text{number of observations}}} = \sqrt{\frac{.515 * .485}{500}} \cong .022$$

[Note that *p*(1-*p*) is the standard deviation of a binary (0 or 1) variable. See p. 128, where, in fact, they don't mention this. Instead, they do it the long way that you actually have to do if you're dealing with the mean of a non-binary variable, but can be used for a binary variable just the same way.)]

Because we know the true proportion (.515 female), we can talk confidently about where repeated samples will fall. (But remember: Usually we don't have this luxury of knowing the true value.)

So we know that because the sampling distribution of a binomial random variable is normally distributed (see middle of p. 126) and 95% of the outcomes of a normal random variable are withing 1.96 standard deviations of the mean, then 95 percent of our samples of this size should give us proportion female within roughly 2 standard errors of the true mean: .515 ± 2(.022). That's between .471 and .559; or, 47.1% to 55.9%. How many of your class' samples were outside of that range? I got 4, not 1 as we'd expect. It's possible, though extreeeeeemly unlikely, to get ALL of your samples outside that range, right?

Now, since we wouldn't be sampling if we really knew the true population mean, we are doing something conceptually quite different with the one sample we take when we do research like this. We tell people what the sample mean is, because it's the best estimate we've got of the population mean. And then we say, separately, that we would expect to be within a certain distance from the mean in a certain proportion of samples. So with this example, we would say we expect to be within .044 [.022 times 2] (or 4.4% if you're talking percentages) of the population mean in 95% of random samples.

4.3 Repeated Samples of a MEAN

We just did that for a proportion, now let's do it for a mean. It's really the same thing even though it's generated from an underlying continuous variable. So go to Analyze-Descriptive Statistics-Descriptives. Let's put in two variables: Average Number of drinks per week (qfvolwk) and number of times using Cannabis per month (CannabisMo). Do the same thing as above: report your results to the TA in that order. Once around the room to collect average number of drinks per week and then around again to collect cannabis per month.

And again, write down how far off you were from the full sample means: They are 3.3 drinks/week and .66 times using cannabis/month.

For the male/female example above, we actually knew the population proportion female thanks to the full Canadian Census, so it was easier to think of our samples as being 'off' from a true value. Now, we're using a sample survey to estimate something we only know from a bigger sample. We don't have Census information on drinks-per-week! If we use the full sample (13,909) to estimate these quantities, we get:

	N	Minimum	Maximum	Mean	Std. Deviation
Number of drinks / week	13705	.00	147.40	3.3052	7.81701
Cannabis per Month	13899	.00	20.00	.6607	3.19227

Using the formula for the standard error of the mean (text p. 126), we calculate

$$\frac{\sigma}{\sqrt{n}} = \frac{\text{standard deviation}}{\text{sqrt(number of observations)}} = \frac{7.81}{\sqrt{13705}} \cong .067$$

So we *estimate* that if we were to repeatedly take samples of 13705 from the Canadian population and take the average number of drinks per week, we'll get 68% of our samples within ±.067 of the true average number of drinks per week for all Canadians. And we'll get 95% of our samples within two times this number (2 X .067 = .134). And we'll get only 1 sample in a thousand below three times this number below the population mean (3 X .067 = .2) and only 1 in a thousand above three times this number above the population mean. The bottom line is that we're *probably* not going to be out by much if we have a sample this large.

But what about your smaller samples of around 500. We'll, let's just plug 500 into the formula above. Our answer is .35; that's five times bigger! So if we have samples of only 500, we're going to be less confident that we are as close to the true value as if we have a sample as large as 13,705. 95% of our samples of 500 will be within .7 of the true mean in the population (exercise: why .7?); and only two samples in a thousand will be further than 1.05 drinks away from the true Canadian Drinking Average. (I wonder if Molson has a copyright on phrases like that.)

Finally, have a look at how far off you were from the proportion female and compare that to how far off you were on the other two variables.

EXERCISE: Are the female, alcohol, and cannabis ones 'off' in the same direction? Why might they be 'off' in the same direction?

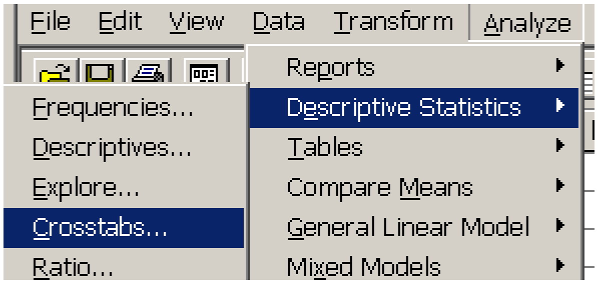
LESSION 5 – Crosstabulations (Percentage Tables)

Objectives: Understand percentage tables (crosstabulations) to analyze the strength of a relationship (association) between two variables.

Commands: Crosstabs

Again, the data is the National Election Study, 2000 (nes2000) and the file is Nes2000depr.sav.

You can use the one you used for your assignment. Or get a fresh version from http://faculty.arts.ubc.ca/fcutler/teaching/POLI380/data. Just double-click on Nes2000depr.sav and it will bring it up in the SPSS data editor.



We're going to make a bunch of crosstabs.

Remember, crosstabs only work for variables with a limited number of categories (<6 or so).

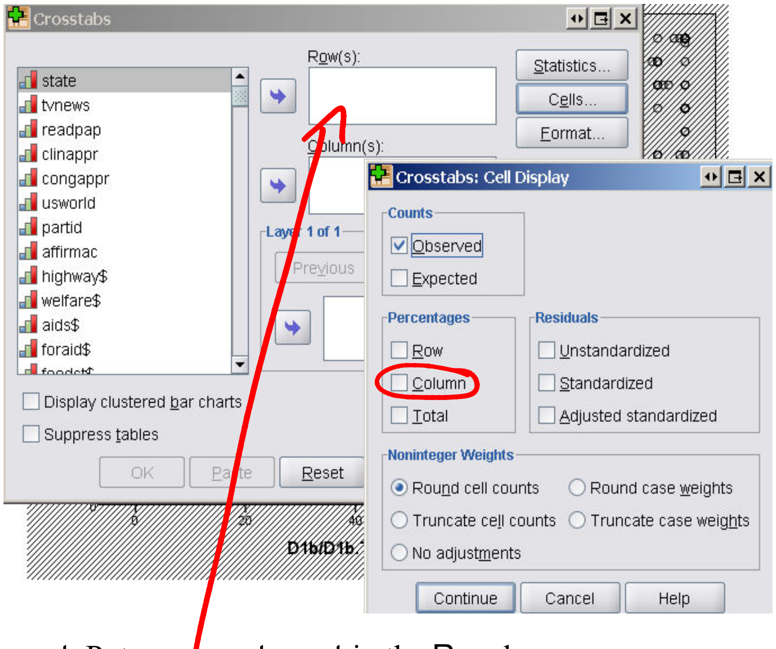
5.1 2 X 2 Table

Let's start small. Let's see if more men or women gave money to a US political party.

Go to Analyze-Descriptive Statistics-Crosstabs.

You'll get a screen like this, with the Crosstabs: Cell Display box coming up if you click Cells at the bottom. Don't do so yet.

Pay attention to the row and column variables. The Row variable is generally the dependent variable in your analysis.



We want to see if men or women behave differently, so the **dependent variable** is the behaviour: giving money to a party, prtycont. Put prtycont in the Row box.

Sex is the **independent variable** we put in the Column box.

Then click on the Cells box. By default you get the Observed Count. This is just the number of cases that fall into each cell. Leave the box checked. But we also want percentages. Column Percentages if the Column variable is our independent variable. So click the Column box circled in this picture.

prtycont B7. Did R give money to party * sex ZZ1. IWR obs: R gender Crosstabulation

			sex ZZ1. IWR obs: R gender		Total
			1 1. MALE	2 2. FEMALE	
prtycont B7. Did R give money to party	1 1. YES	Count	49	30	79
		% within sex ZZ1. IWR obs: R gender	9.7%	4.6%	6.8%
	5 5. NO	Count	454	621	1075
		% within sex ZZ1. IWR obs: R gender	90.3%	95.4%	93.2%
Total		Count	503	651	1154
		% within sex ZZ1. IWR obs: R gender	100.0%	100.0%	100.0%

Compare these two percentages

Here's your crosstabulation. Now take a minute to READ the table out loud.

Look across the rows to *compare* the rates of party giving among men and women.

Read out loud: "9.7% of Men (49/503) gave money to a party, while only half that proportion of women did so (30/651 = 4.6%)."

Notice that the other percentages just make the columns add up to 100 by listing the other behaviour: not giving to a party.

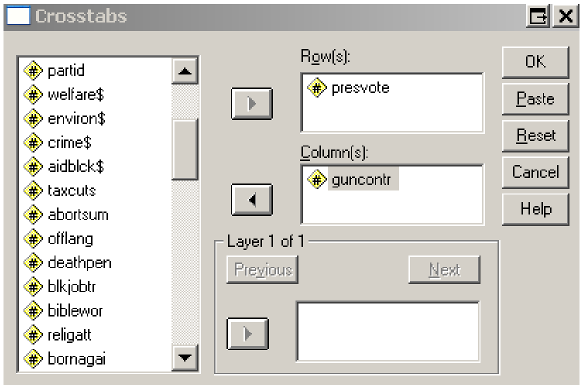
You can also use the right-hand Total column as a simple frequencies report for the row variable: Overall, 6.8% said they gave money to a party.

5.2 A More Complicated Table

Let's find out about how opinions on political issues related to the choice of President. Did people who want more gun control favour the Democratic Party candidate, Mr. Inconvenient, Al Gore?

Go to Analyze-Descriptive Statistics-Crosstabs.

Now we want the vote choice to be the dependent variable, so it goes in the Row box. Put presvote in the Row(s) box and guncontr in the Column box. You don't need to go to the Cells box because it remembers your settings from the last time. Click OK.



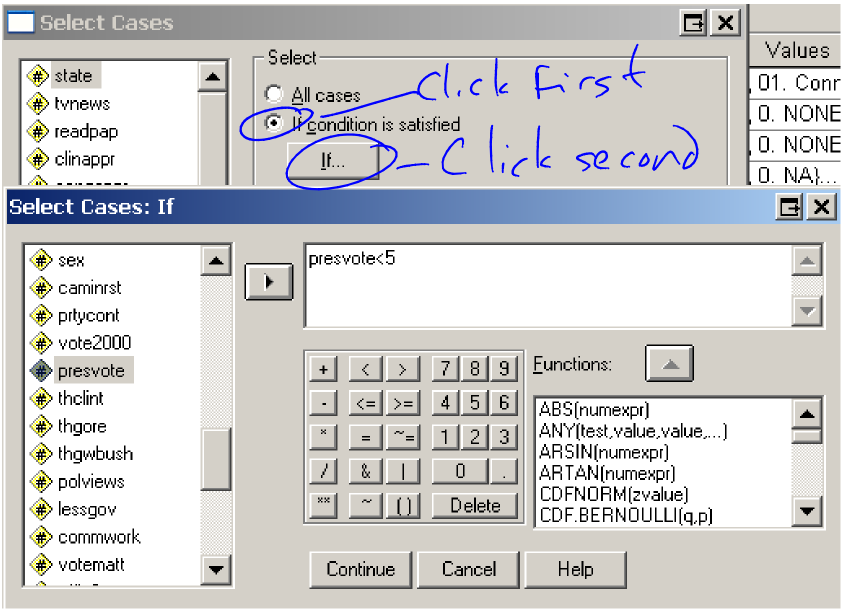
Row(s): presvote

Column(s): guncontr

presvote C6. R vote cast for President * guncontr M6ax. Summary gun control Crosstabulation

			guncontr M6ax. Summary gun control					
			1 1. A lot more difficult	2 2. Somewhat more difficult	3 3. Keep rules about the same	4 4. Somewhat easier	5 5. A lot easier	Total
presvote C6. R vote cast for President	1 1. AL GORE	Count	275	46	98	4	3	426
		% within guncontr M6ax. Summary gun control	66.7%	52.9%	31.3%	17.4%	16.7%	49.9%
	3 3. GEORGE W. BUSH	Count	118	35	208	17	13	391
		% within guncontr M6ax. Summary gun control	28.6%	40.2%	66.5%	73.9%	72.2%	45.8%
	5 5. PAT BUCHANAN	Count	1	0	1	0	0	2
		% within guncontr M6ax. Summary gun control	.2%	.0%	.3%	.0%	.0%	.2%
	6 6. RALPH NADER	Count	16	5	3	1	1	26
		% within guncontr M6ax. Summary gun control	3.9%	5.7%	1.0%	4.3%	5.6%	3.0%
	7 7. OTHER (SPECIFY)	Count	2	1	3	1	1	8
		% within guncontr M6ax. Summary gun control	.5%	1.1%	1.0%	4.3%	5.6%	.9%
Total		Count	412	87	313	23	18	853
		% within guncontr M6ax. Summary gun control	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

This is no good. Notice how few cases there are in the Buchanan, Nader, and Other rows. We can learn from this by just looking at Gore and Bush, but it's not suitable to put in a report.



presvote<5

Let's fix it. Go to Data-Select Cases. Click on the If radio button and then the If button. Then just type in presvote<5 into the empty box. You're selecting the cases where the presvote variable is less than 5: meaning only those who said they voted for Bush or Gore because they have codes less than 5.

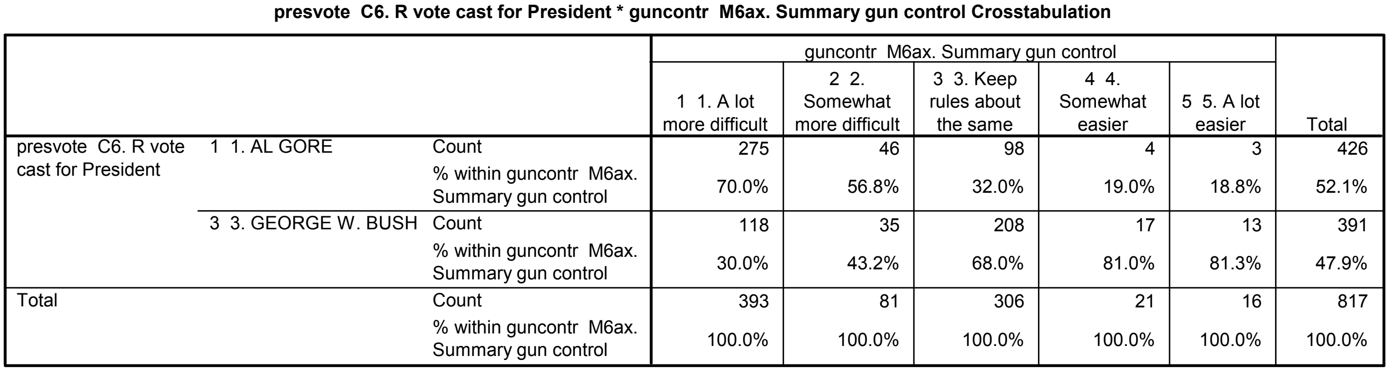
This picture is both the original Select Cases screen, and the If screen that comes up after you click the If button.

Then click Continue.
Then click OK.

Now go back to Analyze-Descriptive Statistics-Crosstabs.

You don't need to do anything! The same variables are involved and they're still there (I hope), so just click OK.

That's better, and simpler.



presvote C6. R vote cast for President * guncontr M6ax. Summary gun control Crosstabulation

			guncontr M6ax. Summary gun control					
			1 1. A lot more difficult	2 2. Somewhat more difficult	3 3. Keep rules about the same	4 4. Somewhat easier	5 5. A lot easier	Total
presvote C6. R vote cast for President	1 1. AL GORE	Count	275	46	98	4	3	426
		% within guncontr M6ax. Summary gun control	70.0%	56.8%	32.0%	19.0%	18.8%	52.1%
	3 3. GEORGE W. BUSH	Count	118	35	208	17	13	391
		% within guncontr M6ax. Summary gun control	30.0%	43.2%	68.0%	81.0%	81.3%	47.9%
Total		Count	393	81	306	21	16	817
		% within guncontr M6ax. Summary gun control	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

So read it. Read down the columns within each option on the gun control question.

"Americans who wanted access to guns to be 'a lot more difficult' were overwhelmingly in favour of Al Gore: 70% of these voters chose Gore. At the opposite end on this issue were those who wanted access to guns be 'a lot easier'; these voters chose Bush by more than a 4-to-1 margin, 81% to 19%. Even among the many voters who wanted to keep the rules about the same, Bush came out on top, with 68% support among this group."

Note, however, that the sample sizes in the 2, 4, and 5 categories are so small that it's best to report the contrast between "a lot more difficult" and "keep rules about the same".

One of the main objectives of the course is for you to be able to look at a table like this and write up the results in interesting, accurate prose, as I hope I've done here. EXERCISE for after the lab: If you want to do well on the assignments and exam—and learn the material!—you should generate a couple of these kind of tables (with different variables), and practice writing up the results at the end of this lab.

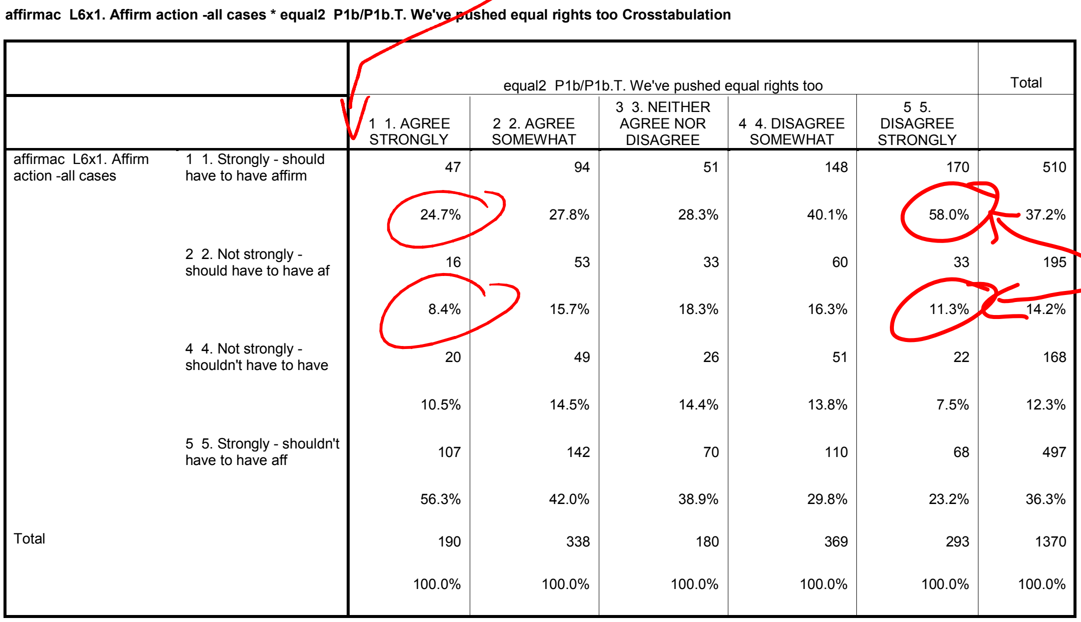
Now go back to Analyze-Descriptive Statistics-Crosstabs again. Leave the variables as they are and click the little box at the bottom left called Display Clustered Bar Charts. Then hit OK. Check out what you get. It's actually a bit of a reversed version of the table. *Within each candidate* you get counts of the gun control opinions. Is it useful? You be the judge.

5.3 Ordinal Variables Crosstabulation

Crosstabulations are most common for looking at the relationship between two ordinal variables. So let's do that with Analyze-Descriptive Statistics-Crosstabs again. The two variables to try here are affirmac (Agree or Disagree with affirmative action) and equal2 ("We've pushed equal rights too far"). Do people's views on these two things 'go together'.

Put those in the row and column variables boxes respectively and leave the Cells stuff the same as before. The result is:

(I've taken out the column that was here).



affirmac L6x1. Affirm action -all cases * equal2 P1b/P1b.T. We've pushed equal rights too Crosstabulation

		equal2 P1b/P1b.T. We've pushed equal rights too					Total
		1 1. AGREE STRONGLY	2 2. AGREE SOMEWHAT	3 3. NEITHER AGREE NOR DISAGREE	4 4. DISAGREE SOMEWHAT	5 5. DISAGREE STRONGLY	
affirmac L6x1. Affirm action -all cases	1 1. Strongly - should have to have affirm	47	94	51	148	170	510
		24.7%	27.8%	28.3%	40.1%	58.0%	37.2%
	2 2. Not strongly - should have to have af	16	53	33	60	33	195
		8.4%	15.7%	18.3%	16.3%	11.3%	14.2%
	4 4. Not strongly - shouldn't have to have	20	49	26	51	22	168
		10.5%	14.5%	14.4%	13.8%	7.5%	12.3%
	5 5. Strongly - shouldn't have to have aff	107	142	70	110	68	497
		56.3%	42.0%	38.9%	29.8%	23.2%	36.3%
Total		190	338	180	369	293	1370
		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

So the affirmative action variable is on the left, from Strongly Should Have Affirmative Action (1) down to Strongly Shouldn't Have Affirmative Action (5). Across the top we have Agree Strongly (1) to Disagree Strongly (5) with the statement "We have pushed equal rights too far". So you're looking at 4 X 5 = 20 cells here. Look for patterns. Are the variables positively or negatively related?

Well, among people who Disagree Strongly that we've pushed equal rights too far, 58% + 11.3% = 69.3% think 'we' should have affirmative action. (Where the "we" is people in the USA!). Among people who Agree Strongly that we've pushed equal rights too far, only 24.7% + 8.4% = 33.1% think we should have affirmative action. That's a big difference, 36%, dontcha think? Another way to put this is that a person who disagrees that equal rights have been pushed too far is 36% *more likely* to want affirmative action than a person who agrees that rights have been pushed too far.

So this is how you have to look at crosstabulations. Find the patterns by saying out loud: "Among people who [have a certain characteristic], ??% think [your dependent variable], as opposed to people who [have a different score on the same characteristic], among these people ??% think [your dependent variable]". Or, "Among _______ countries, ??% have [a certain value or values on the dependent variable], but among _______ countries, ??% have [that value or values (or the opposite ones)]."

Roughly this same approach applies whatever you're analyzing.

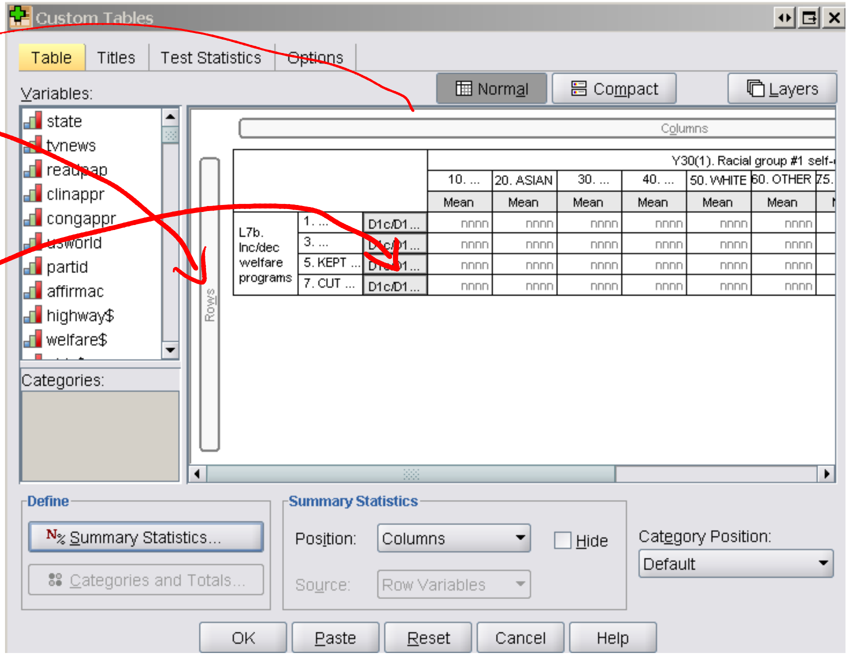
5.4 Means in a Table

Tables can show you a lot of things.
Now we're going to put the Bush thermometer (0 to 100 degrees) into a table with the person's race on one axis and their welfare spending attitudes on the other axis. Wow, three-variable analysis!

But first, we have to tell SPSS to ignore the categories of both variables that have few cases. So go to Data-Select Cases and Click on IF again. Enter this in the If box:

Race<60 and welfare$<7. So

Now go to Analyze-Tables-Custom Tables (we haven't done this before)



Specify the subgroups: drag race (into the Columns) and Welfare$ (into the rows). Before you do, you may have to right-click the variables and set them to "Scale" type variables.

Then drag the Bush thermometer thgwbush just to the right of the categories of the welfare variable. (The column header under the race categories should change to "Mean")

Now you have to tell it what summary you want. Click on the Summary Statistics button toward the bottom left.

Mean is already there. So don't do anything, just notice all the stuff you *could* put in there.

Then hit Continue (or Close in v.18). And then OK.

This is good; we're getting average Bush thermometer within categories of race and welfare spending attitudes. That's 15 separate *averages* of feeling about GW Bush to compare. Reading down the columns, there's a predictable pattern in the Hispanic/Latino and White groups, but not so much among Blacks. Blacks like Bush a lot less than others, but among blacks, the Bush rating doesn't differ between those who want welfare funding decreased and those who want it kept the same.

			Y30(1). Racial group #1 self-description				
			10 10. BLACK	20 20. ASIAN	30 30. NATIVE AMERICAN	40 40. HISPANIC OR LATINO	50 50. WHITE
			Mean	Mean	Mean	Mean	Mean
L7b. Inc/dec welfare programs	1 1. INCREASED	D1c/D1c.T. Thermometer GW Bush	37	63	70	48	51
	3 3. DECREASED	D1c/D1c.T. Thermometer GW Bush	42	80	60	58	63
	5 5. KEPT ABOUT THE SAME	D1c/D1c.T. Thermometer GW Bush	42	49	63	54	56

What's wrong with this table? The welfare variable is coded funny; it isn't in the right order. So the patterns aren't obvious. To see whatever patterns are there more clearly, we'd want to … . What? What would you recode?

LESSON 6 – Graphics

Objectives: Make and INTERPRET a Bar Chart, Histogram, Boxplot, Scatterplot.

Commands: GRAPH

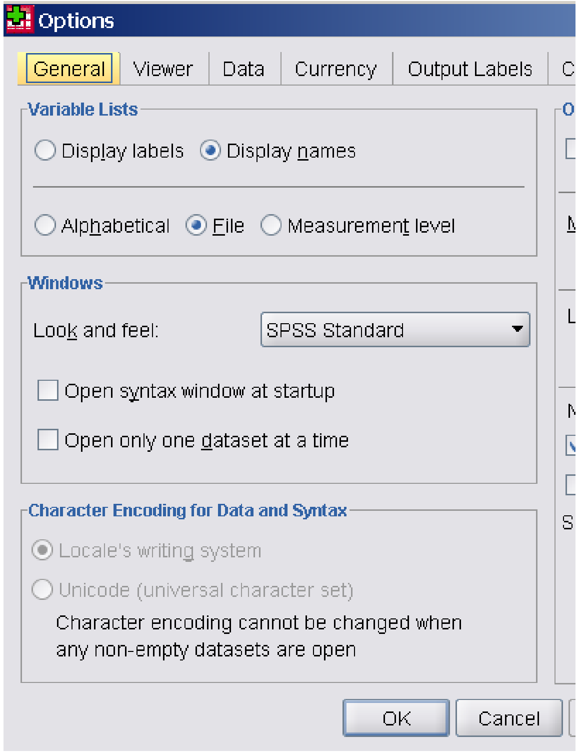
Again, the data is the National Election Study, 2000 so you should have a saved version from before. If not, go to the website and get it again.
Just double-click on Nes2000depr.sav and it will bring it up in the SPSS data editor.

NOTE: There is a chance that you are using a version of SPSS different from the one used to produce these pictures. If you are, you will have to look a bit harder to find the same options.

Right away, go to the file menu and Save As. Save it to a disk or a USB drive or to the My Documents directory on the computer you're using. IF YOU DO THE LATTER, make sure you email it to yourself at the end of the lab so you can use the changed dataset for your next assignment and for next week's lab.

PRELIMINARY: Let's all look at the same display for variables. Let's make it the variable names, not the labels. So go to the menu Edit-Options. And you'll get this box. You should be in the General tab, so just click the two buttons for Display Names and the one below that, File.

That gives you variable lists with names and in the order in the file.



Variable Lists

○ Display labels ● Display names

○ Alphabetical ● File ○ Measurement level

Windows

Look and feel: SPSS Standard

□ Open syntax window at startup

□ Open only one dataset at a time

Character Encoding for Data and Syntax

● Locale's writing system

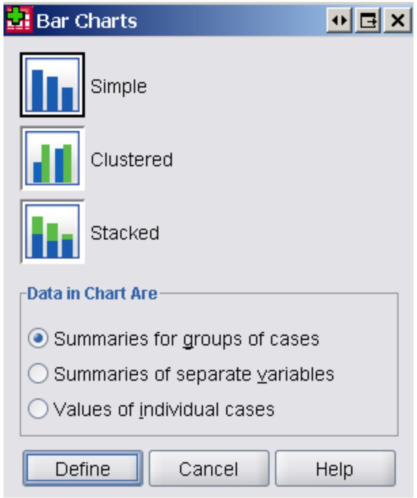
○ Unicode (universal character set)

Character encoding cannot be changed when any non-empty datasets are open

6.1 Bar Chart

The first two things we'll do involve only one variable. We use bar charts for categorical (nominal) variables like religion or party or regime type. All of the things we'll do in this lab involve the Graphs menu. So go there now :Graphs-LEGACY DIALOGS-Bar.

This is what you get. Now this can be confusing. If you're graphing multiple variables, or want separate graphs for different values of another variable, you choose clustered or stacked. But for now we'll just choose Simple.



Data in Chart Are

● Summaries for groups of cases

○ Summaries of separate variables

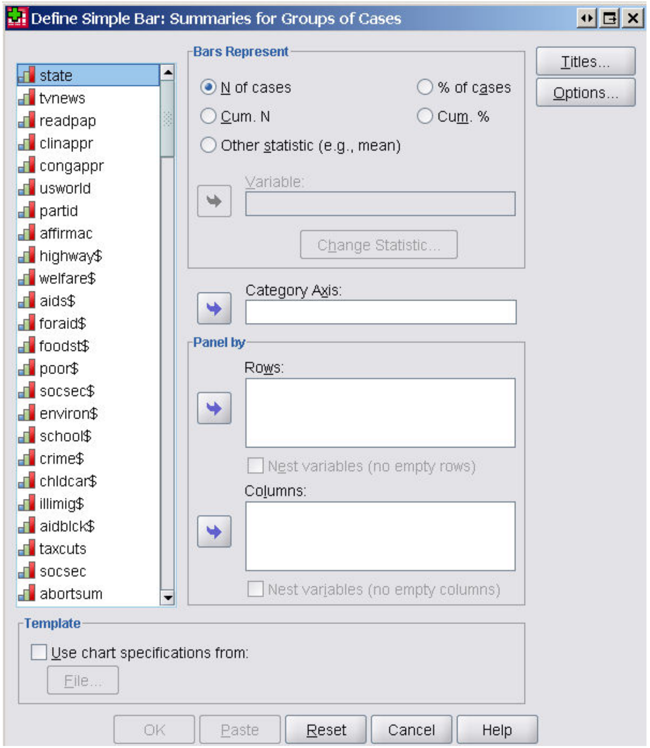
○ Values of individual cases

Down at the bottom you have to make another choice. You can get summaries for groups of cases, which means categories of one variable.

You can get summaries of separate variables.

Or, rarely, you can get values of individual cases (like countries).

Don't change anything for now, just hit Define.



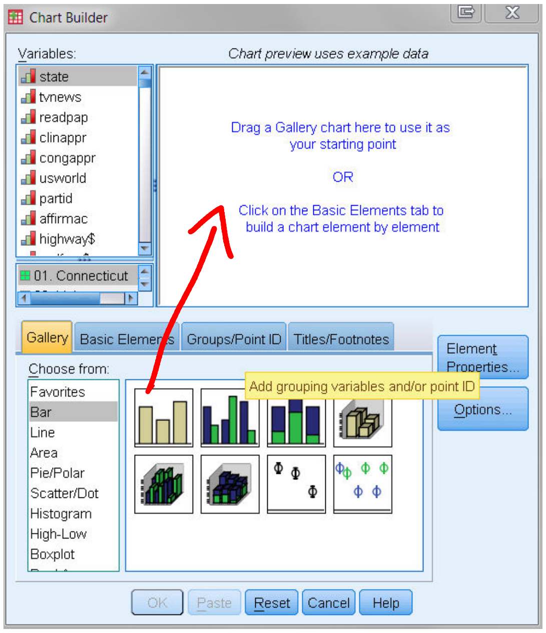
You'll get another box like this one. The variable you want bars for goes into the middle box (Category Axis). Put partid in here.

Then you go up and choose what the bars represent. You'll get the same result whether you use N of cases or % of cases, but the y-axis will be labelled diferently. For now use % of cases.

You should normally click the Titles box and put a title on every graph. But for now just hit OK.

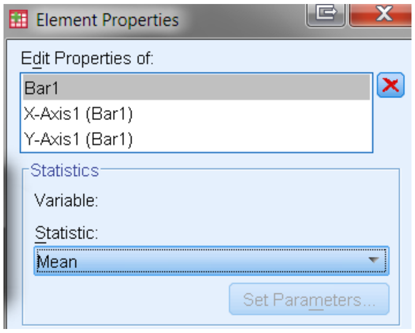
Have a look at the output. Now you know that the Democrats have a stronger base in the electorate. How do they manage to lose both the Presidency and both chambers of Congress? As you do this lab, they may be about to lose or hold on to the Congress.

6.2 Bar Chart of Means



This is JUST LIKE the Compare Means procedure, except graphically. So go to Graphs-Chart Builder. You'll get this box or something like it. First, drag the simpe bar picture from the bottom area up to the empty area in the middle-right (see arrow). Then you'll get the Element Properties box and you have to change the Statistic area to be Mean instead of Count then hit Apply and Close that box. Now drag the thgwbush variable (Bush Thermometer) into the "Y-axis?" box. And drag aidblck$ (increase or decrease aid to blacks) into the other "X-Axis?" box. Notice that it tells you that you've got Mean:thgwbush as what the bars will represent. Then hit OK.

You get a pretty simple bar graph of the mean Bush rating for people with different ideas about whether "aid to blacks" should be increased or decreased. Stop and think and interpret it to yourself.



Notice that the bars don't go up evenly. Why? Would you want to recode the aid to blacks variable so it works as an ordered (ordinal) scale? This is the same question as at the end of the last lab.

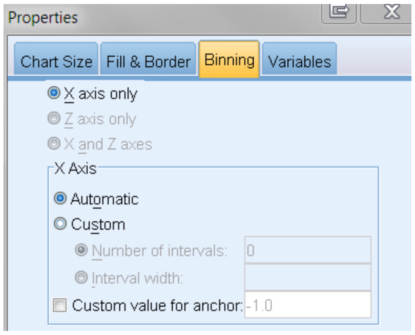
6.3 Histogram

I know you've done this already, but it's SO important to get to know your data, let's do it again. Go back to Graphs-Chart Builder and just do what you did a second ago, but take the aidblack$ variable out of the x-axis area and move the thgwbush in there, with nothing in the y-axis area. Just click OK.

Now you get a nice histogram showing the distribution of feelings toward George W. Bush. (Wonder if it looks a little more polarized now than it did in 2000? Me too.)

Double click somewhere on the histogram itself and you'll unleash the wonders of interactive graphing. It should come up on top of the original in a new window called "Chart Editor".

First, right-click one of the bars and select Properties Window. You'll get this. →
Choose the Binning tab. Move this properties window to the side of the original graph window so you can see the graph change when you do the following…



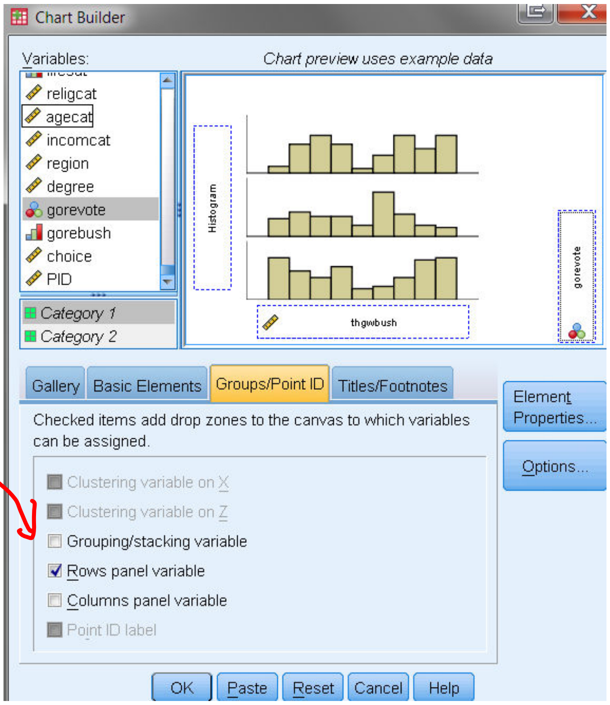
● X axis only
○ Z axis only
○ X and Z axes

X Axis
● Automatic
○ Custom
○ Number of intervals: 0
○ Interval width:
□ Custom value for anchor: -1.0

Now you can choose the best number of categories to convey the distribution of the variable. Click the button for Custom and then specify a number of intervals. Try 12. Then click Apply. Kind of silly with the little bar above 50. Try 11… Much better. Using 20 shows clearly how people use round numbers in their answers to this variable. Leave it at 11 and close the Interval tool. Have a look and then get back to the Data window.

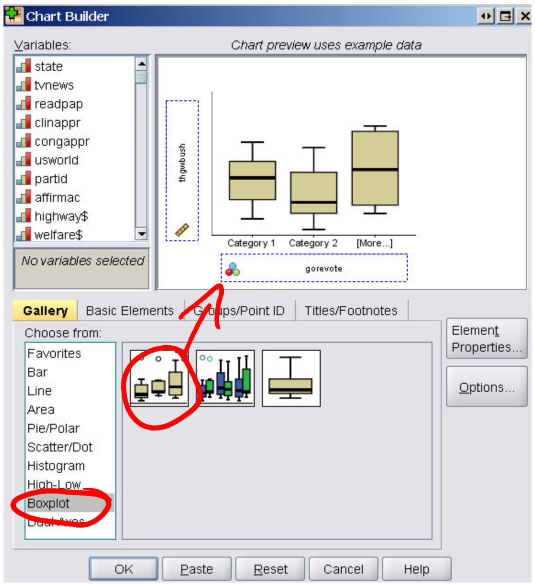
Now let's do another way to display how different groups of people felt about George W Bush. Go to GRAPHS-CHART BUILDER. Drag the simple histogram up to the empty preview space again. Get thgwbush into the x-axis box again. (If the Element Properties box comes up, just exit it with the usual x at the top-right of the window). If you hit OK at this point, you'd get the basic histogram you got before. But we're going to get two histograms. So go to the Groups/Point ID tab in the middle. Then check the Rows Panel box.

Now slide way down toward the bottom of the variable list and find gorevote. (I made this up: a dummy variable indicating Gore voters.) See the ruler beside the name? That means SPSS thinks it's a Scale (continuous) variable . We need to tell SPSS it's really nominal (i.e. categories). So right-click on gorevote and then click Nominal in the menu that appears. Finally, drag gorevote into the little Panel? box on the side. Hit OK and Presto (after a second or two)! You now have in your Output window separate histograms for Gore voters and everyone else. They're different on feelings about Bush, right?



Checked items add drop zones to the canvas to which variables can be assigned.

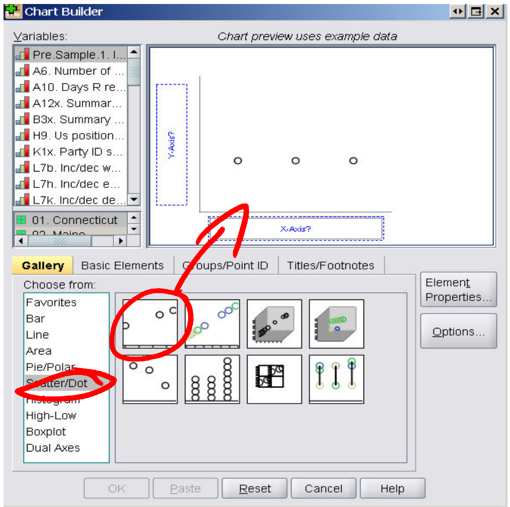
□ Clustering variable on X
□ Clustering variable on Z
□ Grouping/stacking variable
☑ Rows panel variable
□ Columns panel variable
□ Point ID label



Finally, do a boxplot to get the same information. Go to the Graphs-Chart Builder again, click Reset at the bottom. Then make it look like the one here by getting Boxplot in the Choose From area, drag the first boxplot up, then put thgwbush on the y-axis and gorevote (make it nominal again) in the x-axis. Hit OK. Wait for the result and then… Notice that the boxplot gives you the same (but less) information as the histogram. Make sure you understand how the two graphs depict the same information. Print 'em out and turn one on its side if you have to.

6.4 Scatterplot

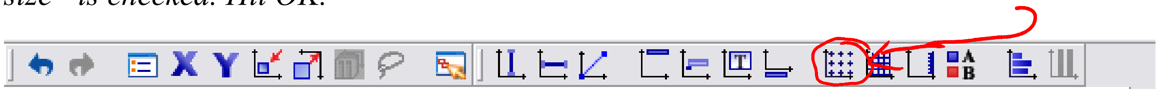
Finally, let's do a scatterplot of two variables. This is the foundation for regression analysis in the second half of the course.



Go to the Graphs-Chart Builder again, and click Reset at the bottom. Click Scatter/Dot and then drag the top left of the little graph pictures up to the preview window. (Close the Element Properties window that pops up). Let's just see how close people's ratings of Bill Clinton and Al Gore were—how much these two feelings go together. So put thclint in the Y axis box (arrow up) and thgore in the X axis box (arrow sideways). Then just hit OK.

This graph just places each person's ratings of Clinton and Gore on the two dimensions. Someone who gave them both a rating of 100 would be in the top right, while someone giving them both zero would be at the bottom left. So you see that there are more people along the line from bottom left to top right than in the top-left or bottom-right quadrants. That means a person's ratings of Clinton and Gore were similar. Few people gave a high rating to one and a low rating to the other. Try to make yourself see this in the graph.

But there's a problem. The graph has only one dot for all the people who gave the two guys the same rating. Let's fix this. Double click on the graph to pull up the chart editor. On the toolbar, click the icon circled below. It's called "binning". Make sure "marker size" is checked. Hit OK.



Now the graph shows you when there are many people at one point by making the symbols bigger, like cities with more people on a map.. It should be even more apparent that ratings of Clinton and Gore tend to be similar.

LESSON 7 – Difference of Means & Regression

Objectives: Difference of Means. Regression t-tests.

The data is the US election file (Nes2000depr.sav) from http://faculty.arts.ubc.ca/fcutler/teaching/POLI380/data or a saved version you already have.
We're going to do three things today.

The most important thing to know is that the first two are really THE SAME THING.

One is a Difference of Means test, using the t-test.

The other is a Regression, with a t-test for the significance of the coefficient.

They give you the same answers! They tell you the same thing.

7.1 Difference of Means

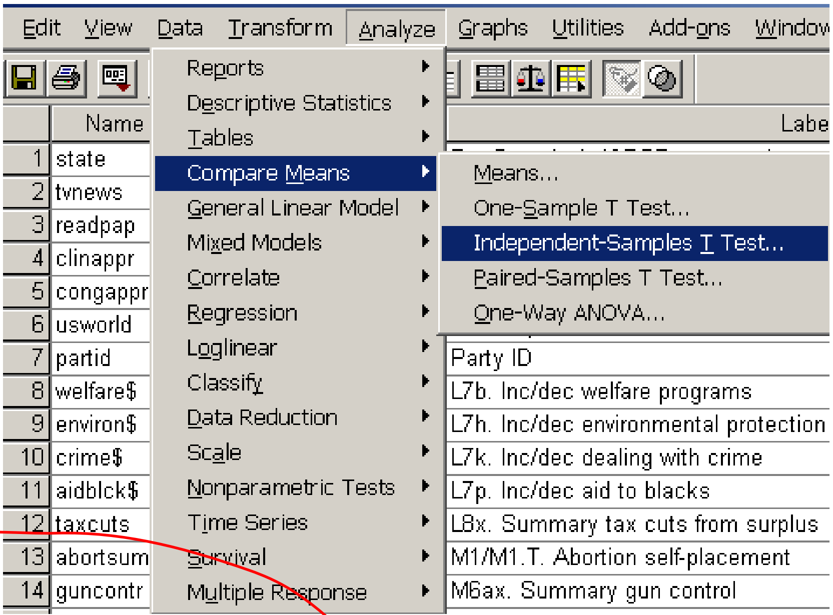
Go to Analyze-Compare Means-Independent Samples T test.

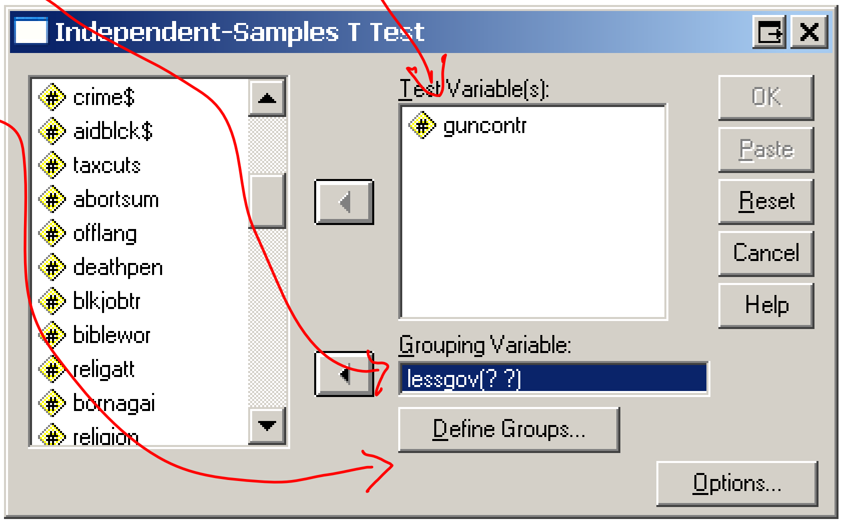
It is "Independent Samples" because in the election study data, everyone is separate. Each person is a random draw from the population. So if we compare two groups, the sample for one does not depend on the sample for the other (e.g. men and women).

Let's just get a p-value on the difference between the more and less government people on gun control. That way we can express our certainty about the result.

The Test Variable is the variable whose mean you are testing: gun control (guncontr).

The Grouping Variable is lessgov. But first you have to tell SPSS which groups of that variable you want. So after you put it in the box you have to click the Define Groups button. Before you do that, you might need to right-click on the variable to get Variable Information to see how it's coded. But for now, just click the button and put 1 and 2 in the Group 1 and Group 2 boxes. Then continue, then hit OK.

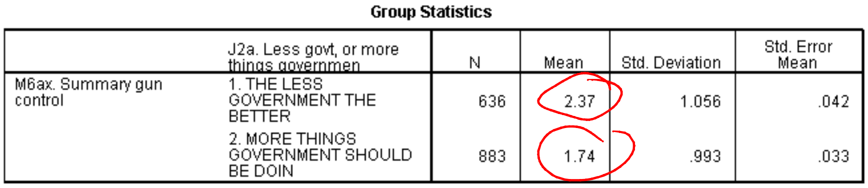




Test Variable(s): guncontr

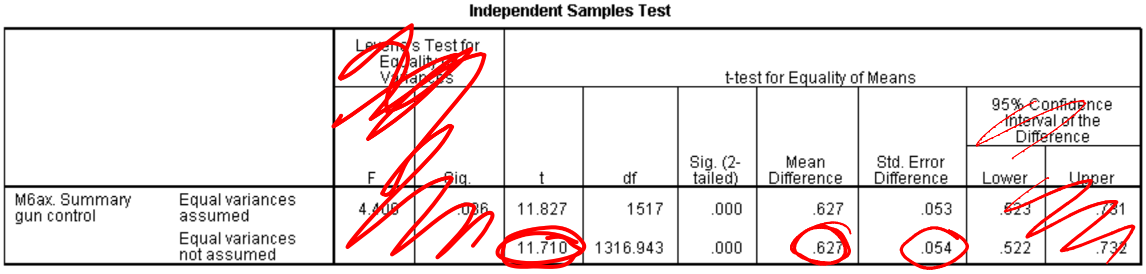
Grouping Variable: lessgov(? ?)

You get two boxes of output. The top one, Group Statistics, gives you the breakdown of Means and tells you the standard deviation and the standard error of the mean (i.e. one standard deviation of the sampling distribution). The Independent Samples test gives you two lines of output. One is for Equal Variances Assumed, the other Equal Variances Not Assumed. In general, you want the bottom line of this (Not Assumed): you have no grounds to assume that your two groups would have the same variation in the means variable (here, guncont) in the population. Note that you might have slightly different numbers from the ones here. Please think through what your own numbers mean after you've looked at these.



Group Statistics

	J2a. Less govt, or more things governmen	N	Mean	Std. Deviation	Std. Error Mean
M6ax. Summary gun control	1. THE LESS GOVERNMENT THE BETTER	636	2.37	1.056	.042
	2. MORE THINGS GOVERNMENT SHOULD BE DOIN	883	1.74	.993	.033



Independent Samples Test

		~~Levene's Test for Equality of Variances~~		t-test for Equality of Means					~~95% Confidence Interval of the Difference~~	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
M6ax. Summary gun control	Equal variances assumed	~~4.406~~	~~.036~~	11.827	1517	.000	.627	.053	~~.523~~	~~.731~~
	Equal variances not assumed			11.710	1316.943	.000	.627	.054	.522	~~.732~~

The first thing to look at is the mean difference: it's just the second group's mean subtracted from the first (2.37 – 1.74 = .627). Find these three numbers and see what they mean.

The next thing is to look at the t-statistic, which is actually the absolute value of the mean difference (.627) divided by the standard error of that difference (.054). Remember, the standard error is the range from the population mean that we would expect sample means to be in 68% of random samples. So here, the *t* statistic is a whopping 11.710. The probability we'd get this big a t-statistic just by chance is less than .000 (the Sig column, which is another name for the p-value). In fact, it is .000000000000000000000000000001175. That's the chance you'd get a difference of means that big if there's really no difference between the groups in the population. Wow!

Now, it's possible that you got the wacky one-in-a-decillion sample that could produce this difference of .627 with Ns of 636 and 883 just by chance *when there's really no difference between groups* (null hypothesis). (Notice that I repeated this language; get used to it.) I think you can take this difference to the bank.

Let's do one more and get something a little less impressive. Go back and do guncontrol by unionmem. The groups for the "define groups" button are 1 (Yes) and 5 (No). Say out loud what we are testing.

The difference turns out to be an unimpressive -.034 on the 1 to 5 gun control scale. And the *t*-statistic is -.478. The Sig. (p-value) is .633. What does this mean? Think it through and write your answer in this space: [TAs: STOP and let the students write an answer down].

Don't look now, but the answer appears in a footnote.[2]

More importantly, and substantively, this isn't much of a difference anyway. You can safely tell your audience that union members are no different on gun control than everyone else.

7.2 Regression With a Dummy Variable (a Means test!)

Dummy variables are variables that take on just two values.
Usually these are 0 and 1, but they don't have to be. And in this example, they aren't, so this is a lot more confusing than the usual case. You'll learn something from that!

Go to Analyze-Regression-Linear. Put guncontr in the Dependent box, and lessgov in the Independent box. So the regression equation would be? Write it down.

Then just hit OK to run the regression. You get four separate boxes of output. For now, just focus on the bottom one: "Coefficients". IGNORE THE STANDARDIZED COEFFICIENTS (BETA) column. And note that SPSS calls the intercept the "constant" – they are synonyms.

(Remember the language: This is "running a regression of gun control (dependent var) ***on*** less government (independent var)").

What's the regression coefficient on the lessgov variable (column labelled "B")? Hey, we've seen that before: .627. But whereas the difference above was positive .627, here the coefficient is -.627. Why? Well, above, it was the first group (1. Less Government) *minus* the second group (2. More Government). So since the first group had a higher mean value on gun control, the .631 was positive.

Here, we're doing a regression. And it means just what it always means: when you increase the independent variable by 1, you get a change of *b* in the dependent variable. So when you go *up* from 1 to 2 on the less government variable (cuz those are the only two values), you actually go down .627 on the gun control variable, so *b* is -.627. [You probably have to read these last two paragraphs 2 or 3 times so you get it. It's weird, but shows something fundamental about what regression is doing.]

[2] *It means that there is a 63.3% chance of getting a difference this big in samples of these sizes with variances like these when there is, in truth, no difference between the two groups in the population. Understand that, or else. Or else, ask and think and ask again!*

[Pretty annoying gun control coding, eh? The gun control variable is coded backwards so it would be easier if we actually called it 'access to guns' or something.]

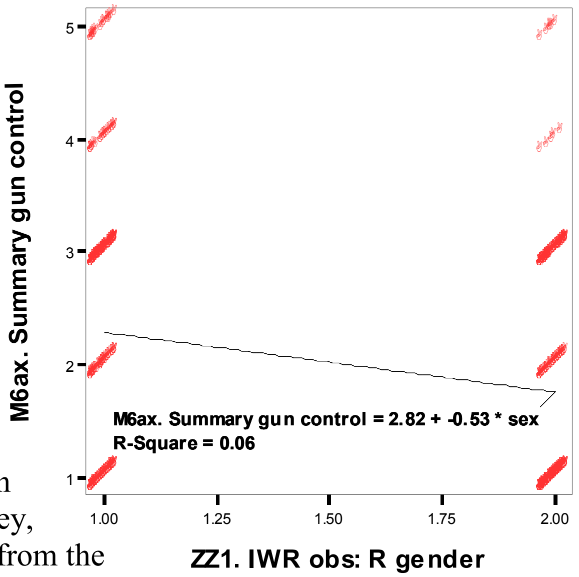
Notice one more thing. The t-statistic is -11.827. Since it's the absolute value that matters, this is big, and what-do-you-know, it's nearly identical to what we got with the Compare Means thing. Unfortunately, it turns out that it's the same as the first line, the Equal Variances line, in the Compare Means output above. Don't worry about the difference, it's always teeny. Just remember that you can get a test of the difference of means between groups with Compare Means *or* with a regression on a dummy variable that indicates the two groups. The Sig column gives you the p-value, just like above.

Let's just check this out as we see if gun control attitudes differ between men and women. Go back to the Analyze-Regression-Linear Regression and just replace lessgov with sex.

What is the difference between Men (coded 1) and Women (coded 2) on gun control (or, as we're now calling it 'access to guns'). And what is the t-statistic and p-value (Sig). What does this mean?

Everyone pause and answer that: what does it mean?

Just for the **thrill** of it, have a look at this graph. (You don't need to run it). I told SPSS that gun control and gender were scale variables to trick it into letting me do a scatterplot with a regression line. There it is. The points are "jittered" to give you a sense of how many people are at each point. Notice there are a lot fewer women at 4 and 5 on gun control. This is why we get a negative slope. And, hey, the coefficent -.53 should be what you got just now from the regression table you generated in the previous two paragraphs.



7.3 Bivariate Regression Practice

Now load up the countries dataset (from the usual place): countries.sav.
Let's see if countries with more democratic "Voice and Accountability" spend less on the military.

First, get a Descriptives on the two variables: Military expenditures as a percent of GDP (milexgdp) and Voice and Accountability (voicacct). Have a look at how they're measured. Think about those measurement scales.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Military expenditures as a percent of GDP, various years	153	.2	29.0	3.059	4.0178
Voice and Accountability, 1999	171	-1.789	1.694	-.00312	.964735
Valid N (listwise)	150				

Now go to Analyze-Regression-Linear. In the dialog box, put Military Expenditures in the Dependent box and Voice and Accountability in the independent box. Write out this regression equation for practice. Then just click OK.

You get those pesky four boxes of output again. Again, just go to the last one and look at the regression coefficient for the variable "Voice and Accountability". It's -1.156.

Coefficients(a)

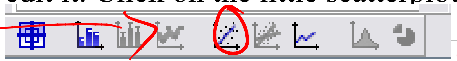
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.069	.320		9.586	.000
	Voice and Accountability, 1999	-1.156	.345	-.266	-3.356	.001

a Dependent Variable: Military expenditures as a percent of GDP, various years

So what does it mean. Again, read from right (indep var) to left.[3]

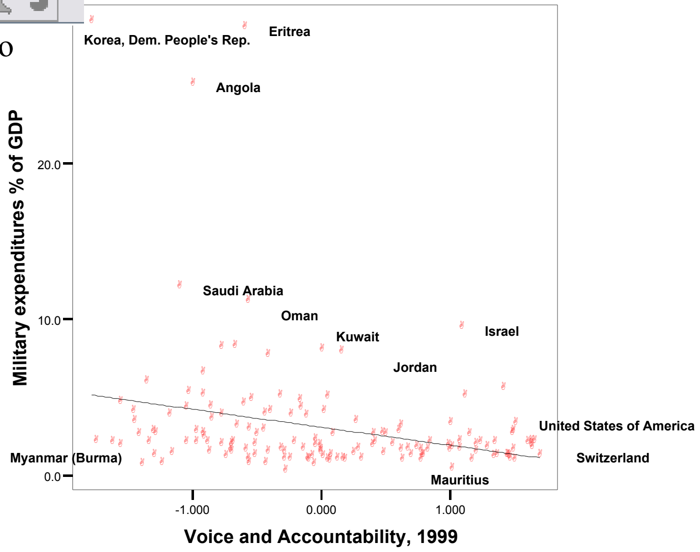
IMPORTANT: This dependent variable was measured in percent. So you can talk about changes as changes in the percent of GDP devoted to military spending. DO NOT TALK ABOUT PERCENT CHANGES IF YOUR DEPENDENT VARIABLE IS NOT MEASURED IN PER CENT!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! Every year I teach this course, people talk about percentage changes in variables that are not measured as percents.

Finally, let's look at a scatterplot of the same variables. Go to Graphs-Chart Builder. Select Scatter/Dot from the list at bottom left. Then drag the top-left one up into the preview area. Then drag Military Expenditures into the y-axis box and Voice and Accountability into the x-axis box. Click OK. When the graph appears in your output, double click on it to edit it. Click on the little scatterplot plus regression icon in one of the toolbars.



When the box comes up, just click Apply, then go and look at what you got.

You should get a graph a little like this one. I've played with it to show some countries and put the regression equation on the bottom. Same result as before.



Military expenditures as a percent of GDP, various years = 3.07 + -1.16 * voicacct
R-Square = 0.07

[3] Answer: "**A one point increase in Voice and Accountability is associated with a decrease in military spending of 1.15% of GDP. A relationship this strong or stronger would occur in .1% of samples just by chance if there really is no relationship.**"

LESSON 8: Multivariate Regression Analysis

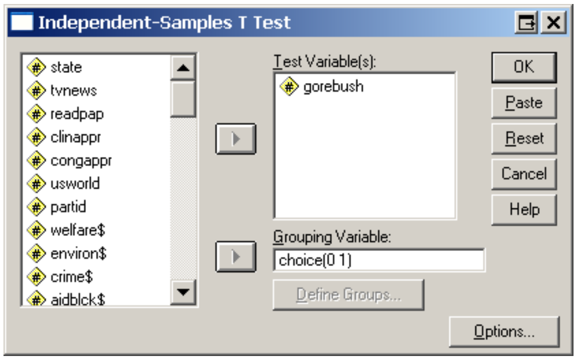
Objectives: Multivariate Crosstabs/Means. Multiple Regression.

Commands: Basic Tables. Regression. Data is the US Election file Nes2000depr.sav.

8.1 Tables of Means

We're going to jump right into tables of means because multivariable crosstabulations are really clunky and you will almost never use them. They're half-century-old technology.

Instead, we're going to check out the relationship between abortion views and voting for Gore vs. Bush. First, just run a comparison of means Analyze-Compare Means-Independent Samples T-test. Make gorebush the test variable and choice the Grouping Variable (define 0 and 1 as the Groups for the Grouping Variable. You should find that support for Gore was about 41% among those who want restrictions on abortion and 67% among those who are pro-choice.

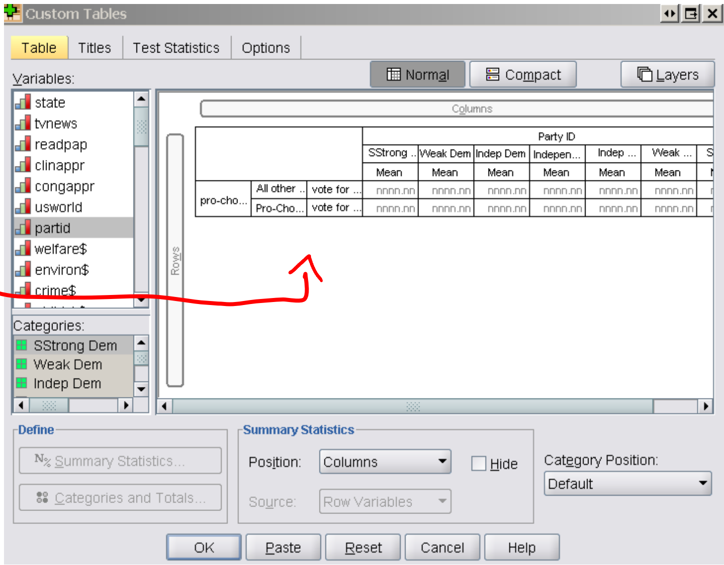

Independent-Samples T Test

Test Variable(s): gorebush

Grouping Variable: choice(0 1)

But how much of this is just due to partisans of each party having abortion views consistent with their party identification? Let's add partyid as a control variable by using the Analyze-Tables-Custom Tables procedure. First, set the variables' measurement scales properly by right-clicking on the variable names in the list and changing their scale. partyid and choice have to be "nominal" or "ordinal" and gorebush has to be a "Scale" so that SPSS will calculate a mean.

Once you've done that, make the box look like this one by **dragging** partid into the Columns area, choice into the rows area, and then gorebush just to the right of where it says "all other…" and "Pro-Choice". Hit OK.


Custom Tables

Summary Statistics Position: Columns

Source: Row Variables

Category Position: Default

You get this output or something like it depending on your sample. Now what do you think of how abortion views affect the Gore vs. Bush decision? It seems to affect the choices of Independent Democrats through to Weak Republicans. Among Strong Democrats and Republicans, abortion views don't help voters make up their minds – they've already decided. Only for independents does abortion tip the balance from 37% to 55% support. This makes a good deal of sense.

So we've controlled for a variable. It hasn't made the relationship go away, but it has clarified it a bit.

Table 1

			Party ID						
			SStrong Dem	Weak Dem	Indep Dem	Independent	Indep Repub	Weak Repub	Strong Repub
			Mean	Mean	Mean	Mean	Mean	Mean	Mean
pro-choice	All other responses	vote for gore or bush	.98	.89	.71	.37	.10	.06	.03
	Pro-Choice	vote for gore or bush	.97	.91	.85	.55	.22	.26	.00

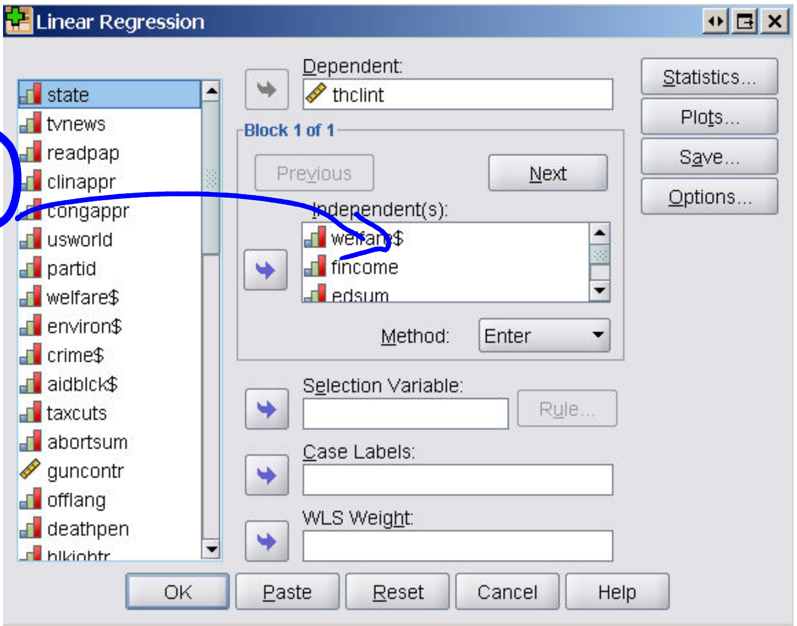
8.2 Multiple Regression

Now, let's try to explain feelings about Bill Clinton on the basis of a bunch of variables. Our model is:

$$\text{Clinton Thermometer} = a + b_1(\text{Welfare\$}) + b_2(\text{aidblack\$}) + b_3(\text{taxcuts\$}) + b_4(\text{abortsum\$}) + b_5(\text{guncontr\$}) + b_6(\text{religatt\$}) + b_7(\text{education\$}) + b_8(\text{unionmem\$}) + b_9(\text{sex\$}) + b_{10}(\text{lessgov\$}) + b_{11}(\text{fincome\$}) + e$$

Wow. Do we really think that all these variables are going to be related, separately, independently to how someone feels about Clinton? Well, maybe. Let's find out.

To estimate this equation and get regression coefficients on all 11 of these variables, you just need to go to Analyze-Regression-Linear like you did when you were using one lonely variable.



Dependent: thclint

Independent(s): welfare$ fincome edsum

Method: Enter

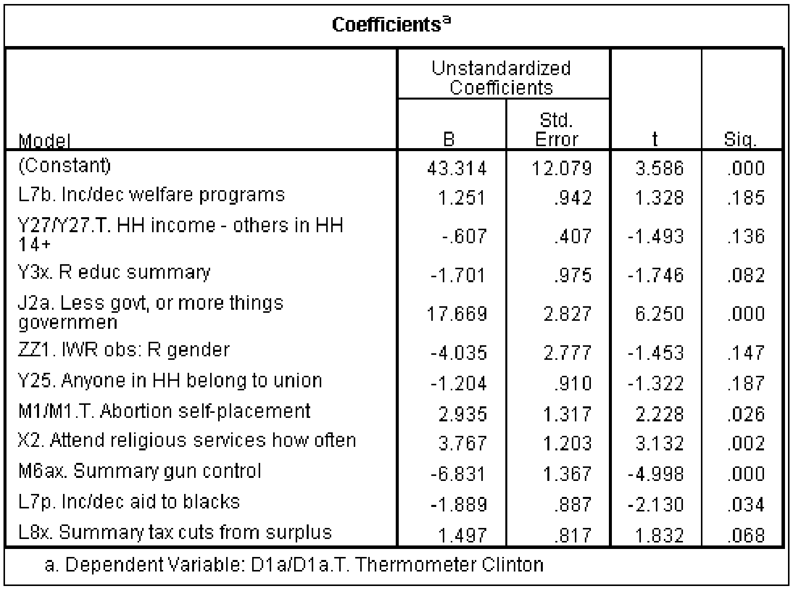
Make your Linear Regression box look like this, with the following variables in the Independent(s) box:

welfare$ aidblck$ taxcuts abortsum guncontr religatt edsum unionmem sex lessgov fincome

It doesn't really matter that you get all of these. If you run out of time in the lab just hit OK when you've got a few of these variables in there.

The Output, remember, comes in four boxes.

The Coefficients box is the one that you want to look at. It will look a bit like the one below. (I double-clicked on it and changed the width of some columns, and got rid of the Beta column entirely by narrowing it to zero width).



Coefficients^a

Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	t	Sig.
(Constant)	43.314	12.079	3.586	.000
L7b. Inc/dec welfare programs	1.251	.942	1.328	.185
Y27/Y27.T. HH income - others in HH 14+	-.607	.407	-1.493	.136
Y3x. R educ summary	-1.701	.975	-1.746	.082
J2a. Less govt, or more things governmen	17.669	2.827	6.250	.000
ZZ1. IWR obs: R gender	-4.035	2.777	-1.453	.147
Y25. Anyone in HH belong to union	-1.204	.910	-1.322	.187
M1/M1.T. Abortion self-placement	2.935	1.317	2.228	.026
X2. Attend religious services how often	3.767	1.203	3.132	.002
M6ax. Summary gun control	-6.831	1.367	-4.998	.000
L7p. Inc/dec aid to blacks	-1.889	.887	-2.130	.034
L8x. Summary tax cuts from surplus	1.497	.817	1.832	.068

a. Dependent Variable: D1a/D1a.T. Thermometer Clinton

So we get 11 regression coefficients (the *slope* if you think of it as a 2-D scatterplot for each variable separately). Remember, each one represents the separate *b* for that variable, *with the other variables held constant*.

Intepret this just like with only one independent variable. The measurement of the dependent variable is where you must start: it's a 0-100 'Feeling Thermometer' about Bill Clinton. So all of the coefficients indicate how this thermometer feeling changes when the independent variables change by 1.

Or, perhaps it's more useful to think of comparing two people who are otherwise equal on all the other variables, but just one point different on the variable you're looking at.

Before you do this, you need to realize that these 11 variables all have a particular scale so that a change of one means very different things on some of them. Unless you know each variable intimately, you might as well get a report on them to help you interpret the regression table.

So you might want to run Analyze-Descriptives and make sure you get the maximum and the minimum in the Options box. But don't do this for now. Take my word for it. The output would look like this, giving you a quick reference when you're talking about each variable's relationship (*b*) to the Clinton Thermometer.

Descriptive Statistics

	N	Min	Max	Mean	Std. Deviation
thclint D1a/D1a.T. Thermometer Clinton	1144	0	100	53.58	30.239
welfare$ L7b. Inc/dec welfare programs	1324	1	4	2.74	.733
aidblck$ L7p. Inc/dec aid to blacks	1278	1	4	2.87	.709
taxcuts L8x. Summary tax cuts from surplus	1303	1	5	2.51	1.695
abortsum M1/M1.T. Abortion self-placement	1316	1	4	2.91	1.099
guncontr M6ax. Summary gun control	1337	1	5	1.99	1.056
religatt X2. Attend religious services how often	937	1	5	2.26	1.217
edsum Y3x. R educ summary	1341	1	7	4.27	1.627
unionmem Y25. Anyone in HH belong to union	1337	1	5	4.43	1.400
fincome Y27/Y27.T. HH income - others in HH 14+	795	1	22	7.48	3.592
sex ZZ1. IWR obs: R gender	1346	1	2	1.56	.497
Valid N (listwise)	420				

So what attitudes and characteristics make people feel positive or negative about Clinton? Well, probably the most interesting thing in this table is that the welfare$ variable has only a tiny effect. It's coded 1 (Increase) to 4 (Cut out Entirely). So even going from one extreme to the other on this scale, a difference of 3, translates into only a 3.4 degree[4] difference on the 0-to-100 feeling about Clinton (-1.251 * 3 = 3.75). That's not very big. NOW, look at the Sig (p-value): .185. So not only is the effect small, but we would expect an effect this big just by chance in nearly 20% of the time. Based on both of these facts it's safe to say that we're pretty confident that perhaps only a tiny, unreliable relationship between welfare spending attitudes and Clinton feeling.
(No jokes about Clinton feeling are permitted in this class! If you don't get this, google: Clinton Lewinsky)

Let's have a look at a much more powerful determinant of feelings about Clinton: the less government variable. Remember this is a two-category variable, so a difference of one

[4] It is a thermometer, after all, so we can talk about *degree* differences!

compares one group (less government) to the other group (more government). The *b* is 17.669. A pretty huge difference on the 0-100 scale. On average, and controlling for all our other independent variables, those who think there is more government could be doing are 18 degrees warmer (don't say 'hotter'!) for Clinton than those who think "the less government the better". And the p-value is .000, so this effect is really unlikely to be the result of a really unusual sample from a population (all Americans) where size-of-government attitudes are not in fact related to feeling about Clinton.

Note finally that fincome, HHINCOME, is weakly related to Clinton feelings, and the p-value is.136.

Have a look at some of the other coefficients and practice interpreting them. A table like this, but obviously with different subject matter, WILL appear on the final exam.

The most important thing I want you to get from this is that you *can* include a whole bunch of variables in a regression analysis. But you have to do so intelligently. Give as much thought to the inclusion of each variable and the interpretation of each *b* as you would if you only had one variable. For every one you need to think through exactly what it measures, how it might be related to the other independent variables, and then interpret the result with an awareness of these two things.

And to *really understand regression* you'll need to take another course in grad school! Sorry I could only bring you this far.

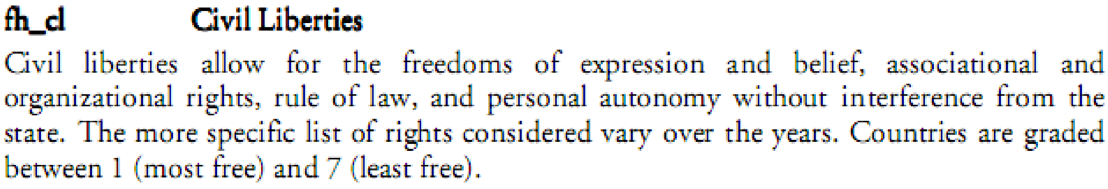
LESSON 9 – WRITING UP RESULTS

Data is taken from the Quality of Government dataset. Codebook entries are pasted here.

9.1 Frequencies: The distribution of a categorical variable.

fh_cl Civil Liberties

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	37	19.3	19.3	19.3
	2	38	19.8	19.8	39.1
	3	29	15.1	15.1	54.2
	4	27	14.1	14.1	68.2
	5	34	17.7	17.7	85.9
	6	17	8.9	8.9	94.8
	7	10	5.2	5.2	100.0
	Total	192	100.0	100.0	



fh_cl Civil Liberties

Civil liberties allow for the freedoms of expression and belief, associational and organizational rights, rule of law, and personal autonomy without interference from the state. The more specific list of rights considered vary over the years. Countries are graded between 1 (most free) and 7 (least free).

Write up the results:

9.2 Descriptives: Summary information from any variable

fh_press Freedom of the press

The operative word for this survey is "everyone." All states, from the most democratic to the most authoritarian, are through the UN system (Article 19 of the Universal Declaration of Human Rights) committed to universality of information freedom-a basic human right. We recognize that cultural distinctions or economic underdevelopment may limit the volume of news flows within a country, but these and other arguments are not acceptable explanations for outright centralized control of the content of news and information. Some poor countries allow for the exchange of diverse views, while some developed countries restrict content diversity. We seek to recognize press freedom wherever it exists, in poor and rich countries as well as in countries of various ethnic, religious, and cultural backgrounds. The scale ranges from 0 (Most free) to 100 (Least Free).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
fh_press Freedom of the press	192	8	96	44.79	25.322
Valid N (listwise)	192				

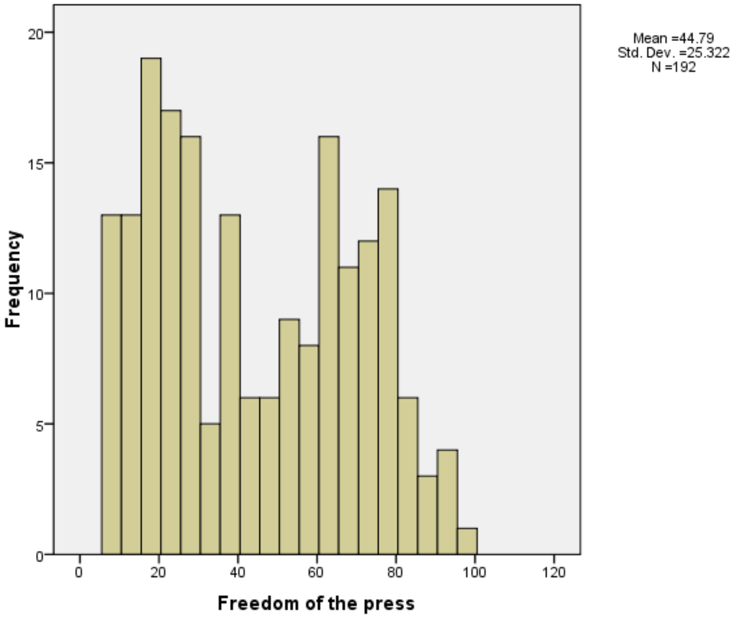
Statistics

fh_press Freedom of the press

N	Valid	192
	Missing	0
Percentiles	25	21.25
	50	40.50
	75	67.00

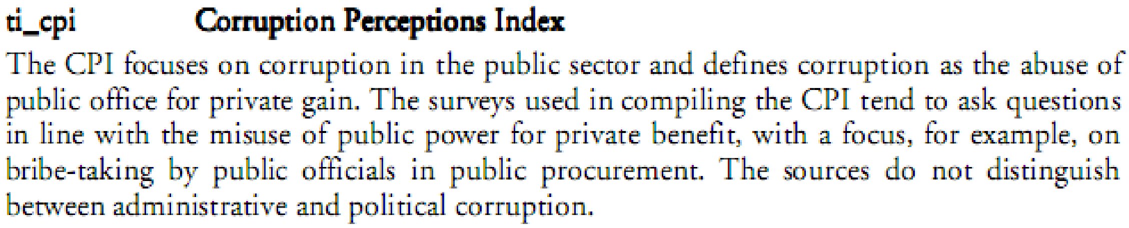
And this is from Frequencies, with the Frequencies table turned off but some statistics (i.e. quartiles) displayed.

And below is a histogram.



Write up these results:

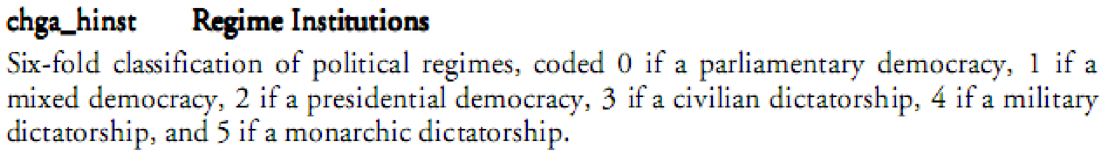
9.3 Compare Means



ti_cpi **Corruption Perceptions Index**

The CPI focuses on corruption in the public sector and defines corruption as the abuse of public office for private gain. The surveys used in compiling the CPI tend to ask questions in line with the misuse of public power for private benefit, with a focus, for example, on bribe-taking by public officials in public procurement. The sources do not distinguish between administrative and political corruption.

by



chga_hinst **Regime Institutions**

Six-fold classification of political regimes, coded 0 if a parliamentary democracy, 1 if a mixed democracy, 2 if a presidential democracy, 3 if a civilian dictatorship, 4 if a military dictatorship, and 5 if a monarchic dictatorship.

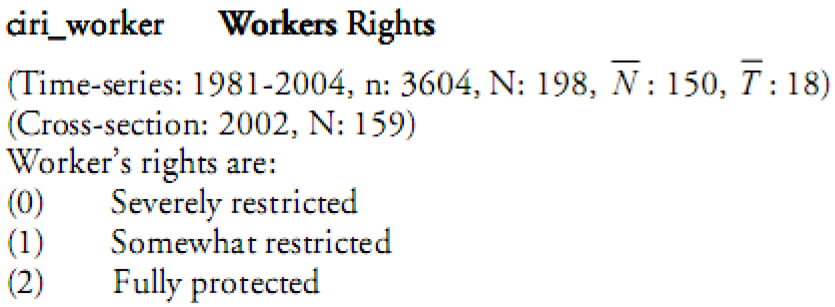
Report

ti_cpi Corruption Perceptions Index

chga_hinst Regime Institutions	Mean	N	Std. Deviation
0 0. Parliamentary Democracy	5.6082	49	2.38195
1 1. Mixed Democracy	3.6667	21	1.87599
2 2. Presidential Democracy	3.6588	34	1.71590
3 3. Civilian Dictatorship	3.0342	38	1.47085
4 4. Military Dictatorship	2.6545	22	.71696
5 5. Monarchic Dictatorship	4.5308	13	1.44302
Total	4.0045	177	2.08846

Write up the results:

9.4 Crosstabulation

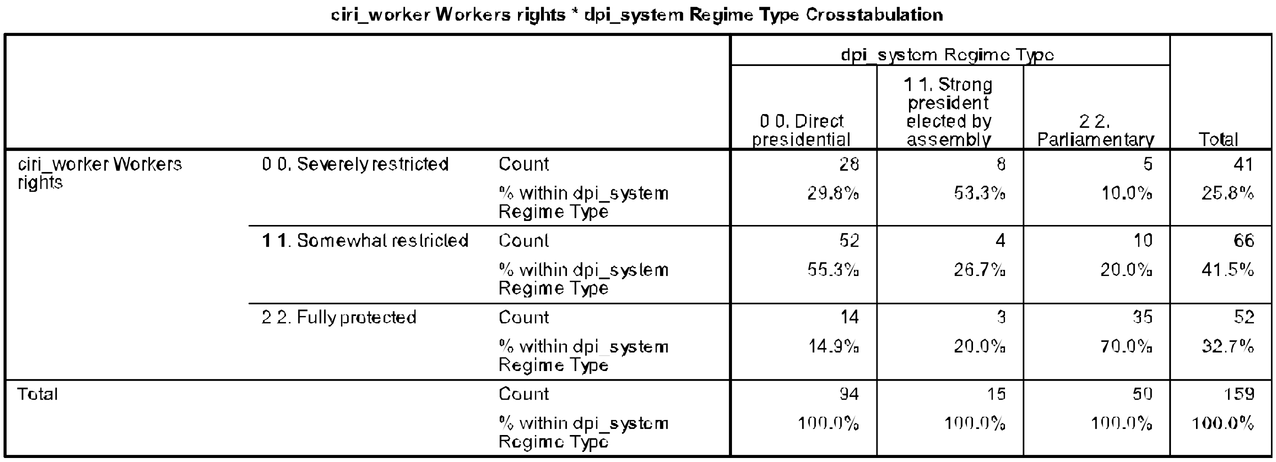


ciri_worker Workers Rights

(Time-series: 1981-2004, n: 3604, N: 198, $\overline{N}$: 150, $\overline{T}$: 18)
(Cross-section: 2002, N: 159)
Worker's rights are:
(0) Severely restricted
(1) Somewhat restricted
(2) Fully protected

by

Regime type. (Self explanatory)



ciri_worker Workers rights * dpi_system Regime Type Crosstabulation

			dpi_system Regime Type			
			0 0. Direct presidential	1 1. Strong president elected by assembly	2 2. Parliamentary	Total
ciri_worker Workers rights	0 0. Severely restricted	Count	28	8	5	41
		% within dpi_system Regime Type	29.8%	53.3%	10.0%	25.8%
	1 1. Somewhat restricted	Count	52	4	10	66
		% within dpi_system Regime Type	55.3%	26.7%	20.0%	41.5%
	2 2. Fully protected	Count	14	3	35	52
		% within dpi_system Regime Type	14.9%	20.0%	70.0%	32.7%
Total		Count	94	15	50	159
		% within dpi_system Regime Type	100.0%	100.0%	100.0%	100.0%

Write up the results:

9.5 t-test

epi_co2pc Carbon Emissions per Capita

by

sgps_oneparty Single Party System

Group Statistics

	jw_oneparty Single Party System	N	Mean	Std. Deviation	Std. Error Mean
epi_co2pc Carbon Emissions per Captia	0 0. Not a single-party system	96	10.0068	8.32516	.84968
	1 1. Single-party system	21	11.9557	13.36080	2.91557

Independent Samples Test

Dependent variables=epi_co2pc Carbon Emissions per Captia,Assumptions=Equal variances not assumed

t-test for Equality of Means						
					95% Confidence Interval of the Difference	
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
-.642	23.506	.527	-1.94897	3.03685	-8.22371	4.32576

Write up the results:

9.6 Multiple Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	37.143	4.274		8.691	.000
	r_elf85 Ethnolinguistic Fractionalization 1985	-4.669	4.320	-.126	-1.081	.285
	undp_gem Gender Empowerment Measure	25.107	9.355	.479	2.684	.010
	bl_asyt25 Average Schooling Years (Total)	.491	.684	.129	.718	.476

a. Dependent Variable: esi Environmental Sustainability Index

Here is a regression of the Environmental Sustainability Index on:

Ethnoreligious fractionalization

Gender Empowerment

Total Years of Schooling of the population over 25

Look at the codebook for how these are measured. The Descriptives procedure gives you:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
esi Environmental Sustainability Index	146	29.20	75.10	49.8795	8.47676
r_elf85 Ethnolinguistic Fractionalization 1985	171	.00	.98	.4581	.27276
undp_gem Gender Empowerment Measure	78	.12	.91	.5514	.16902
bl_asyt25 Average Schooling Years (Total)	103	.76	12.25	6.0250	2.89688
Valid N (listwise)	51				

Write up your results. This is hard because the first three of these variables are not simple.

Over and out. I hope you've found this workbook useful. Please send me any suggestions for making it a more effective learning tool.