Session 8 - SPSS[1]

More on using help in SPSS. Aggregate. Dates. A little about weights. Functions.

USING HELP and AGGREGATE.

The help within in SPSS can be very helpful to you. Bring up SPSS, and click on the help button at the upper right of the data screen. Click on Topics and then on the Index tab. Type in Aggregate. SPSS will present you with options that you can learn about. If you look at Aggregate Command Syntax you will see all of the syntax options for aggregate.

It is also possible to use SPSS's point and click process and then paste to generated command lines into a syntax window. This can be helpful when using a new command. From the Data Editor Window, I clicked on Data, Aggregate, then specified the variables commid and hhid as break variables, and then chose the variable "h5p1q6" to put into the variable window. I then clicked on "Paste" in the dialogue box, and it put the following text into my open syntax window. This is not exactly what I want, but it gave me the information I need to make the command as I want it.

```
AGGREGATE
  /OUTFILE='C:\all\spssclass\vn95\data\AGGR.SAV'
  /BREAK=commid hhid
  /h5p1q6_1 = MEAN(h5p1q6).
```

Another source of help are the .pdf files that comes on the CD with SPSS. The contain all of the SPSS manuals. If you have a UW ID, you may also access these files at:

https://depts.washington.edu/csde1/

I copied the following from the "base.pdf".

Overview

AGGREGATE aggregates groups of cases in the working data file into single cases and cre-ates a new, aggregated file. The values of one or more variables in the working file define the case groups. These variables are called **break variables.** A set of cases with identical values for each break variable is called a **break group.** A series of aggregate functions are applied to **source variables** in the working file to create new, aggregated variables that have one value for each break group.

AGGREGATE is often used with MATCH FILES to add variables with summary measures (sum, mean, etc.) to a file. Transformations performed on the combined file can create com-posite summary measures. With the REPORT procedure, the composite variables can be used to write reports with nested composite information.

AGGREGATE can be used to get sums, means, or other statistics for a group of cases. DESC and other procedures can do this too, but AGGREGATE does not print these (by default) but can be used to put them in a new data set. Why would you want to do this? As an example, think of school data. Suppose you believe that the context of the school makes a difference in the likelihood of a student to use drugs, even when controlling for the student's own characteristics. Your hypothesis is that students attending schools where there is higher rates of drug use will be more likely to use drugs than those with lower rates, even controlling for their own GPA, SES, aspirations for college, family structure, etc. To test this, you need information at the school level. You have a data set that has information on students in many schools. You can use proc summary to get information about the prevalence of drug use at the school level. You can them merge this data so that the school level information is attached to each student. Then, you can run analysis to determine if you hypothesis is correct (using proc mixed which we will not cover in class).

[1] This document has been prepared by Patty Glynn, University of Washington.

Now, as an example, we will look at a program that uses AGGREGATE on the Viet Nam data that we have been using.

```
*** PROGRAM BEGINS HERE *** .
* vn95u6.sps .
** Use AGGREGATE .
title "c:\all\SPSSclass\vn95\vn95u5.sps" .
* Documentation and data from following sites .
* http://csde.washington.edu/csde/vietnam/documents.html .
* http://csde.washington.edu/csde/vietnam/data.html .
** Read SPSS data set created by vn95.SPSS .

get file = "c:\all\SPSSclass\vn95\data\vn95.sav"
 / keep = commid hhid I5Q1B I5Q1B I5C7A I5A24 I5E11 .

* create variables with value of 1 so that we can count how many.
if I5Q1B = 1     male = 1 .
if I5Q1B = 2     female = 1 .
if female = 1 and I5E11 ge 1   womwch = 1 .
variable label womwch  'Women with children' .

if I5C7A = 1   gocoll = 1 .
if I5C7A = 2   gocoll = 1 .
variable label gocoll  'Attended college' .

if I5C7A = 3   nocoll = 1 .
if nocoll = 1  gocoll = 0 .
variable label nocoll  'NOT Attended college' .

if I5A24 = 1   Buddhist   = 1 .
if I5A24 = 2   Catholic   = 1 .
if I5A24 = 3   Protstnt   = 1 .
if I5A24 = 4   OthRel     = 1 .
if I5A24 = 5 or I5A24 = 9   NoRelig    = 1 .
compute sumrel = sum(Buddhist,Catholic,Protstnt,OthRel,NoRelig) .
variable label sumrel 'N with valid data for relig-I5A24' .

FREQUENCIES  VARIABLES=male .
DESCRIPTIVES
  VARIABLES=male
  /STATISTICS=MEAN SUM STDDEV MIN MAX .

** AGGREGATE can be used to add up all of the values across cases.
** If one or more break variables are used, cases are summarized within those variables.

** outfile = * makes the new file the active file .
** I could save it to disk instead (as shown above).
** After the aggregate, we will have fewer cases - as many cases as there are households.
AGGREGATE
  /OUTFILE= *
  /BREAK=commid hhid
  /males   = sum(male)
  /females = sum(female)
  /womwchs = sum(womwch)
  /gocolls = sum(gocoll)
  /Buddhis = sum(Buddhist)
  /Catholis= sum(Catholic)
  /Protstns= sum(Protstnt)
  /OthRels = sum(OthRel)
  /NoRelis = sum(NoRelig)
  /sumrels = sum(sumrel)
FREQUENCIES  VARIABLES=males .
DESCRIPTIVES   VARIABLES=males   /STATISTICS=MEAN SUM STDDEV MIN MAX .
```

Examine the following (abbreviated) output from the program.

MALE

					Cumulative
		Frequency	Percent	Valid Percent	Percent
Valid	1.00	2089	46.8	100.0	100.0
Missing	System	2375	53.2		
Total		4464	100.0		

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean	Std. Deviation
MALE	2089	1.00	1.00	2089.00	1.0000	.0000
Valid N (listwise)	2089					

BEFORE THE AGGREGATE, there are 4,464 cases, and 2,089 males.

MALES

					Cumulative
		Frequency	Percent	Valid Percent	Percent
Valid	1.00	1274	68.7	76.4	76.4
	2.00	370	19.9	22.2	98.6
	3.00	18	1.0	1.1	99.7
	4.00	4	.2	.2	99.9
	5.00	1	.1	.1	100.0
	Total	1667	89.9	100.0	
Missing	System	188	10.1		
Total		1855	100.0		

AFTER THE AGGREGATE: there are 1,667 cases (each HOUSEHOLD is a case now, rather than each person) and 2,089 males. The frequencies show that the number of males in each household range from 0 (shown as missing) to 5.

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean	Std. Deviation
MALES	1667	1.00	5.00	2089.00	1.2531	.4821
Valid N (listwise)	1667					

DATES

Please see the document:

http://staff.washington.edu/glynn/stat.html/spssdate.pdf

It provides information on how to calculate the difference between dates. And, following is a program that demonstrates the process.

```
title 'dates.sps' .

DATA LIST FREE / ID * name (A9) mnthb * dayb * yearb * .
BEGIN DATA
1 George    04 26 1951
2 Frank     07 24 1951
3 Sally     05 05 1968
4 Michelle  04 29 1920
END DATA.
variable label
 name   'First name'
 mnthb  'Month of birth'
 dayb   'Day of birth'
 yearb  'Year of birth'

compute mnow = 5 .
compute dnow = 15 .
compute ynow = 2000 .

* YRMODA is a date function that SPSS provides .
** It computes the number of days since October 14, 1582.
** Once a date is translated into days from any date, differences between.
** dates can be calculated .
compute gdob   = yrmoda(yearb,mnthb,dayb) .
compute gnow   = yrmoda(ynow,mnow,dnow) .
variable labels
 gdob ‘Day of Birth - Gregorian’
 gnow ‘Day now - Gregorian’ .

compute agenow = ( (gnow - gdob) / 365.25 ) .
variable label agenow ‘Age in Years’.

compute agemo = agenow * 12 .
variable label agemo ‘Age in Months, now’ .

list var = name agenow agemo.
```

Weights - the short answer: Sometimes samples are not randomly drawn. For example, in the Adolescent Health Survey, researchers wanted to over-sample middle class African American adolescents. (For some samples, such as the Adolescent Health Survey, weighting is a far more complex issue than I will deal with. There are “clusters” that also must be dealt with.) In the 1910 PUMS sample, there was an over-sample of northern African Americans. Some of the 1990 IPUMS samples are not random, and the same is true for 1940 and 1950. If a certain type of person is over-sampled, they are over-represented, and contribute more to the data set than they do to the population The values for means for the overall sample will not represent the means for the overall population unless adjustments are made. Cases that are over-sampled need to be adjusted so that they contribute less, and cases that are under-sampled need to be adjusted so that they contribute more – so that the true population values are achieved again. If analysis is done separately for the under-represented and over-represented populations, then no adjustment is necessary.

There is an additional issue. In 1% PUMS files, each person represents 100 people. The mean weight is 100. Depending on the procedure that you use, SPSS may or may not “normalize” the weights. To normalize weights means to adjust them so that the mean of weights is 1, rather than 100. This is important for inferential statistics. If your standard errors are calculated with an inflated N, the likelihood of finding a statistically significant difference is higher – and the inferences you make will be incorrect. If you are going to use inferential statistics, you weighted N should be the same as your unweighted N.

Homework: For next week, read and understand document on weighting: http://staff.washington.edu/glynn/stat.html “Adjusting, or Normalizing Weights "On the Fly" in SPSS”. Make a sample data set to practice what is covered in that document. Practice weighting separately by group, and adjusting the weights for the overall sample. Use SPSS help, work with others. Document your programs well. Understand what you are doing. Bring the programs to class next week. Run frequencies with a chi-square test unweighted and weighted – having the weights have a mean of about 100, and having the weights have a mean of 1. Examine the differences in N and chi-square values.