Session 5 - SPSS[1]

Display labels. Using the TO keyword so that every variable does not have to be listed. Some SPSS functions (truncate, max, sum), asking for descriptive statistics on a variable. Creating dichotomous variables from a categorical variable. Check frequencies of variables before using them. Value labels. An example of multiple regression, correlations, frequencies and crosstabs with chi-square.

The syntax:
display labels .

Will show the variable names, the position, and the variable labels for all variables in a file. The variables will be listed in the order of they are stored in the data set. The "position" tells which variable number it is. Knowing the order of variables is important if you want to use "to" so that you do not have to write out each variable name.

*** MODIFY THE FOLLOWING SO THAT IT WILL RUN ON YOUR MACHINE .
*** EXAMINE THE RESULTS IN THE OUTPUT WINDOW .
** vn95disp.sps .
** use display labels on the file created by vn95.sps.

```
GET
  FILE='C:\all\spssclass\vn95\data\vn95.sav'.
display labels.
```

The word "to" can be used to specify a list of variable. The variables in the file beginning with the variable preceding the "TO" and ending with the variable after the "TO" will be included. The "TO" keyword can be used in a variety of settings. In some settings, such as DO REPEAT loops, it is VERY important that you test to make sure that you KNOW what variables are being included. An error here can make a big difference.

```
*** PROGRAM BEGINS HERE *** YOU MAY MODIFY THIS AND RUN IT.
* vn95u3 sps .
title 'c:\all\SPSSclass\vn95\vn95u3.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 .
* Select only ever-married women, calculate age .
* Use "DO REPEAT"  .

get file = 'c:\all\SPSSclass\vn95\data\vn95.sav' .

* Create an ID number for all people in HH so that we can use it .
* for selection and identification .
compute idvar = 1 .
create idvar = csum(idvar) .
variable label idvar 'Identification Variable' .
execute .
* test variable - should range from 1 to 4464.

************ Selection begins *****************  .
* Select only ever-married women who are heads of households  .
* or wives of heads of households  .

* Select only ever-married women  .
* women only  .
select if I5Q1B = 2  .
* ever married only  .
select if I5C15 = 1  .

* The respondent is not asked about her relationship to head  .
```

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

```
* in the individual questionnaire.  We must get that information from  .
* the household questionnaire, and then look at the line number of the  .
* respondent.  Info for 14 people in HH is given, much check all line numbers  .

* ln is for line number ;
* rl is for relationship to head 1 = Household Head, 2 = Wife of head  .
** A "to" keyword cannot be used to list the variables in these arrays.
** A "display labels" on the file will show that these variables are not .
** in consecutive order in the file.
do repeat
  ln = H5P1Q1 H5P2Q1 H5P3Q1 H5P4Q1 H5P5Q1 H5P6Q1 H5P7Q1 H5P8Q1 H5P9Q1
       H5P10Q1 H5P11Q01 H5P12Q01 H5P13Q01 H5P14Q01 /
  rl = H5P1Q3 H5P2Q3 H5P3Q3 H5P4Q3 H5P5Q3 H5P6Q3 H5P7Q3 H5P8Q3 H5P9Q3
       H5P10Q3 H5P11Q03 H5P12Q03 H5P13Q03 H5P14Q03  .
* create a variable called target if head or relationship to head .
if rl = 1 or rl = 2  target = ln  .
end repeat .

* Select only cases where line number = target  .
* These should be household heads and wives of head  .
select if I5Q1A2 = target  .
* test .
************ Selection ends *****************  .

************* calculating age begins ************ .
* month of birth = I5C1A  .
* year of birth  = I5C1B  .
* date of interview = I5Q8  .

* The dates of interviews are in the form of 291095  .
* We need to extract the year from it  .
compute yrint = (I5Q8 - (trunc(I5Q8/100)*100)) + 1900  .

* Set I5C1B to missing if value = 9999  .
if I5C1B = 9999   I5C1B = -99  .
missing values I5C1B (-99) .
compute age = yrint - I5C1B  .
variable label age = 'Age, from  yr of birth and year of interview'  .
************* calculating age ends ************ .

***** calculate age at first marriage begins ***********  .
* I5C22 - What year did your first marriage begin - value = 98 if current marriage .
compute yr1stmr = I5C22  .
* I5C16 - is year current marriage began .
missing values I5C16 (9998) .
if yr1stmr = 98   yr1stmr = I5C16  .
compute age1stmr = yr1stmr - I5C1B  .
variable label
  yr1stmr  = 'Year of 1st marriage'
  age1stmr = 'Age of 1st marriage'  .
***** calculate age at first marriage ends   ***********  .

** The results of the command "display labels" shows the.
* var name   position  label .
* I5E4C1        773  sex of child 1
* I5E5AC1       774  month of birth
* I5E5BC1       775  Year of birth of 1st child
* I5E6C1        776  Did you want to become pregnant then?
* I5E7C1        777  Place of residence
* I5E8C1        778  Age at death (d/m/y)
*** It is the POSITION that is important in using the keyword "TO".
*** We cannot use the keyword TO in the following lists.
*** When using "DO REPEAT" loops it is IMPERITIVE that the variables be .
*** listed in the proper order !!! .
*** BUT, we are also creating variables.  If we name .
*** them ending with numbers, we can use the TO keyword .
```

```
*** BUT, since we are creating boy1 to boy14 and girl1 to girl14 .
*** in this do repeat, we will not be able to use the TO keyword .
*** to refer to these variables after the do repeat - they will not .
*** be in the proper order -- check.
do repeat
   sexch = I5E4C1   I5E4C2   I5E4C3   I5E4C4   I5E4C5   I5E4C6   I5E4C7
           I5E4C8   I5E4C9   I5E4C10  I5E4C11  I5E4C12  I5E4C13  I5E4C14 /
   boys  = boy1    to boy14  /
   girls = girl1   to girl14 /
   want1 = I5E6C1   I5E6C2   I5E6C3   I5E6C4   I5E6C5   I5E6C6   I5E6C7
           I5E6C8   I5E6C9   I5E6C10  I5E6C11  I5E6C12  I5E6C13  I5E6C14 /
   wantth= wantth1 to wantth14 .

compute boys = 0 .
if sexch = 1   boys  = 1 .
compute girls = 0 .
if sexch = 2   girls = 1 .

if want1 =  1  wantth = 1 .
end repeat .

** calculate age of children - since only these variables are .
** being created in the DO REPEAT loop, we will be able to use the .
** TO keyword to refer to them later.
do repeat
   yobch = I5E5BC1  I5E5BC2  I5E5BC3  I5E5BC4  I5E5BC5  I5E5BC6  I5E5BC7
           I5E5BC8  I5E5BC9  I5E5BC10 I5E5BC11 I5E5BC12 I5E5BC13 I5E5BC14 /
      agech = agech1  to agech14 .
if yobch gt 1996   yobch = -99 .
missing values I5E5BC1  I5E5BC2  I5E5BC3  I5E5BC4  I5E5BC5  I5E5BC6  I5E5BC7
               I5E5BC8  I5E5BC9  I5E5BC10 I5E5BC11 I5E5BC12 I5E5BC13 I5E5BC14
               (-99) .
compute agech = 1995 - yobch .
end repeat .

* Oldest child, and number of children, boys, and girls? .
** A display labels showed that the variables we cannot use a "TO" keyword .
** with the variables created in these do repeat loops .

compute numboys   = sum(boy1,boy2,boy3 ,boy4, boy5, boy6, boy7,
                        boy8,boy9,boy10,boy11,boy12,boy13,boy14,0).
variable label numboys  'number of boys' .

compute  numgirls = sum(girl1,girl2,girl3 ,girl4, girl5, girl6, girl7,
                        girl8,girl9,girl10,girl11,girl12,girl13,girl14,0).
variable label numgirls  "number of girls".

compute numchild = sum(numgirls, numboys).
variable label numchild  'sum(numgirls, numboys)'.

** since we created only one set of variables in a DO REPEAT loop .
** we can use the TO keyword with this set of variables .
compute oldestch =  max(agech1 to agech14).
variable label oldestch  'Oldest child' .

compute age1brth = age - oldestch .
variable label age1brth  'Age of woman at 1st birth' .

freq var = numboys to age1brth .
desc var = all.

** You can get information about variables with the "DESCRIPTIVES" command.
** (shortcut - "desc") You may select all or some of these statistics .
** To get MEDIAN, you must use the FREQ command .
** You may specify "var = all" to get all numeric data .
DESC
  VARIABLES=age1brth
```

```
  /STATISTICS=MEAN SUM STDDEV VARIANCE RANGE MIN MAX SEMEAN KURTOSIS SKEWNESS
  .
FREQUENCIES
  VARIABLES=age1brth
  /NTILES=  4
  /NTILES= 10
  /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
  SUM SKEWNESS SESKEW KURTOSIS SEKURT
  /ORDER=  ANALYSIS .
```

Now we will create some variables that might be interesting predictors.

```
*** PROGRAM BEGINS HERE *** .
* vn95u4 sps .
title 'c:\all\SPSSclass\vn95\vn95u4.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 .
* Select only ever-married women, calculate age .
* Use "DO REPEAT"  .

get file = 'c:\all\SPSSclass\vn95\data\vn95.sav' .

* Create an ID number for all people in HH so that we can use it .
* for selection and identification .
compute idvar = 1 .
create idvar = csum(idvar) .
variable label idvar 'Identification Variable' .
execute .
* test variable - should range from 1 to 4464.

************ Selection begins *****************  .
* Select only ever-married women who are heads of households  .
* or wives of heads of households  .

* Select only ever-married women  .
* women only  .
select if I5Q1B = 2  .
* ever married only  .
select if I5C15 = 1  .

* The respondent is not asked about her relationship to head  .
* in the individual questionnaire.  We must get that information from  .
* the household questionnaire, and then look at the line number of the  .
* respondent.  Info for 14 people in HH is given, much check all.  .
* Line numbers  .

* ln is for line number ;
* rl is for relationship to head 1 = Household Head, 2 = Wife of head  .
** A "to" keyword cannot be used to list the variables in these arrays.
** A "display labels" on the file will show that these variables are not .
** in consecutive order in the file.
do repeat
  ln = H5P1Q1 H5P2Q1 H5P3Q1 H5P4Q1 H5P5Q1 H5P6Q1 H5P7Q1 H5P8Q1 H5P9Q1
       H5P10Q1 H5P11Q01 H5P12Q01 H5P13Q01 H5P14Q01 /
  rl = H5P1Q3 H5P2Q3 H5P3Q3 H5P4Q3 H5P5Q3 H5P6Q3 H5P7Q3 H5P8Q3 H5P9Q3
       H5P10Q3 H5P11Q03 H5P12Q03 H5P13Q03 H5P14Q03  .
* create a variable called target if head or relationship to head .
if rl = 1 or rl = 2  target = ln  .
end repeat .

* Select only cases where line number = target  .
* These should be household heads and wives of head  .
select if I5Q1A2 = target  .
* test .
************ Selection ends *****************  .
```

```
************* calculating age begins ************ .
* month of birth = I5C1A  .
* year of birth  = I5C1B  .
* date of interview = I5Q8  .

* The dates of interviews are in the form of 291095  .
* We need to extract the year from it  .
compute yrint = (I5Q8 - (trunc(I5Q8/100)*100)) + 1900  .

* Set I5C1B to missing if value = 9999  .
if I5C1B = 9999   I5C1B = -99  .
missing values I5C1B (-99) .
compute age = yrint - I5C1B  .
variable label age = 'Age, from  yr of birth and year of interview'  .
************* calculating age ends ************ .

***** calculate age at first marriage begins ***********  .
* I5C22 - What year did your first marriage begin - value = 98 if current marriage .
compute yr1stmr = I5C22  .
* I5C16 - is year current marriage began .
missing values I5C16 (9998) .
if yr1stmr = 98   yr1stmr = I5C16  .
compute age1stmr = yr1stmr - I5C1B  .
variable label
  yr1stmr  = 'Year of 1st marriage'
  age1stmr = 'Age of 1st marriage'  .
***** calculate age at first marriage ends   ***********  .


** The results of the command "display labels" shows the.
* var name   position  label .
* I5E4C1        773  sex of child 1
* I5E5AC1       774  month of birth
* I5E5BC1       775  Year of birth of 1st child
* I5E6C1        776  Did you want to become pregnant then?
* I5E7C1        777  Place of residence
* I5E8C1        778  Age at death (d/m/y)
* I5E9AC1       779  Currently enrolled in what level
* I5E9BC1       780  Not currently enrolled in what level?
* I5E10C1       781  Current occupation .
*** It is the POSITION that is important in using the keyword "TO".
*** We cannot use the keyword TO in the following lists.
*** When using "DO REPEAT" loops it is IMPERITIVE that the variables be .
*** listed in the proper order !!! .
*** BUT, we are also creating variables.  If we name .
*** them ending with numbers, we can use the TO keyword .

*** BUT, since we are creating boy1 to boy14 and girl1 to girl14 .
*** in this do repeat, we will not be able to use the TO keyword .
*** to refer to these variables after the do repeat - they will not .
*** be in the proper order -- check.
do repeat
   sexch = I5E4C1   I5E4C2   I5E4C3   I5E4C4   I5E4C5   I5E4C6   I5E4C7
           I5E4C8   I5E4C9   I5E4C10  I5E4C11  I5E4C12  I5E4C13  I5E4C14 /
   boys  = boy1    to boy14  /
   girls = girl1   to girl14 /
   want1 = I5E6C1   I5E6C2   I5E6C3   I5E6C4   I5E6C5   I5E6C6   I5E6C7
           I5E6C8   I5E6C9   I5E6C10  I5E6C11  I5E6C12  I5E6C13  I5E6C14 /
   wantth= wantth1 to wantth14 .

compute boys = 0 .
if sexch = 1   boys  = 1 .
compute girls = 0 .
if sexch = 2   girls = 1 .

if want1 =  1  wantth = 1 .
end repeat .
```

```
** calculate age of children - since only these variables are .
** being created in the DO REPEAT loop, we will be able to use the .
** TO keyword to refer to them later.
do repeat
   yobch = I5E5BC1  I5E5BC2  I5E5BC3  I5E5BC4  I5E5BC5  I5E5BC6  I5E5BC7
           I5E5BC8  I5E5BC9  I5E5BC10 I5E5BC11 I5E5BC12 I5E5BC13 I5E5BC14 /
   agech = agech1  to agech14 .
if yobch gt 1996   yobch = -99 .
missing values I5E5BC1  I5E5BC2  I5E5BC3  I5E5BC4  I5E5BC5  I5E5BC6  I5E5BC7
               I5E5BC8  I5E5BC9  I5E5BC10 I5E5BC11 I5E5BC12 I5E5BC13 I5E5BC14
               (-99) .
compute agech = 1995 - yobch .
end repeat .

* Oldest child, and number of children, boys, and girls? .
** A display labels showed that the variables we cannot use a "TO" keyword .
** with the variables created in these do repeat loops .

compute numboys   = sum(boy1,boy2,boy3 ,boy4, boy5, boy6, boy7,
                        boy8,boy9,boy10,boy11,boy12,boy13,boy14,0).
variable label numboys  'number of boys' .

compute  numgirls = sum(girl1,girl2,girl3 ,girl4, girl5, girl6, girl7,
                        girl8,girl9,girl10,girl11,girl12,girl13,girl14,0).
variable label numgirls  "number of girls".

compute numchild = sum(numgirls, numboys).
variable label numchild  'sum(numgirls, numboys)'.

** since we created only one set of variables in a DO REPEAT loop .
** we can use the TO keyword with this set of variables .
compute oldestch =  max(agech1 to agech14).
variable label oldestch  'Oldest child' .

compute age1brth = age - oldestch .
variable label age1brth  'Age of woman at 1st birth' .

********** NEW PART BEGINS HERE ********************** .
************************************
* BEGIN Create dichotomous variable for religion that can be used in a regression
* Does religion of family of origin have an impact on age of first birth? ;
* 68. I5A24 What religion did your family follow when you were growing up?
* 1. Buddhist
* 2. Catholic
* 3. Protestant
* 4. Other
* 5. Did not follow any formal religion
* 9. Do not know
*                              Family religion
*                                          Cumulative    Cumulative
*       I5A24    Frequency     Percent     Frequency      Percent
*           1         335       22.19           335        22.19
*           2         285       18.87           620        41.06
*           4           4        0.26           624        41.32
*           5         882       58.41          1506        99.74
*           9           4        0.26          1510       100.00
*                        Frequency Missing = 1
************************************ .

** LESSON: Be careful when creating variables.
** Are missing coded as missing? .

compute Buddhist = 0 .
if I5A24 = 1   Buddhist   = 1 .
compute Catholic = 0 .
```

```
if I5A24 = 2   Catholic   = 1 .
compute Protstnt = 0 .
if I5A24 = 3   Protstnt   = 1 .
compute OthRel = 0 .
if I5A24 = 4   OthRel     = 1 .
compute NoRelig = 0 .
if I5A24 = 5 or I5A24 = 9 NoRelig    = 1 .

** Because there is a missing value for relig, must assign missing to new variables .
do repeat
  relig = Buddhist to NoRelig .
if missing(I5A24) relig = -9 .
end repeat .

missing values Buddhist to NoRelig (-9).
**************************** .

 ** Note that there are no protestants in our sample.
 ** For the regression, we must enter three not four dummy variables;
 ** NoRelig is the omitted category - the reference group ;
 ** END Create dichotomous variable for religion that can be used in a regression ;

****** BEGIN: Create variable for literacy ********
 253. I5C4 Are you able to read and write?
 1. Yes
 2. No
 9. DK .

compute readwrit = 0 .
if I5C4 = 1   readwrit = 1 .
variable label readwrit 'Can read and write' .
****** END: Create variable for literacy ******** .

* LESSON:  CHECK FREQUENCIES AND/OR DISTRIBUTIONS OF VARIABLES BEFORE USING THEM.
* EDUCATION --
* 254 I5C5A What is the highest grade of formal education you have completed?
*      _____ Grade (0-12)
*     99 DK
*        I5C5A    Frequency
*            0          17
*            1          19
*            2          46
*            3          66
*            4          76
*            5         119
*            6          95
*            7          86
*            8         633
*            9          82
*           10          15
*           11          15
*           12         203
*           99           2     Frequency Missing = 37  Must set 99 to missing .

missing values I5C5A (99) .

*  Create dichotomous variable - ever go to college? ;
 258. I5C7A Did you ever attend or complete a College or University
      education after finishing regular schooling?
  1. Yes, attended and completed
  2. Attended, but did not complete
  3. No, never attended
  9. DK
  * Hint.  When creating dichomous variables, name them so that the name;
  * tells you the value.  For example, instead of naming a variable "gender" ;
  * with male coded 1 and female coded 0, name the variable "male".
```

```
if I5C7A = 1  gocoll = 1 .
if I5C7A = 2  gocoll = 0 .
if I5C7A = 3  gocoll = 0 .
variable label gocoll  'Attended College' ;

** LESSON ON VALUE LABELS .

value label relig
 1='1 Buddhist'              2='2 Catholic'
 3='3 Protestant'            4='4 Other'
 5='5 No formal religion'    9='9 Do not know' .
value label gocoll readwrit  0 'No' 1 'Yes' .
value collq 1 'Completed college' 2 'Attended, no finish' 3 'Never attend' 9 'DK' .

freq var = I5A24 Buddhist to NoRelig I5C4 readwrit I5C5A gocoll I5C7A .

** Create a variable to use as filter to select only cases with non-missing values .
** for all variables that will be used in the analysis .

* age1brth Buddhist Catholic OthRel gocoll .
compute excl = 0 .
if   missing(age1brth)  or missing(Buddhist)
  or missing(Catholic)  or missing(OthRel)    or missing(gocoll)  excl = 1 .
variable label excl 'Cases to be excluded - missing values' .
freq var = excl .

**** The following code creates a filter variable.
**** Cases with a value of 1 on the variable excl will be excluded .
**** from the analysis until the command "USE ALL" is executed again .
**** I used the point and click method to create this, and pasted it into.
**** this syntax file .
USE ALL.
COMPUTE filter_$=(excl = 0 ).
VARIABLE LABEL filter_$ 'excl = 0  (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

**** I used the point and click method to create this, and pasted it into
**** this syntax file, then modified the text created to add second model .
REGRESSION
  /MISSING LISTWISE                  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)       /NOORIGIN
  /DEPENDENT age1brth
  /METHOD=ENTER buddhist catholic othrel
  /METHOD=ENTER buddhist catholic othrel gocoll .

CORRELATIONS
  /VARIABLES= age1brth buddhist catholic othrel gocoll
  /MISSING LISTWISE         /PRINT=TWOTAIL NOSIG    /MISSING= LISTWISE .

CROSSTABS
  /TABLES= I5A24 buddhist catholic protstnt othrel norelig readwrit BY gocoll
  /FORMAT= AVALUE TABLES   /STATISTIC=CHISQ   /CELLS= COUNT ROW COLUMN TOTAL .
```

Homework for next week:

As an exercise for next week, use this data set to create more variables to use in a regression equation. Write a program that runs one or more equations that you will put into a table (simulating a paper for publication).

IMPORTANT HINT: Write your program so that the equations and variables are in the same order that you will want to present them. This makes preparing the tables and proof reading much easier. Bring the programs to class. Be sure to have a program that runs from start to finish, and that creates output that you will want to put into a table. We will run the job in batch mode.