Sociology 2200: Statistics **Instructor: Natasha Sarkisian**

Introduction to Stata

[To get started, you need to complete the Stata Prep assignment. That includes downloading GSS 2012 dataset and placing it into your Stata folder.].

Opening data files

Once you have a data file saved in your Stata folder on AppStream, you can open the data file from inside Stata program using the following command:

```
. use gss2012.dta, clear
```

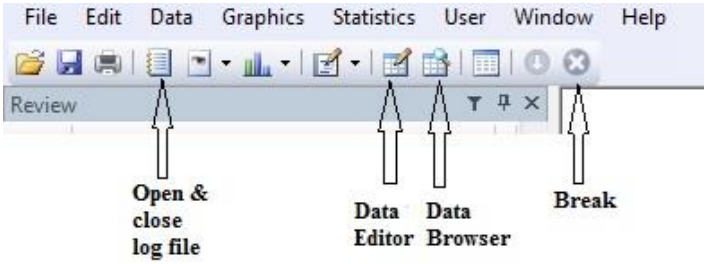
"Clear" is an option in this command that clears any other data from the memory without saving; it is not actually necessary in our situation because we do not have another dataset open, but we use it just in case.

Keeping a record of your work: Log files

To keep a record of all the commands and output when working in Stata, start by opening a log file:

```
log using learn_stata.log, replace
```

We include replace option so that you don't get an error if you already used that log file name before – it will get replaced by the new one; if you want to add to the existing log, use "append" instead of "replace. To see the log, you can at any time press the button and view a snapshot of the log. (You can also close or suspend log using that same button.) If a log is already open and you try to open another one, you will get an error message that a log is already open. You need to type "log close" to close it before opening another one.



Stata has two types of log files – they have extensions .log and .scml. I choose .log rather than .scml type of file so it can be read in any text editor or word processor. I recommend that you always use .log format for now. To ensure that, always type .log extension when typing log using command – if you just specify the name of the file but not .log, the default log type will be created, which is scml.

Note that if you are opening a Stata log file in a Word processor, you should change the font to a fixed width font, such as Courier New (otherwise the output looks misaligned). Courier New 10 or 9 point usually works the best. Otherwise things won't be aligned.

Examining the data

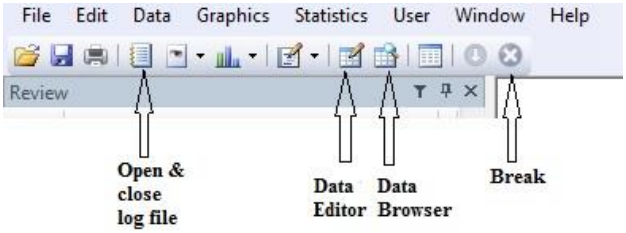
Describing the dataset (this will list all variables with their labels and basic info, as well as some overall info on the dataset characteristics, such as number of observations and number of variables):

```
. des

Contains data from gss2012.dta
  obs:         1,974
```

```
 vars:           800                          11 Sep 2013 06:50
 size:     1,717,380
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
year            int     %8.0g                 GSS YEAR FOR THIS RESPONDENT
id              int     %8.0g                 RESPONDNT ID NUMBER
wtss            double  %12.0g     WTSS       WEIGHT VARIABLE
vpsu            byte    %8.0g      LABA       Variance primary sampling unit
vstrat          int     %8.0g      LABA       Variance stratum
abany           byte    %8.0g      LABB       ABORTION IF WOMAN WANTS FOR ANY
                                                REASON
abdefect        byte    %8.0g      LABB       STRONG CHANCE OF SERIOUS DEFECT
abhlth          byte    %8.0g      LABB       WOMANS HEALTH SERIOUSLY
                                                ENDANGERED
abnomore        byte    %8.0g      LABB       MARRIED--WANTS NO MORE CHILDREN
abpoor          byte    %8.0g      LABB       LOW INCOME--CANT AFFORD MORE
                                                CHILDREN
abrape          byte    %8.0g      LABB       PREGNANT AS RESULT OF RAPE
absingle        byte    %8.0g      LABB       NOT MARRIED
accntsci        byte    %8.0g      LABC       HOW SCIENTIFIC: ACCOUNTING
accptoth        byte    %8.0g      LABD       R ACCEPT OTHERS EVEN WHEN THEY DO
                                                THINGS WRONG
acqntsex        byte    %8.0g      ACQNTSEX   R HAD SEX WITH ACQUAINTANCE LAST
                                                YEAR
actupset        byte    %8.0g      LABE       PPL AT WORK THROW THINGS WHEN
                                                UPSET WITH R
--Break--
r(1);
```

I used Break button to stop Stata from producing more output. If you do want to see all the output, either click on the more link on the bottom of the output viewer, or press space key. For some of you, more link doesn't appear and the output appears all at once rather than one page at a time. That is regulated with "set more off" and "set more on" commands in Stata.



Get codebook info (this will show variable label, numeric codes, and value labels):

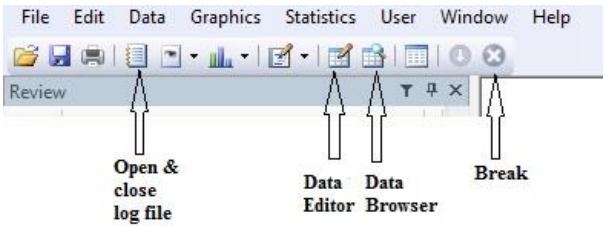
```
. codebook class
-------------------------------------------------------------------------------
class                                          SUBJECTIVE CLASS IDENTIFICATION
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  CLASS

                 range:  [1,4]                        units:  1
         unique values:  4                        missing .:  17/1974

            tabulation:  Freq.   Numeric  Label
                           200         1  LOWER CLASS
                           853         2  WORKING CLASS
                           839         3  MIDDLE CLASS
                            65         4  UPPER CLASS
                            17         .
```

Use data browser to look at the data (or type "browse" command):



Descriptive Statistics in Stata

Frequency tables – tabulate (shortened to tab) command:

```
. tab class
   SUBJECTIVE |
        CLASS |
IDENTIFICATIO |
            N |      Freq.     Percent        Cum.
--------------+-----------------------------------
  LOWER CLASS |        200       10.22       10.22
WORKING CLASS |        853       43.59       53.81
 MIDDLE CLASS |        839       42.87       96.68
  UPPER CLASS |         65        3.32      100.00
--------------+-----------------------------------
        Total |      1,957      100.00
```

This also allows us to identify the mode – here, WORKING CLASS is the mode.

Including missing values (coded in Stata as a dot .):

```
. tab class, miss

   SUBJECTIVE |
        CLASS |
IDENTIFICATIO |
            N |      Freq.     Percent        Cum.
--------------+-----------------------------------
  LOWER CLASS |        200       10.13       10.13
WORKING CLASS |        853       43.21       53.34
 MIDDLE CLASS |        839       42.50       95.85
  UPPER CLASS |         65        3.29       99.14
            . |         17        0.86      100.00
--------------+-----------------------------------
        Total |      1,974      100.00
```

To suppress labels and see numeric values:

```
. tab class, nol

 SUBJECTIVE |
      CLASS |
IDENTIFICAT |
        ION |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        200       10.22       10.22
          2 |        853       43.59       53.81
          3 |        839       42.87       96.68
          4 |         65        3.32      100.00
------------+-----------------------------------
      Total |      1,957      100.00
```

Multiple univariate tables of frequencies are obtained using tab1 command:

```
. tab1 marital class
```

```
-> tabulation of marital

      MARITAL |
       STATUS |      Freq.     Percent        Cum.
--------------+-----------------------------------
      married |        900       45.59       45.59
      widowed |        163        8.26       53.85
     divorced |        317       16.06       69.91
    separated |         68        3.44       73.35
NEVER MARRIED |        526       26.65      100.00
--------------+-----------------------------------
        Total |      1,974      100.00

-> tabulation of class

   SUBJECTIVE |
        CLASS |
IDENTIFICATIO |
            N |      Freq.     Percent        Cum.
--------------+-----------------------------------
  LOWER CLASS |        200       10.22       10.22
WORKING CLASS |        853       43.59       53.81
 MIDDLE CLASS |        839       42.87       96.68
  UPPER CLASS |         65        3.32      100.00
--------------+-----------------------------------
        Total |      1,957      100.00
```

Measures of central tendency and variability:

```
. sum tvhours
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     tvhours |      1298    3.088598      2.8651          0         24


. sum tvhours, detail
                  HOURS PER DAY WATCHING TV
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            1              0       Obs                1298
25%            1              0       Sum of Wgt.        1298

50%            2                      Mean           3.088598
                        Largest       Std. Dev.        2.8651
75%            4             24
90%            6             24       Variance       8.208798
95%            8             24       Skewness       3.123997
99%           15             24       Kurtosis       18.48296

. tabstat tvhours, stats(mean median min max p25 p75 range iqr sd variance)
    variable |      mean       p50       min       max       p25       p75
-------------+--------------------------------------------------------------
     tvhours |  3.088598         2         0        24         1         4
----------------------------------------------------------------------------

    variable |     range       iqr        sd  variance
-------------+----------------------------------------
     tvhours |        24         3    2.8651  8.208798
----------------------------------------------------------
```

Note: p50 is median, p25 is the 1st quartile, p75 is the 3rd quartile. IQR=interquartile range; SD=standard deviation. As for mode, we don't have an option for that, but you can just use tab to get a frequency distribution and identify the category with the highest frequency – that's your mode.

Help in Stata – help and search commands

If you know the name of the command and want to learn more, use help command; if you don't know the command and want to just search by keyword, use search command:

```
. help tabulate
. search median
```

Closing log and exiting Stata

Your log is automatically saved by Stata, you do not need to save it. If you use "Save as" menu, you will be saving a copy of the data file, NOT your log. We just close the log and exit Stata:

```
. log close
. exit, clear
```