Unit 5

Stata for Data Description

version 16
Also fine for earlier versions

"It is difficult to understand why statisticians commonly limit their enquiries to Averages and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances could be got rid of at once."

- Sir Frances Galton (1822-1911)

Data description is done for at least two reasons – **data management** (***e.g. – to ensure that the data clean and correct***) and **describing a sample** (***e.g – to report on who is actually represented, and what they "look like" with respect to the variables being studied?***).

Data description for data management involves the production of descriptive statistics for ***every*** study variable: (1) to explore the distributions themselves (frequencies, shape, etc); and (2) to identify missing values, errors, and extremes.

Data description for reporting describes the analysis cohort itself. It also provides a sense of the extent to which the available sample is representative of the population of interest. It is used in intervention studies for the comparison of consenters and non-consenters and in retrospective studies for the comparison of cases and controls.

Suggestion – follow along!

These notes have been written so that you can follow along and practice the commands given.
Consider: (1) Downloading from the course website two data sets: (i) **relate100obs.dta**, (ii) **wws1000.dta**.
(2) Printing out a hard copy of these notes to follow during a Stata session; and
(3) Launching Stata and following along as you read these notes….

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

Table of Contents

Learning Objectives

When you have finished this unit, you should be able to use Stata to:

- List individual values of selected variables for selected individuals in a data set**;**
- Construct frequency, relative frequency, cumulative frequency, and cumulative relative frequency tables;
- Obtain standard summary statistics (eg – mode, median, mean, sd, se, percentiles) for selected variables for selected individuals in a data set;
- Construct cross-tabulations of discrete variable distributions, overall and for selections of individuals;
- Obtain correlations among multiple continuous variables, overall and for selections of individuals;
- Produce, using a do-file, variable by variable descriptions of a sample of data;

and you should be able to:

- Produce descriptive statistics in formats that are suitable for publication.

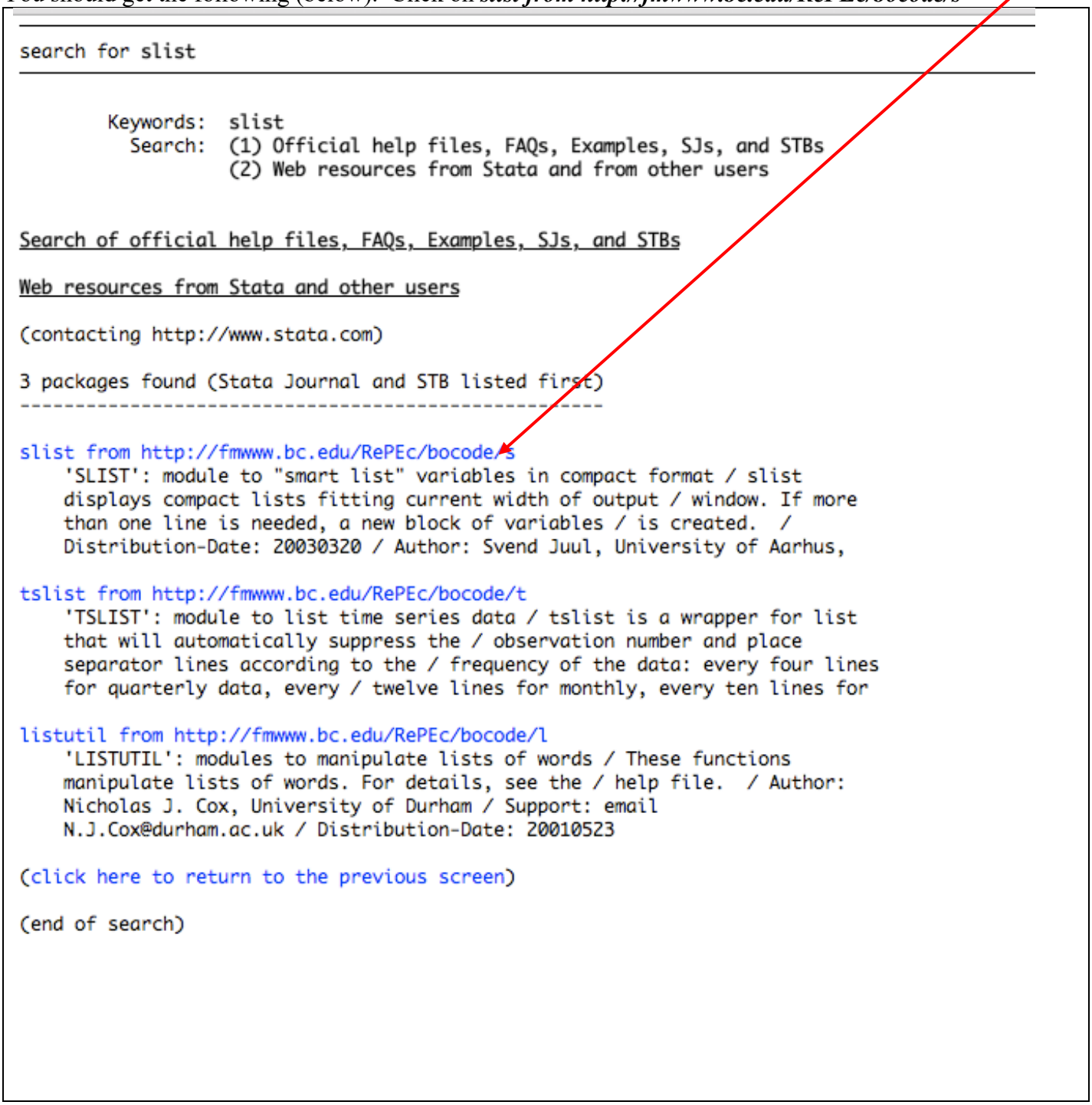
Suggestion – follow along!

These notes have been written so that you can follow along and practice the commands given.
Consider: (1) Downloading from the course website two data sets: (i) **relate100obs.dta**, (ii) **wws1000.dta**.
(2) Printing out a hard copy of these notes to follow during a Stata session; and
(3) Launching Stata and following along as you read these notes….

Preliminaries A – How to Download the Command slist

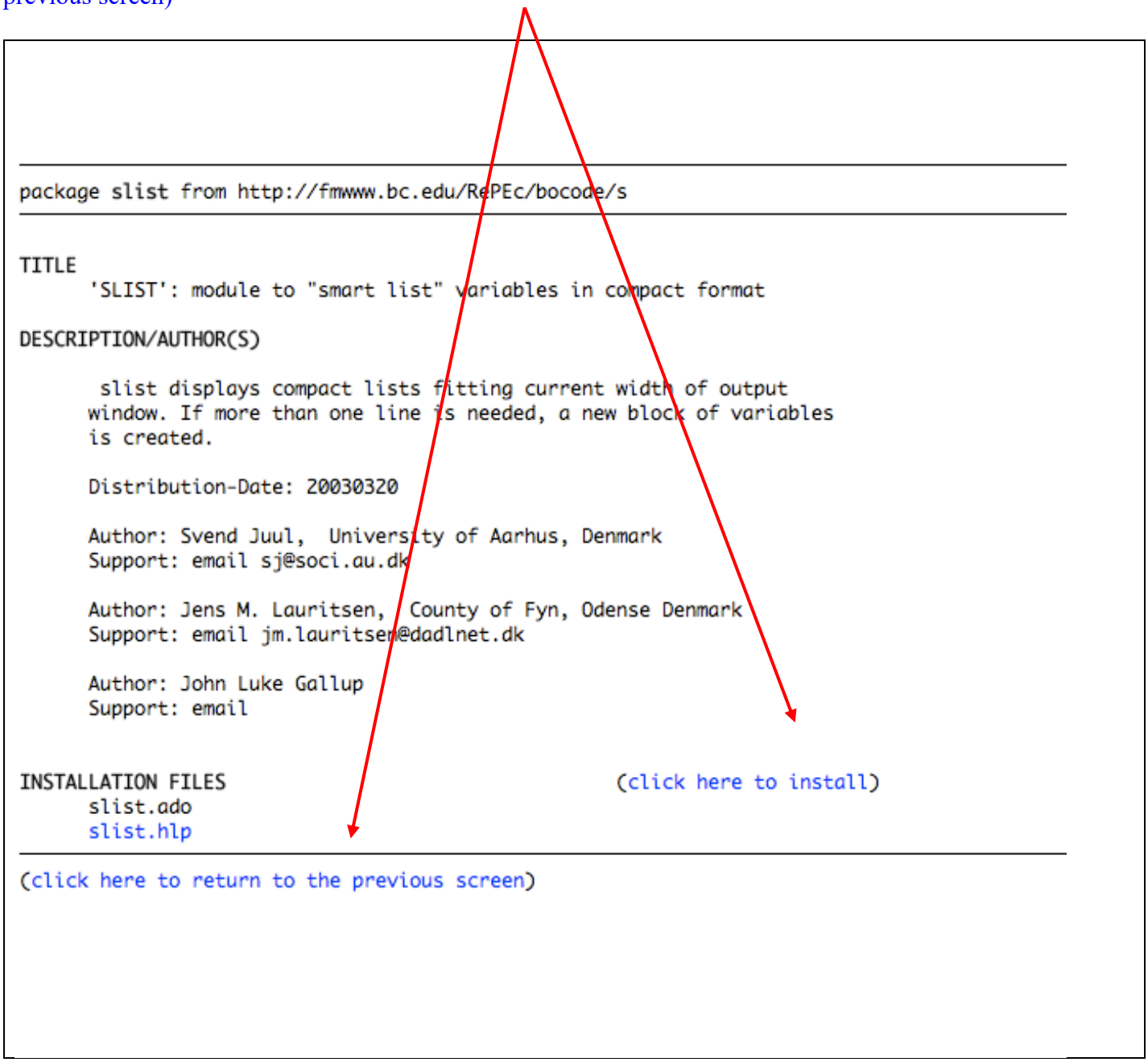
Step 1. At the command line, type findit slist

You should get the following (below). Click on ***slist from http://fmwww.bc.edu/RePEc/bocode/s***



```
search for slist

        Keywords:  slist
          Search:  (1) Official help files, FAQs, Examples, SJs, and STBs
                   (2) Web resources from Stata and from other users

Search of official help files, FAQs, Examples, SJs, and STBs

Web resources from Stata and other users

(contacting http://www.stata.com)

3 packages found (Stata Journal and STB listed first)
-----------------------------------------------------

slist from http://fmwww.bc.edu/RePEc/bocode/s
    'SLIST': module to "smart list" variables in compact format / slist
    displays compact lists fitting current width of output / window. If more
    than one line is needed, a new block of variables / is created.  /
    Distribution-Date: 20030320 / Author: Svend Juul, University of Aarhus,

tslist from http://fmwww.bc.edu/RePEc/bocode/t
    'TSLIST': module to list time series data / tslist is a wrapper for list
    that will automatically suppress the / observation number and place
    separator lines according to the / frequency of the data: every four lines
    for quarterly data, every / twelve lines for monthly, every ten lines for

listutil from http://fmwww.bc.edu/RePEc/bocode/l
    'LISTUTIL': modules to manipulate lists of words / These functions
    manipulate lists of words. For details, see the / help file.  / Author:
    Nicholas J. Cox, University of Durham / Support: email
    N.J.Cox@durham.ac.uk / Distribution-Date: 20010523

(click here to return to the previous screen)

(end of search)
```

Step 2. Click on (click here to install). When the installation is complete click on (click here to return to previous screen)

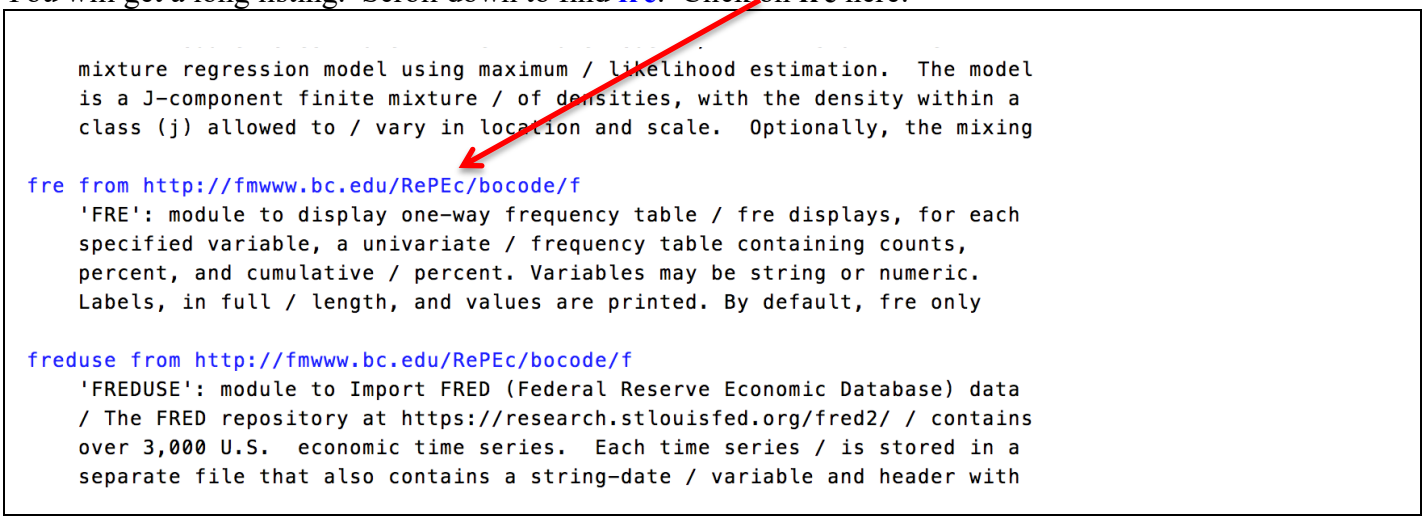


```
package slist from http://fmwww.bc.edu/RePEc/bocode/s

TITLE
      'SLIST': module to "smart list" variables in compact format

DESCRIPTION/AUTHOR(S)

       slist displays compact lists fitting current width of output
      window. If more than one line is needed, a new block of variables
      is created.

      Distribution-Date: 20030320

      Author: Svend Juul,  University of Aarhus, Denmark
      Support: email sj@soci.au.dk

      Author: Jens M. Lauritsen,  County of Fyn, Odense Denmark
      Support: email jm.lauritsen@dadlnet.dk

      Author: John Luke Gallup
      Support: email


INSTALLATION FILES                                (click here to install)
      slist.ado
      slist.hlp

(click here to return to the previous screen)
```

Preliminaries B – How to Download the Command fre

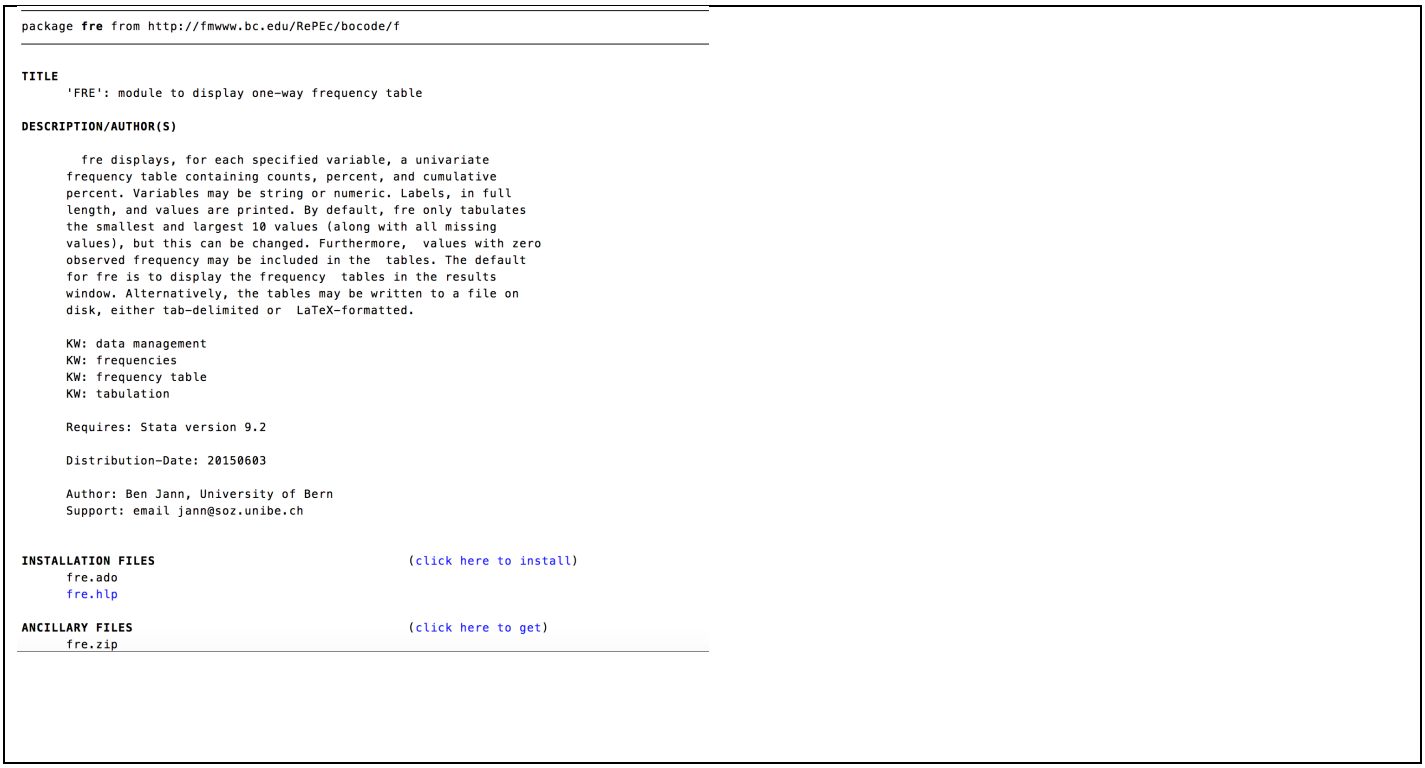
Step 1. At the command line, type findit fre

You will get a long listing. Scroll down to find **fre**. Click on **fre** here:



```
    mixture regression model using maximum / likelihood estimation.  The model
    is a J-component finite mixture / of densities, with the density within a
    class (j) allowed to / vary in location and scale.  Optionally, the mixing

fre from http://fmwww.bc.edu/RePEc/bocode/f
    'FRE': module to display one-way frequency table / fre displays, for each
    specified variable, a univariate / frequency table containing counts,
    percent, and cumulative / percent. Variables may be string or numeric.
    Labels, in full / length, and values are printed. By default, fre only

freduse from http://fmwww.bc.edu/RePEc/bocode/f
    'FREDUSE': module to Import FRED (Federal Reserve Economic Database) data
    / The FRED repository at https://research.stlouisfed.org/fred2/ / contains
    over 3,000 U.S.  economic time series.  Each time series / is stored in a
    separate file that also contains a string-date / variable and header with
```

Step 2. Click on (click here to install). When the installation is complete click on (click here to return to previous screen)



package **fre** from http://fmwww.bc.edu/RePEc/bocode/f

TITLE

'FRE': module to display one-way frequency table

DESCRIPTION/AUTHOR(S)

fre displays, for each specified variable, a univariate frequency table containing counts, percent, and cumulative percent. Variables may be string or numeric. Labels, in full length, and values are printed. By default, fre only tabulates the smallest and largest 10 values (along with all missing values), but this can be changed. Furthermore, values with zero observed frequency may be included in the tables. The default for fre is to display the frequency tables in the results window. Alternatively, the tables may be written to a file on disk, either tab-delimited or LaTeX-formatted.

KW: data management
KW: frequencies
KW: frequency table
KW: tabulation

Requires: Stata version 9.2

Distribution-Date: 20150603

Author: Ben Jann, University of Bern
Support: email jann@soz.unibe.ch

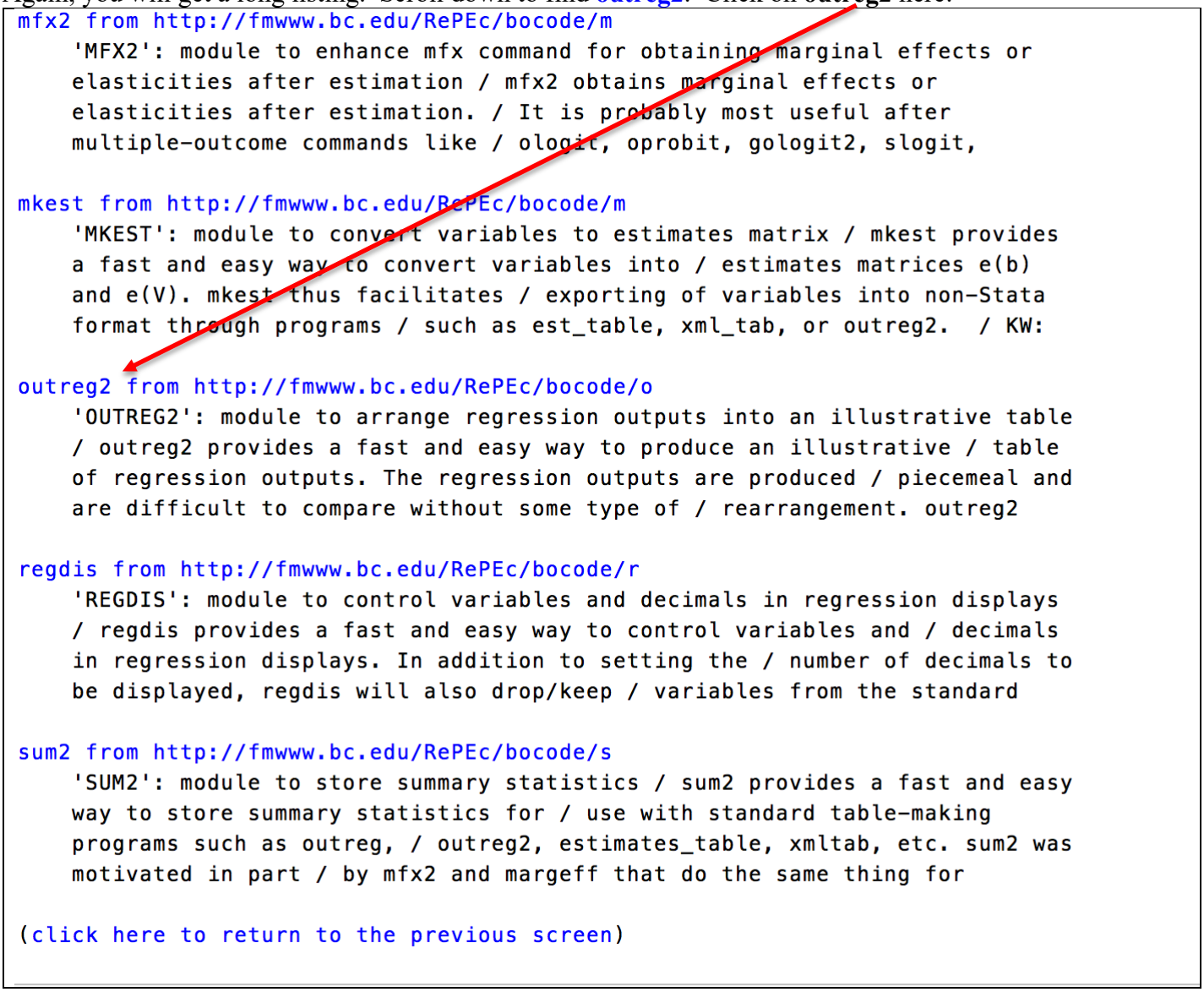
INSTALLATION FILES (click here to install)
fre.ado
fre.hlp

ANCILLARY FILES (click here to get)
fre.zip

Preliminaries C – How to Download the Command outreg2

Step 1. At the command line, type findit outreg2

Again, you will get a long listing. Scroll down to find **outreg2**. Click on **outreg2** here:



```
mfx2 from http://fmwww.bc.edu/RePEc/bocode/m
    'MFX2': module to enhance mfx command for obtaining marginal effects or
    elasticities after estimation / mfx2 obtains marginal effects or
    elasticities after estimation. / It is probably most useful after
    multiple-outcome commands like / ologit, oprobit, gologit2, slogit,

mkest from http://fmwww.bc.edu/RePEc/bocode/m
    'MKEST': module to convert variables to estimates matrix / mkest provides
    a fast and easy way to convert variables into / estimates matrices e(b)
    and e(V). mkest thus facilitates / exporting of variables into non-Stata
    format through programs / such as est_table, xml_tab, or outreg2.  / KW:

outreg2 from http://fmwww.bc.edu/RePEc/bocode/o
    'OUTREG2': module to arrange regression outputs into an illustrative table
    / outreg2 provides a fast and easy way to produce an illustrative / table
    of regression outputs. The regression outputs are produced / piecemeal and
    are difficult to compare without some type of / rearrangement. outreg2

regdis from http://fmwww.bc.edu/RePEc/bocode/r
    'REGDIS': module to control variables and decimals in regression displays
    / regdis provides a fast and easy way to control variables and / decimals
    in regression displays. In addition to setting the / number of decimals to
    be displayed, regdis will also drop/keep / variables from the standard

sum2 from http://fmwww.bc.edu/RePEc/bocode/s
    'SUM2': module to store summary statistics / sum2 provides a fast and easy
    way to store summary statistics for / use with standard table-making
    programs such as outreg, / outreg2, estimates_table, xmltab, etc. sum2 was
    motivated in part / by mfx2 and margeff that do the same thing for

(click here to return to the previous screen)
```

Step 2. Again, click on (click here to install). When the installation is complete click on (click here to return to previous screen)

Sample Session

Want to follow along?
Before you Begin – Download data from course website

(1) relate100obs.dta
(2) wws1000.dta

Sample session green-comments black-commands blue-results

```
. set more off

. * Note to reader – User edits highlighted yellow according to their computer
. use "/Users/cbigelow/Desktop/relate100obs.dta"

 . * Rename variables to be more meaningful
 . * rename OLDVARIABLENAME NEWVARIABLENAME
. rename R3483600 m_praise
. rename R3485200 f_praise
. rename R3828100 age

. * Command mvdecode.  Convert missing value codes to Stata missing values
. * mvdecode VARIABLENAME, mv(ORIGINALVALUE=MISSINGVALUE\ORIGINALVALUE=MISSINGVALUE\etc)
. mvdecode m_praise,mv(-1=.r\-2=.d\-4=.s\-5=.)
    m_praise: 43 missing values generated

. mvdecode f_praise,mv(-1=.r\-2=.d\-4=.s\-5=.)
    f_praise: 66 missing values generated

. mvdecode age,mv(-1=.r\-2=.d\-4=.s\-5=.)
    age: 4 missing values generated

. * Command label.  Create variable labels
. * label variable YOURVARIABLE "YOURVARIABLELABLE"

. label variable m_praise "m_praise: Mother praises R for doing well"
. label variable f_praise "f_praise: Father praises R for doing well"
. label variable age "age: Age of R (years)"


. * Command label define.  Create value labels.
. * label define YOURDEFINENAME value "label" value "label" etc.
. label define PRAISEF 0 "never" 1 "rarely" 2 "sometimes" 3 "usually" 4 "always"


. * Command label values.  Apply value labels to variables.
. * label values YOURVARIABLE YOURDEFINENAME
. label values m_praise PRAISEF
. label values f_praise PRAISEF.
```

Sample session, continued: green-comments black-commands blue-results

```
.* Command keep.  Keep only the variables of interest.
.* Command save.  Save under new name.
. keep m_praise f_praise age
. save relatenew100, replace
(note: file relatenew100.dta not found)
file relatenew100.dta saved

. * ---  One Variable Descriptives.  Discrete variable --*
. use relatenew100,clear

. * Before producing a one way table, instruct Stata to display number labels
. * Command is numlabel, add
. * Command tab1.  Obtain one way frequency tables, with option missing
. numlabel, add
. tab1 m_praise f_praise, missing

-> tabulation of m_praise
   m_praise: |
      Mother |
   praises R |
   for doing |
        well |      Freq.     Percent        Cum.
-------------+-----------------------------------
    0. never |          3        3.00        3.00
   1. rarely |          1        1.00        4.00
2. sometimes |         16       16.00       20.00
  3. usually |         13       13.00       33.00
   4. always |         24       24.00       57.00
           . |          4        4.00       61.00
          .s |         39       39.00      100.00
-------------+-----------------------------------
       Total |        100      100.00

-> tabulation of f_praise
   f_praise: |
      Father |
   praises R |
   for doing |
        well |      Freq.     Percent        Cum.
-------------+-----------------------------------
    0. never |          2        2.00        2.00
   1. rarely |          5        5.00        7.00
2. sometimes |          8        8.00       15.00
  3. usually |          6        6.00       21.00
   4. always |         13       13.00       34.00
           . |          4        4.00       38.00
          .s |         62       62.00      100.00
-------------+-----------------------------------
       Total |        100      100.00
```

Sample session, continued: green-comments black-commands blue-results

```
. * ---  One variable descriptives. Continuous variable --*
. * Command summarize.  After comma, option detail.
. summarize age, detail

                    age: Age of R (years)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           14              14
 5%           15              15
10%           15              15       Obs                  96
25%           16              15       Sum of Wgt.          96

50%           17                       Mean           16.91667
                      Largest          Std. Dev.      1.389181
75%           18              19
90%           19              19       Variance       1.929825
95%           19              19       Skewness       .0074556
99%           20              20       Kurtosis        2.01358


. * Command tabstat, with various options.
. tabstat age, statistics(n mean sd semean min p25 median p75 max cv)

    variable |         N      mean        sd  se(mean)       min       p25       p50       p75       max
-------------+--------------------------------------------------------------------------------------------
         age |        96  16.91667  1.389181  .1417827        14        16        17        18        20
-------------------------------------------------------------------------------------------------------

    variable |        cv
-------------+----------
         age |  .0821191
-------------------------
```

1. Data Set Description and Listing Observations

PRELIMINARY - Introduction to Stata System Datasets

Stata includes in its software several datasets that you can use. These are accessible through the system command **sysuse**. The commands that are introduced in this section are illustrated using the Stata system data set **bplong.dta**.

```
*   Preliminary – See a listing of all the Stata system datasets
. sysuse dir, all
```

Version 16 users will see:

```
. sysuse dir, all
  __i10v2003.dta  __i10v2013.dta  bplong.dta      icd9_cop.dta    surface.dta
  __i10v2004.dta  __i10v2014.dta  bpwide.dta      lifeexp.dta     tsline1.dta
  __i10v2006.dta  __i10v2016.dta  cancer.dta      network1.dta    tsline2.dta
  __i10v2007.dta  __icd10.dta     census.dta      network1a.dta   uslifeexp.dta
  __i10v2008.dta  __icd10cm.dta   citytemp.dta    nlsw88.dta      uslifeexp2.dta
  __i10v2009.dta  __icd10pcs.dta  citytemp4.dta   nlswide1.dta    voter.dta
  __i10v2010.dta  auto.dta        educ99gdp.dta   pop2000.dta     xtline1.dta
  __i10v2011.dta  auto2.dta       gnp96.dta       sandstone.dta
  __i10v2012.dta  autornd.dta     icd9_cod.dta    sp500.dta

.
```

Version 15 users will see:

```
.
. sysuse dir, all
  __i10v2003.dta    __i10v2014.dta    auto2.dta         couart2.dta       nlsw88.dta        surface.dta
  __i10v2004.dta    __i10v2016.dta    autornd.dta       educ99gdp.dta     nlswide1.dta      travel2.dta
  __i10v2006.dta    __icd10.dta       binlfp2.dta       gnp96.dta         nomocc2.dta       travel2case.dta
  __i10v2007.dta    __icd10cm.dta     bplong.dta        hormone.dta       pop2000.dta       tsline1.dta
  __i10v2008.dta    __icd10pcs.dta    bpwide.dta        icd9_cod.dta      pop2000mf.dta     tsline2.dta
  __i10v2009.dta    allstates.dta     cancer.dta        icd9_cop.dta      sandstone.dta     uis_gof.dta
  __i10v2010.dta    allstates3.dta    census.dta        lifeexp.dta       sp2001.dta        uslifeexp.dta
  __i10v2011.dta    allstatesdc.dta   citytemp.dta      network1.dta      sp2001ts.dta      uslifeexp2.dta
  __i10v2012.dta    allstatesn.dta    citytemp4.dta     network1a.dta     sp500.dta         voter.dta
  __i10v2013.dta    auto.dta          comp2001ts.dta    nlsw.dta          spjanfeb2001.dta  xtline1.dta

.
```

Introduction to Dataset Description

In dataset description we learn such things as:

- Dataset name (and any notes)
- Number and names of variables
- Variable types and, importantly, storage
- Variable value labels (if provided)
- Number of observations

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

1.1. Illustration

green-comments **black-commands** **blue-results**

```
.*  Command sysuse.   Open a Stata system data set.
.  sysuse bplong, clear
(fictional blood-pressure data)

.*  Command describe with option short.   Obtain minimal description of the data set.
. describe, short

Contains data from /Applications/Stata/ado/base/b/bplong.dta
  obs:           240                          fictional blood-pressure data
 vars:             5                          1 May 2014 11:28
 size:         1,680
Sorted by: patient


.*  Command describe with option simple.   Obtain list of variables.
. describe, simple
patient  sex      agegrp   when     bp

.
. .*  Command describe.   Obtain more detailed description of the data set.
. describe

Contains data from /Applications/Stata/ado/base/b/bplong.dta
  obs:           240                          fictional blood-pressure data
 vars:             5                          1 May 2014 11:28
 size:         1,680
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
patient         int     %8.0g                 Patient ID
sex             byte    %9.0g      sex        Sex
agegrp          byte    %9.0g      agegrp     Age Group
when            byte    %8.0g      when       Status
bp              int     %9.0g                 Blood pressure
-------------------------------------------------------------------------------
Sorted by: patient


. .*  Command label list.   Display correspondence of values and value labels.
. label list
agegrp:
           1 30-45
           2 46-59
           3 60+
sex:
           0 Male
           1 Female
when:
           1 Before
           2 After
```

```
. .*  Command codebook with option compact.   Obtain minimal numerical summaries.
. codebook, compact

Variable   Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------
patient    240    120      60.5    1  120  Patient ID
sex        240      2        .5    0    1  Sex
agegrp     240      3         2    1    3  Age Group
when       240      2       1.5    1    2  Status
bp         240     55  153.9042  125  185  Blood pressure
```

```
.*  Command order _all, alphabetic.   Alphabetize the variables in the data set.

. order _all, alphabetic

. codebook, compact

Variable   Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------
agegrp     240      3         2    1    3  Age Group
bp         240     55  153.9042  125  185  Blood pressure
patient    240    120      60.5    1  120  Patient ID
sex        240      2        .5    0    1  Sex
when       240      2       1.5    1    2  Status
-------------------------------------------------------------------------------
```

1.2 Data Set Description Commands

Command	Example
describe The command **describe** tells you basic information about the data in memory or in a file: date of creation, variable names, type, labels, etc. ***Options after a comma*** **short** Produces minimal information: sample size, variables, creation date. **simple** Produces just a list of the variable names.	**.describe** **.describe, short** **.describe, simple**
label list Displays, for each categorical variable ***for which the values have been labeled***, the numerical values and the associated labels.	**.label list**
note: ***usersupplies*** Attach a note to a data set. Use command **`note:`** (don't forget colon) to attach a note to the data set. Handy! **notes** Display notes that have been attached to this data set.	**.note: mph project data** **. notes**
codebook The command **codebook** produces very detailed information about a variable or about every variable in your data set ***Options after a comma*** **compact** The option **compact** produces a nice summary for each variable including: # non missing values, # unique values, mean, min, max	**.codebook** **This produces the very detailed codebook for every variable in the data set.** **.codebook agegrp bp** **This produces the very detailed codebook for the two variables agegrp and bp** **.codebook, compact** **This produces a very nice overview of selected information for all the variables in your data set.**

1.3 Listing Individual Observations

Command	Example
list The command **list** will produce a listing of all your observations for all your variables.	**.list** **This produces a listing of the entire data set. It may or may not be handy. Be careful here – lest you get a huge listing that you don't want!**
list *variable variable* in #/# This more detailed command will produce a listing of the observations for the particular variables you choose and for the observations you specify. See example at right.	**.list agegrp bp in 1/5** **This produces the list of values of agegrp and bp for observations 1 through 5**
Tip! The list command has some wonderful options.	**.list agegrp bp in f/5** **"f" is for "first". This ALSO produces the list of values of agegrp and bp for the first 5 observations.**
	.list agegrp bp in 5/l **Again, be careful here – lest you get a huge listing that you don't want! "l" after the slash is the letter "el". It is for "last". This command produces the list of values for observation #5 to the last observation.**
	.list agegrp bp in -5/l **Again, "l" after the slash is the letter "el". It is for "last". Note the "minus" sign before the "5". This command produces the last 5 values of agegrp and bp.**
Options after a comma (partial listing)	
nolabel The **nolabel** option will produce a listing of actual values instead of labels.	**.list, nolabel** **This produces a listing of the entire data set with the actual data values. Data value labels are suppressed.**
separator(#) or **sep(#)** The option **sep** stands for "separator". By default, Stata draws a separator line after every 5 observations. You can change this. Use separator(0) if you want no separating lines.	**.list, separator(10)** **This produces a listing of the entire data set, with separator lines after every 10 observations instead of after the default of every 5 observations.**

1.2 Listing Observations – *continued*

Command	Example
Options after a comma - continued **sepby(*variablename*)** The option **sepby** produces a line separator by the values of the variable specified. ***NOTE!*** You ***must*** have sorted the data by this variable. ***Tip!*** Include the sorting variable among the variables you are listing.	**.sort sex** **. list sex agegrp bp, sepby(sex)** **This produces a listing of sex agegrp and bp with a separator line between the listing for males and females.**
clean The **clean** option will eliminate the borders that Stata would produce otherwise	**.list sex agegrp bp in 1/5, clean** **This produces a tidy looking listing, with no box around the data values.**
noobs Use the **noobs** option to eliminate the listing of observation number	**.list sex agegrp bp in 1/5, clean noobs** **Another tidy listing, this time with observation number omitted; again- no box around the data values.**
HIGHLY RECOMMENDED for many variables … **slist, id(*variablename*) decimal(2)** **slist** is a special command that is useful when doing listings from large data sets. Use this when you have too many variables to appear across the page. The command **slist** lets you produce a split list. In order to use this command you need to install it using the **findit** command. Follow the instructions on page 28. **id** – Inside the parentheses, you can choose the variable that you want to appear in each block **decimal(#)** – choose the number of significant digits you want to display	***Tip!* – Depending on the version of Stata you have, you may have to download the special command slist. See Preliminary A, page 4.** **. slist, id(patient) decimal(2)** **.Another example of using slist that illustrates a slightly different syntax can be found on page 24.**

2. One Variable Descriptions

2.1 Illustration

Try it yourself!
This illustration also uses the data set *bplong.dta*. Again, it is not a data set that you have to download. Instead, you will use the command sysuse for access.

green-comments black-commands blue-results

```
. *
. * ------------------  Preliminaries  ----------------
. set more off
. sysuse bplong
(fictional blood-pressure data)
. numlabel, add

. *
. * --- Reorder variables so that they are alphabetical ------
. order _all, alphabetic


. *
. * Compact codebook
. codebook, compact

Variable   Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------------------
agegrp     240      3         2    1    3  Age Group
bp         240     55  153.9042  125  185  Blood pressure
patient    240    120      60.5    1  120  Patient ID
sex        240      2        .5    0    1  Sex
when       240      2       1.5    1    2  Status
-------------------------------------------------------------------------------------------


. *
. *------------  Summaries, variable by variable, in alphabetic order -------


. *--------- agegrp -----------------
. tab1 agegrp, missing

-> tabulation of agegrp

  Age Group |      Freq.     Percent        Cum.
------------+-----------------------------------
   1. 30-45 |         80       33.33       33.33
   2. 46-59 |         80       33.33       66.67
     3. 60+ |         80       33.33      100.00
------------+-----------------------------------
      Total |        240      100.00
```

```
. *------ I actually prefer the command fre (See page 6, Preliminary B for download)-
---
. fre agegrp

agegrp -- Age Group
-------------------------------------------------------------
                |      Freq.    Percent      Valid       Cum.
----------------+--------------------------------------------
Valid   1 30-45 |          80      33.33      33.33      33.33
        2 46-59 |          80      33.33      33.33      66.67
        3 60+   |          80      33.33      33.33     100.00
        Total   |         240     100.00     100.00
-------------------------------------------------------------



. *------------ bp --------------
. tabstat bp, stat(n mean sd sem q min max cv) missing

    variable |        N      mean        sd  se(mean)       p25       p50       p75       min       max
-------------+--------------------------------------------------------------------------------------------
          bp |      240  153.9042   13.0837  .8445493       144       152     162.5       125       185
----------------------------------------------------------------------------------------------------------

    variable |       cv
-------------+----------
          bp |   .085012
----------------------



. *-------------- sex -------------
. tab1 sex, missing
-> tabulation of sex

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
      0. Male |        120       50.00       50.00
    1. Female |        120       50.00      100.00
------------+-----------------------------------
      Total |        240      100.00



. *----------------- when -----------------
. tab1 when, missing
> tabulation of when

     Status |      Freq.     Percent        Cum.
------------+-----------------------------------
   1. Before |        120       50.00       50.00
    2. After |        120       50.00      100.00
------------+-----------------------------------
      Total |        240      100.00
```

2.2 One Discrete Variable

Command	Example
tab1 ***variable variable*** **tab1** produces <u>one way</u> frequency distributions for the variables listed. Stata returns frequencies, relative frequencies, and cumulative relative frequencies. **tab1** ***variablefirst-variablelast*** Note the dashed line. The result will be one way tabulations for ALL of the variables between the first and last variable, inclusive. **Tip!** You can see the order of your variables by having the variables window open. **Tip!** Alternatively, you can sort your variable order by using the command **order _all, alphabetic**. ***<u>Options after a comma</u>*** **nolabel** Use the option **nolabel** to display data values **missing** Use **missing** to include missing values. **sort** This will display the results in order of decreasing frequency. **plot** This will display a simple bar chart.	**.tab1 sex agegrp** **This produces frequency distributions for two variables: sex and agegrp.** **Tip! – In the example below, I have used the commands preserve and restore to set aside the full data set just for now. These two commands were introduced on page 52 of the unit 4 notes, *Introduction to Stata*.** **. preserve** **. order _all, alphabetic** **. drop bp patient** **. tab1 agegrp-when** **. restore** **Notes - The command order _all, alphabetic reorders the variables so that they are alphabetic. The command drop drops the variables bp and patient (so that I don't get a tabulation of patient id and bp values. The command tab1 agegrp-when produces the frequency distributions for all three variables remaining: agegrp, sex, and when.** **.tab1 sex, nolabel** **Table shows data values instead of their labels** **.tab1 sex, missing** **Missing values are included in the tabulation**
tabulate ***variable*** **tabulate** produces <u>one way</u> frequency distributions for the ONE variable listed. It also frequencies, relative frequencies, and cumulative relative frequencies. ***<u>Options after a comma are the same</u>*** **nolabel, missing, sort, plot**	**.tabulate sex**

2.2 One Discrete Variable - *continued*

Command	Example
HIGHLY RECOMMENDED! **fre** ***variablename variablename*** This has some nice advantages compared to the commands tab1 and tabulate. **fre** produces: 1) Both sample sizes and valid sample sizes; 2) Both values and value labels; 3) Missing value codes; plus 4) ***LOTS of options!*** Check them out by **help fre**	***Tip!*** **– This requires that you have downloaded the command fre. See again Preliminary B, page 6.** **. fre agegrp**

2.3 One Continuous Variable

Command	Example
summarize ***variable variable*** The command **summarize** produces the following selected summary statistics for the variables listed: n, mean, standard deviation, minimum, and maximum **summarize** This will produce the same selected statistics but for EVERY variable in the data set, regardless of whether they are discrete or continuous. Be careful in the use of this command. ***Options after a comma*** **detail** Use the option **detail** to obtain additional descriptive statistics! **Tip!** Type **help summarize** for more options.	**. summarize bp** **. summarize bp, detail**
HIGHLY RECOMMENDED! **tabstat** ***variable variable*** The command **tabstat**, in the absence of any options, produces just the mean. **Tip!** ***The command tabstat is actually quite wonderful when you combine it with its options*** ***Options after a comma (partial listing)*** **stat(*option option option option*)** Use the option stat to request the summary statistics you want. Choices are shown at right **by(*groupingvariable*)** Use the option by( ) to obtain stratified summary statistics. **Note -** You must sort the data by the grouping variable first.	**.tabstat bp** **.tabstat bp, stat(n mean sd med)** *statname* Definition mean — mean count — count of nonmissing observations n — same as **count** sum — sum max — maximum min — minimum range — range = **max** − **min** sd — standard deviation variance — variance cv — coefficient of variation (**sd/mean**) semean — standard error of mean (**sd/sqrt(n)**) skewness — skewness kurtosis — kurtosis p1 — 1st percentile p5 — 5th percentile p10 — 10th percentile p25 — 25th percentile median — median (same as **p50**) p50 — 50th percentile (same as **median**) p75 — 75th percentile p90 — 90th percentile p95 — 95th percentile p99 — 99th percentile iqr — interquartile range = **p75** − **p25** q — equivalent to specifying **p25 p50 p75**

2.2 One Continuous Variable – *continued*

Command	Example
tabstat - continued ***Options after a comma (partial listing) - continued*** **longstub** display the name of the statistics that are reported **col(stat)** display the statistics in columns instead of in rows **format(%8.2f)** This is handy in that it limits the number of significant digits displayed to 2. **Tip!** Type **help tabstat** for more options.	**. sort sex** **. tabstat bp, by(sex) stat(n mean sd) longstub** **. tabstat bp, by(sex) stat(n mean sd) col(stat)** **. tabstat bp, by(sex) stat(n mean sd) format(%8.2f)** **. tabstat bp, by(sex) col(stat) stat(n mean sd) format(%8.2f)**

2.3 One Continuous Variable Descriptions using outreg2

outreg2 is a module of Stata commands developed primarily for the output of regression output in formats (Word or Excel) that are suitable for presentation. But luckily for us, it can also be used for the output of descriptive statistics. How it works. Briefly, the results of outreg2 are sent to two places: 1) the Stata results window; and 2) a MS WORD or MS EXCEL file that you specify.

Preliminaries:

__1. Be sure to have downloaded the command **outreg2**. See again Preliminary C, page 7.
__2. Be sure to issue the command **cd** to specify the directory that will be the destination of your fancy output so that you can find it later! For example, I would type: `cd /Users/cbigelow/Desktop`
__3. Be sure to explore using the command `help outreg2`

Command	Example
Dear class – for some reason, when I try this the word file MUST have extension ".doc". If it has extension ".docx", I get an error message – cb (10/8/2019).	
Basic summary statistics for ALL variables (continuous or discrete) The basic summary statistics are n, mean, sd, min, and max **outreg2 using *file.doc*, replace sum(log)**	**. outreg2 using carol1.doc, replace sum(log)**
Detailed summary statistics for ALL variables (continuous or discrete) It's a lot. **outreg2 using *file.doc*, replace sum(detail)**	**. outreg2 using carol1.doc, replace sum(detail)**
Basic summary statistics for SELECTED VARIABLES ONLY) Note – You can select the variables you want either via keep( ) or via drop () **outreg2 using *file.doc*, replace sum(log)** **keep(*var1 var2 var3 etc*)** **outreg2 using *file.doc*, replace sum(log)** **drop(*var1 var2 var3 etc*)**	**. outreg2 using carol1.doc, replace sum(log)** **keep(agegrp bp)**
SELECTED summary statistics for SELECTED VARIABLES ONLY) **outreg2 using *file.doc*, replace sum(log)** **keep(*var1 var2 var3 etc*)** **eqkeep(*statistic statistic statistic etc*)**	**. outreg2 using carol1.doc, replace sum(log)** **keep(agegrp bp)** **eqkeep(n mean sd min max)**

Summary statistics options (case sensitive)

N	sum_w	p1	p75
mean	Var	p5	p90
sd	skewness	p10	p95
min	kurtosis	p25	p99
max	sum	p50	

Examples:

```
. * Example 1:  obtain basic stats on all variables send to word document name.doc
. outreg2 using name.doc, replace sum(log)
```

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
agegrp	240	2	0.818	1	3
bp	240	153.9	13.08	125	185
patient	240	60.50	34.71	1	120
sex	240	0.500	0.501	0	1
when	240	1.500	0.501	1	2

```
. * Example 2:  obtain selected stats on selected variables send to word document name2.doc
. outreg2 using name2.doc, replace sum(log) keep(agegrp bp) eqkeep(N mean p50 min max)
```

VARIABLES	(1) N	(2) mean	(3) p50	(4) min	(5) max
agegrp	240	2	2	1	3
bp	240	153.9	153.9	125	185

3. Multiple Variable Descriptions

3.1 Illustration

Want to follow along?
Before you Begin – Download data from course website

wws1000.dta

green-comments **black-commands** **blue-results**

```
. * User edits yellow highlighted to match their computer specifications.
. . use "/Users/cbigelow/Desktop/wws1000.dta", clear
(Working Women Survey)


. * Command order _all, alphabetic.  Reorder variables so that they are in alphabetic order
. order _all, alphabetic

. * Command codebook with option compact.  Quick look at codebook in compact form
. codebook, compact

Variable       Obs Unique      Mean    Min       Max  Label
-------------------------------------------------------------------------------------------
age           1000     28    36.276     21        83  age in current year
ccity         1000      2      .297      0         1  Does woman live a city center?
collgrad      1000      2      .241      0         1  college graduate
currexp        994     26  5.114688      0        26  Years worked at current job
everworked    1000      2      .972      0         1  Has woman ever worked?
fwt           1000     10     4.356      0         9  Frequency weight
grade          998     15  13.12425      4        18  current grade completed
grade4         998      4  2.533066      1         4  4 level Current Grade Completed
hours          998     59  37.40481      1        80  usual hours worked
idcode        1000   1000  2590.679      1      5159  Unique ID
industry       991     12  8.088799      1        12  industry
kidage1        765     21  10.34771      0        21  Age of first child
kidage2        521     15  7.047985      0        14  Age of second child
kidage3        244      8  3.430328      0         7  Age of third child
married       1000      2       .64      0         1  married
marriedyrs    1000     12     3.558      0        11  Years married (rounded to nearest year)
metro         1000      2      .704      0         1  Does woman live in metro area?
networth      1000    575  817.8472  -7000  33198.08  Net worth
nevermarried  1000      2      .104      0         1  Woman never been married
numkids       1000      4      1.53      0         3  Number of children
occupation     994     13  4.592555      1        13  occupation
prevexp        994     23  6.031187      0        25  Years worked at previous job
race          1000      3     1.275      1         3  race
south         1000      2      .422      0         1  lives in south
unempins      1000    145    30.121      0       298  Under/Unemployment insur. received last week
union          842      2   .236342      0         1  union worker
uniondues      997     30   5.47342      0        29  Union Dues paid last week
wage          1000    576  387.8163      0    380000  hourly wage
wage2         1000    513   7.82131      0      40.2  Wages, rounded to 2 digits
yrschool       998     11  13.15731      8        18  Years of school completed
-------------------------------------------------------------------------------------------
```

```
. * Command slist with options id( ) and decimal.
. *  -------  Illustration of SLIST command for the first 10 observations ----*
. slist in 1/10, id(idcode) decimal(2)

     idcode   age   ccity   collgrad   currexp   everworked   fwt   grade   grade4   hours   industry   kidage1   kidage2
  1.   4350    35       1          1        13            1     7      16        4      50         11         .         .
  2.   2709    37       1          0         8            1     6      13        3      40         12         .         .
  3.   3563    41       1          0         3            1     9      12        2      40          4         9         7
  4.    956    41       0          0         1            1     1      10        1      20         12        12         9
  5.   2710    35       1          0         6            1     8      12        2      40         11        11         7
  6.   4096    43       1          0         6            1     5      12        2      40         11         5         .
  7.   2598    27       1          0         6            1     2      14        3      40          7        13         .
  8.   1436    43       0          1         1            1     5      16        4      40         11         .         .
  9.   4125    42       1          1        15            1     5      16        4      40          4        12        11
 10.   4056    41       1          0         3            1     9      13        3      40         11         2         .

     idcode   kidage3   married   marriedyrs   metro   networth   nevermarried   numkids   occupation   prevexp   race
  1.   4350         .         1           11       1    3531.40              0         0            1         1      1
  2.   2709         .         1            6       1    6550.72              0         0            1         9      1
  3.   3563         .         1           10       1    3466.99              0         2            6         8      1
  4.    956         7         1            2       0       4.83              0         3            5         5      1
  5.   2710         1         1           11       1   -1702.09              0         3            8         6      1
  6.   4096         .         0            0       1    -212.57              0         1            3         0      1
  7.   2598         .         0            0       1     552.33              0         1            2         5      2
  8.   1436         .         0            0       1   -2740.74              0         0           13         7      1
  9.   4125         4         0            0       1    5391.30              0         3            1         2      1
 10.   4056         .         0            0       1     -27.38              0         1            2         6      1

     idcode   south   unempins   union   uniondues   wage   wage2   yrschool
  1.   4350       1          0       0           0   10.53   10.53         16
  2.   2709       1          0       0           0   13.55   13.55         13
  3.   3563       0          0       1          19   10.47   10.47         12
  4.    956       0        128       1          10    7.00    7.00         10
  5.   2710       1          0       0           0    5.30    5.30         12
  6.   4096       0          0       1          10    6.79    6.79         12
  7.   2598       1          0       0           0    7.55    7.55         14
  8.   1436       1          0       0           0    4.26    4.26         16
  9.   4125       0          0       0           0   12.39   12.39         16
 10.   4056       0          0       0           0    6.97    6.97         13


. * -----  TWO DISCRETE VARIABLES - Crosstab -----.
 * Command tab2 with options row and exact
. tab2 race numkids, row exact

-> tabulation of race by numkids

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
```

```
           |              Number of children
      race |         0          1          2          3 |     Total
-----------+--------------------------------------------+----------
         1 |       183        169        204        181 |       737
           |     24.83      22.93      27.68      24.56 |    100.00
-----------+--------------------------------------------+----------
         2 |        49         70         71         61 |       251
           |     19.52      27.89      28.29      24.30 |    100.00
-----------+--------------------------------------------+----------
         3 |         3          5          2          2 |        12
           |     25.00      41.67      16.67      16.67 |    100.00
-----------+--------------------------------------------+----------
     Total |       235        244        277        244 |     1,000
           |     23.50      24.40      27.70      24.40 |    100.00

          Fisher's exact =                 0.378


. * -----  MULTIPLE CONTINUOUS VARIABLES -----.
. *  Command tabstat with options col, stat( ) and format( )
. tabstat age uniondues wage, col(stat) stat(n mean sd min max) format(%8.2f)

    variable |         N      mean        sd       min       max
-------------+----------------------------------------------------
         age |   1000.00     36.28      5.62     21.00     83.00
   uniondues |    997.00      5.47      8.95      0.00     29.00
        wage |   1000.00    387.82  12016.41      0.00   3.8e+05
----------------------------------------------------------------



. *
. *-------- ONE DISCRETE and ONE CONTINUOUS VARIABLE --------*
. *------- Don't forget to sort first!! --------*
. sort race
. *  Command tabstat with options by( ), col, stat( ) and format( )
. tabstat age uniondues wage, by(race) col(stat) stat(n mean sd min max) format(%8.2f) long

race         variable |         N      mean        sd       min       max
----------------------+----------------------------------------------------
1                 age |    737.00     36.58      5.59     21.00     83.00
            uniondues |    736.00      5.35      8.96      0.00     29.00
                 wage |    737.00      8.18      6.10      1.03     40.20
----------------------+----------------------------------------------------
2                 age |    251.00     35.39      5.63     22.00     45.00
            uniondues |    249.00      5.99      9.06      0.00     29.00
                 wage |    251.00   1520.68  23984.96      0.00   3.8e+05
----------------------+----------------------------------------------------
3                 age |     12.00     36.33      5.87     25.00     43.00
            uniondues |     12.00      2.00      4.84      0.00     15.00
                 wage |     12.00      7.79      4.41      1.81     17.53
----------------------+----------------------------------------------------
Total             age |   1000.00     36.28      5.62     21.00     83.00
            uniondues |    997.00      5.47      8.95      0.00     29.00
                 wage |   1000.00    387.82  12016.41      0.00   3.8e+05
------------------------------------------------------------------------
```

```
. *
. *---------  MULTIPLE CONTINUOUS VARIABLES - correlations ------
. * Command correlate with option covariance will produce a variance-covariance matrix
*
. correlate age uniondues wage, covariance
(obs=997)

             |      age uniond~s     wage
-------------+---------------------------
         age |  31.6838                          key:  variance(age) = 31.6838
   uniondues | -.771014  80.1592                       covariance(uniondues, age) = -0.771014
        wage | -102.753 -2082.86  1.4e+08


. *  Command pwcorr with options obs and sig will produce a correlation matrix
. pwcorr age uniondues wage, obs sig

             |      age uniond~s     wage
-------------+---------------------------
         age |   1.0000
             |
             |     1000
             |
   uniondues |  -0.0153   1.0000
             |   0.6295
             |      997      997
             |
        wage |  -0.0016  -0.0193   1.0000
             |   0.9605   0.5421
             |     1000      997     1000
             |
```

3.2 Two Discrete Variables

Want to follow along?

Multivariable descriptions are also illustrated using the data set ***bplong.dta***. Recall that, on page 19, an example was done using only a selection of the variables. If you did not use **preserve** and **restore** when you were on page 19, enter the following at the command line before you follow the examples in this section:

. sysuse bplong, clear

Command	Example
tabulate ***rowvariable columnvariable*** **tab2** ***rowvariable columnvariable*** The command **tab2** produces a two way cross-tabulation. In Stata (unlike SAS), you have to request the percentages you want. The default is a cross-tabulations of frequency counts only. See options below. **Tip!** The command **tab2** does not have options for obtaining risk ratios or odds ratios. For these, consider the commands **cs, cc,** or **tabodds** (**More on this in Unit 8, "*Stata For Categorical Data Analysis*"**) ***Options after a comma*** **row - d**isplay row percentages **column** or **col** - display column percentages. **cell -**display cell percentages. **Exact -** Display p-value for Fisher's exact test of null hypothesis of zero association **nolog** Use with option exact ONLY. The nolog option restricts the output displayed with the option exact **chi2** - Display p-value for Pearson's chi square test of null hypothesis of zero association. **nolabel -** Display actual data values instead of data value labels. **missing -** Include missing values in tabulation.	**.tab2 sex agegrp** **This produces a cross-tabulation of the frequencies for each joint combination of sex and agegrp.** **.tab2 sex when, row col exact nolog nolabel** **This produces a cross-tabulation of the frequencies for each joint combination of sex and when, together with both row and column percentages plus a fisher exact test of the null hypothesis of no association.**

3.2 Two Discrete Variables - *continued*

Command	Example
tabi #11 #12 …\#21 #22 …\ etc, options Very handy! This is an example of an "immediate" command. Specifically, you type in the cell entries of the table. You then request the analysis you want. To do this, you list these as options after the comma. ***Options after a comma (partial listing)*** **row** – display row percents **col** – display column percents **cell** – display cell percent **exact** – fisher's exact test **chi2** – Pearson chi square test **Tip!** Type **help tabi** for more options..	**.tabi 20 125 98\100 612 514, exact** **This produces a fisher's exact test for the following 2x3 table of cell counts. Handy!** 20 125 98 100 612 514

3.3 One Discrete Variable and Multiple Continuous Variables

Command	Example
bysort *discretevariable*: summarize *continuousvariable* **bysort *discretevariable*: summarize *continuousvariable*, detail** **Tip!** Don't forget the colon. It's hard to see here.	**.bysort sex: summarize bp, detail** **discrete** **continuous**
table *discretevariable*, contents(mean *variablename* sd *variablename* min *variablename* etc) **Tip!** Be careful. Here, the separator is a comma, not a colon . ***Options for contents (partial listing)*** **n** – number of non missing observations **count** – number of nonmissing observations **mean** – mean **sd** – standard deviation **semean** – standard error of the mean **min** – minimum **max** – maximum **median** – median **iqr** – interquartile range **semean** – standard error of the mean **p1** – 1st percentile **p15** – 15th percentile etc. **Tip!** Type **help table** for more options.	**.sort sex** **.table sex, contents(mean bp sd bp)** **discrete** **continuous**

3.3 One Discrete Variable and Multiple Continuous Variables - continued

Command	Example
tabstat ***variable variable,*** **by(*groupingvariable*) stat(*option option option option*)** Note - this was introduced on page 21. ***Options after a comma (partial listing)*** **by(*groupingvariable*)** Use the option by( ) to obtain stratified summary statistics. **Note -** You must sort the data by the grouping variable first. **long** Use this with the option so that you will see the variable names **stat(*option option option option*)** Use the option stat to request the summary statisticis you want. Choices include, but are not limited to: **n, mean, sd, min, max, sem, med.** **longstub -** display the name of the statistics that are reported **col(stat)** display the statistics in columns instead of in rows **format(%8.2f)** This is handy in that it limits the number of significant digits displayed to 2. **Tip!** Type **help tabstat** for more options.	**. sort sex** **. tabstat bp, by(sex) stat(n mean sd)** **. sort sex** **.tabstat bp, by(sex) stat(n mean sd) col(stat)** **. sort sex** **.tabstat bp, by(sex) stat(n mean sd) format(%8.2f)** **. sort sex** **.tabstat bp, by(sex) col(stat) stat(n mean sd) format(%8.2f)**

3.4 Multiple Continuous Variables

Command	Example
correlate ***variable variable variable*** This will produce a correlation matrix. If you want to see the variance-covariance matrix, be sure to request the option covariance. ***Options after a comma (partial listing)*** **covariance –**display covariance matrix instead of correlation matrix **means –**display means, sd, min and max **Tip!** Type **help correlate** for more options	**. correlate bp age, covariance**
pwcorr ***variable variable variable*** This will produce all pairwise correlations. ***Options after a comma (partial listing)*** **obs -** display number of observations used **sig –** display significance level **print(#) –** display significance level ONLY if it is less than the threshold indicated by "#" **Tip!** Type help pwcorr for more options.	**Note – It is not possible to do the example below using bplong.dta because it does not contain the variable chol** **. pwcorr bp age chol, obs print(.10)**

3.5 One Continuous Variable by Group Using `outreg2`

Note - You CANNOT specify both **`keep(`*`variable 1 variable2 etc`*`)`** and **`eqkeep(`*`statistic1 statistic2 etc`*`)`** at the same time. Soooooo ….

Hack - To get around this, use **`preserve`** and **`restore`** to keep just the variables you want: (1) issue the command **`preserve`** (2) issue the command **`keep`** ***`variable1 variable2 etc`*** (3) issue your **`outreg2`** command and finally (4) issue the command **`restore`**

Hack - Take care not to specify too many statistics as these will become column headings in your WORD file

Command	Example

preserve
 	
keep *variable1 variable2 etc*
 	
bysort *groupvariable*: outreg2 using *file.doc*, replace sum(log) eqkeep(*statistic statistic statistic etc*)
 	
restore
 	
Summary statistics options (case sensitive)
 	
`N sum_w p1 p75`
`mean Var p5 p90`
`sd skewness p10 p95`
`min kurtosis p25 p99`
`max sum p50`
	. preserve
	
	. keep married uniondues networth
	
	. bysort married: outreg2 using carol2.doc, replace sum(log) eqkeep(N mean sd)
	
	. restore

Example:

```
. * Example
. preserve
. keep married uniondues networth
. bysort married: outreg2 using name3.doc, replace sum(log) eqkeep(N mean sd)
. restore
```

	(1)	(2)	(3)	(4)	(5)	(6)
	married 0			married 1		
VARIABLES	N	mean	sd	N	mean	sd
uniondues	360	6.267	9.381	637	5.025	8.677
networth	360	1,082	6,532	640	669.3	5,491