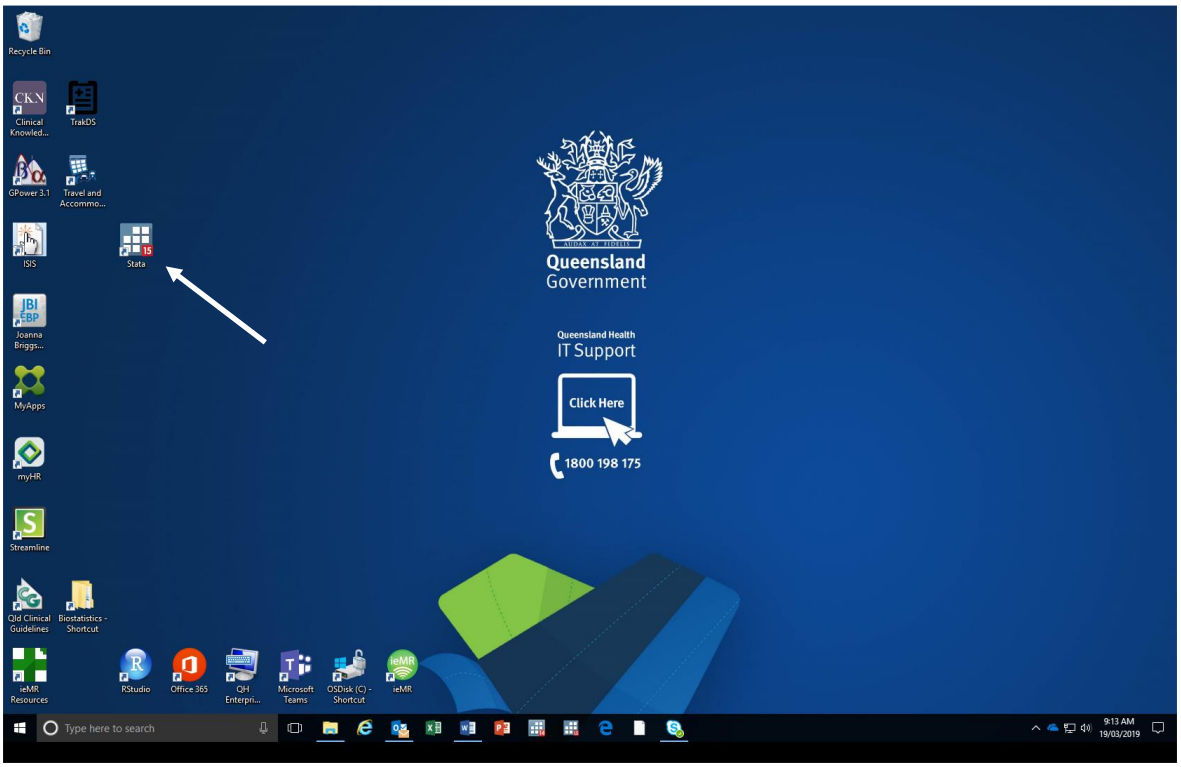
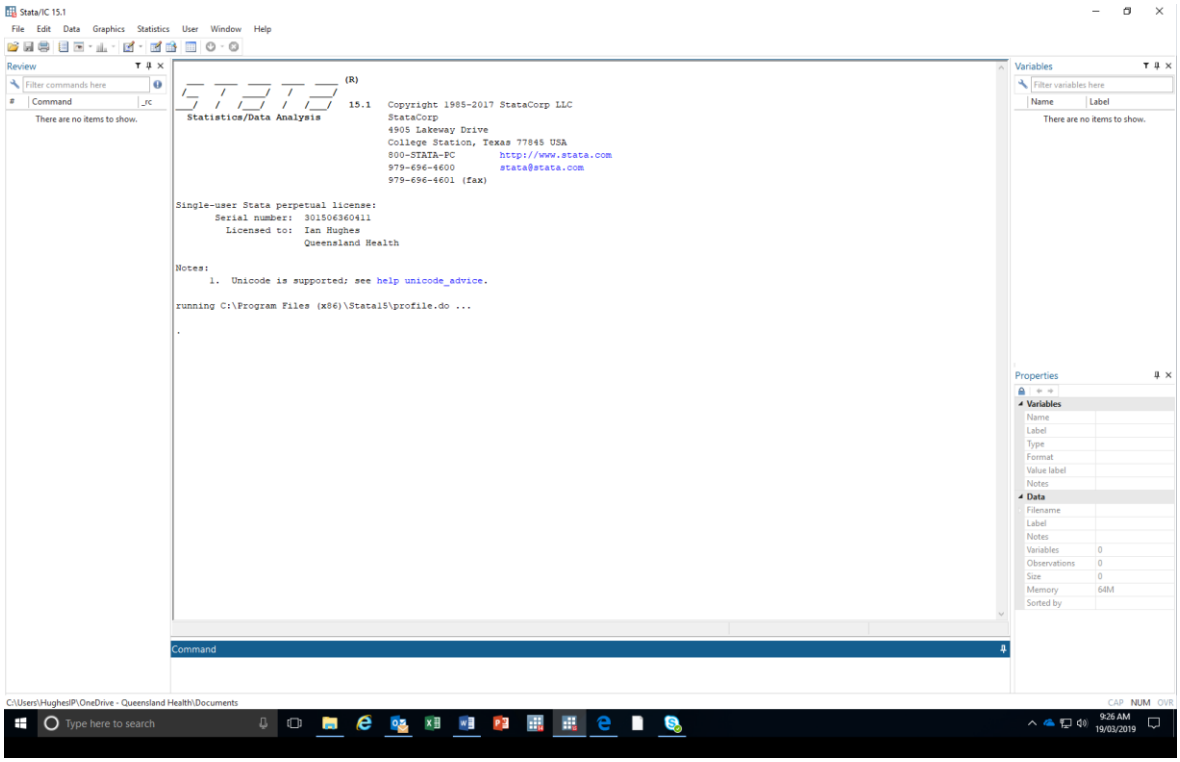
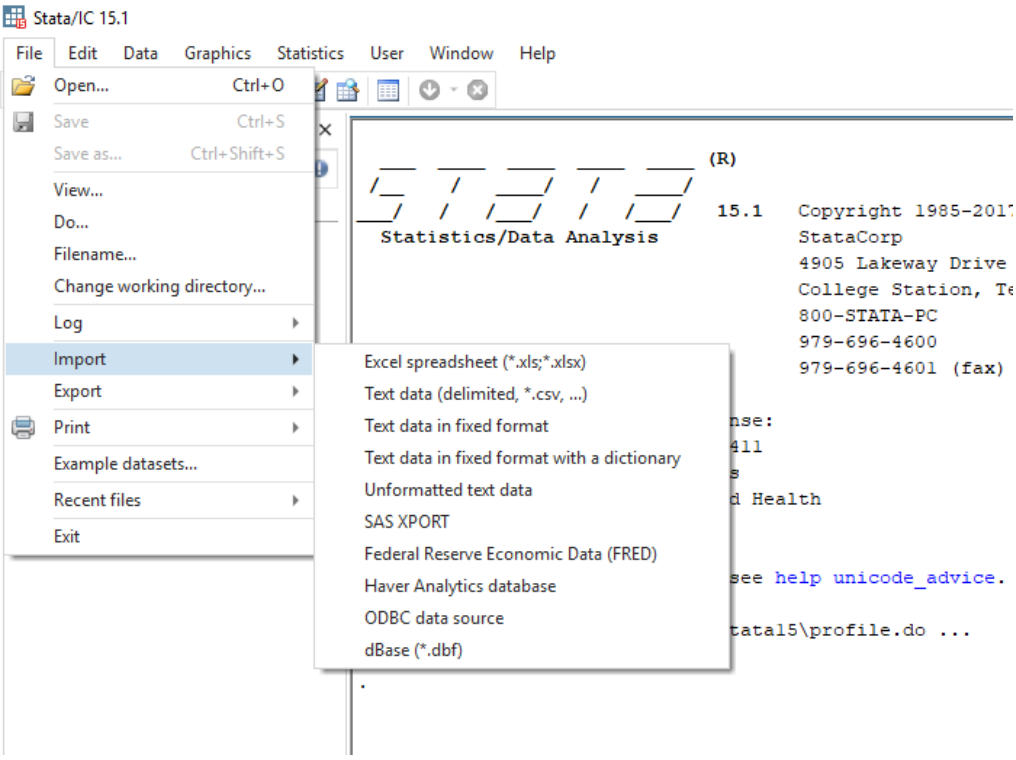
Introduction to Using Stata15 at Gold Coast Health

The Stata 15 icon should be located in the start menu and can be moved to the desk top. Double click on the icon to open Stata.

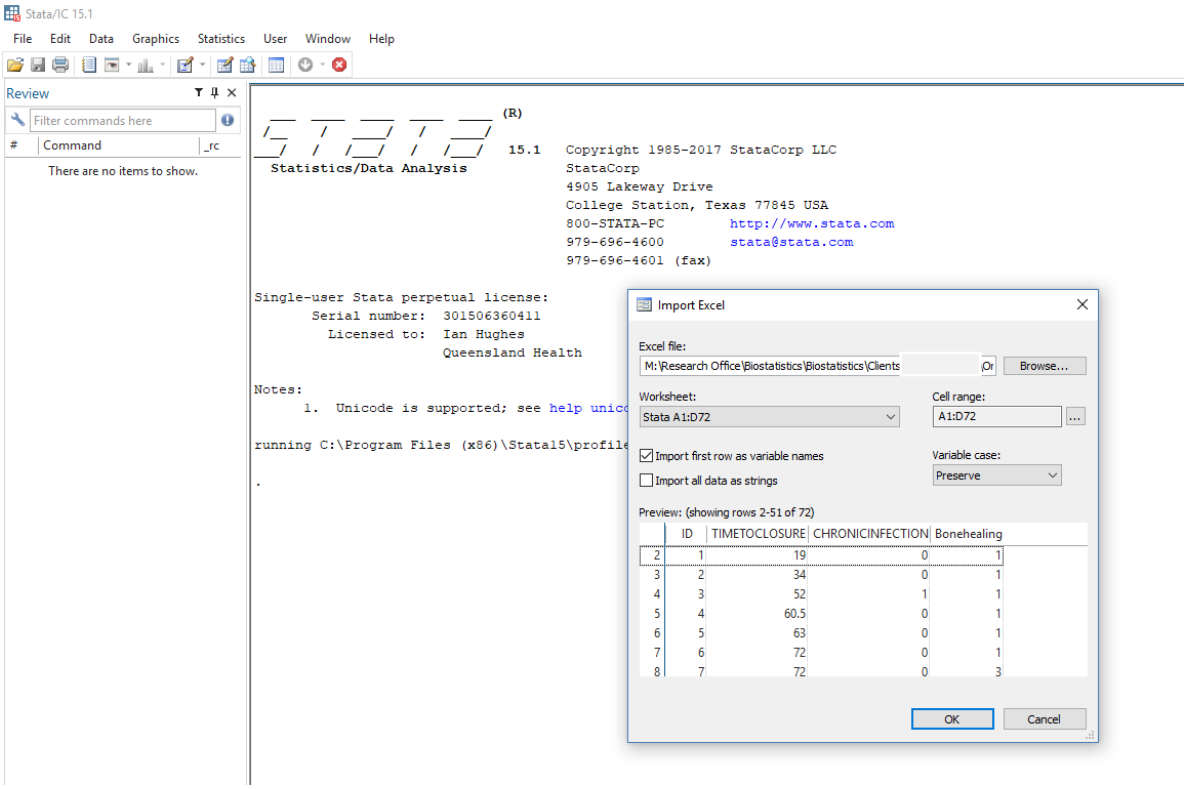




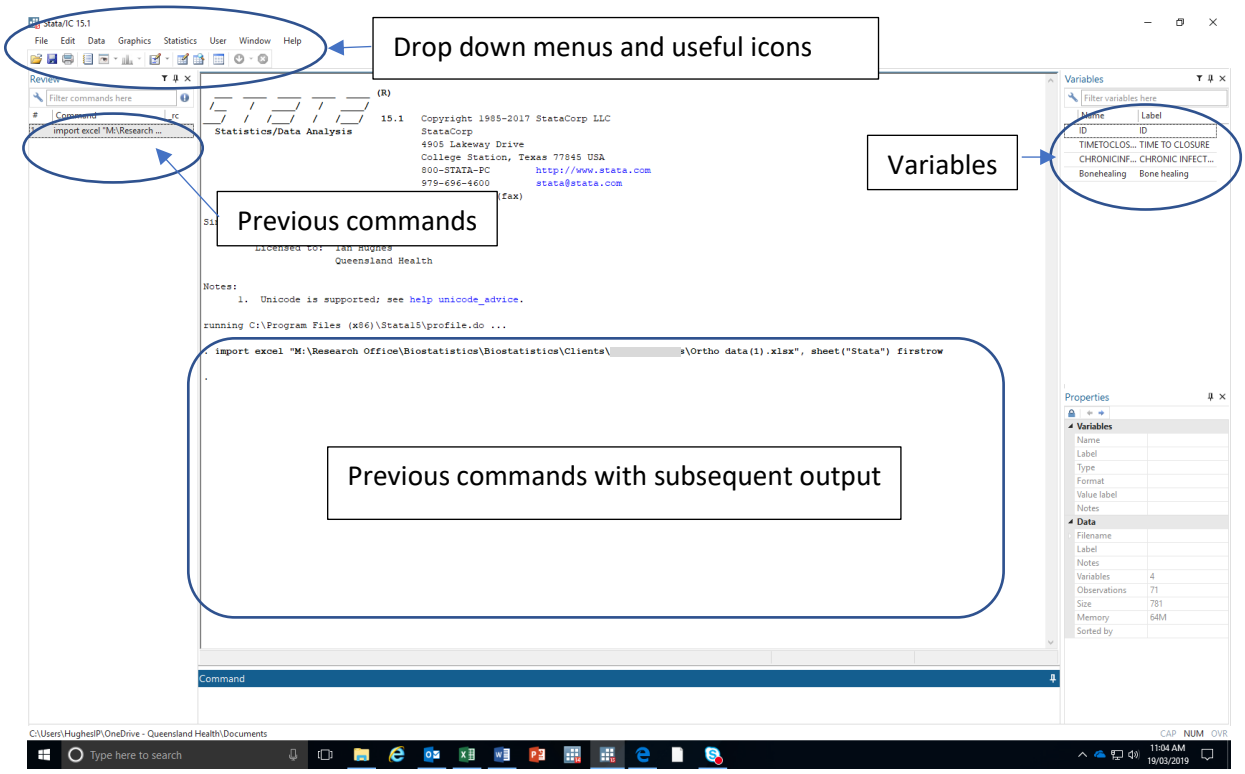
The Stata screen should appear as above. To import data from an Excel file click the File tab, Import, Excel spreadsheet as shown below.



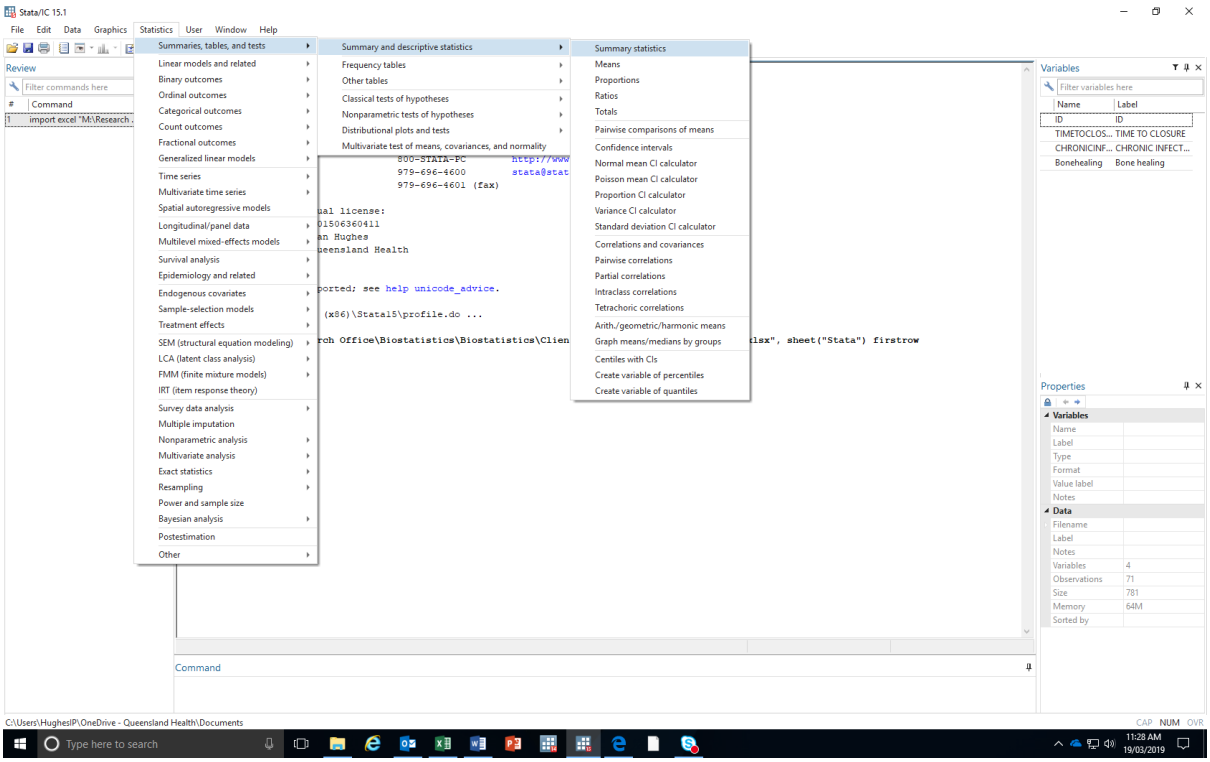
Browse to the file you want to import. Pick the worksheet you want, tick “Import first row as variable names”, click ok.



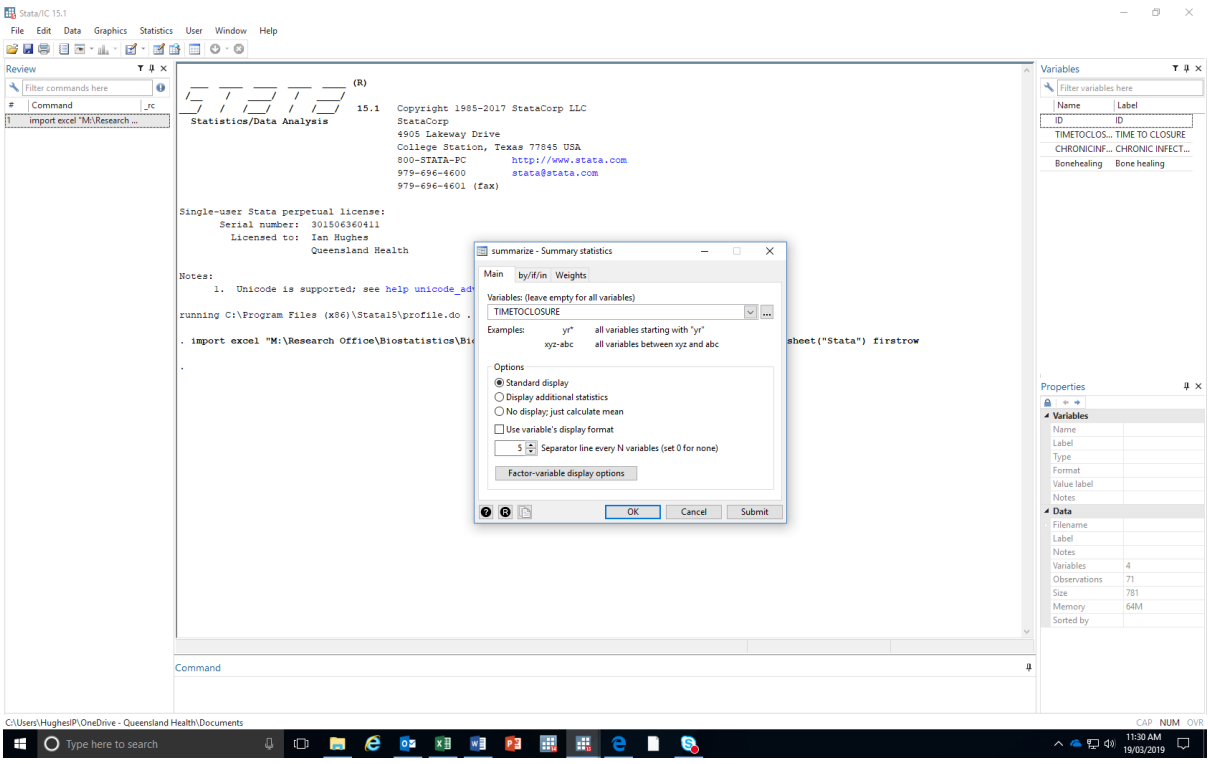
The figure below shows the layout of the Stata home screen once data has been imported.



The drop-down menus can be used to select statistics, graphics etc you may be interested in. To investigate the data we can do some summary statistics. Click on Statistics, Summaries…, Summary and descriptive, Summary statistics as shown below.

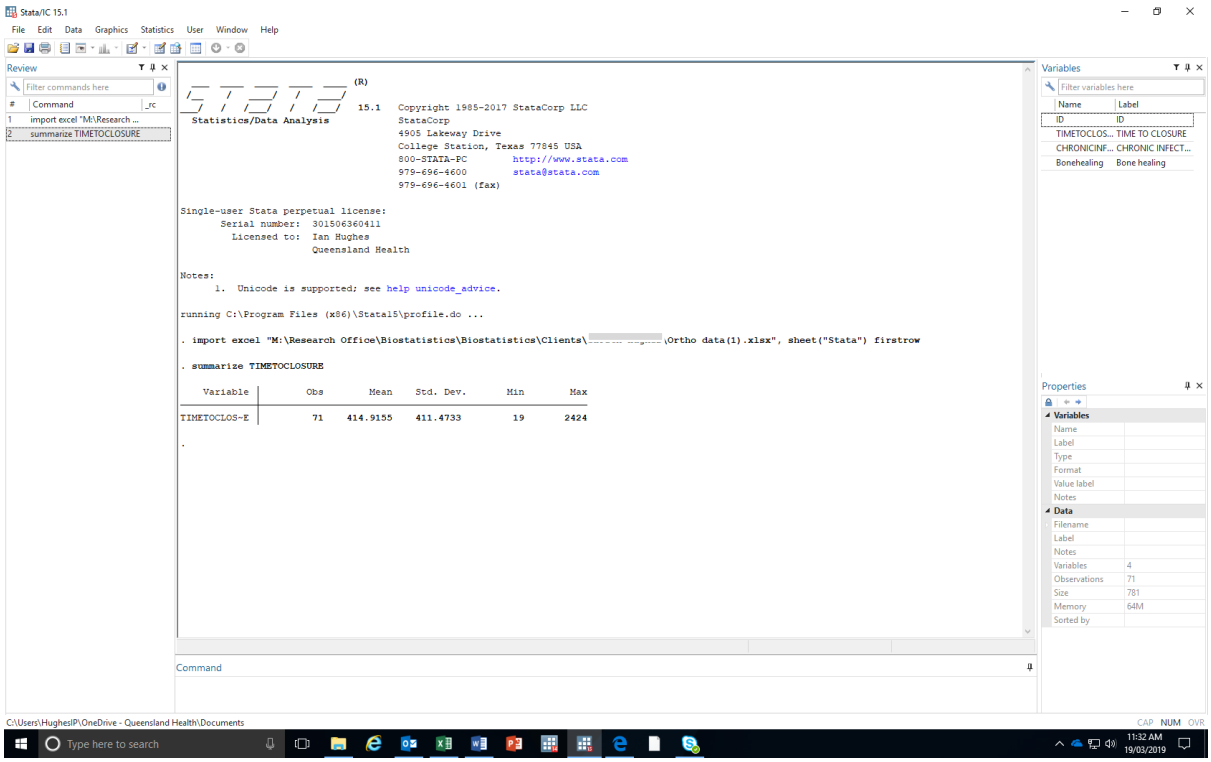


A dialog box appears. Click on the arrow button of the Variables box and select the variable you would like to find the summary statistics of. Click ok.

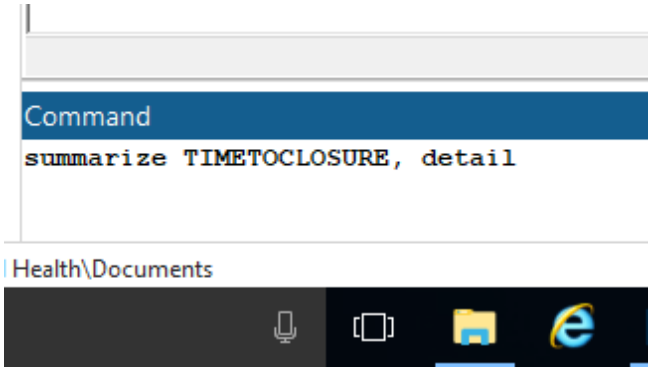


The results are displayed in the main window. Immediately above the results is the syntax that produced the results. You could type this into the Command window to get the same

results as from using the dropdown menus. The syntax is also shown in the left-hand panel which shows a history of commands. Clicking on a command here will copy it to the command window.



Click on the last command in the left-hand panel. Go to the Command window and add a comma and “detail”, as shown below. Click enter.



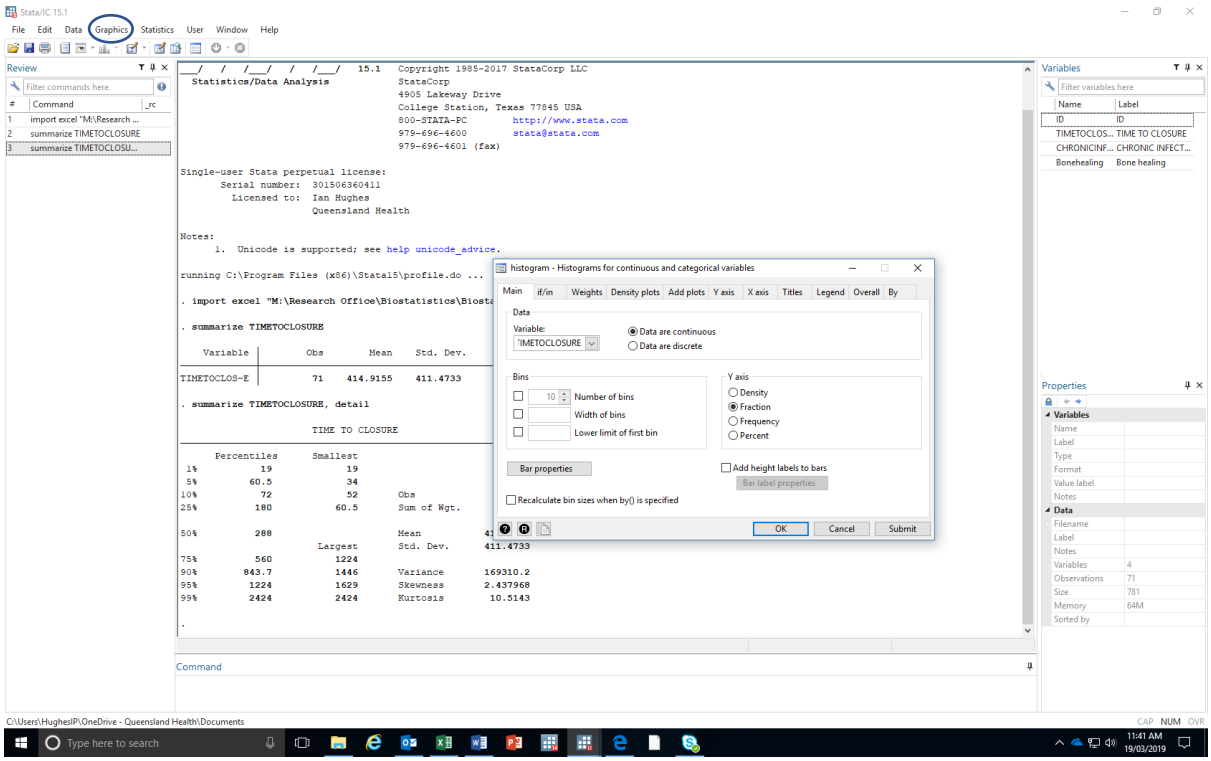
We now have a more complete set of summary statistics.

```
. summarize TIMETOCLOSURE

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
TIMETOCLOS~E |         71    414.9155    411.4733         19       2424

. summarize TIMETOCLOSURE, detail

                        TIME TO CLOSURE
-------------------------------------------------------------
      Percentiles      Smallest
 1%           19            19
 5%         60.5            34
10%           72            52       Obs                  71
25%          180          60.5       Sum of Wgt.          71

50%          288                     Mean            414.9155
                      Largest       Std. Dev.       411.4733
75%          560          1224
90%        843.7          1446       Variance        169310.2
95%         1224          1629       Skewness        2.437968
99%         2424          2424       Kurtosis         10.5143

.
```

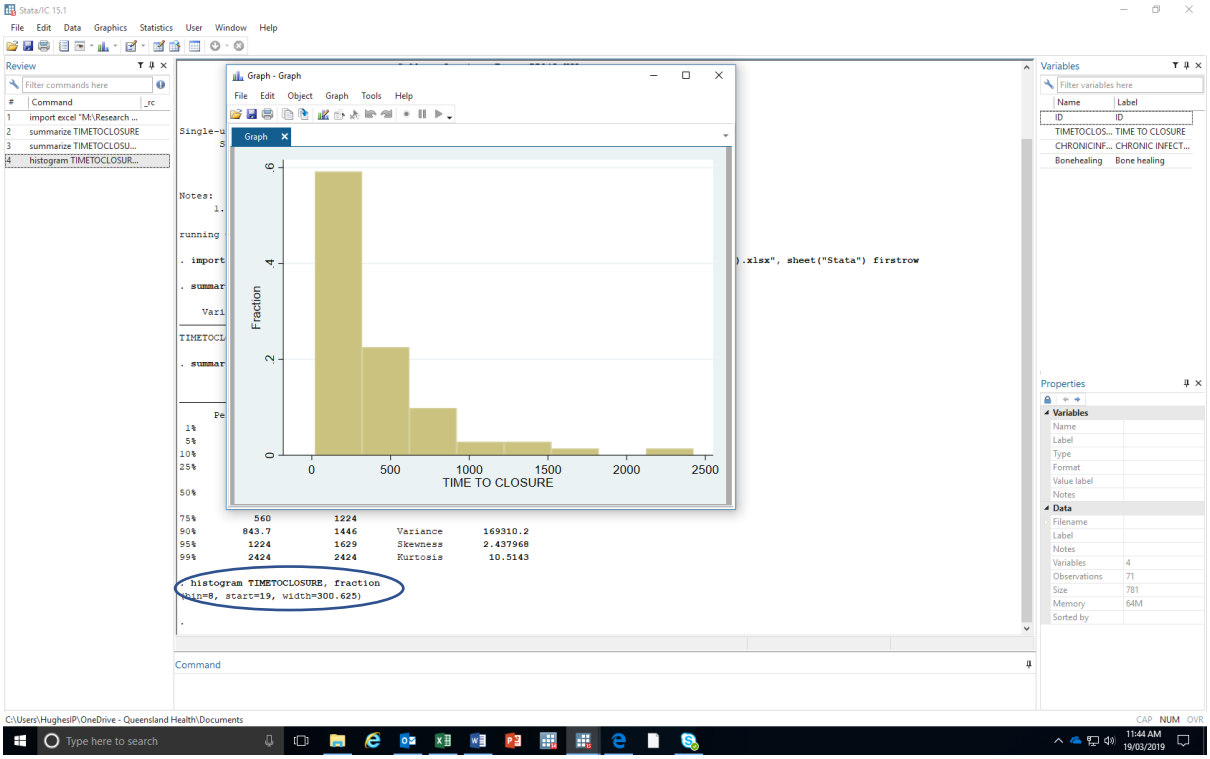
Command

nd Health\Documents

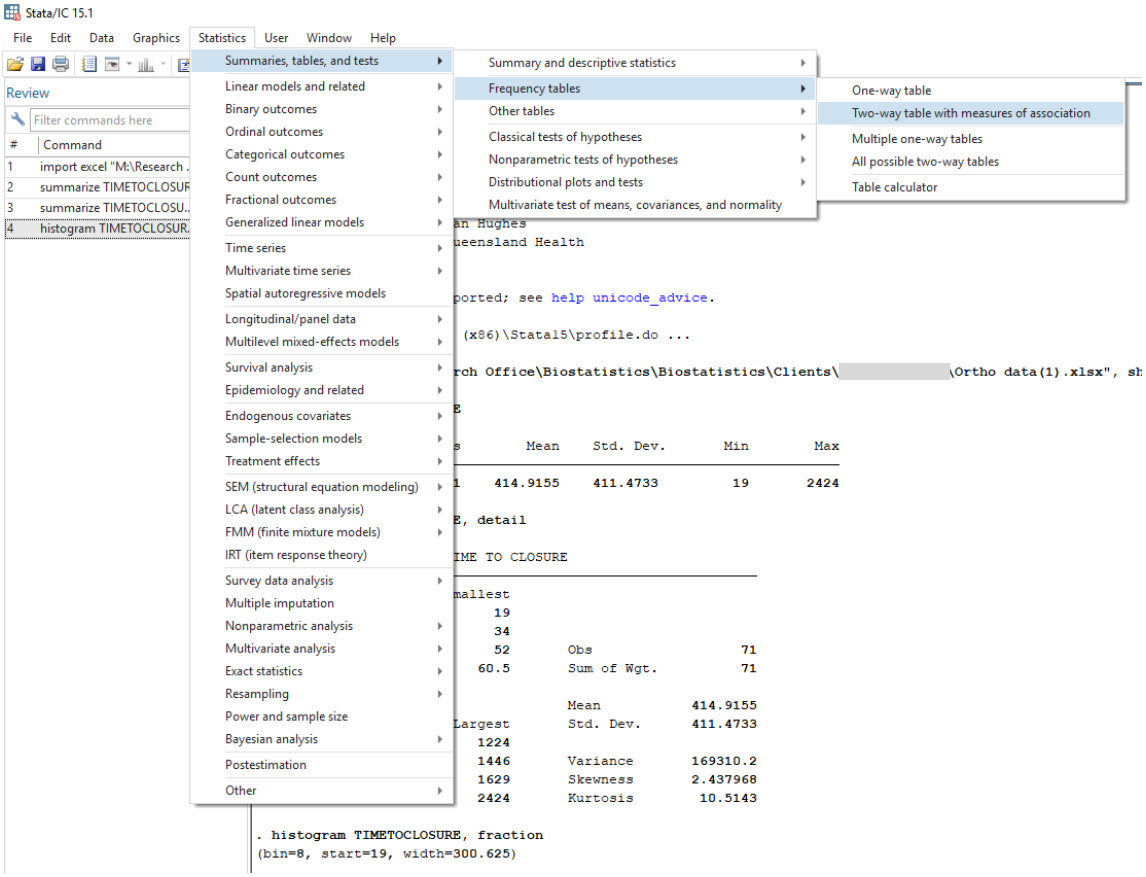
Click on the Graphics tab. Select histogram. Choose a variable and select whether the data is continuous or discrete. Select want you want on the Y axis

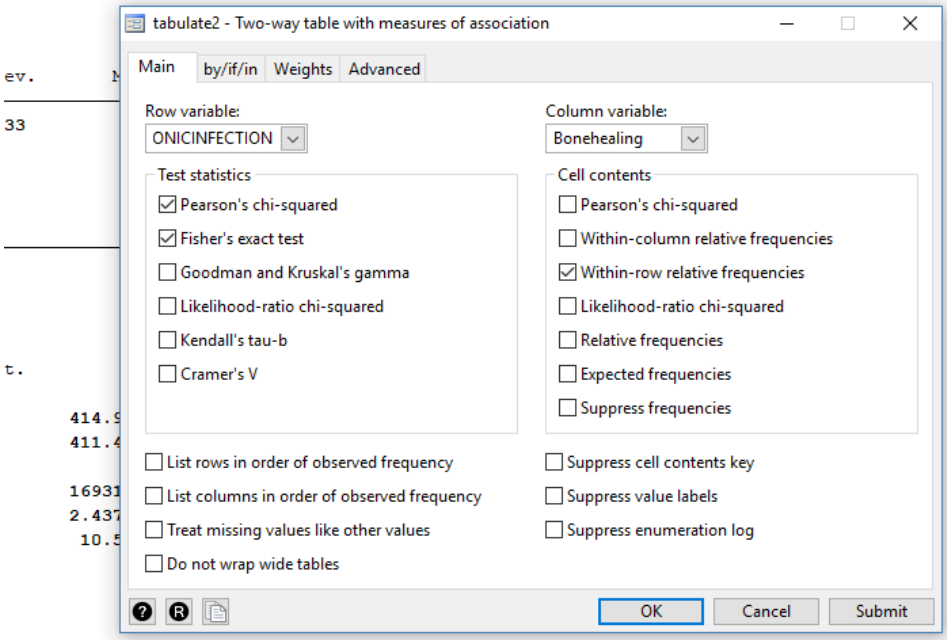


The graphic is shown in a separate window. It can be copied to a Word file. There are many options available to make graphics look really good.

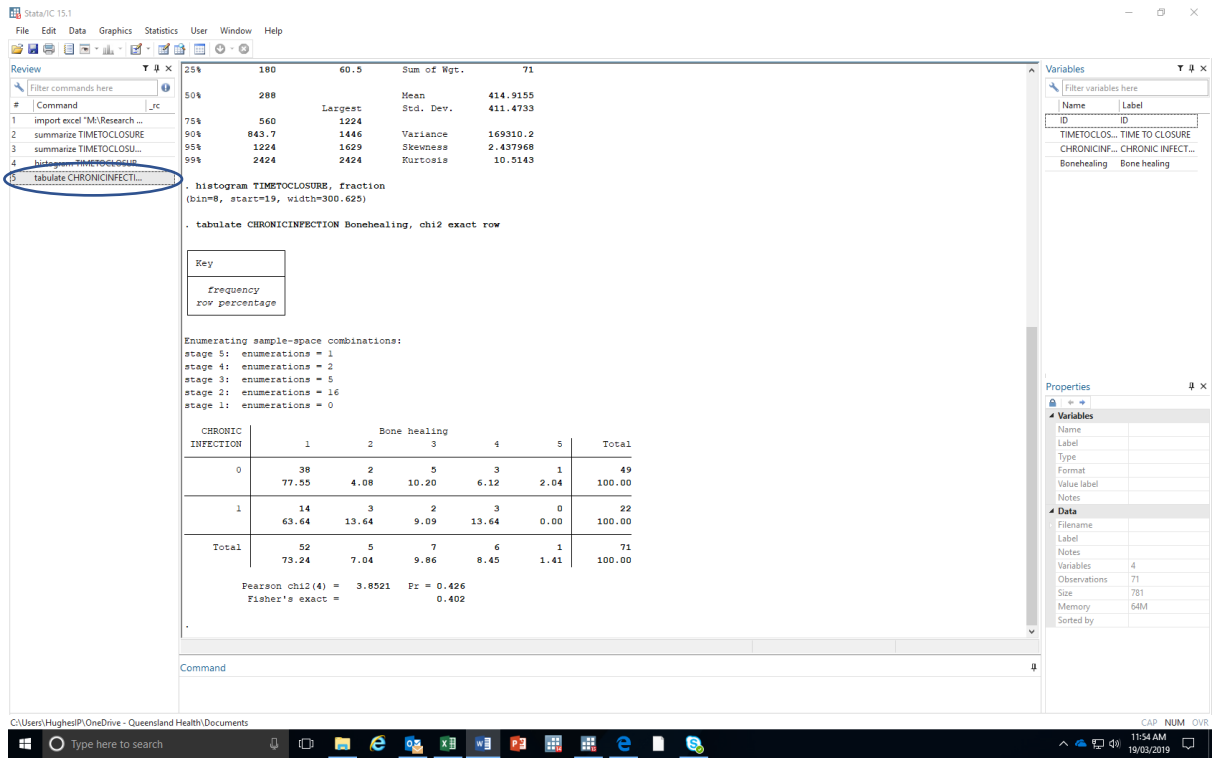


To look for an association between two categorical variables follow the tabs below.

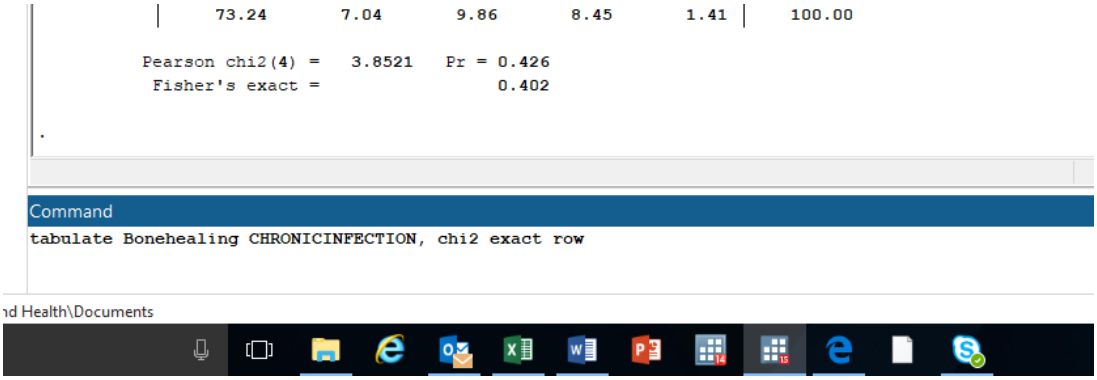


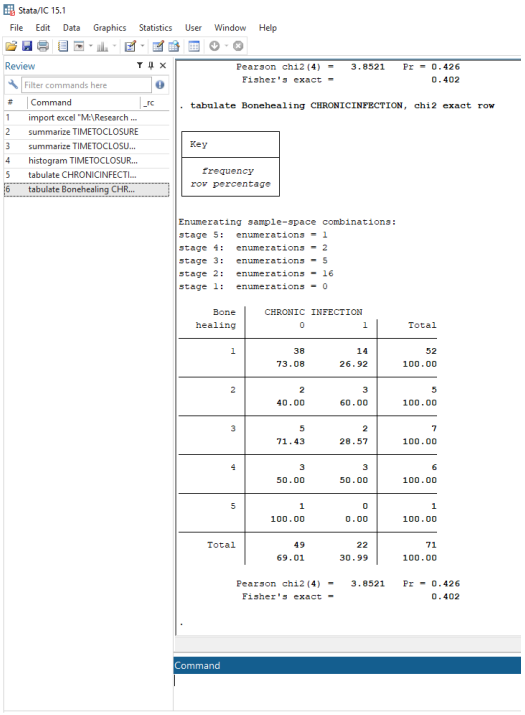


The results are shown below.

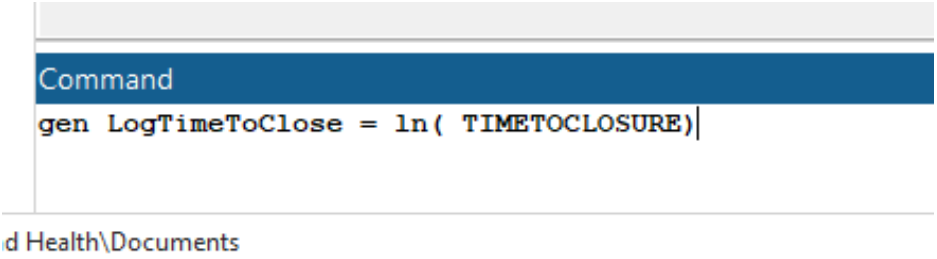


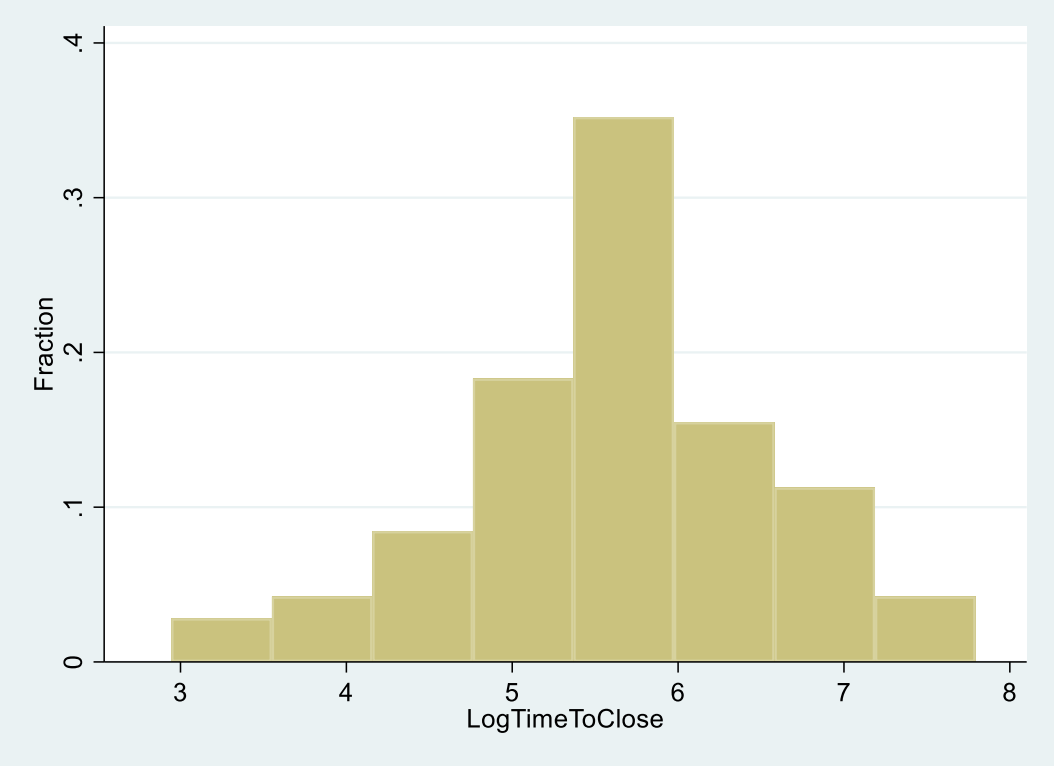
Say we actually wanted the table the other way around. It is easiest to click on the last command (left panel) and then change the order of the variables in the command window as shown below.



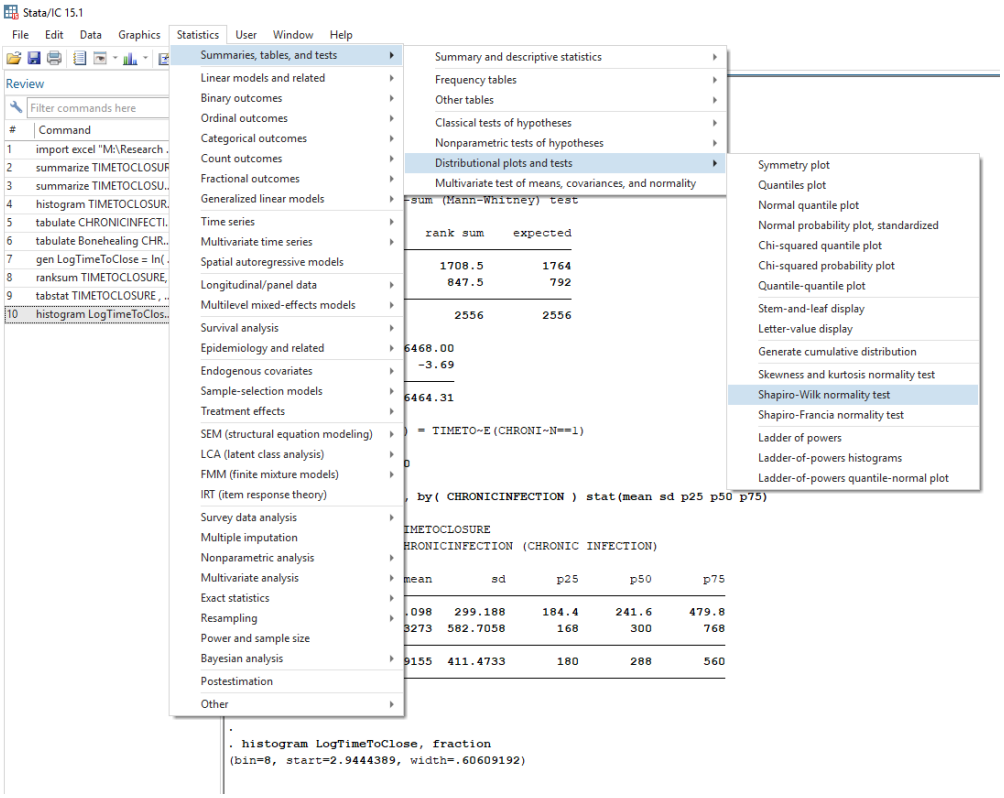


To generate a new variable you can follow the Data tab then select “create or change data”. Or, you can go to the Command window and write generate (or gen) the name you choose for the new variable, =, and then some function that creates the new variable. In the example below, as TIME TO CLOSURE was a skewed distribution I decided to log transform it.





The LogTimeToClose variable looks quite symmetrical. We can use a formal test of normality. Eg the Shapiro-Wilk test.

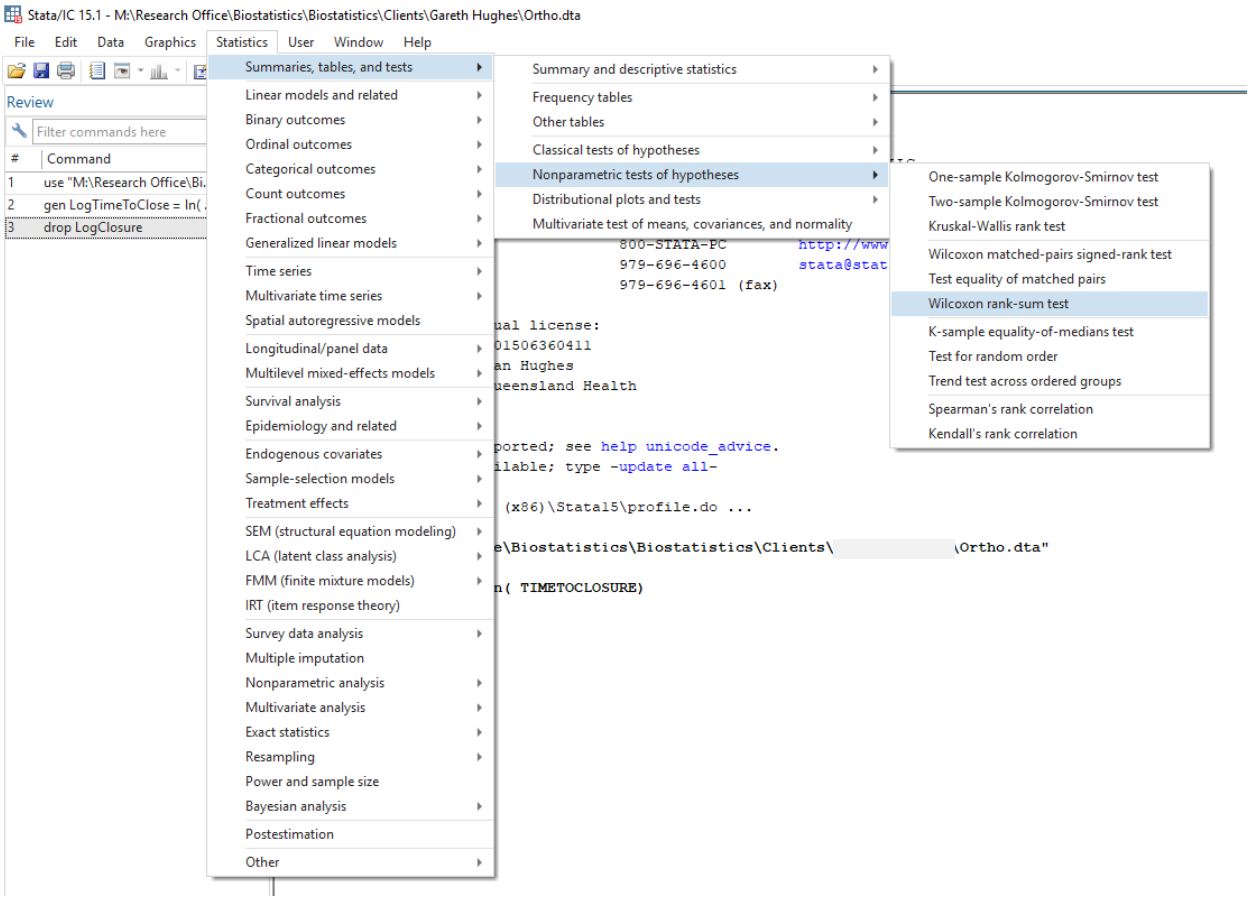


```
.
. histogram LogTimeToClose, fraction
(bin=8, start=2.9444389, width=.60609192)

. swilk LogTimeToClose

                    Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V          z       Prob>z
-------------+-------------------------------------------------------
LogTimeToC~e |         71    0.98790       0.753     -0.616    0.73112

.
```

Command

As suspected, the transformed variable is not significantly different from normal.

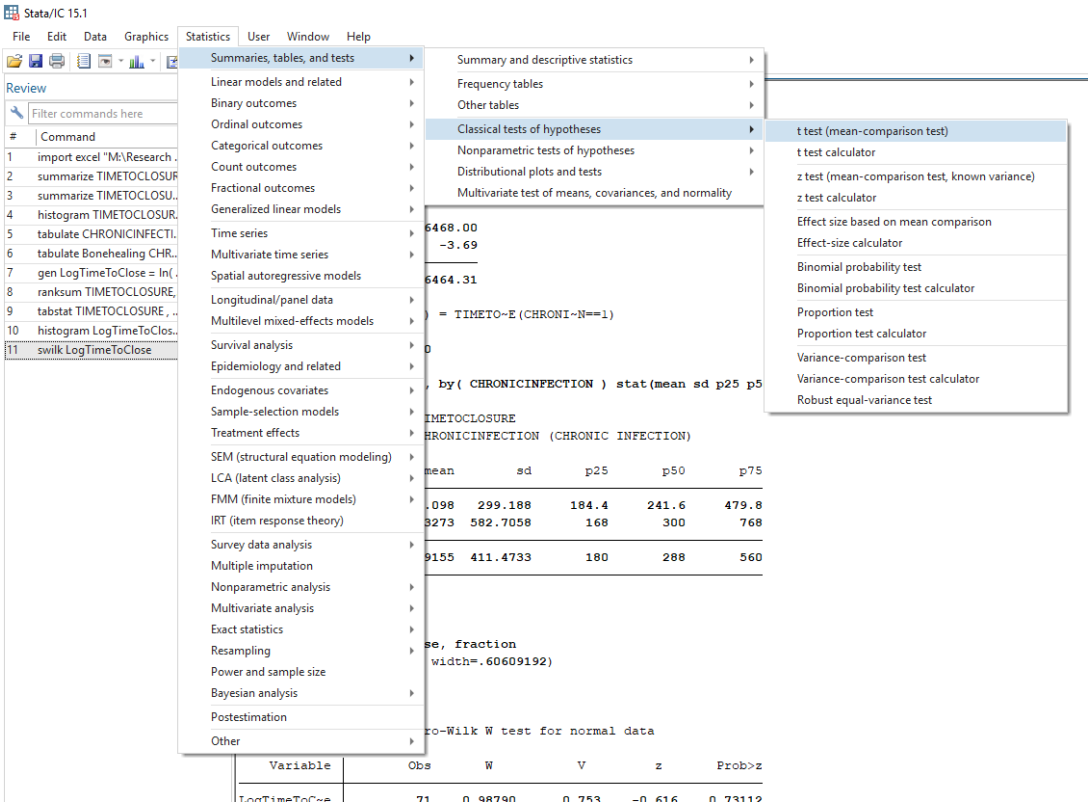
We can thus test to see if time to closure is different between the chronic infection group and the non-chronic infection group by either doing a non-parametric test (Wilcoxon rank-sum test, also called the Mann-Whitney U test) on the original data or we can do a t-test to compare the means of the log-transformed time to closure.



```
. gen LogTimeToClose = ln( TIMETOCLOSURE)

. ranksum TIMETOCLOSURE, by(CHRONICINFECTION)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

CHRONICINF~N |       obs    rank sum    expected
-------------+--------------------------------
           0 |        49      1708.5        1764
           1 |        22       847.5         792
-------------+--------------------------------
    combined |        71        2556        2556

unadjusted variance     6468.00
adjustment for ties       -3.69
                        --------
adjusted variance       6464.31

Ho: TIMETO~E(CHRONI~N==0) = TIMETO~E(CHRONI~N==1)
             z =  -0.690
    Prob > |z| =   0.4900

.
```

The Mann-Whitney U test suggests there is no difference in time to closure between the groups.

A t-test can be performed on the log transformed data as shown (or by typing the command directly, which is quicker when you have done it a few times).



```
. ttest LogTimeToClose, by(CHRONICINFECTION)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      49    5.563024    .1269281    .8884966    5.307818    5.81823
       1 |      22    5.802506    .2113022    .9910952    5.363079    6.241933
---------+--------------------------------------------------------------------
combined |      71     5.63723    .1093155    .9211087    5.419207    5.855253
---------+--------------------------------------------------------------------
    diff |           -.2394812    .2363458               -.7109781    .2320157
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.0133
Ho: diff = 0                     degrees of freedom =       69

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1572         Pr(|T| > |t|) = 0.3145          Pr(T > t) = 0.8428

.
```

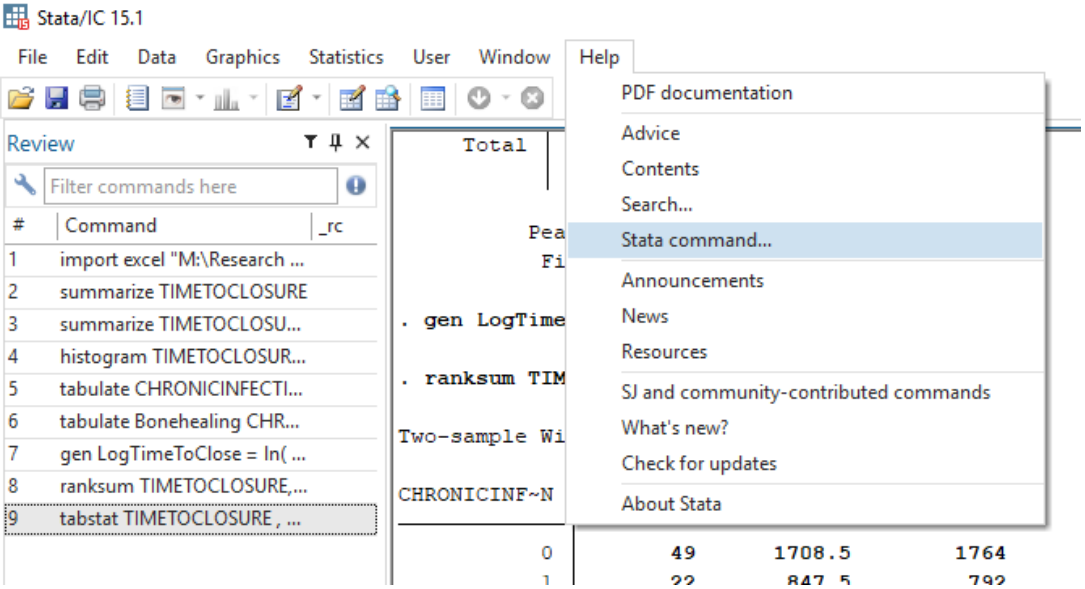
Command

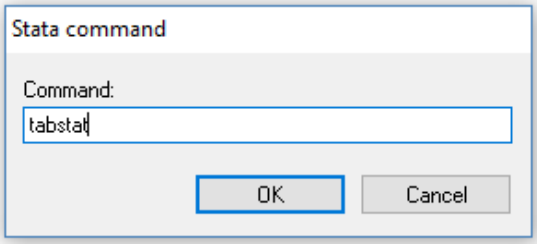
There is no significant difference between the means.

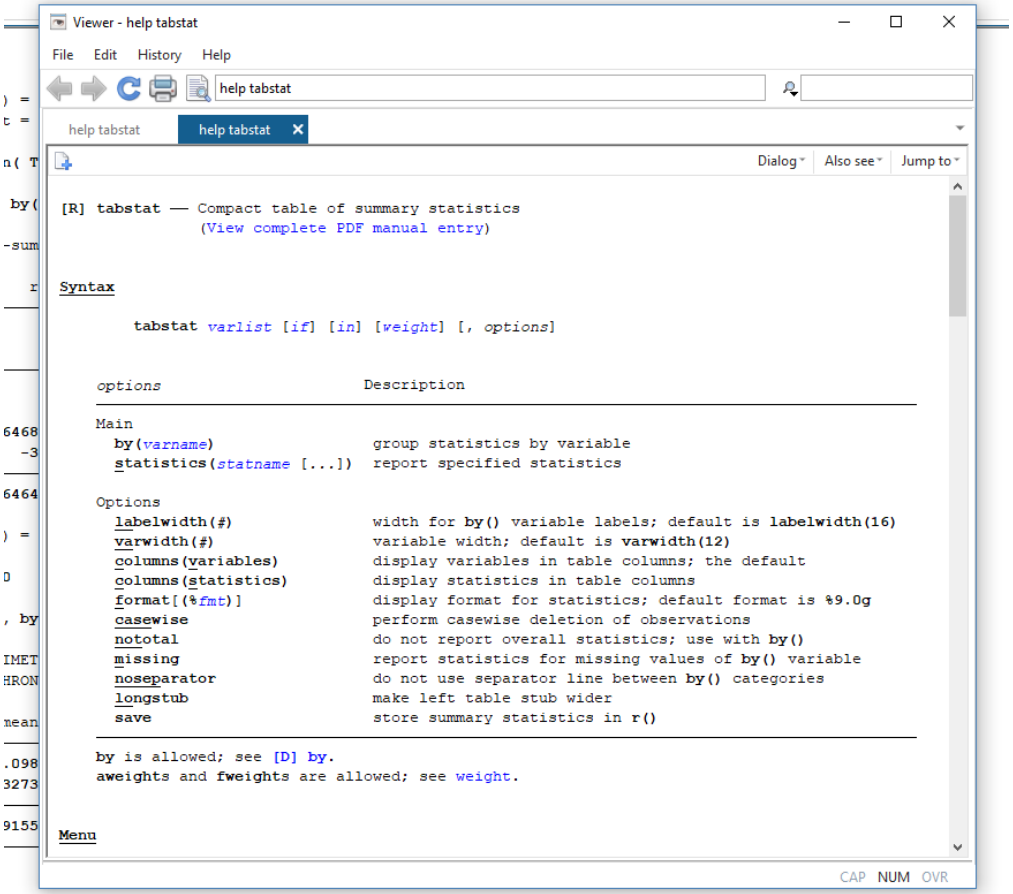
We have seen that the `summarize` command will give summary statistics for a variable but to compare summary statistics of a variable by grouping it is best to use the `tabstat` command.

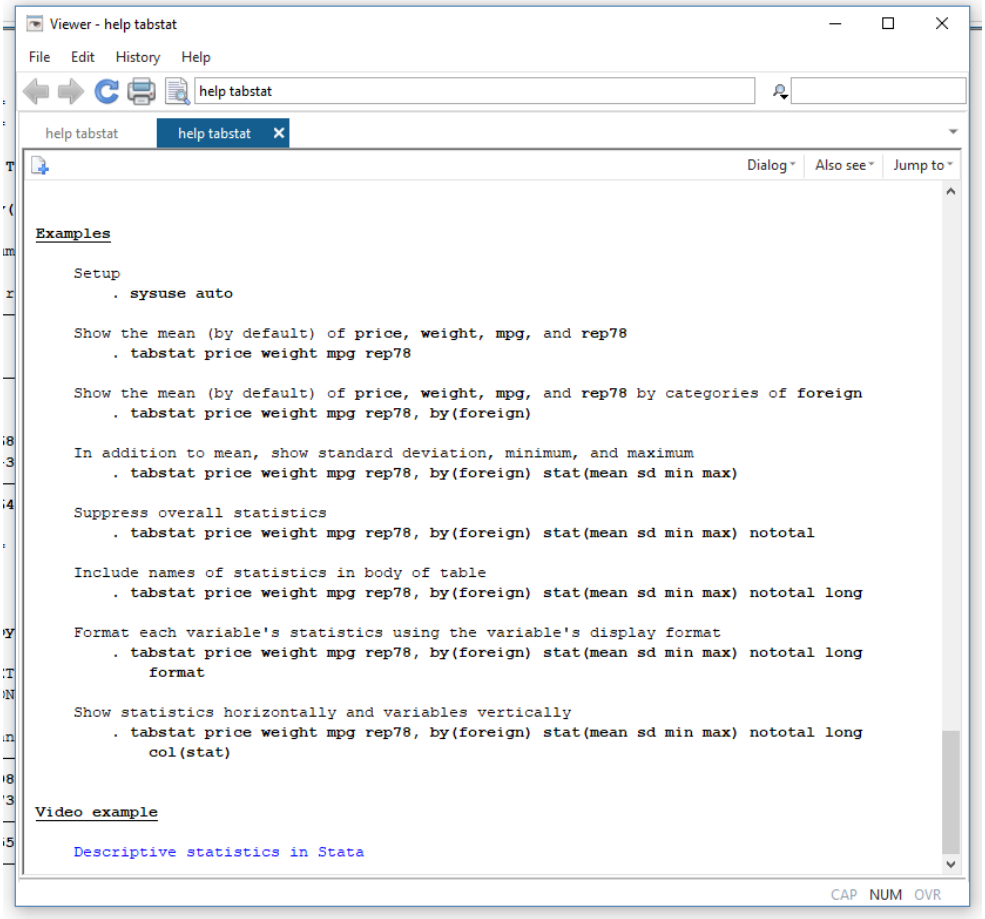
```
. tabstat TIMETOCLOSURE , by( CHRONICINFECTION ) stat(mean sd p25 p50 p75)

Summary for variables: TIMETOCLOSURE
     by categories of: CHRONICINFECTION (CHRONIC INFECTION)

CHRONICINFECTION |      mean        sd       p25       p50       p75
-----------------+--------------------------------------------------
               0 |   363.098   299.188     184.4     241.6     479.8
               1 |  530.3273  582.7058       168       300       768
-----------------+--------------------------------------------------
           Total |  414.9155  411.4733       180       288       560
-------------------------------------------------------------------
```

The command can be found from the drop-down menus but if you know the name of the command you can find all the information about how to use it by clicking on the Help tab then Stata command. Examples are usually given at the end of the help file.

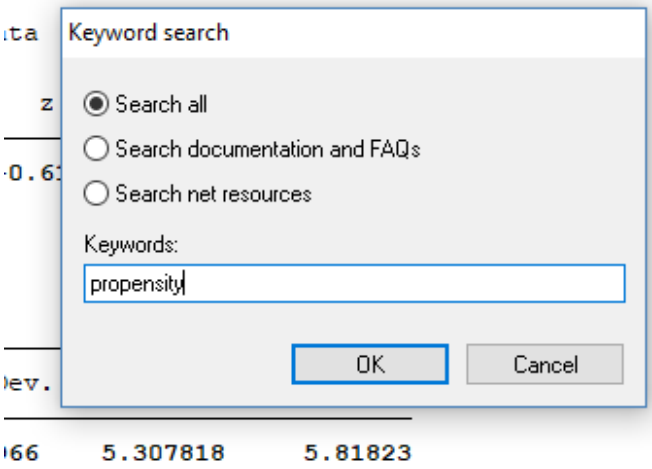

Stata/IC 15.1
File Edit Data Graphics Statistics User Window Help
Review
Filter commands here
Command _rc
1 import excel "M:\Research ...
2 summarize TIMETOCLOSURE
3 summarize TIMETOCLOSU...
4 histogram TIMETOCLOSUR...
5 tabulate CHRONICINFECTI...
6 tabulate Bonehealing CHR...
7 gen LogTimeToClose = ln(...
8 ranksum TIMETOCLOSURE,...
9 tabstat TIMETOCLOSURE , ...
Total
. gen LogTime
. ranksum TIM
Two-sample Wi
CHRONICINF~N
0 49 1708.5 1764
PDF documentation
Advice
Contents
Search...
Stata command...
Announcements
News
Resources
SJ and community-contributed commands
What's new?
Check for updates
About Stata


Stata command
Command:
tabstat
OK
Cancel

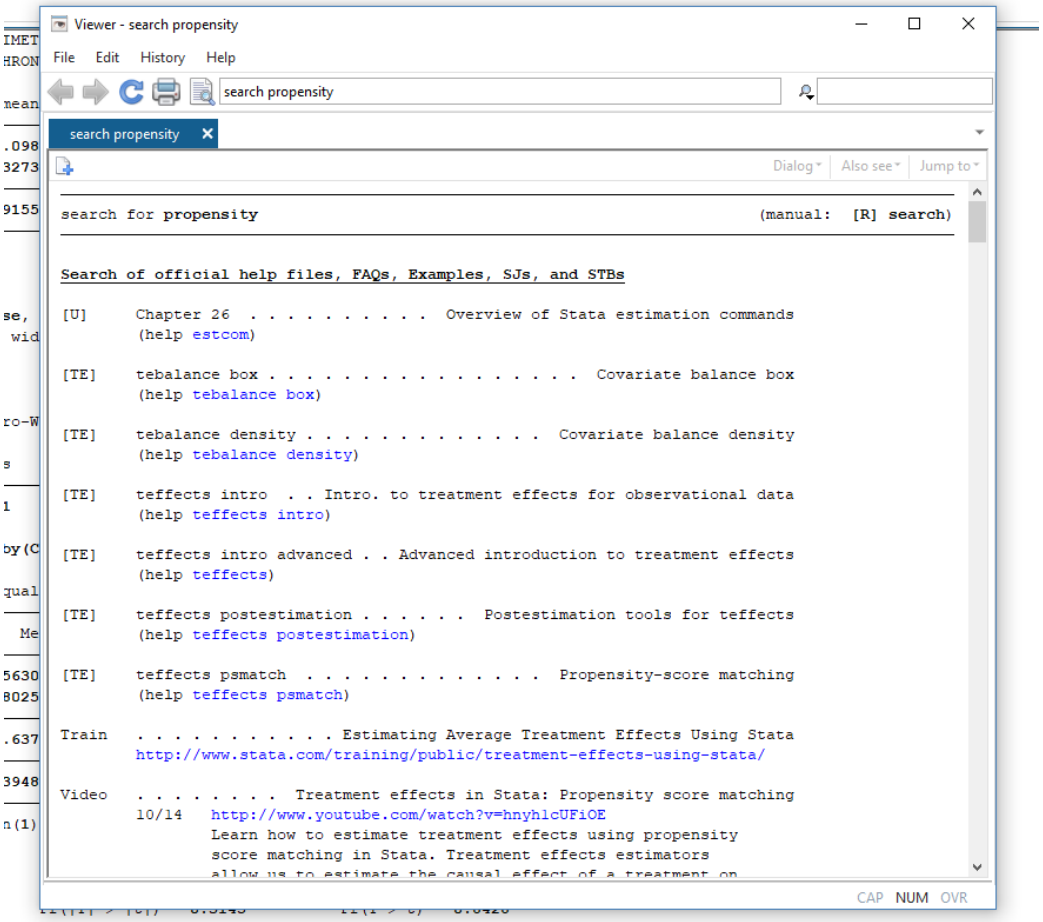

Viewer - help tabstat
File Edit History Help
help tabstat
help tabstat
help tabstat
Dialog
Also see
Jump to
[R] tabstat — Compact table of summary statistics
(View complete PDF manual entry)
Syntax
tabstat varlist [if] [in] [weight] [, options]
options Description
Main
by(varname) group statistics by variable
statistics(statname [...]) report specified statistics
Options
labelwidth(#) width for by() variable labels; default is labelwidth(16)
varwidth(#) variable width; default is varwidth(12)
columns(variables) display variables in table columns; the default
columns(statistics) display statistics in table columns
format[(%fmt)] display format for statistics; default format is %9.0g
casewise perform casewise deletion of observations
nototal do not report overall statistics; use with by()
missing report statistics for missing values of by() variable
noseparator do not use separator line between by() categories
longstub make left table stub wider
save store summary statistics in r()
by is allowed; see [D] by.
aweights and fweights are allowed; see weight.
Menu
CAP NUM OVR



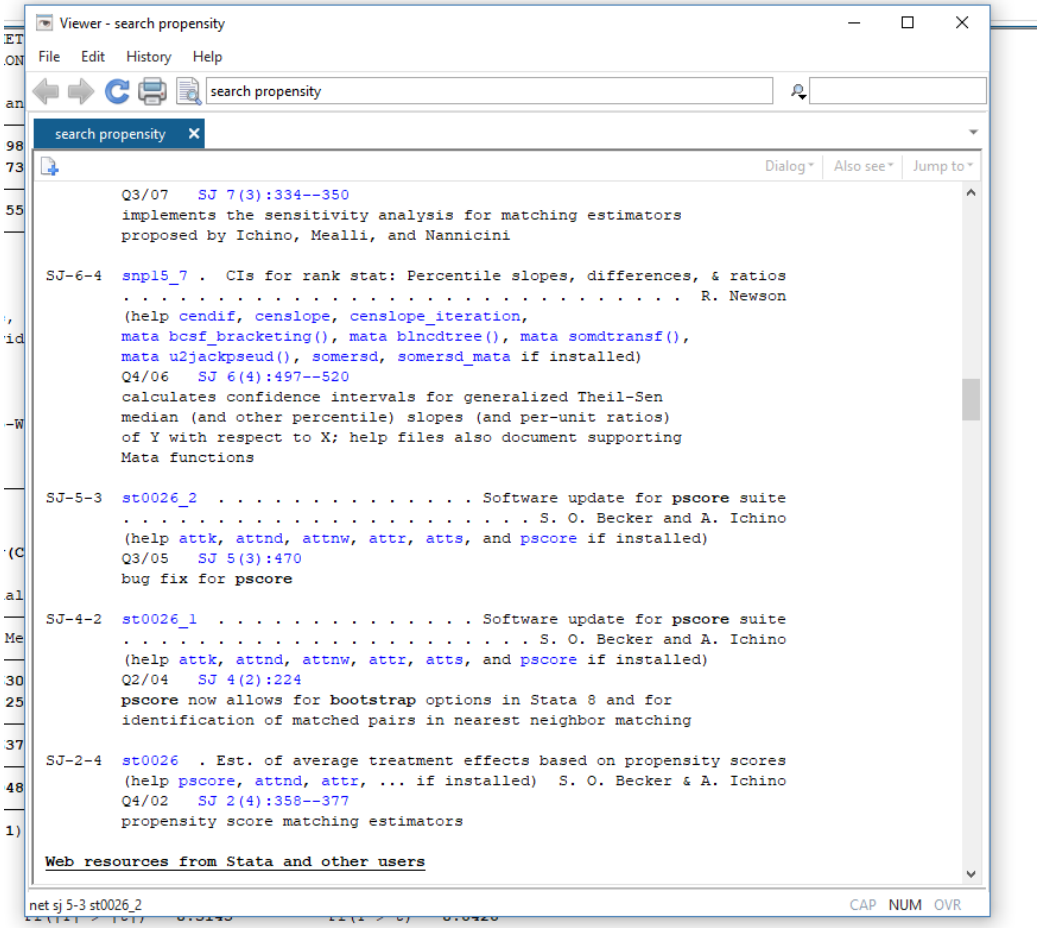
If you don’t know the name of the command but want to find out if Stata has any commands that might be useful for a particular topic you can click on Help then search. For example, say we are interested in propensity scores.



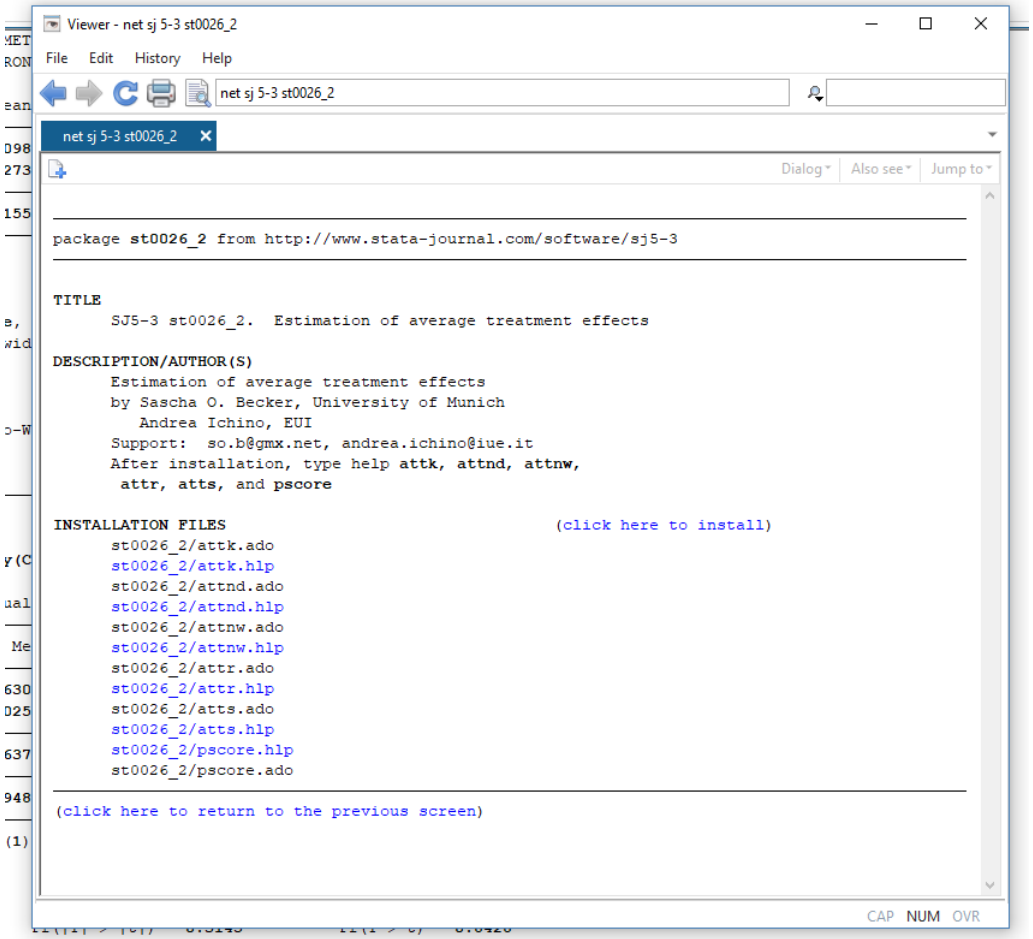
Various sources of information are found.



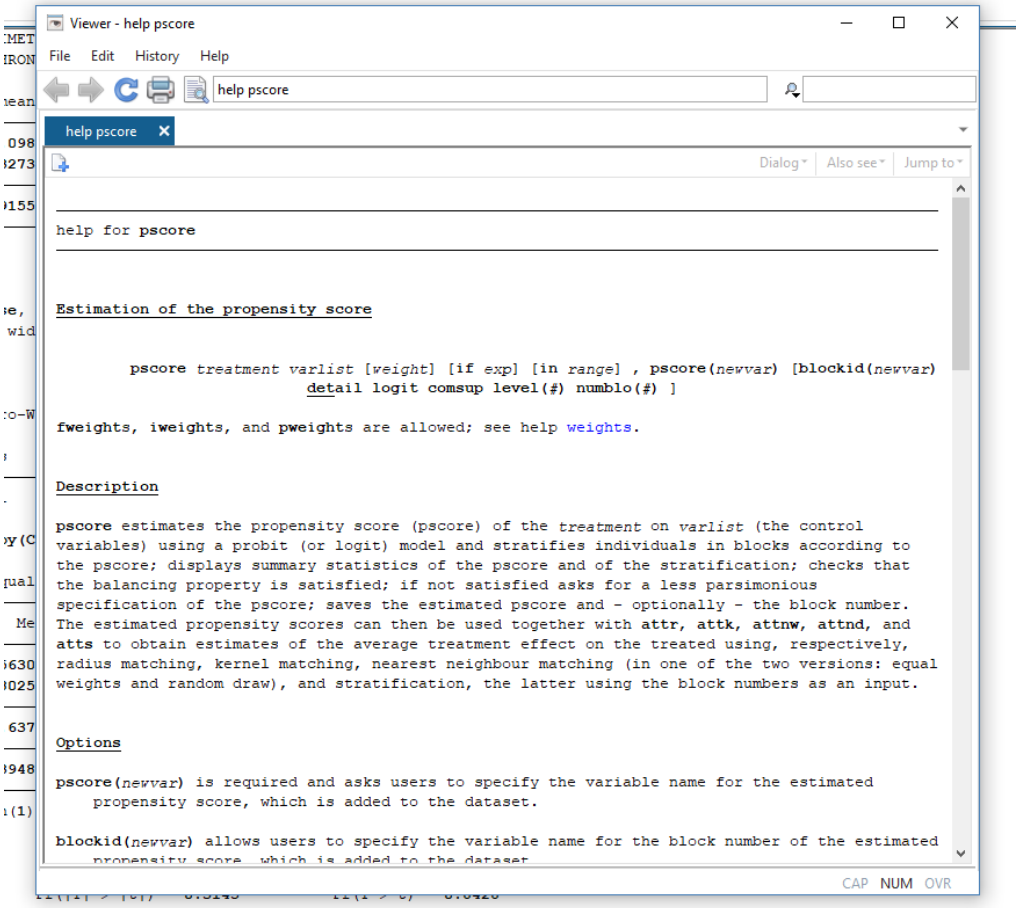
Further down you will find links to Stata packages that may be useful for calculation of propensity scores.



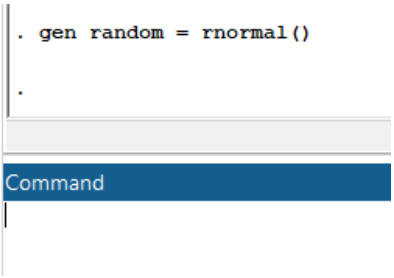
The latest update for the pscore suite looks like it is the best option. Click on st0026_2.



Click on “click here to install”. To check if it has installed, click on Help, Stata command.., and type pscore. Now the help file should come up telling you how to use the command you just downloaded.

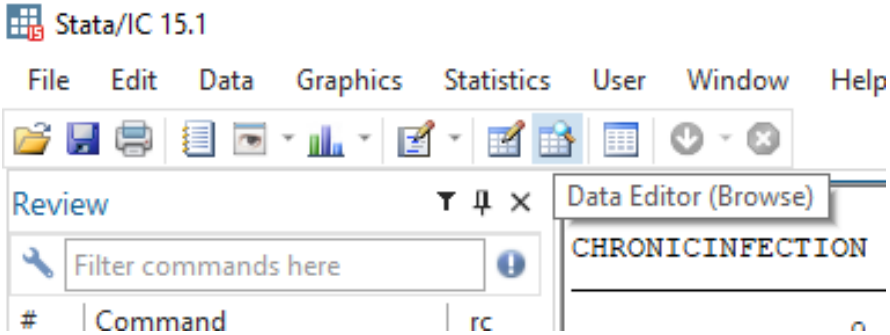


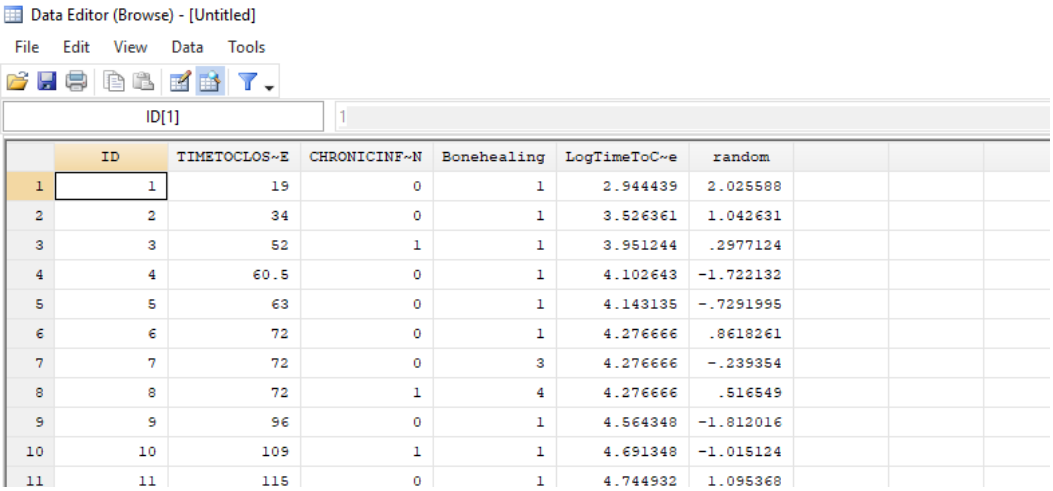
The following illustrates how to create, manipulate, and view variables. You can create a variable that is a random selection of values from a normal distribution.



The r is for “random”. Other distributions could be used too. For details search random number generation.

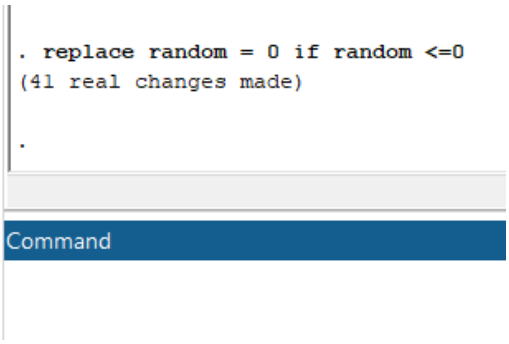
We can view the data, including the new variable we called “random” by clicking on the Data Editor (Browse) icon.



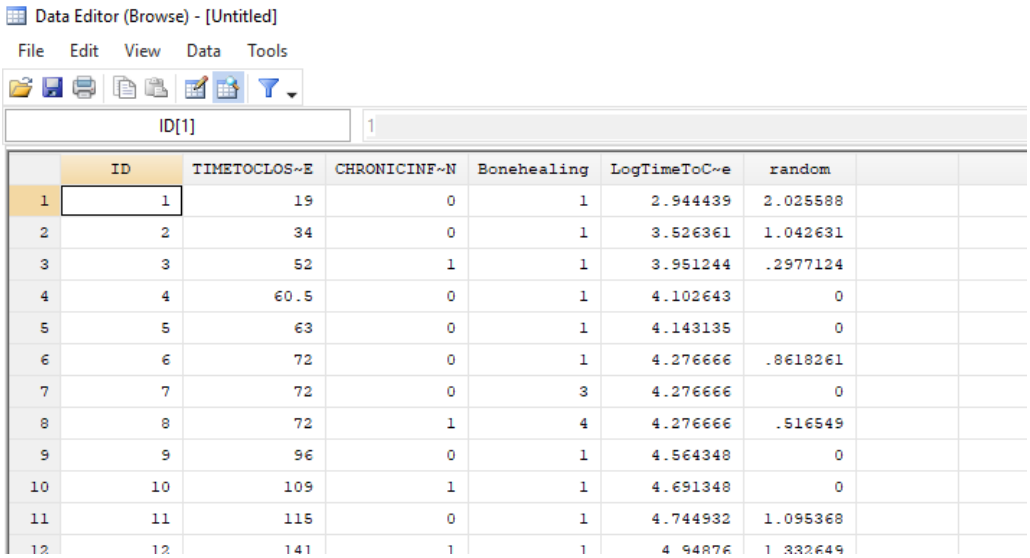


	ID	TIMETOCLOS~E	CHRONICINF~N	Bonehealing	LogTimeToC~e	random
1	1	19	0	1	2.944439	2.025588
2	2	34	0	1	3.526361	1.042631
3	3	52	1	1	3.951244	.2977124
4	4	60.5	0	1	4.102643	-1.722132
5	5	63	0	1	4.143135	-.7291995
6	6	72	0	1	4.276666	.8618261
7	7	72	0	3	4.276666	-.239354
8	8	72	1	4	4.276666	.516549
9	9	96	0	1	4.564348	-1.812016
10	10	109	1	1	4.691348	-1.015124
11	11	115	0	1	4.744932	1.095368

We can change a variable by using the replace command. In this case we are going to change all values in random to 0 if the original value was less than or equal to 0.

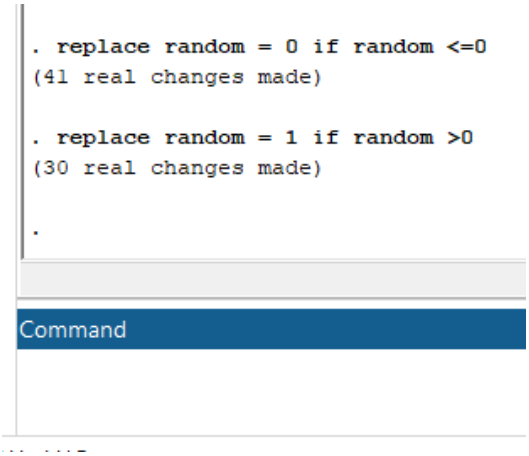


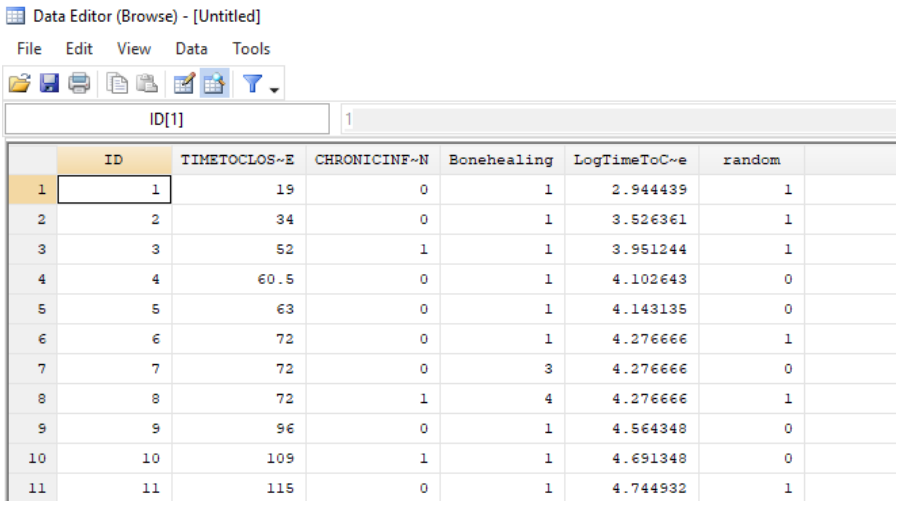
Check to see what happened.



	ID	TIMETOCLOS~E	CHRONICINF~N	Bonehealing	LogTimeToC~e	random
1	1	19	0	1	2.944439	2.025588
2	2	34	0	1	3.526361	1.042631
3	3	52	1	1	3.951244	.2977124
4	4	60.5	0	1	4.102643	0
5	5	63	0	1	4.143135	0
6	6	72	0	1	4.276666	.8618261
7	7	72	0	3	4.276666	0
8	8	72	1	4	4.276666	.516549
9	9	96	0	1	4.564348	0
10	10	109	1	1	4.691348	0
11	11	115	0	1	4.744932	1.095368
12	12	141	1	1	4.94876	1.332649

Now replace all of the positive values with 1.





	ID	TIMETOCLOS~E	CHRONICINF~N	Bonehealing	LogTimeToC~e	random
1	1	19	0	1	2.944439	1
2	2	34	0	1	3.526361	1
3	3	52	1	1	3.951244	1
4	4	60.5	0	1	4.102643	0
5	5	63	0	1	4.143135	0
6	6	72	0	1	4.276666	1
7	7	72	0	3	4.276666	0
8	8	72	1	4	4.276666	1
9	9	96	0	1	4.564348	0
10	10	109	1	1	4.691348	0
11	11	115	0	1	4.744932	1

And rename the variable “sex”. Obviously we would usually not be randomly assigning sex to participants but it gives me the excuse to demonstrate how to assign labels to categorical variables.

```
. replace random = 1 if random >0
(30 real changes made)

. rename random sex

.
```

Command

Here we create a label file called Gender which assigns the label “Female” to the 0s and “male” to the 1s. In the next line I tell Stata to label the values of the variable sex according to the definitions in the label file Gender.

```
. label define Gender 0 "Female" 1 "Male"

. label values sex Gender

.
```

Data Editor (Browse) - [Untitled]

File Edit View Data Tools

var11[17]

	ID	TIMETOCLOS~E	CHRONICINF~N	Bonehealing	LogTimeToC~e	sex
1	1	19	0	1	2.944439	Male
2	2	34	0	1	3.526361	Male
3	3	52	1	1	3.951244	Male
4	4	60.5	0	1	4.102643	Female
5	5	63	0	1	4.143135	Female
6	6	72	0	1	4.276666	Male
7	7	72	0	3	4.276666	Female
8	8	72	1	4	4.276666	Male

Now, when you produce tables for example they have the labels.

```
. tabulate CHRONICINFECTION sex
```

CHRONIC INFECTION	sex Female	Male	Total
0	29	20	49
1	12	10	22
Total	41	30	71

This should be enough to get you started with Stata. There are many resources available beginning with the Help icon in Stata. Below are some really useful resources. The last one,

the UCLA site is very good, it gives the same examples in Stata, SPSS, and R. And of course Googling “How to ….. in Stata” works!

Free resources – webinars (https://www.stata.com/training/webinar/); manuals (https://surveydesign.com.au/manuals.html); or YouTube videos (https://www.youtube.com/channel/UCVk4G4nEtBS4tLOyHqustDA)

Paid resources – Stata Press books (https://surveydesign.com.au/statabooks.html); Stata online NetCourses (https://surveydesign.com.au/events.html) or other public training events (https://www.stata-au.com/events)

https://stats.idre.ucla.edu/stata/

Good luck with Stata!

Ian Hughes

Biostatistician GCHHS