Introducing Stata

1.1 STARTING STATA

Stata can be started several ways. First, there may be shortcut on the desktop that you can double-click. For the Stata/SE Release 10 it will look like

Earlier versions of Stata have a similar looking Icon, but of course with a different number. Alternatively, using the Windows menu, click the **Start > All Programs > Stata 10**.

A second way is to simply locate a Stata data file, with ***.dta** extension, and **double-click**.

1.2 THE OPENING DISPLAY

Once Stata is started a display will appear that contains windows titled

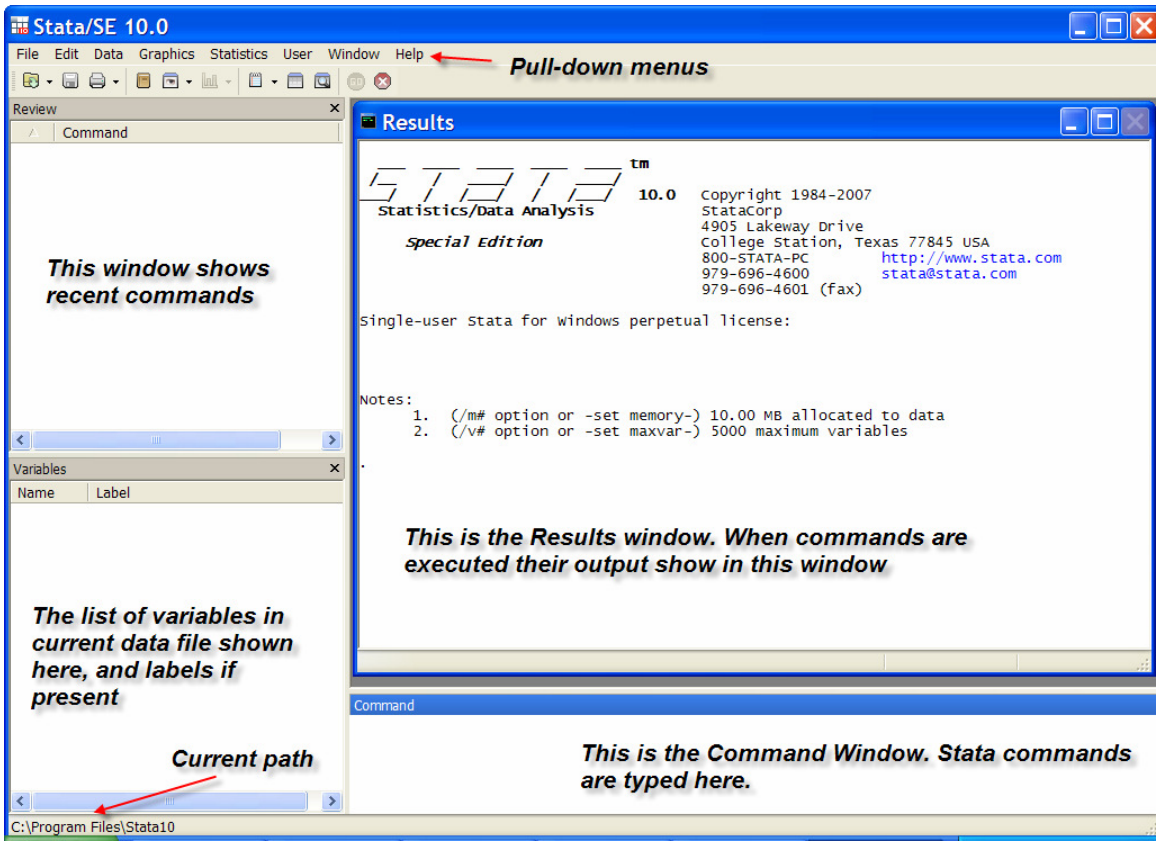
Command—this is where Stata command are typed

Results—output from commands, and error messages, appear here

Review—a listing of commands recently executed

Variables—names of variables in data and labels (if created)

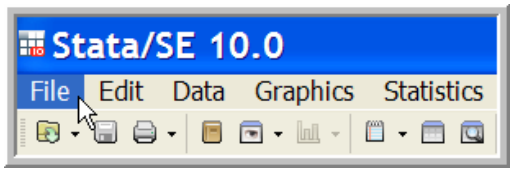
It should look something like



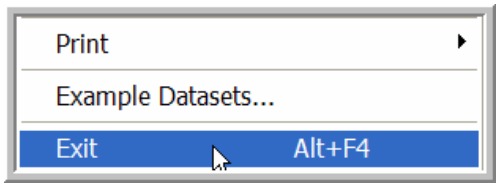
Across the top are Stata **pull-down menus**. We will explore the use of many of these. In the lower left-hand corner is the **current path** to a working directory where Stata saves graphs, data files, etc. We will change this in a moment.

1.3 EXITING STATA

To end a Stata session click on **File**



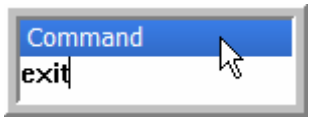
Select **Exit**



We will denote sequential clicking commands like this as **File > Exit**. Alternatively, simply type

```
exit
```

in the **Command** window and press **Enter**.



1.4 STATA DATA FILES FOR STOCK AND WATSON

Stata data files have the extension ***.dta**. These files should not be opened with any program but Stata. If you locate a ***.dta** file using double-click it will also start Stata.

You can obtain the data files from Stock and Watson's web site (follow the link on my homepage). The can also be found at:

http://fmwww.bc.edu/ec-p/data/stockwatson/datasets.list.html

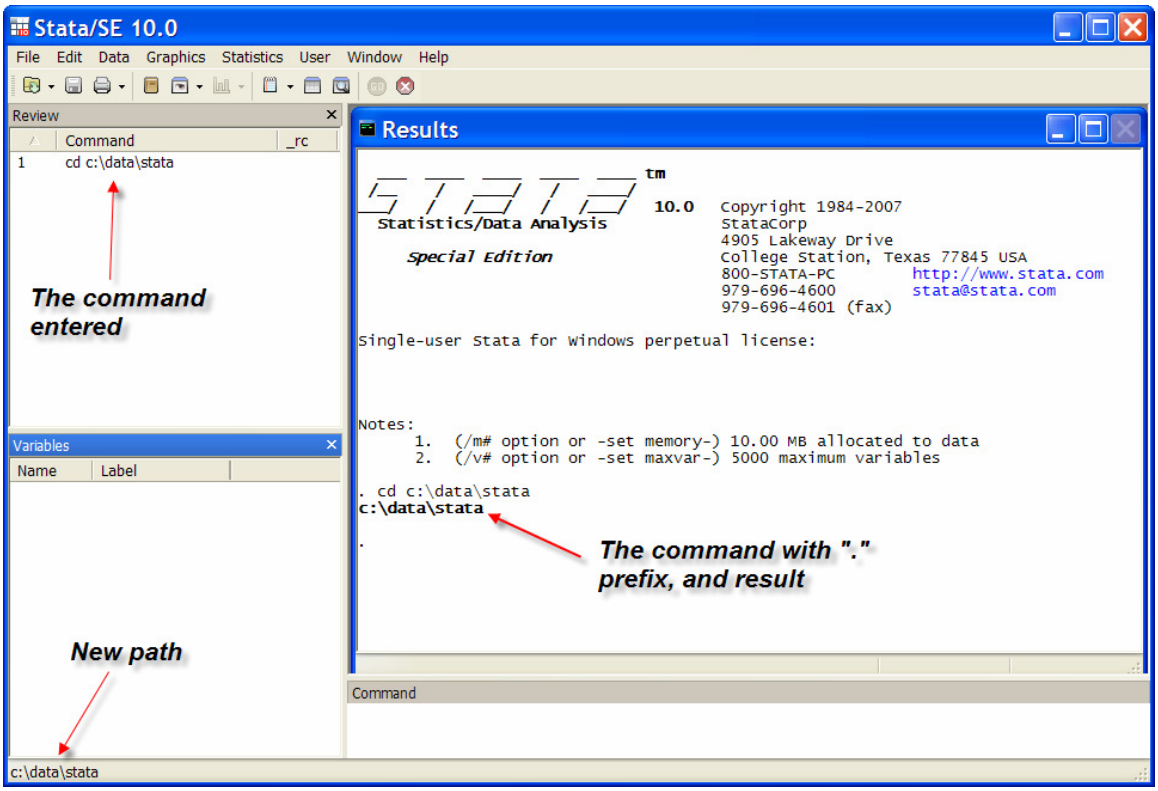
1.4.1 A working directory

You should copy the data into a convenient directory. How to accomplish this will depend on your computer system. In this Windows-based book we will use the subdirectory **`c:\data\stata`** for all our data and result files. To change the working directory to this location type

```
cd c:\data\stata
```

into the **Command** window and press **Enter**.

The result of this command is



Note that in the **Results** window the command is echoed, and it appears in the **Review** window as well. The new path is indicated at the bottom left of the screen.

If you are working in a computer laboratory, you may want to have a storage device such as a "flash" or "travel" drive. These are large enough to hold the Stata data files, definition files and your class work. Make a subdirectory on the device. Calling it **`X:\DATA`** where **`X:\`** is the path to your device, would be convenient.

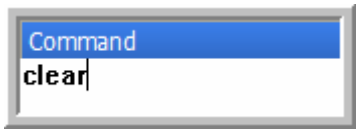
1.5 OPENING STATA DATA FILES

There are several ways to open, or load, Stata data files. We will explain a couple of them.

1.5.1 The use command

With Stata started, change your working directory to the where you have stored the Stata data files. On the **Command** window type use **`caschool`** and press **Enter**.

This feature will prevent you from losing changes to a data file you may wish to save. If this happens, you can either **`save`** the previous data file [more on this below], or enter **`clear`** in the **Command** window.

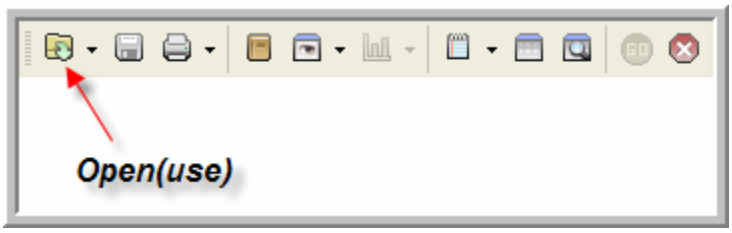


The **`clear`** command will clear what is in Stata's memory. If you want to open the data file and clear memory, enter

```
use caschool, clear
```

1.5.2 Using the toolbar

To open a Stata data file click the **Open (use)** icon on the toolbar



Locate the file you wish to open, select it, and click **Open.**

In the **Review** window the implied command is shown.

```
use caschool
```

1.5.3 Using files on the internet

Stata offers a nice option if you are connected to the internet. Files can be loaded from a web site. The Stata data files are stored at

http://fmwww.bc.edu/ec-p/data/stockwatson/datasets.list.html.

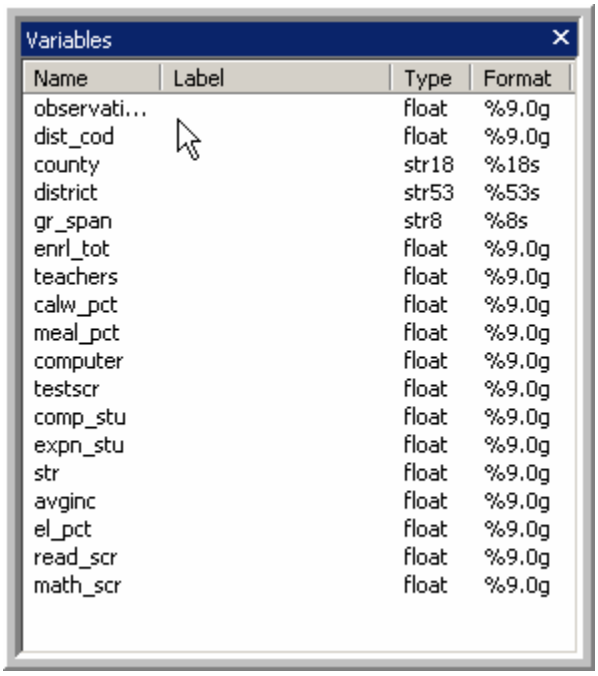
For example, to load *caschool*, after saving previous data and/or clearing memory, enter in the **Command** window

```
use http://fmwww.bc.edu/ec-p/data/stockwatson/caschool
```

Once the data are loaded onto your machine, you can save it using **File > Save as** and filling in the resulting dialog box.

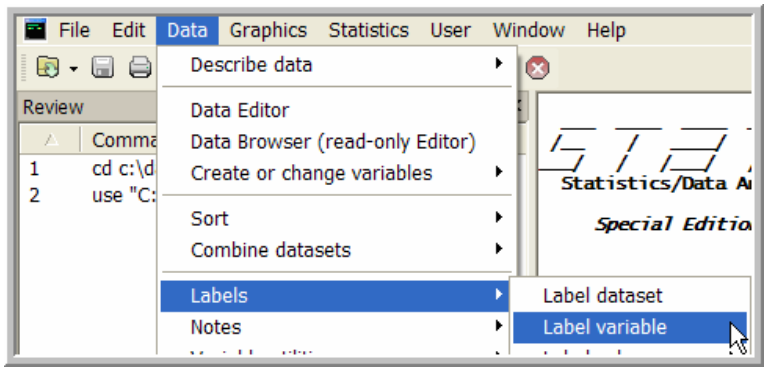
1.6 THE VARIABLES WINDOW

In the **Variables** window the data file variables are listed

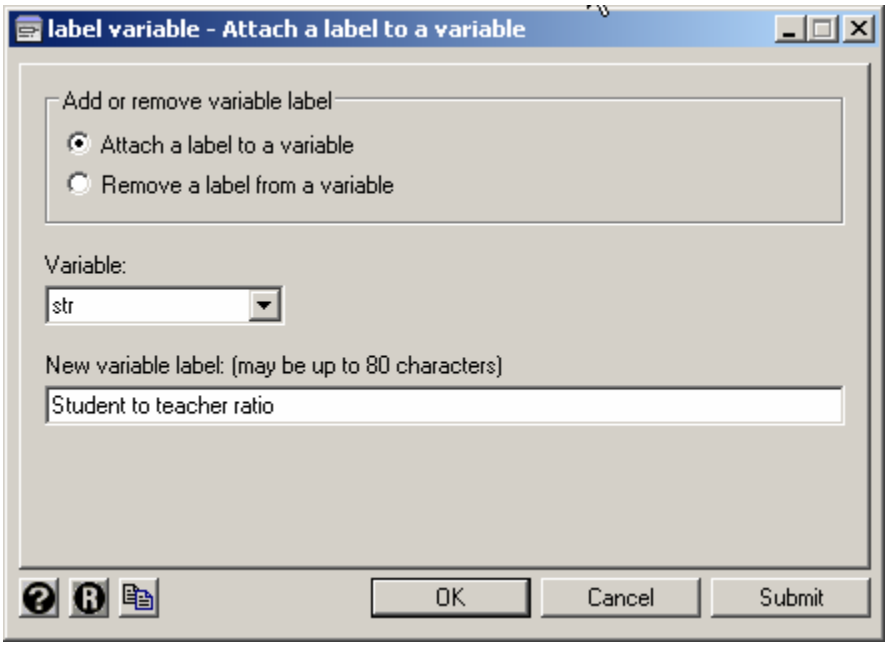


Also shown are variable **Labels**, if they are present, along with the **Type** of variable and its **Format**. We will only display the variable **Name** and **Label** in future screen shots.

Labels are useful and can be easily added, changed or deleted. On the Stata pull-down menu select **Data > Labels > Label Variable**. That is,



In the resulting dialog box, you can alter the existing label by choosing **Attach** a label to a variable, choosing the variable from the **Variable:** drop-down list and typing in the **New variable label**. Click **OK**.



Instead of the dialog box approach, type the following line in the **Command** window and press **Enter**

```
label variable str "Student to teacher ratio"
```

This command will create the label, and it will write over an already existing label for **`str`**. In the dialog box you can also choose to **Remove** a label.

1.7 DESCRIBING DATA AND OBTAINING SUMMARY STATISTICS

There are a few things you should do each time a data file is opened, or when you begin a new problem. First, enter into the **Command** window

```
describe
```

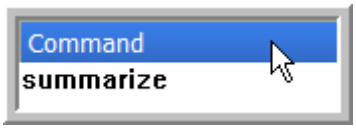
This produces a summary of the variables in the data file, information about them, and their labels.

```
. describe

Contains data from caschool.dta
  obs:           420
 vars:            18                          22 Aug 2007 14:14
 size:        60,060 (99.4% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
observation_n~r float  %9.0g
dist_cod        float  %9.0g
county          str18  %18s
district        str53  %53s
gr_span         str8   %8s
enrl_tot        float  %9.0g
teachers        float  %9.0g
calw_pct        float  %9.0g
meal_pct        float  %9.0g
computer        float  %9.0g
testscr         float  %9.0g
comp_stu        float  %9.0g
expn_stu        float  %9.0g
str             float  %9.0g                 Student to teacher ratio
avginc          float  %9.0g
el_pct          float  %9.0g
read_scr        float  %9.0g
math_scr        float  %9.0g
-------------------------------------------------------------------------------
```

Next, in the **Command** window, type

```
summarize
```

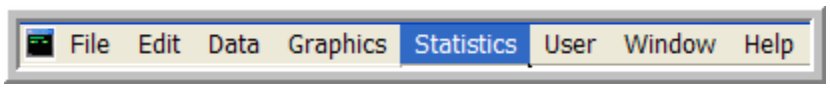
and press **Enter**.



In the **Results** window we find the summary statistics

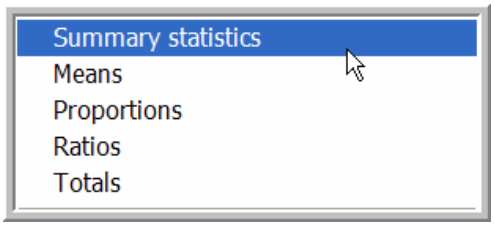
```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
observatio~r |       420       210.5    121.3878          1        420
    dist_cod |       420    67472.81    3466.995      61382      75440
      county |         0
    district |         0
     gr_span |         0
-------------+--------------------------------------------------------
    enrl_tot |       420    2628.793    3913.105         81      27176
    teachers |       420    129.0674    187.9127       4.85       1429
    calw_pct |       420    13.24604    11.45482          0    78.9942
    meal_pct |       420    44.70524    27.12338          0        100
    computer |       420    303.3833    441.3413          0       3324
-------------+--------------------------------------------------------
     testscr |       420    654.1565    19.05335     605.55     706.75
    comp_stu |       420    .1359266    .0649558          0   .4208333
    expn_stu |       420    5312.408    633.9371    3926.07   7711.507
         str |       420    19.64043    1.891812         14       25.8
      avginc |       420    15.31659     7.22589      5.335     55.328
-------------+--------------------------------------------------------
      el_pct |       420    15.76816    18.28593          0   85.53972
    read_scr |       420    654.9705    20.10798      604.5        704
    math_scr |       420    653.3426     18.7542      605.4      709.5
```

Should you forget a Stata command, the pull-down menus virtually assure that with enough clicking you can obtain the desired result. To illustrate, click on **Statistics** on the Stata menu list

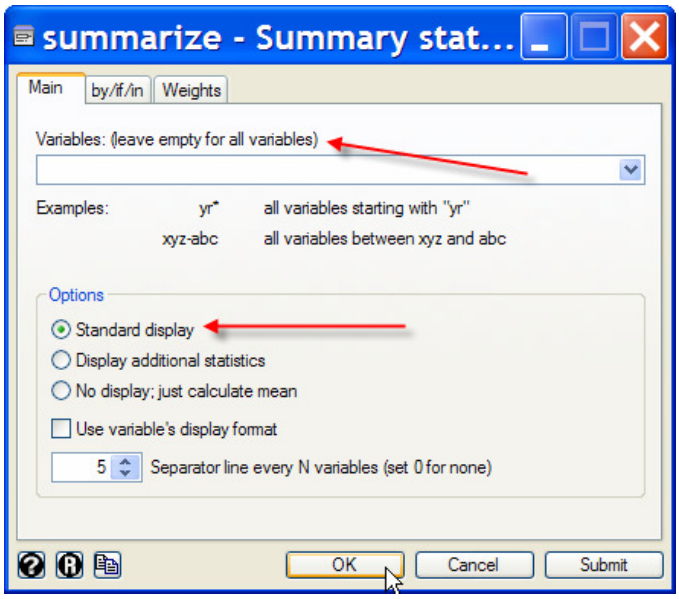


You will find a long list of possible statistical analyses, some of which we will use. For now select **Summaries, tables, and tests**

Select **Summary and descriptive statistics** and then **Summary statistics**

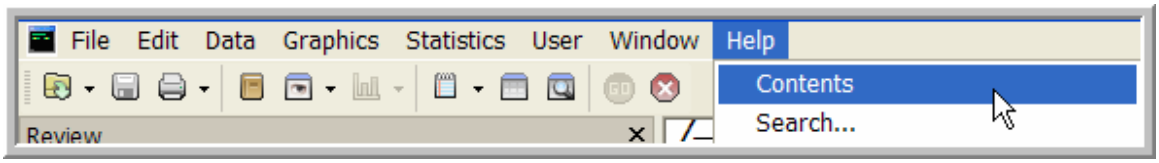


A **dialog box** will open that shows many options. For the basic summary statistics table no options are required. Stata automatically will provide the summary statistics for all the variables in the data set. You can select individual variables by typing their names in the **Variables** box. The **Standard display** will produce the number of observations, the arithmetic mean, the standard deviation, minimum and maximum of the data values

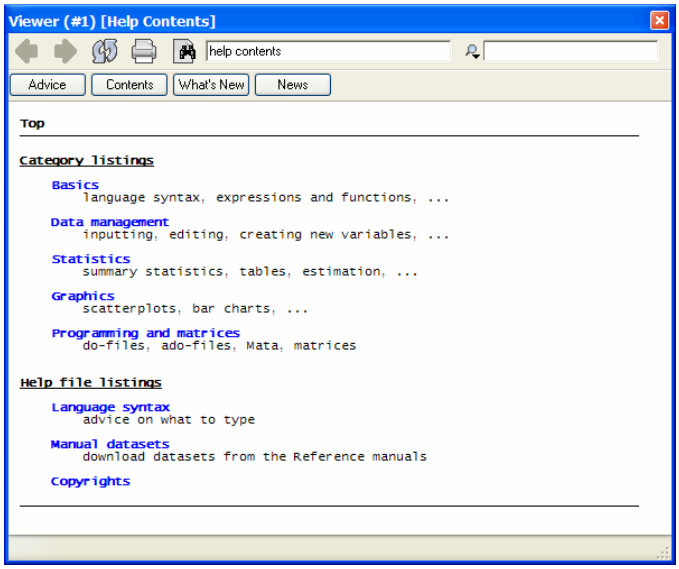


1.8 THE STATA HELP SYSTEM

The Stata help system is one if its most powerful features. Click on **Help** on the menu.



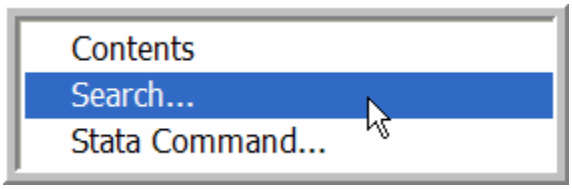
Select **Contents**.



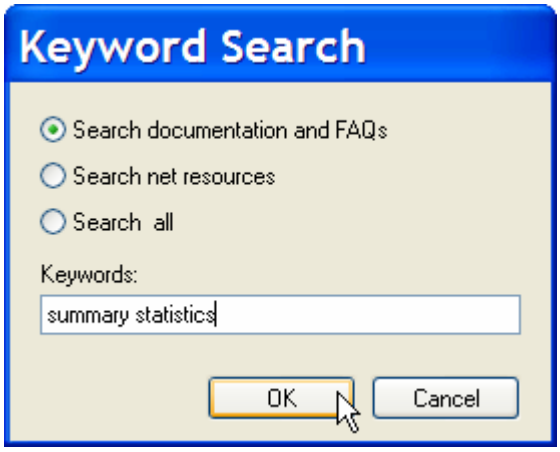
Each of the **blue** words is linked to further screens. You should explore these to get a feel for what is available.

1.8.1 Using keyword search

Now click on **Help > Search**

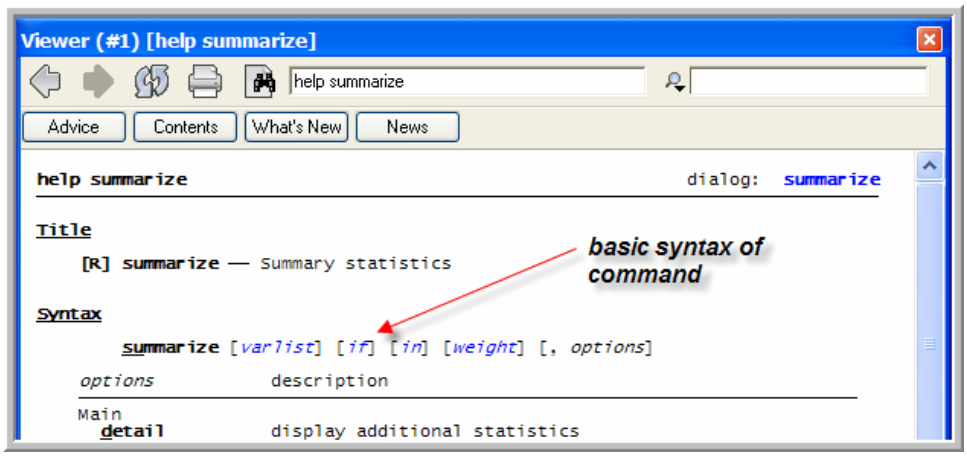


In the **Dialog** box that opens there are several search options. To search all the Stata documentation and **Frequently Asked Questions (FAQs)** simply type in phrase describing what you want to find. It does not have to be a specific Stata command. For example, let's search for **Summary Statistics**.



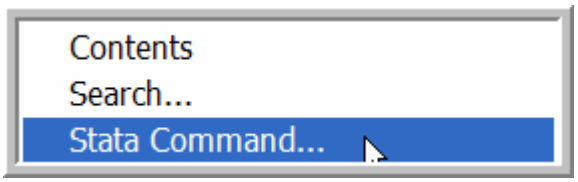
Up comes a list of topics that might be of interest. Once again **blue** terms are links. Click on **Summarize**.

The resulting **Viewer** box shows the command syntax, which can be used when typing commands in the **Command** window, and many options.

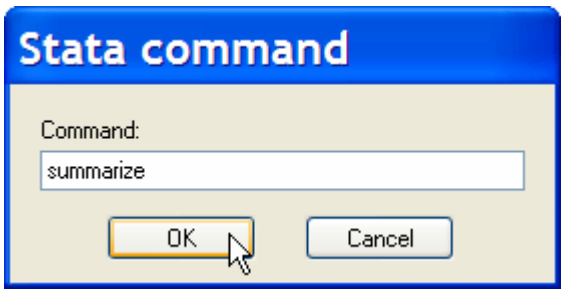


1.8.2 Using command search

If you know the name of the Stata command you want help with, click **Help > Stata Command**



In the resulting dialog box type in the name of the command and click **OK**.



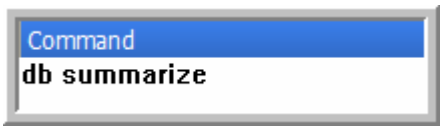
Alternatively, on the command line type

```
help summarize
```

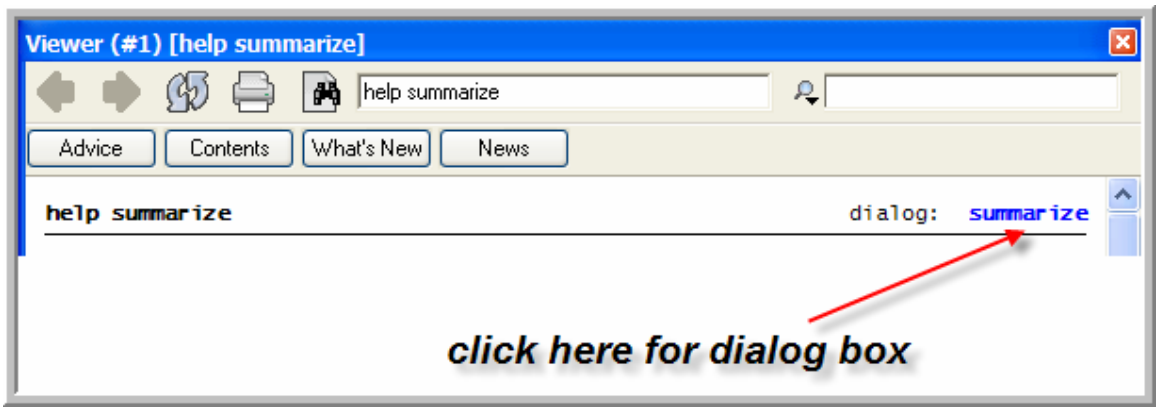
and press **Enter**

1.8.3 Opening a dialog box

If you know the name of the command you want, but do not recall details and options, a dialog box can be opened from the **Command** window. For example, if you wish to **`summarize`** the data using the dialog box, enter **`db summarize`**



Or, enter **`help summarize`**, and click on the **blue** link to the dialog box.



1.9 STATA COMMAND SYNTAX

Stata commands have a common syntax. The name of the command, such as **`summarize`** is first.

```
command [varlist] [if] [in] [weight] [, options]
```

The terms in brackets **`[ ]`** are various optional command components that could be used.

- **`[varlist]`** is the list of variables for which the command is used.
- **`[if]`** is a condition imposed on the command.
- **`[in]`** specifies range of observations for the command.
- **`[weight]`** when some sample observations are to be weighted differently than others.
- **`[, options]`** command options go here.

For more on these options use a **Keyword Search** for **Command syntax**, then click **Language**.

> **Remark:** An important fact to keep in mind when using Stata is that its commands are **case sensitive**. This means that lower case and capital letters have different meanings. Since Stata considers **`x`** to be different from **`X`**, it is easy to make programming errors.

1.9.1 Syntax of summarize

Consider the following examples using the syntax features. In each case type the command into the **Command** window and press **Enter**. For example,

```
Command
summarize str, detail
```

`summarize str, detail` computes detailed summary statistics for the variable wage. The percentiles of wage from smallest to largest are shown, along with additional summary statistics (e.g., skewness and kurtosis) that you will learn about. Note that Stata echoes the command you have issued with a preceding period (.).

```
. summarize str, detail

                    Student to teacher ratio
-------------------------------------------------------------
      Percentiles      Smallest
 1%     15.13898             14
 5%     16.41658       14.20176
10%     17.34573       14.54214       Obs                 420
25%     18.58179       14.70588       Sum of Wgt.         420

50%     19.72321                      Mean           19.64043
                        Largest       Std. Dev.      1.891812
75%     20.87183          24.95
90%     21.87561       25.05263       Variance       3.578952
95%     22.64514       25.78512       Skewness      -.0253655
99%     24.88889           25.8       Kurtosis       3.609597
```

`summarize str if testscr>=650` computes the simple summary statistics of the student teacher ratio for those classes having test scores above 650. In the "**if statement**" [called an "if qualifier" by Stata] equality is indicated by "`==`".

`summarize in 1/50` computes summary statistics for observations 1 through 50.

`summarize str in 1/50, detail` computes detailed summary statistics for the variable **`str`** in the first 50 observations.

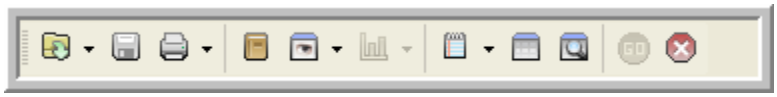
> If you notice at bottom left of the Results window **`—more—`**: when the **Results** window is full it pauses and you must click **`—more—`** in order for more results to appear, or press the space bar.

1.10 SAVING YOUR WORK

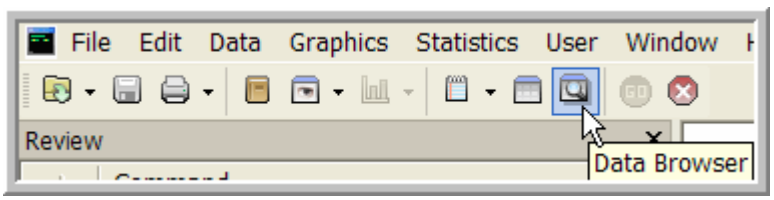
When you carry out a long Stata session you will want to save your work.

1.11 USING THE DATA BROWSER

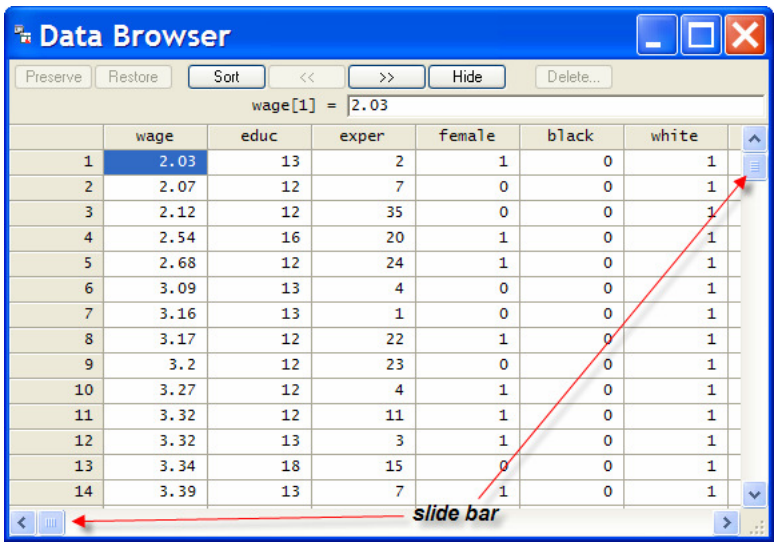
It is a good idea to examine the data to see the magnitudes of the variables and how they appear in the data file. On the Stata toolbar are a number of icons



Sliding the mouse pointer over each icon reveals its use. Click on **Data Browser**



The data browser is a spreadsheet view



	wage	educ	exper	female	black	white
1	2.03	13	2	1	0	1
2	2.07	12	7	0	0	1
3	2.12	12	35	0	0	1
4	2.54	16	20	1	0	1
5	2.68	12	24	1	0	1
6	3.09	13	4	0	0	1
7	3.16	13	1	0	0	1
8	3.17	12	22	1	0	1
9	3.2	12	23	0	0	1
10	3.27	12	4	1	0	1
11	3.32	12	11	1	0	1
12	3.32	13	3	1	0	1
13	3.34	18	15	0	0	1
14	3.39	13	7	1	0	1

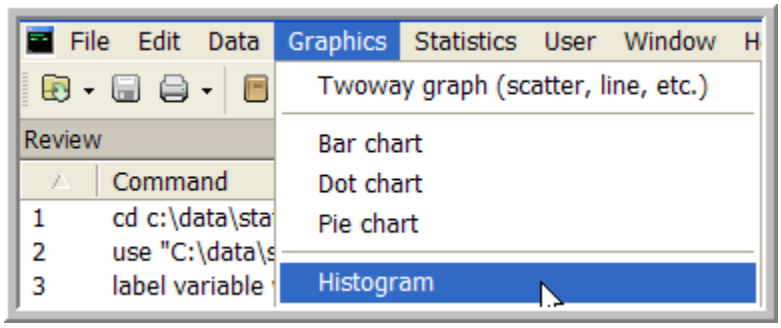
Use the **slide bar** at the bottom and the one on the right to view the entire data array. The browser allows you to scroll through the data, but not to edit any of the entries. This is a good feature that ensures we do not accidentally change a data value. Be sure to close the data browser when finished. Stata will not accept any new commands when the browser is open.

1.12 USING STATA GRAPHICS

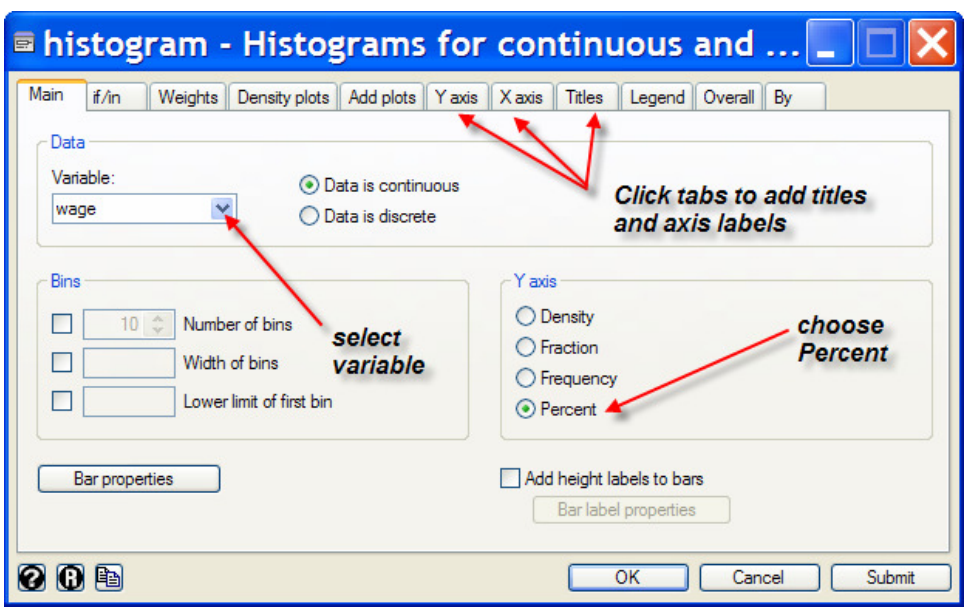
Stata does very nice graphics. We will illustrate a **Histogram** and a **Scatter Plot**.

1.12.1 Histograms

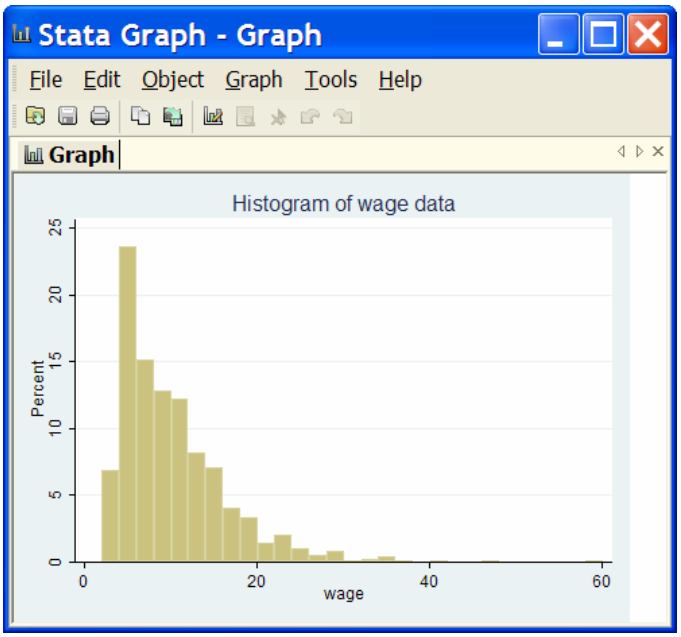
Click on **Graphics > Histogram** on the Stata menu



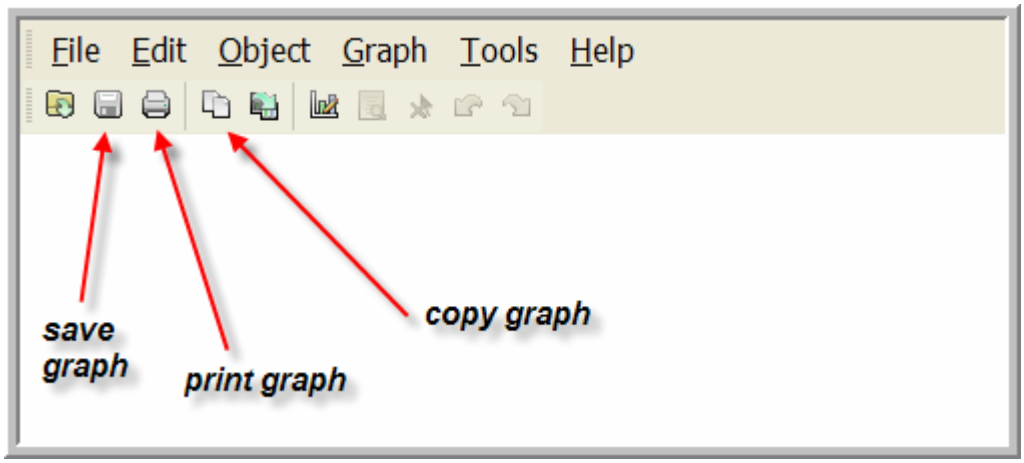
In the resulting dialog box there are again many options. For a simple histogram all you need to is select is the variable from the pull-down list. For illustration, we have entered a title by clicking the **Titles** tab and filling in a box. Click **OK**.



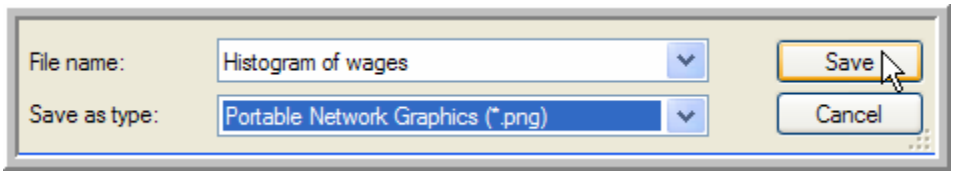
The resulting figure is



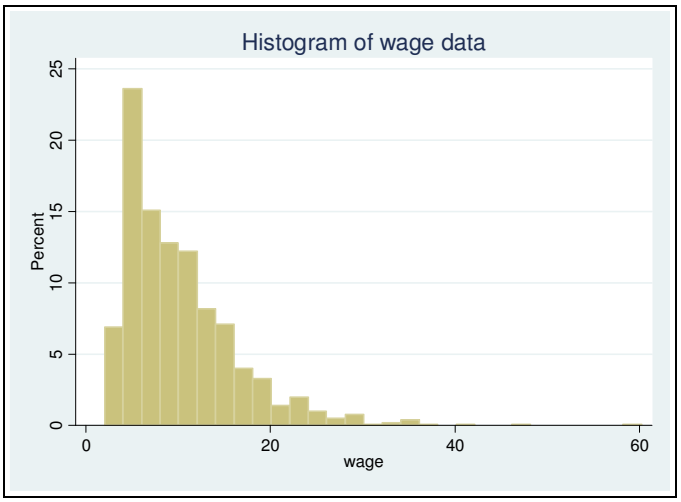
On the graph toolbar you have several options.



Click on the **Save graph** icon. The resulting dialog box shows that the graphics file will be saved into **C:\data\stata**. Attach a file name and choose the type of graphics file from the drop-down list. This book uses **png** files.



Having saved the file, in your word processor you can insert the image as a figure into a document. Alternatively, if you choose the **Copy graph** icon the figure will be copied to the clipboard, and then the figure can be pasted (**Ctrl+V**) into an open document.

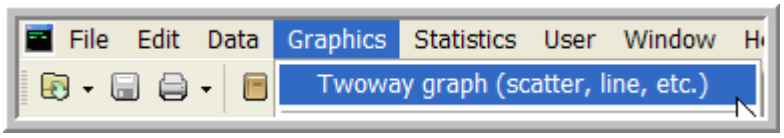


Note that out pointing and clicking could have been replaced by the command

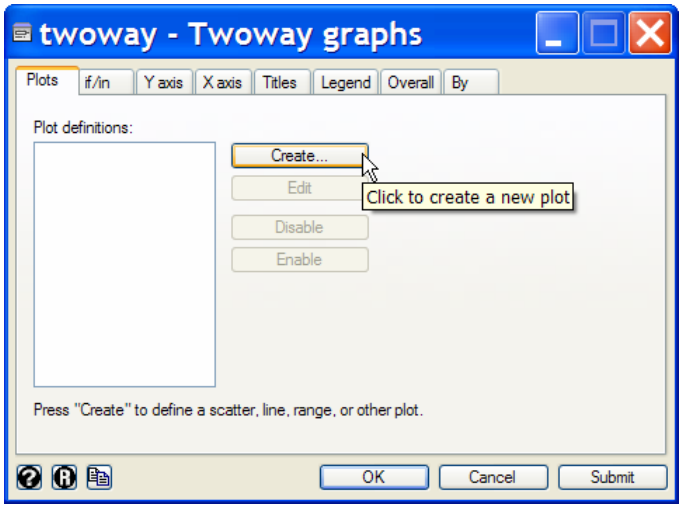
```
histogram wage, percent title(Histogram of wage data)
```

1.12.2 Scatter diagrams

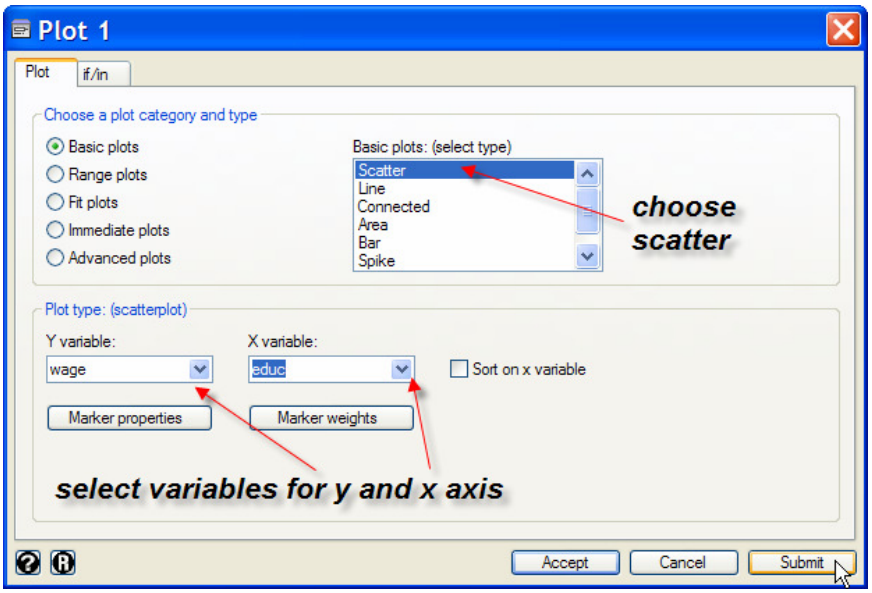
A scatter diagram is a **Two-way Graph**. From the graphics menu select this option



In the dialog box, click **Create**



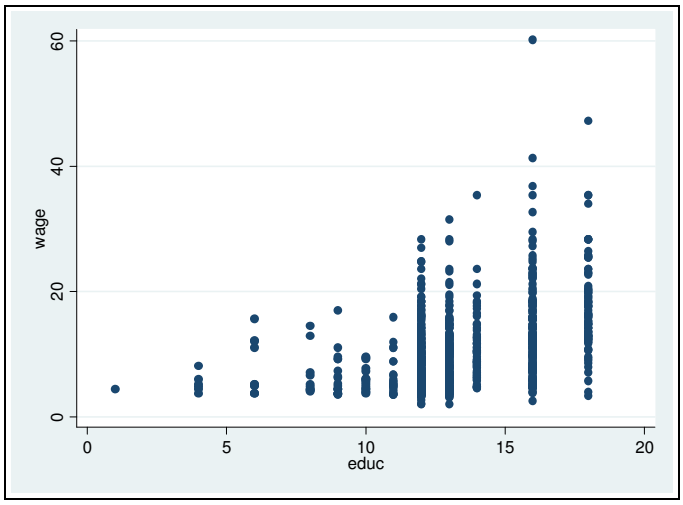
A dialog box opens.



Choose the Y variable (vertical axis) and X variable (horizontal axis). Select the **Scatter** plot, and click **Submit**. The resulting graph can be saved to a file, or copied and pasted into a document, as with the histogram. The result shows "dots" for each data pair (educ, wage), and by casual inspection we see that more education usually leads to higher wages. Aren't you glad.

The Stata command used to create this scatter plot is

```
twoway (scatter wage educ)
```



1.13 USING STATA DO-FILES

While it is possible to point and click your way to success such an approach requires a new pointing and clicking odyssey each time you do a new problem. In our view it is more convenient

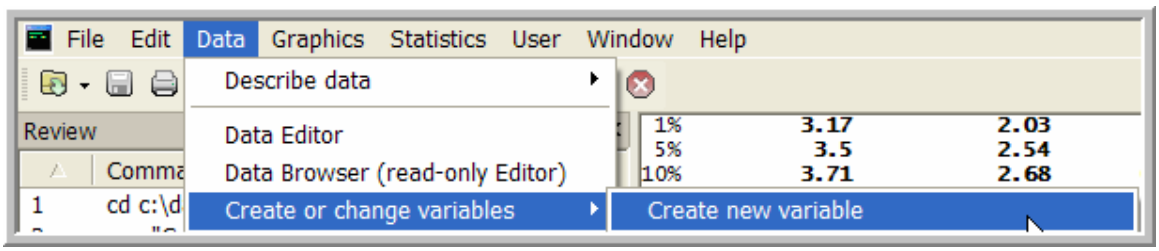
is to use Stata's **Do-files** as a method for executing commands. These are files containing lists of commands that will be executed as a batch.

1.14 CREATING AND MANAGING VARIABLES

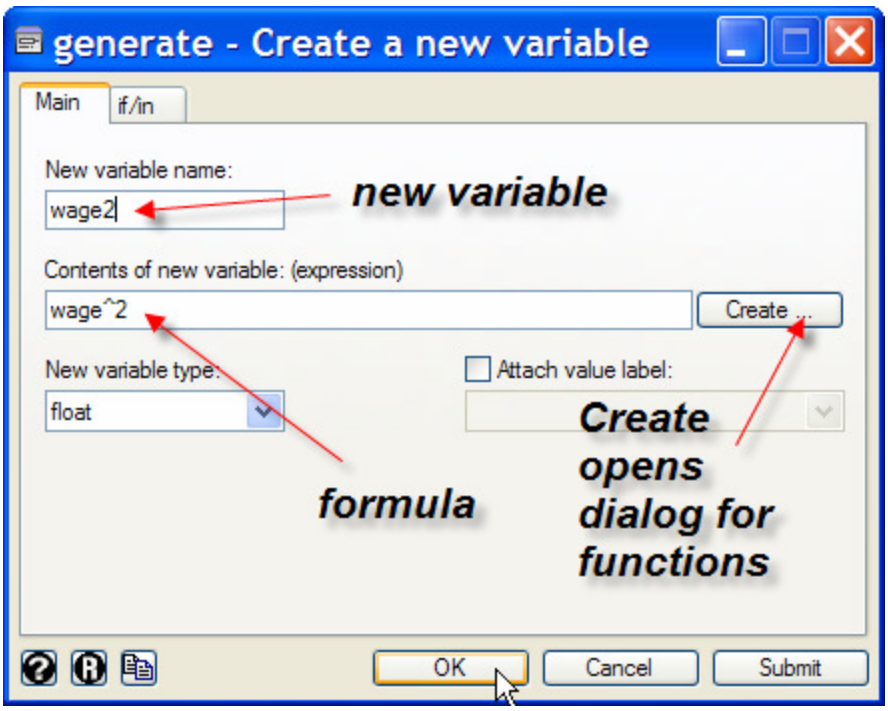
Stata offers a wide variety of functions that can be used to create new variables, and commands that let you alter the variables you have created. In this section we examine some of these capabilities.

1.14.1 Creating (generating) new variables

To create a new variable use the **`generate`** command. Let's start with the pull-down menu. Click on **Data > Create or change variables > Create new variable** on the Stata menu.



A dialog box will open.



Alternatively, in the **Command** window, enter **`db generate`** to open the dialog box. In the dialog box you must fill in

New variable name: choose something logical, informative and not too long.

Contents of new variable: this is a formula (no equal sign required) that is a mathematical expression. In the example above **wage2** is a new variable that will be the square of wage. The operator "^" is the symbol Stata uses for "raise to a power, so **`wage^2`** is the square of wage, **`wage^3`** would be wage cubed, and so on.

Click **OK**. In the **Results** window (and **Review** window) we see that the command implied by the menu process is

```
generate float wage2 = wage^2
```

In this command **`float`** is automatically added by the menu driven process and is a description of the type of variable being created. It stands for **floating point**. Type **`help data type`** if you are curious. It is an **option** and is not required. We can enter

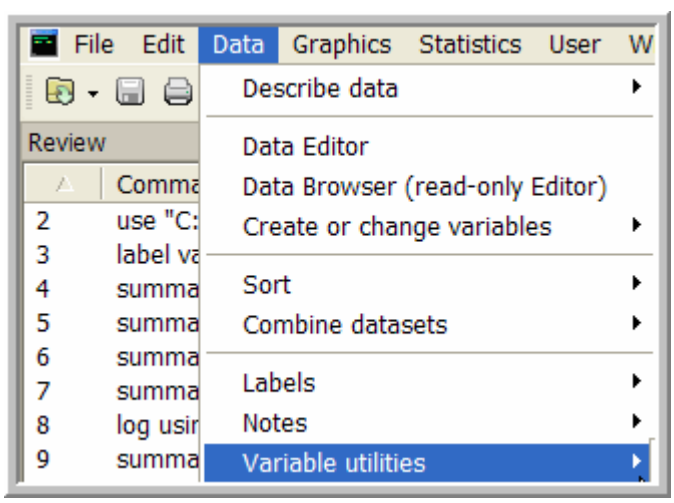
```
generate wage2 = wage^2
```

The command can also be shortened to

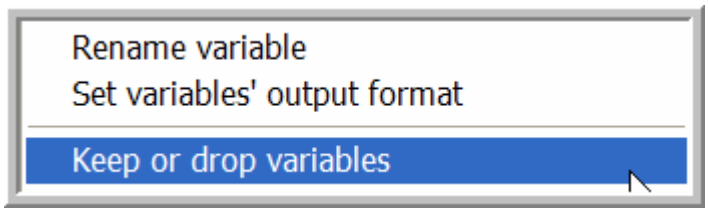
```
gen wage2 = wage^2
```

1.14.2 Dropping or renaming a variable

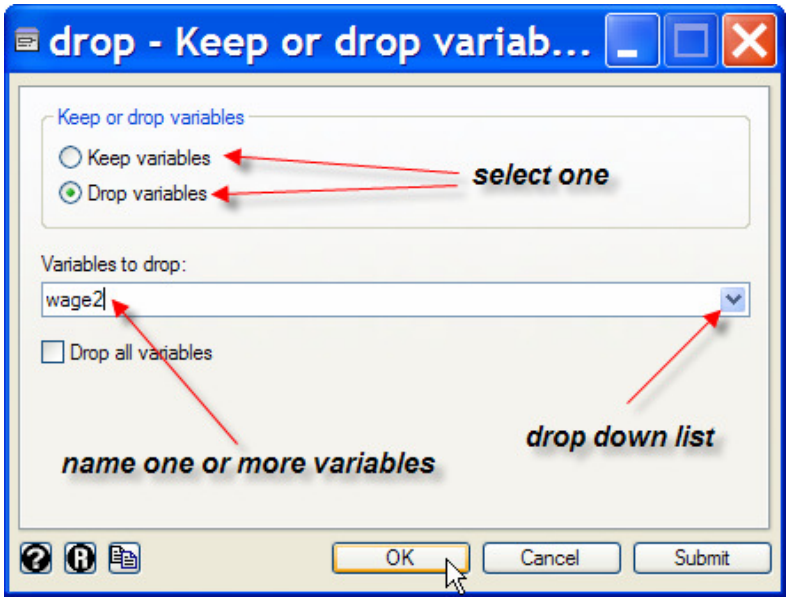
To drop or rename a variable in the variable list, click on **Data > Variable utilities**.



Then

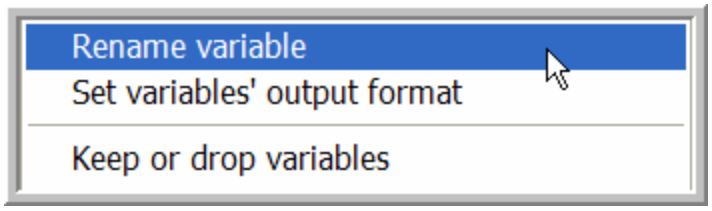


The command choice is **Keep** or **Drop**.

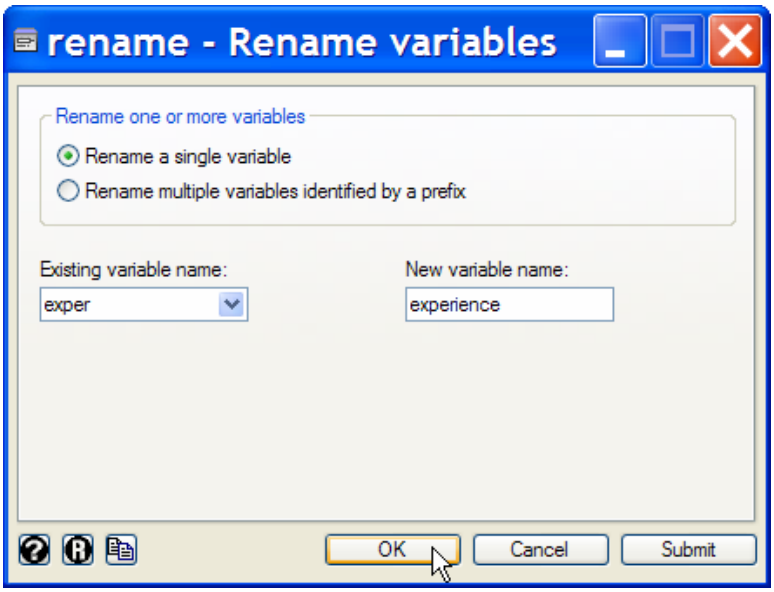


- **`Drop`** deletes the selected variables from the data file.
- **`Keep`** deletes **all** variables from the data file **except** the ones selected.

To **Rename** a variable, click **Data > Variable utilities > Rename variable**.



Suppose we want to rename **exper** as **experience**. Then fill in the dialog box as shown below



The **drop** and **rename** commands are simple to enter directly, and are

```
drop wage2
rename exper experience
```

1.14.3 Using arithmetic operators

The **Arithemetic operators** are:

- **+** addition
- **-** subtraction (or create negative of value, or negation)
- ***** multiplication
- **/** division
- **^** raise to a power

To illustrate these operators consider the following generate statements:

generate wage1 = wage+1 (addition)
generate negwage = -wage (negative or negation)
generate blackeduc = black*educ (multiplication)
generate blackeduc_south = black*educ*south (multiplication)
generate blackeduc_west = blackeduc*west (multiplication with created variable)
generate wage_yr = wage/educ (division)
generate blackeduc_midwest = (black*educ)*midwest (multiplication)

The last line shows the use of parentheses. Like regular algebra parentheses control the order of operations, with expressions in parentheses being performed first.

Several of these constructions were for demonstration purposes only. We'll **drop** them using

```
drop blackeduc_west blackeduc_midwest wage1 negwage wage_yr
```

> **Stata shortcut:** With a list of variables to type it is easier to type the command name, here **`drop`**, and then click on the names of the variables in the **Variables** window. When selected they appear in the **Command** window.

1.14.4 Using Stata math functions

Stata has a long list of mathematical and statistical functions that are easy to use. Type **`help functions`** in the Command window. We will be using **math functions** and **density functions** extensively.

```
help functions
---------------------------------------------------------------------------

Title

    [D] functions — Functions in expressions

    Quick references are available for the following types of functions:

        | Type of function                      | See help                |
        |---------------------------------------|-------------------------|
        | Mathematical functions                | math functions          |
        | Probability distributions and         |                         |
        |   density functions                   | density functions       |
```

Click on **math functions**. Scrolling down the list you will see many functions that are new to you. A few examples of the ones we will be using are:

`generate lwage = log(wage)` (natural logarithm)
`generate elwage = exp(lwage)` (exponential function is antilog of natural log)
`generate rootexper = sqrt(exper)` (square root)

Note that the exponential function is e^x. Use the Stata **browser** to compare the values of **wage** and **elwage**. These are identical because the exponential function is the antilog of the natural logarithm. The variable **lwage** is the logarithm of **wage**, and **elwage** is the antilog of **lwage**. The function **`log(wage)`** is the natural logarithm and so is **`ln(wage)`**. In *Principles of Econometrics* the notation $\ln(x)$ is used to denote the natural logarithm.

Example of standard normal cdf

To illustrate, lets compute the probability that a standard normal random variable Z takes a values less than or equal to 1.27. This is computed using the *cdf* **`normal`**. Type the following **`commands`** into the **Command** window.

`scalar phi = normal(1.27)` computes a scalar variable that is the desired probability.

`display phi` reports the value of the computed probability on the next line.

```
.89795768
```

`display "Prob (Z <= 1.27) = " phi` illustrates inserting text into display.

```
Prob (Z <= 1.27) = .89795768
```

`di "Prob (Z <= 1.27) = " phi` shows that **display** can be abbreviated **di**.

```
Prob (Z <= 1.27) = .89795768
```

We do not have to first create **`phi`** at all. We can simply **`display`** the value by including the function to be evaluated in the display statement.

```
di "Prob (Z <= 1.27) = " normal(1.27)
Prob (Z <= 1.27) = .89795768
```

Example of t-distribution tail-cdf

Compute the probability that a t-random variable with $n = 20$ degrees of freedom takes a value greater than 1.27.

```
scalar p = ttail(20,1.27)

di "Prob (t(20) > 1.27) = " p
      Prob (t(20) > 1.27) = .1093311
```

or

```
di "Prob (t(20) > 1.27) = " ttail(20,1.27)
Prob (t(20) > 1.27) = .1093311
```

Example computing percentile of the standard normal

Compute the value of the standard normal distribution z such that $p = .90$ of the probability falls to its left, so that $P(Z < z) = .90$. In this case z is the 90th percentile of the standard normal distribution.

```
scalar z = invnormal(.90)

di "90th percentile value of standard normal " z
90th percentile value of standard normal 1.2815516
```

Example computing percentile of the t-distribution

Compute the value t of the t-distribution with $n = 20$ degrees of freedom such that $p = .90$ of the probability falls to its left, so that $P(t_{(20)} < t) = .90$. In this case t is the 90th percentile of the t distribution with 20 degrees of freedom. This problem is complicated by the fact that Stata provides only the "tail" function for the t-distribution, so the 90th percentile value is found by locating the point such that $p = .10$ of the probability lies in the upper-tail of the distribution, that is $P(t_{(20)} > t) = .10$.

```
scalar t = invttail(20,.10)

di "90th percentile value of t(20) distribution " t
   90th percentile value of t(20) distribution 1.3253407
```

You will note that the 90th percentile of the $t_{(20)}$ distribution is larger than the 90th percentile of the standard normal distribution. This is as it should be, as the t-distribution is "wider" than the standard normal. As noted earlier the **invttail** function can go into the **display** statement

```
di "90th percentile value of t(20) distribution " invttail(20,.10)
   90th percentile value of t(20) distribution 1.3253407
```