

# A Gentle Introduction to Stata

5th Edition

ALAN C. ACOCK  
*Oregon State University*



A Stata Press Publication  
StataCorp LP  
College Station, Texas



Copyright © 2006, 2008, 2010, 2012, 2014, 2016 by StataCorp LP  
All rights reserved. First edition 2006  
Second edition 2008  
Third edition 2010  
Revised third edition 2012  
Fourth edition 2014  
Fifth edition 2016

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845  
Typeset in L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub>  
Printed in the United States of America  
10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-185-4  
Print ISBN-13: 978-1-59718-185-3

Library of Congress Control Number: 2016935690

No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means—electronic, mechanical, photocopy, recording, or otherwise—without the prior written permission of StataCorp LP.

Stata, **stata**, Stata Press, Mata, **mata**, and NetCourse are registered trademarks of StataCorp LP.

Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations.

L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub> is a trademark of the American Mathematical Society.

## Contents

List of figures	xv
List of tables	xxiii
List of boxed tips	xxv
Preface	xxix
Support materials for the book	xxxv
<b>1 Getting started</b>	<b>1</b>
1.1 Conventions . . . . .	1
1.2 Introduction . . . . .	4
1.3 The Stata screen . . . . .	7
1.4 Using an existing dataset . . . . .	9
1.5 An example of a short Stata session . . . . .	11
1.6 Video aids to learning Stata . . . . .	18
1.7 Summary . . . . .	19
1.8 Exercises . . . . .	19
<b>2 Entering data</b>	<b>21</b>
2.1 Creating a dataset . . . . .	21
2.2 An example questionnaire . . . . .	23
2.3 Developing a coding system . . . . .	24
2.4 Entering data using the Data Editor . . . . .	29
2.4.1 Value labels . . . . .	33
2.5 The Variables Manager . . . . .	34
2.6 The Data Editor (Browse) view . . . . .	40
2.7 Saving your dataset . . . . .	41
2.8 Checking the data . . . . .	43

documentation. One of the best aspects of the Stata documentation is that it provides several real-data examples for most commands. An entry will start with a fairly simple example and then give examples that are more complex. Looking at the examples is how I have learned much of what I know about Stata. You will find that the capabilities for many of the commands I discuss far exceed what I was able to cover here.

If you remember the name of a command, you can type `help command_name` in the Command window. For example, typing `help summarize` would display a Viewer window with brief information and examples of how to run the command. If you do not know the exact name of the command, you could just enter the first part. For example, typing `help sum` opens a window with two options, one of which is `summarize`. If you enter the wrong name for a command, say, you type `help summary`, Stata opens a Viewer window with a list of files where the word “summary” was listed as a keyword. You scroll through the list and find the `summarize` command. If you click on `summarize`, the help file for the `summarize` command opens in the Viewer window.

The help file does not give you all the detailed explanation and examples that you get from the PDF documentation, but it is often all you need. You can open the PDF document for a specific command by clicking on the command name in the *Title* section or in the **Also See** menu of the help file.

My hope in writing this book is to give you sufficient background so that you can use the manuals effectively.

# 1 Getting started

---

- 1.1 Conventions
  - 1.2 Introduction
  - 1.3 The Stata screen
  - 1.4 Using an existing dataset
  - 1.5 An example of a short Stata session
  - 1.6 Video aids to learning Stata
  - 1.7 Summary
  - 1.8 Exercises
- 

## 1.1 Conventions

Listed below are the conventions that are used throughout the book. I thought it might be convenient to list them all in one place should you want to refer to them quickly.

**Typewriter font.** I use this font when something would be meaningful to Stata as input. I also use it to indicate Stata output.

I use a typewriter font to indicate the text to type in the Command window. Because Stata commands do not have any special characters at the end, any punctuation mark at the end of a command in this book is not part of the command. Sometimes, to be consistent with Stata manuals, I will put a command on a line by itself with the dot preceding it, as in

```
. sysuse cancer, clear
```

All of Stata's dialog boxes generate commands, which will be displayed in the Review window and in the Results window. In the Results window, each command will be preceded by the dot prompt. If you make a point of looking at the command Stata prints each time you use the dialog boxes, you will quickly learn the commands. I may include the equivalent command in the text after explaining how to navigate to it through the dialog boxes.

### Why do we show the dot prompt with these commands?

When we show a listing of Stata commands, we place a dot and a space in front of each command. When you enter these commands in the Command window, you enter the command itself and not the dot prompt or space. We include these because Stata always shows commands this way in the Results window. Stata manuals and many other books about Stata follow this convention.

When you type a Stata command in the Command window, you execute the command when you press the Enter key. The command may wrap onto more than one line, but if you press the Enter key in the middle of entering a command, Stata will interpret that as the end of the command and will probably generate an error. The rule is that you should just keep typing when entering a command in the Command window, no matter how long the command is. Press Enter only when you want to execute the command.

I also use the typewriter font for variable names, for names of datasets, and to show Stata's output. In general, I use the typewriter font whenever the text is something that can be typed into Stata or when the text is something that Stata might print as output. This approach may seem cumbersome now, but you will catch on quickly.

Folder names, filenames, and filename extensions, as in "The `survey.dta` file is in the `C:\data` directory (or folder)", are also denoted in the typewriter font. Stata assumes that `.dta` will be the extension, so you can use just the filename without an extension, if you prefer.

**Sans serif font.** I use this font to indicate menu items (in conjunction with the ▷ symbol), button names, dialog-box tab names, and particular keys:

- Menu items, such as "Select Data ▷ Data utilities ▷ Rename groups of variables from the Stata menu" (see figure 1.1).

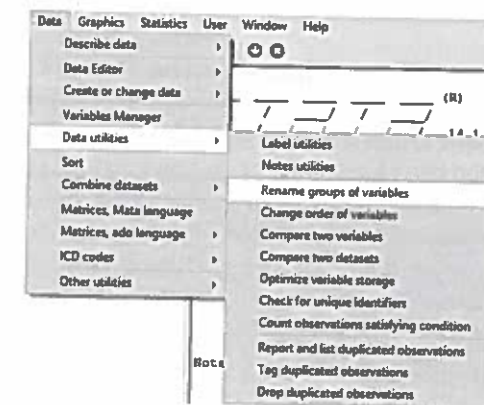


Figure 1.1. Stata menu

- Buttons that can be clicked on, as in "Remember, if you are working on a dialog box, it will now be up to you to click on OK or Submit, whichever you prefer."
- Keys on your keyboard, as in "The Page Up and Page Down keys will move you backward and forward through the commands in the Review window." Some functions require the use of the Shift, Ctrl, or Alt key, which will be held down while the second key is pressed. For example, Alt+f will open the File menu.

**Slant font.** I use this font for dialog-box titles and when I talk about labeled elements of a dialog box, with both items capitalized as they are on the dialog box.

**Italics font.** I use this font when I refer to a word that is to be replaced.

**Quotes.** I use double quotes when I am talking about labels in a general way, but I will use the typewriter font to indicate a specific label in a dataset. For example, if we decided to label the variable `age` "Age at first birth", we would enter `Age at first birth` in the textbox.

**Capitalization.** Stata is case sensitive, so `summarize` is a Stata command, whereas `Summarize` is not and will generate an error if you use it. Stata also recognizes capitalization in variable names, so `agegroup`, `Agegroup`, and `AgeGroup` will be three different variables. Although you can certainly use capital letters in variable names, you will probably find yourself making more typographical errors if you



do. I have found that using all lowercase letters when creating variable names is usually the best practice.

I will capitalize the names of the various Stata windows, but I do not set them off by using a different font. For example, we will type commands in the Command window and look at the output in the Results window.

### Setting how much output is in the Results window

The default size for the scrollbar buffer size for the Results window is 200 kilobytes, approximately 200,000 characters. If you have many results being displayed in the Results window, the default is to drop the oldest lines once you use up the 200 kilobyte buffer. If you want to be able to scroll back further, you can make the buffer size larger, up to 2,000 kilobytes. Select **Edit > Preferences > General preferences...** and click on the Results tab. Stata for Mac users can make this change by selecting **Stata > Preferences > General preferences...** and clicking on the Windows tab. You might change the scrollbar buffer size from the default 200 kilobytes to 500 kilobytes. This change will not take effect until you restart Stata.

Stata for Unix users cannot make this change from the Preferences dialog box; they must type the command `set scrollbarbuffer 500000` directly in the Command window.

Typing the command sets the scrollbar buffer size in bytes by default, whereas using the menu method sets the size in kilobytes.

Many Stata users find having to click on the **—more—** message when it appears in the Results window irritating. It is designed to make it easier to read the results of a single command, but if you do not like this feature, you can type the command `set more off` or `set more off, permanently`. The **permanently** option specifies that the setting be remembered for each future Stata session until you reverse the action by typing `set more on` or `set more on, permanently`.

## 1.2 Introduction

The best way to learn data analysis is to actually do it with real data. These days, doing statistics means doing statistics with a computer and a software package. There is no other software package that can match the internal consistency of Stata, which makes it easy to learn and a joy to use. Stata empowers users more effectively than any other statistical package.

### Work along with the book

Although it is not necessary, you will probably find it helpful to have Stata running while you read this book so that you can follow along and experiment for yourself. Having your hands on a keyboard and replicating the instructions in this book will make the lessons that much more effective, but more importantly, you will get in the habit of just trying something new when you think of it and seeing what happens. In the end, experimentation is how you will really learn how Stata works. The other great advantage to following along is that you can save the examples we do for future use.

Stata is a powerful tool for analyzing data. Stata makes statistics and data analysis fun because it does so much of the tedious work for you. A new Stata user should start by using the dialog boxes. As you learn more about Stata, you will be able to do more sophisticated analyses with Stata commands. Learning Stata well now is an investment that will pay off in saved time later. Stata is constantly being extended with new capabilities, which you can install using the Internet from within Stata. Stata is a program that grows with you.

Stata is a command-driven program. It has a remarkably simple command structure that you use to tell it what you want it to do. You can use a dialog box to generate the commands (this is a great way to learn the commands or prompt yourself if you do not remember one exactly), or you can enter commands directly. If you enter the `summarize` command, you will get a summary of all the variables in your dataset (mean, standard deviation, number of observations, minimum value, and maximum value). Enter the command `tabulate gender`, and Stata will make a frequency distribution of the variable called `gender`, showing you the number and percentage of men and women in your dataset.

After you have used Stata for a while, you may want to skip the dialog box and enter these commands directly. When you are just beginning, however, it is easy to be overwhelmed by all the commands available in Stata. If you were learning a foreign language, you would have no choice but to memorize hundreds of common words right away. This is not necessary when you are learning Stata because the dialog boxes are so easy to use.

## Searching for help

Stata can help when you want to find out how to do something. You can use the `search` command along with a keyword. For example, you believe that a  $t$  test is what you want to use to compare two means. Enter `search t test`; Stata searches its own resources and others that it finds on the Internet. The first entry of the results is

```
[R]      ttest . . . . . t tests (mean-comparison tests)
        (help ttest)
```

The [R] at the beginning of the line means that details and examples can be found in the *Stata Base Reference Manual*. Click on the blue `ttest` to go to the help file for the `ttest` command. If you think this help is too cryptic, repeat the `search t test` command and look farther down the list. Scroll past the lines starting with Video, and look for the lines starting with FAQ (frequently asked questions). One of these is “What statistical analysis should I use?” Click on the blue URL to go to a UCLA webpage that will help you decide whether the  $t$  test is the best choice for what you are doing. You might click on some of the other resources to see how much support you get from a wide variety of resources.

When using the `search` command, you need to pick a keyword that Stata knows. You might have to try different keywords before you get one that works. Searching these Internet locations is a remarkable capability of Stata. If you are reading this book and want to know more about a command, the online help is the first place to start. Suppose that we are discussing the `summarize` command and you want to know more options for this command. Type `help summarize` and you will get an informative help screen. To obtain complete information for a command, you should see the PDF documentation. The PDF documentation can be opened from the Stata menu by selecting Help ▸ PDF documentation. Bookmarks to all the Stata manuals are available; click on the plus sign (+) next to each manual to see bookmarks to sections therein.

Stata has done a lot to make the dialog boxes as friendly as possible so that you feel confident using them. The dialog boxes often show many options, which control the results that are shown and how they are displayed. You will discover that the dialog boxes have default values that are often all you need, so you may be able to do a great deal of work without specifying any options.

As we progress, you will be doing more complex analyses. You can do these using the dialog boxes, but Stata lets you create files that contain a series of commands you can run all at once. These files, called do-files, are essential once you have many commands to run. You can reopen the do-file a week or even several months later and repeat exactly what you did. Keeping a record of what you do is essential; otherwise, you will not be able to replicate results of elaborate analyses. Fortunately, Stata makes this easy.

## 1.3 The Stata screen

You will learn more about replicating results in chapter 4. The do-files that reproduce most of the tables, graphs, and statistics for each chapter are available on the webpage for this book (<http://www.stata-press.com/data/ags5/>).

Because Stata is so powerful and easy to use, I may include some analyses that are not covered in your statistics textbook. If you come to a procedure that you have not already learned in your statistics text, give it a try. If it seems too daunting, you can skip that section and move on. On the other hand, if your statistics textbook covers a procedure that I omit, you might search the dialog boxes yourself. Chances are that you will find it there.

Depending on your needs, you might want to skip around in the book. Most people tend to learn best when they need to know something, so skipping around to the things you do not know may be the best use of the book and your time. Some topics, though, require prior knowledge of other topics, so if you are new to Stata, you may find it best to work through the first four chapters carefully and in order. After that, you will be able to skip around more freely as your needs or interests demand.

## 1.3 The Stata screen

When you open Stata, you will see a screen that looks something like figure 1.2.

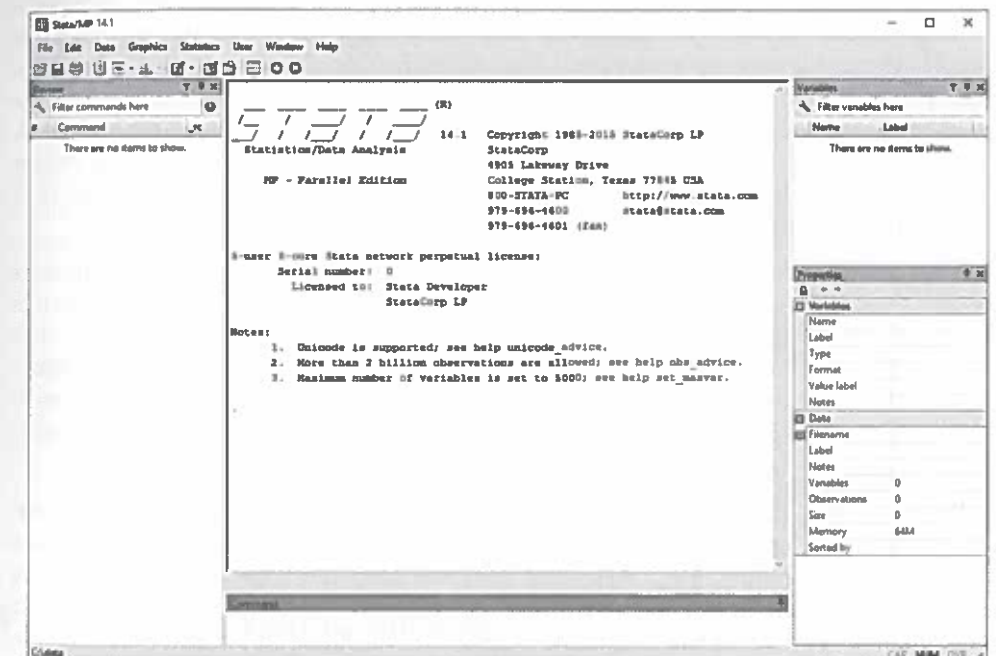


Figure 1.2. Stata's opening screen



You can rearrange the windows to look the way you want them, although many users are happy with the default layout. If you are satisfied with the defaults, you might skip the next couple paragraphs and come back to them if you change your mind later. Many experienced Stata users have particular ways to arrange these screens. Feel free to experiment with the layout.

Selecting **Edit > Preferences** gives you several options. One thing you might want to do is change the size of the buffer for the Results window. The factory default of 200 kilobytes may be too small to be able to scroll through all your results. To change the size of the buffer, select **Edit > Preferences > General preferences...** and then click on the tab labeled **Results**. Depending on how much memory your computer has available, you might want to raise the default value to as much as 500 kilobytes. You can resize the Stata interface as you would any other Windows product. There are other options you can try under **Results** and each of the other tabs. It is nice to personalize your interface in a way that is attractive to you. I will use the generic “factory settings” for this book, however. If you make several changes and want to return to the starting point, select **Edit > Preferences > Load preference set > Widescreen layout (default)**. If you are using Stata for Mac, select **Stata > Preferences > Manage preferences > Factory settings**.

When you open a file that contains Stata data, which we will call a Stata dataset, a list of the variables will appear in the Variables window. The Variables window reports the name of the variable (for example, `abortion`) and a label for the variable (for example, `Attitude toward abortion`). Other information about the variable is shown in the Properties window, such as the type of variable (for example, `float`) and the format of the variable (for example, `%8.0g`). For now, just consider the name and label. You can vary the width of each column in the Variables window by placing your cursor on the vertical line between the name and label, clicking on it, and then dragging your cursor to the right or left.

When Stata executes a command, it prints the results or output in the Results window. First, it prints the command preceded by a `.` (dot) prompt, and then it prints the output. The commands you run are also listed in the Review window. If you click on one of the commands listed in the Review window, it will appear in the Command window. If you double-click on one of the commands listed in the Review window, it will be executed. You will then see the command and its output, if any, in the Results window.

When you are not using the interface, you enter commands in the Command window. You can use the **Page Up** and **Page Down** keys on your keyboard to recall commands from the Review window. On a Mac that does not have the **Page Up** and **Page Down** keys, you can use the **fn** key with the arrow up or arrow down key. You can also edit commands that appear in the Command window. I will illustrate all these methods in the coming chapters.

The gray bar at the bottom of the screen, called the status bar, displays the current working directory (folder). (Note that this status bar looks slightly different on a Mac.) This directory may be different on different computers depending on how Stata was installed. The working directory is where Stata will look for a file or save a file unless you specify the full path to a different directory that contains the file. Stata recognizes the tilde to represent your home directory. If you have a project and want to store all files related to that project in a particular directory, say, `C:\data\thesis`, you could enter the command `cd C:\data\thesis`. This command assumes that this directory already exists on your computer.

To change the working directory, you can either use the `cd` command or select **File > Change working directory...** and click on the directory of choice. It is helpful to have a working directory for each project. On a Mac, if you cannot remember where you saved a file containing data, you can click on the magnifying glass in the upper-right corner of the Mac screen to search the name of the file and then click on the file to open it. If you already have a dataset open, you may need to type the `clear` command in the Command window first.

Stata has the usual Windows title bar across the top, on the right side of which are the three buttons (in order from left to right) to minimize, to expand to full-screen mode, and to close the program. Immediately below the Stata title bar is the menu bar, where the names of the menus appear. Some of the menu items (**File**, **Edit**, and **Window**) will look familiar because they are used in other programs. The **Data**, **Graphics**, and **Statistics** menus are specific to Stata, but their names provide a good idea of what you will find under them.

Figures 1.3 and 1.4 show the Stata toolbar as it appears in Windows and Mac, respectively. The icons provide alternate ways to perform some of the actions you would normally do with the menus. If you hold the cursor over any of these icons for a couple of seconds, a brief description of the function appears. For a complete list of the toolbar icons and their functions, see the *Getting Started with Stata* manual.



Figure 1.3. The toolbar in Stata for Windows



Figure 1.4. The toolbar in Stata for Mac

## 1.4 Using an existing dataset

Chapter 2 discusses how to create your own dataset, save it, and use it again. You will also learn how to use datasets that are on the Internet. For now, we will use a simple dataset that came with Stata. Although we could use the dialog box to do this, we will enter a simple command. Click once in the Command window to put the cursor there,

and then type the command `sysuse cancer, clear`; the Command window should look like the one in figure 1.5.



```
Command
sysuse cancer, clear
```

Figure 1.5. Stata command to open `cancer.dta`

The `sysuse` command we just used will find the sample dataset on your computer by name alone, without the extension; in this case, the dataset name is `cancer`, and the file that is found is actually called `cancer.dta`. The `cancer` dataset was installed with Stata. This particular dataset has 48 observations and 4 variables related to a cancer treatment.

What if you forget the command `sysuse`? You could open a file that comes with Stata by using the menu `File > Example datasets....` A new window opens in which you click on *Example datasets installed with Stata*. The next window then lists all the datasets that come with Stata. You can click on `use` to open the dataset.

Now that we have some data read into Stata, type `describe` in the Command window. That is it: just type `describe` and press the Enter key. `describe` will yield a brief description of the contents of the dataset.

```
. describe
Contains data from C:\Program Files\Stata14\ado\base/c/cancer.dta
  obs:          48      Patient Survival in Drug Trial
  vars:          8      3 Mar 2014 16:09
  size:        384
```

variable name	storage type	display format	value label	variable label
<code>studytime</code>	byte	%8.0g		Months to death or end of exp.
<code>died</code>	byte	%8.0g		1 if patient died
<code>drug</code>	byte	%8.0g		Drug type (1=placebo)
<code>age</code>	byte	%8.0g		Patient's age at start of exp.
<code>_st</code>	byte	%8.0g		1 if record is to be used; 0 otherwise
<code>_d</code>	byte	%8.0g		1 if failure; 0 if censored
<code>_t</code>	byte	%10.0g		analysis time when record ends
<code>_t0</code>	byte	%10.0g		analysis time when record begins

Sorted by:

The description includes a lot of information: the full name of the file, `cancer.dta` (including the path entered to read the file); the number of observations (48); the number of variables (8); the amount of memory the data consume (384 bytes); a brief description of the dataset (*Patient Survival in Drug Trial*); and the date the file was last saved. The body of the table displayed shows the names of the variables on the far left and the labels attached to them on the far right. We will discuss the middle columns later.

Now that you have opened `cancer.dta`, note that the Variables window lists the eight variables `studytime`, `died`, `drug`, `age`, `_st`, `_d`, `_t`, and `_t0`.

#### Internet access to datasets

Stata can use data stored on the Internet just as easily as data stored on your computer. If you did not have the `cancer.dta` file installed on your computer, you could read it by entering `webuse cancer`. However, you are not limited to data stored at the Stata site. Typing `use http://www.ats.ucla.edu/stat/stata/notes/hsb2` will open a dataset stored at the UCLA website.

Stata does not discard changes to the dataset currently in memory unless you tell it to do so. That is, if you have a dataset in memory and you have modified it, you will receive an error message if you try to load another dataset. You need to save the dataset in memory, type the `clear` command to discard the changes, or type the `clear` option of the `use` command to discard the changes. You can then load the new dataset.

Stata provides all the datasets for every example in its manuals. For example, click on `File > Example datasets....` A new window opens in which you click on *Stata 14 manual datasets*. There you might click on *Base Reference Manual [R]*; scroll down to `correlate`, and click on `use` to open any of the datasets or `describe` to see what variables are in the dataset.

## 1.5 An example of a short Stata session

If you do not have `cancer.dta` loaded, type the command `sysuse cancer`. We will execute a basic Stata analysis command. Type `summarize` in the Command window and then press Enter.

Rather than typing in the command directly, you could use the dialog box by selecting `Data > Describe data > Summary statistics` to open the corresponding dialog box. Simply clicking on the OK button located at the bottom of the dialog box will produce the `summarize` command we just entered. Because we did not enter any variables in the dialog box, Stata assumed that we wanted to summarize all the variables in the dataset.

You might want to select specific variables to summarize instead of summarizing them all. Open the dialog box again and click on the pulldown menu within the *Variables* box, located at the top of the dialog box, to display a list of variables. Clicking on a variable name will add it to the list in the box. Dialog boxes allow you to enter a variable more than once, in which case the variable will appear in the output more than once. You can also type variable names in the *Variables* box. Figure 1.6 shows the dialog box with the drop-down variable list displaying the variables in your dataset:



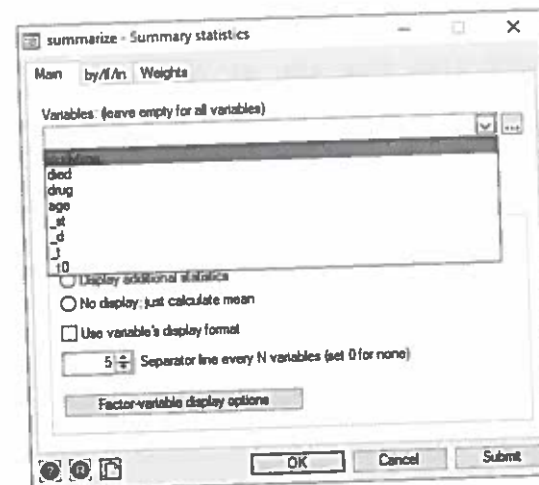








Figure 1.6. The summarize dialog box

In the bottom left corner of the dialog box, there are three icons: , , and . The  icon gives us a help screen explaining the various options. The explanations are brief, but there are examples at the bottom of the Viewer window. The  icon resets the dialog box. Just to the right of the  icon is an icon that looks like two pages. If you click on this icon, the command is copied to the Clipboard.

If you enter the `summarize` command directly in the Command window, simply follow it with the names of the variables for which you want summary statistics. For example, typing `summarize studytime age` will display only statistics for the two variables named `studytime` and `age`.

In the Results window, the `summarize` command will display the number of observations (also called cases or  $N$ ), the mean, the standard deviation, the minimum value, and the maximum value for each variable.

. summarize					
Variable	Obs	Mean	Std. Dev.	Min	Max
studytime	48	15.5	10.25629	1	39
died	48	.6458333	.4833211	0	1
drug	48	1.875	.8410986	1	3
age	48	55.875	5.659205	47	67
_st	48	1	0	1	1
<hr/>					
_d	48	.6458333	.4833211	0	1
_t	48	15.5	10.25629	1	39
_t0	48	0	0	0	0

The first line of output displays the dot prompt followed by the command. After that, the output appears as a table. As you can see, there are 48 observations in this dataset. *Observations* is a generic term. These could be called participants, patients,

subjects, organizations, cities, or countries depending on your field of study. In Stata, each row of data in a dataset is called an observation. The average, or mean, age is 55.875 years with a standard deviation of 5.659,<sup>1</sup> and the subjects are all between 47 (the minimum) and 67 (the maximum) years old.

If you have computed means and standard deviations by hand, you know how long this can take. Stata's virtually instant statistical analysis is what makes Stata so valuable. It takes time and skill to set up a dataset so that you can use Stata to analyze it, but once you learn how to set up a dataset (chapter 2), you will be able to compute a wide variety of statistics in little time.

We will do one more thing in this Stata session: we will make the histogram for the `age` variable, shown in figure 1.7.

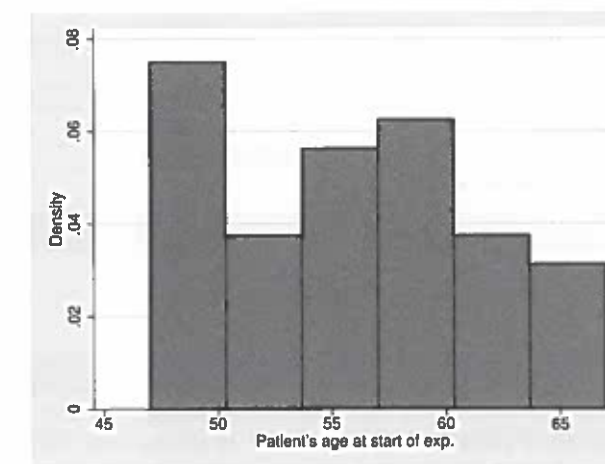


Figure 1.7. Histogram of age

A histogram is just a graph that shows the distribution of a variable, such as `age`, that takes on many values.

Simple graphs are simple to create. Just type the command `histogram age` in the Command window, and Stata will produce a histogram using reasonable assumptions. I will show you how to use the dialog boxes for more complicated graphs shortly.

At first glance, you may be happy with this graph. Stata used a formula to determine that six bars should be displayed, and this is reasonable. However, Stata starts the lowest bar (called a bin) at 47 years old, and each bin is 3.33 years wide (this information is displayed in the Results window) even though we are not accustomed to measuring years in thirds of a year. Also notice that the vertical axis measures density, but we

1. I may round numbers in the text to fewer digits than shown in the output unless it would make finding the corresponding number in the output difficult.

might prefer that it measure the frequency, that is, the number of people represented by each bar.

Using the dialog box can help us customize our histogram. Let's open the histogram dialog box shown in figure 1.8 by selecting Graphics ▸ Histogram from the menu bar.

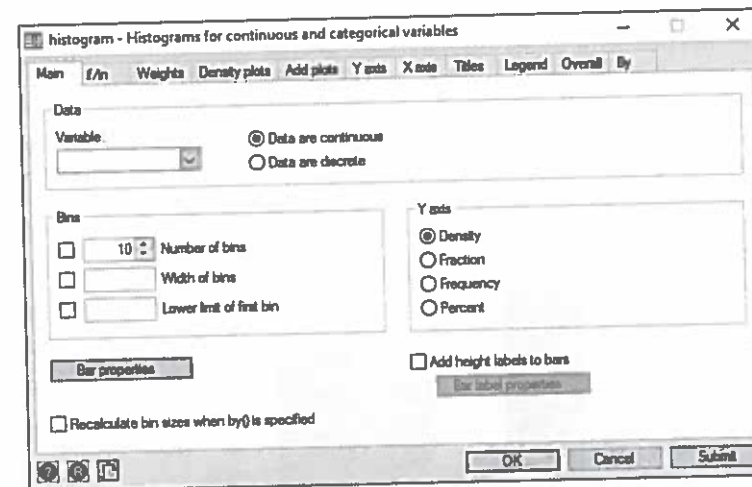


Figure 1.8. The histogram dialog box

Let's quickly go over the parts of the dialog box. There is a textbox labeled *Variable* with a pulldown menu. As we saw on the *summarize* dialog, you can pull down the list of variables and click on a variable name to enter it in the box, or you can type the variable's name yourself. Only one variable can be used for a histogram, and here we want to use *age*. If we stop here and click on OK, we will have re-created the histogram shown in figure 1.7.

There are two radio buttons visible to the right of the *Variable* box: one labeled *Data are continuous* (which is shown selected in figure 1.8) and one labeled *Data are discrete*. Radio buttons indicate mutually exclusive items—you can choose only one of them. Here we are treating *age* as if it were continuous, so make sure that the corresponding radio button is selected. On the right side of the *Main* tab is a section labeled *Y axis*. Click on the radio button for *Frequency* so that the histogram shows the frequency of each interval. In the section labeled *Bins*, check the box labeled *Width of bins* and type 2.5 in the textbox that becomes active (because the variable is *age*, the 2.5 indicates 2.5 years). Also check the box labeled *Lower limit of first bin* and type 45, which will be the smallest age represented by the bar on the left.

The dialog box shows a sequence of tabs just under its title bar, as shown in figure 1.9. Different categories of options will be grouped together, and you make a different set of options visible by clicking on each tab. The options you have set on the current tab will not be canceled by clicking on another tab.



Figure 1.9. The tabs on the histogram dialog box

Graphs are usually clearer when there is a title of some sort, so click on the *Titles* tab and add a title. Here we type *Age Distribution of Participants in Cancer Study* in the *Title* box. Let's add the text *Data: Sample cancer dataset* to the *Note* box so that we know which dataset we used for this graph. Your dialog box should look like figure 1.10.

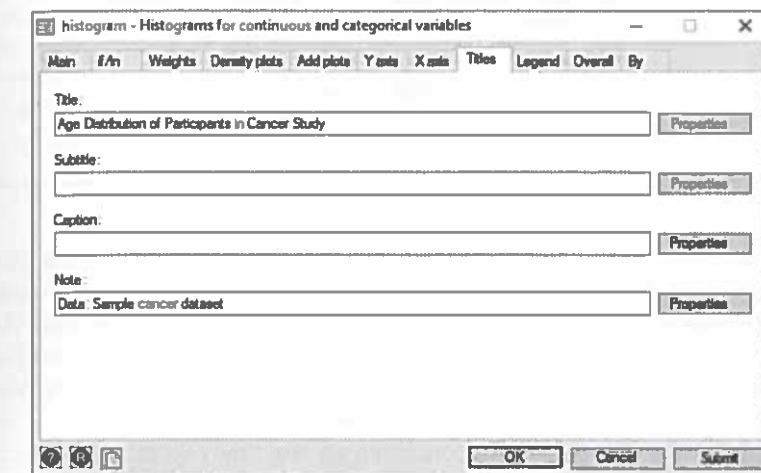


Figure 1.10. The Titles tab of the histogram dialog box

Now click on the *Overall* tab. Let's select *s1 monochrome* from the pulldown menu on the *Scheme* box. Schemes are basically templates that determine the standard attributes of a graph, such as colors, fonts, and size; which elements will be shown; and more.

From the *Legend* tab, under the *Legend behavior* section, click on the radio button for *Show legend*. Whether a legend will be displayed is determined by the scheme that is being used, and if we were to leave *Default* checked, our histogram might have a legend or it might not, depending on the scheme. Choosing *Show legend* or *Hide legend* overrides the scheme, and our selection will always be honored.

Now that we have made these changes, click on *Submit* instead of *OK* to generate the histogram shown in figure 1.11. The dialog box does not close. To close the dialog box, click on the X (close) button in the upper right corner, but we are not ready to do that yet.

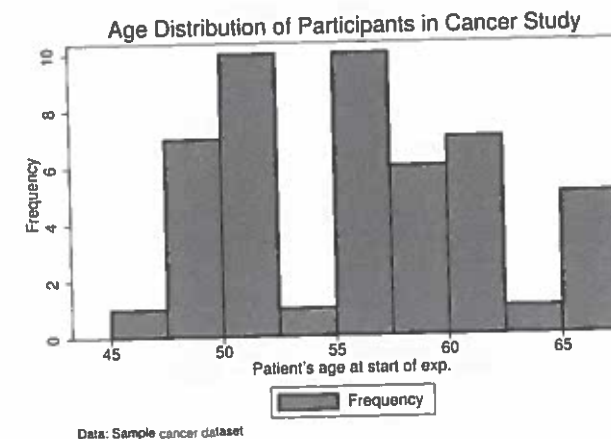


Figure 1.11. First attempt at an improved histogram

If you look at the complex command that the dialog box generated, you will see why even experienced Stata programmers will often rely on the dialog box to create **graph** commands. In reading this command, you will want to ignore the opening dot (Stata prints this in front of commands in the Results window, but the dot is not part of the command and you do not type it). Stata prints the > sign at the start of the second and third line, which might be confusing. Stata uses the Enter key to submit a command. Because of this, Stata sees the entire command as one line. To print the entire line in the confines of the Results window, Stata inserts the > for a line break. If you wanted to enter this command in the Command window, you would simply type the entire thing without the > and let Stata do the wrapping as needed in the Command window. Never press the Enter key until you have entered the entire command.

```
. histogram age, width(2.5) start(45) frequency
> title(Age Distribution of Participants in Cancer Study)
> note(Data: Sample cancer dataset) legend(on) scheme(simple)
(bin=9, start=45, width=2.5)
```

### Clearing the Results window: The cls command

As you run commands, the results are displayed in the Results window. There may be times when you want to clear the Results window, so that, for example, seeing the top of the results of a command is easier, especially if your commands and results are lengthy. Beginning with Stata 13, you can type the **cls** command (with no options) to clear the Results window.

It is much more convenient to use the dialog box to generate that command than to try to remember all its parts and the rules of their use. If you do want to enter a long command in the Command window, remember to type it as one line. Whenever you press Enter, Stata assumes that you have finished the command and are ready to submit it for processing.

### When to use Submit and when to use OK

Stata's dialogs give you two ways to run a command: by clicking on OK or by clicking on Submit. If you click on OK, Stata creates the command from your selections, runs the command, and closes the dialog box. This is just what you want for most tasks. At times, though, you know you will want to make minor adjustments to get things just right, so Stata provides the Submit button, which still runs the command but leaves the dialog open. This way, you can go back to the dialog box and make changes without having to reopen the dialog box.

The resulting histogram in figure 1.11 is an improvement, but we might want fewer bins. Here we are making small changes to a Stata command, then looking at the results, and then trying again. The Submit button is useful for this kind of interactive, iterative work. If the dialog box is hidden, we can use the Alt+Tab (Windows) or Cmd+Tab (Mac) key combination to move through Stata's windows until the one we want is on top again.

Instead of a width of 2.5 years, let's use 5 years, which is a more common way to group ages. If you clicked on OK instead of on Submit, you need to reopen the **histogram** dialog box as you did before. When you return to a dialog that you have already used in the current Stata session, the dialog box reappears with the last values still there. So all you need to do is change 2.5 to 5 in the *Width of bins* box on the Main tab and click on Submit. The result is shown in figure 1.12.



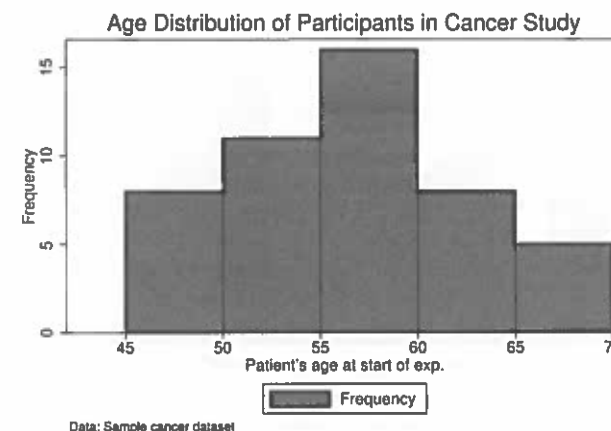


Figure 1.12. Final histogram of age

Notice how different the three graphs appear. You need to use judgment to pick the best combination and avoid using graphs that misrepresent the distribution. A good graph will give the reader a true picture of the distribution, but a poor graph may be quite deceptive. When people say that you can lie with statistics, they are often thinking about graphs that do not provide a fair picture of a distribution or a relationship. Can you think of any more improvements? The legend at the bottom center of the graph is unnecessary. You might want to go back to the dialog box, click on the Legend tab, and click on *Hide legend* to turn off the legend.

To finish our first Stata session, we need to close Stata. Do this with File > Exit. If you are using Stata for Mac, select Stata > Quit Stata.

## 1.6 Video aids to learning Stata

Many Stata users have put videos online that explain aspects of Stata. You can find these videos by searching YouTube or Googling, for example, “video tutorial Stata *t* test”. Stata has also produced an excellent set of tutorials, which you can find at <http://www.stata.com/links/video-tutorials/>. Most of these tutorials are specific and short. At this point, it would be good to view the tutorials listed under *Stata basics*. I will mention a few specific tutorials as we go through the book, but you should look for a tutorial whenever something is not clear or you want to go beyond what is covered in the text.

## 1.7 Summary

We covered the following topics in this chapter:

- The font and punctuation conventions I will use throughout the book
- The Stata interface and how you can customize it
- How to open a sample Stata dataset
- The parts of a dialog box and the use of the OK and Submit buttons
- How to summarize the variables
- How to create and modify a simple histogram

## 1.8 Exercises

Some of these exercises involve little or no writing; they are simply things you can do to make sure you understand the material. Other exercises may require a written answer.

1. You can copy and paste text to and from Stata as you wish. You should try highlighting some text in Stata's Results window, copying it to the Clipboard, and pasting it into another program, such as your word processor. To copy highlighted text, you can use the Edit > Copy menu or, as indicated on the menu, Ctrl+c. You will probably need to change the font to a monospaced font (for example, Courier), and you may need to reduce its font size (for example, to 9 point) after pasting it to prevent the lines from wrapping. You may wish to experiment with copying Stata output into your word processor now so that you know which font size and typeface work best. It may help to use a wider margin, such as 1 inch, on each side.
2. After you highlight material in the Results window, right-click on it. You can copy this output in several formats, including Copy, Copy Table (only works with some commands), Copy Table as HTML, and Copy as Picture (copies a graphic image of what you highlighted). The Copy option works nicely, but you will need to use a monospaced font, such as Courier, and may need to use a smaller font size when you paste it into your word processing document. The Copy Table option is limited because it only works with a few commands. The Copy Table as HTML option will create a table that looks like what you would see on a webpage. Using Microsoft Word, you can edit the table by making columns wider or narrower and by aligning the columns so that each number has the same number of decimal places. Just copy the tabular results and not the command when using the HTML option. The Copy as Picture option works nicely in Windows, but you cannot edit it in Word because it is a graphic image. In Word, you can resize the image.

Run the **summarize** command and copy the results to a Word document by using each of the options. Highlight the table. Right-click on it and then select the option you want. Switch to your word processor. Press Ctrl+v to paste what you copied. In your word processor, make the table as nice as you can by adjusting the font, font size, margins, etc.

3. Stata has posted all the datasets from its manuals that were used to illustrate how to do procedures. You can access the manual datasets from within Stata by going to the **File ▸ Example datasets...** menu, which will open a Viewer window. Click on *Stata 14 manual datasets* and then click on *User's Guide [U]*.

The Viewer window works much like a web browser, so you can click on any of the links in the list of datasets. Scroll down to chapter 25, and select the **use** link for *censusfv.dta*, which opens a dataset that is used for chapter 25 of the *User's Guide*. Run two commands, **describe** and **summarize**. What is the variable *divorcert* and what is the mean (average) divorce rate for the 50 states?

4. Open *cancer.dta*. Create histograms for *age* using bin widths of 1, 3, and 5. Use the right mouse button to copy each graph to the Clipboard, and then paste it into your word processor. Does the overall shape of the histogram change as the bins get wider? How?
5. UCLA has a Stata portal containing a lot of helpful material about Stata. You might want to browse the collection now just to get an idea of the topics covered there. The URL for the main UCLA Stata page is

<http://www.ats.ucla.edu/stat/stata/>

In particular, you might want to look at the links listed under Learning Stata. On the *Stata Starter Kit* page, you will find a link to *Class notes with movies*. These movies demonstrate using Stata's commands rather than the dialog box. The topics we will cover in the first few chapters of this book are also covered on the UCLA webpage using the commands. Each movie is about 25 minutes long. Some of these movies are for older versions of Stata, but they are still useful.