Data Analysis
Stata (version 14)

There are many options for learning Stata (www.Stata.com). Stata's help facility (accessed by a pull-down menu or by command or by clicking on ?) consists of help files for each command (works best if you know the name of the command). Stata installation includes Getting Started with Stata for Windows. It is well-written and a logical place to start learning Stata. A number of books and third party courses are also available. Acock AC. A Gentle Introduction to Stata, 5th Edition. College Station, TX: Stata Press, 2016 is a popular introductory book.

Online resources include:

Stata Tutorials: http://www.stata.com/links/video-tutorials/

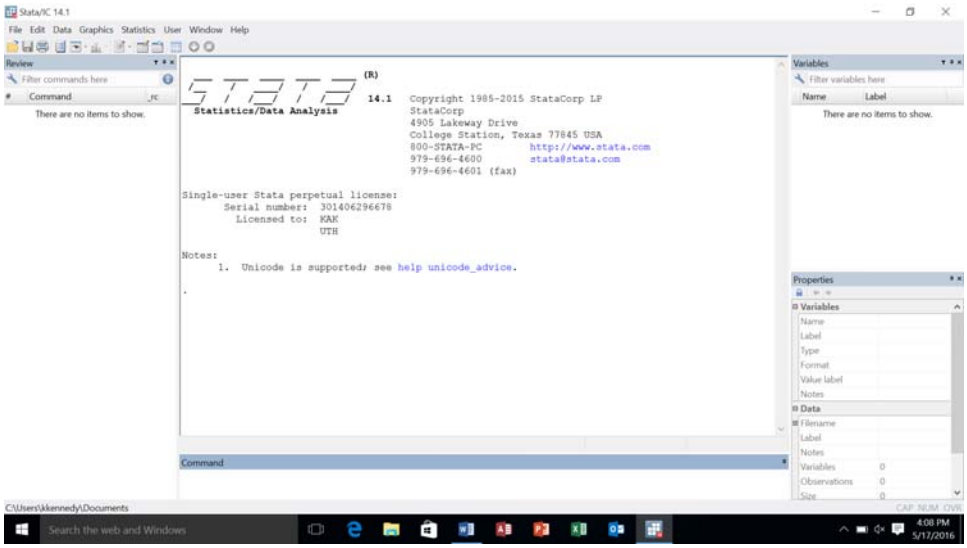
The UCLA Academic Technology Services: http://www.ats.ucla.edu/stat/stata

This exercise is designed as an introduction to navigating through the Stata software so that software hurdles will not get in the way of learning the concepts and their implementation at the beginning of the introductory biostatistics course.

Stata Overview

This exercise uses Intercooled Stata version 14 (StataIC 14) although all version 14 alternatives share the same interface. You may encounter differences if you are using other versions of Stata.

Stata is usually launched by clicking on the Stata icon (e.g., StataIC 14) on the desktop or from the Windows Start button (in the Start popup window or in the All Programs popup window).



At the top of the Stata interface is a row of eight pull-down menus (File, Edit, Data, Graphics, Statistics, User, Window, Help). A toolbar with icons (icons can be identified by hovering your mouse arrow over the icon) that are shortcuts to some of the actions executed from the pull-down menus (or by commands) is just below the menus. The default Stata user interface consists of five separate child windows (the Review, Variables, Results, Properties, and Command windows).

Stata is a command driven program. However, you can use the pull-down menus and associated dialog boxes to automatically generate most of the commands instead of entering them manually. In this exercise, we will use red font to indicate executing commands with the pull-down menus. The advantage of using the pull-down menus is that you don't have to remember the commands and required syntax. Manually writing the commands works better for some applications. The commands can be saved in executable files (*.do files) for future analyses and for documentation (not required for this exercise). That is extremely useful when conducting and revising analyses for publication.

When the pull-down menus are used to generate commands, the typed commands are entered automatically in the Review window as they are executed. That facilitates learning the commands. Stata commands and variable names are case sensitive. In this exercise, we will use green font to indicate the commands typed manually into the Command window.

Directly above the Command window is the default position for the Results window where the commands and the results (output) from executed Stata commands are displayed. On the left side of the Stata user interface is the default position for the Review window that lists all the commands executed in the current session. In the

Command window, commands can be edited if an error was made or you want to repeat the command on other variables. New variables can be inserted in the Command window by typing or by clicking on the variable name in the Variable window (on top of the Properties window to the right of the Results and Command windows).

The following exercise involves reading data into Stata, then manipulating and analyzing the data. This exercise can be done by experienced Stata users by simply executing the tasks (commands in red and green). Less experienced users may require the supporting information provided in this tutorial.

Description of the Data

The data for this exercise are from the Framingham Heart Study (www.framinghamheartstudy.org/about-fhs/index.php). These data are a subset of variables for men and women aged 60-62 years at study initiation. The variable names (bolded), codes, and definitions are listed below.

id	Case identification number
CHD	Coronary heart disease diagnosis
0	No evidence of CHD
1	Pre-existing CHD at study entry
SBP	Systolic blood pressure in mm Hg at study entry
DBP	Diastolic blood pressure in mm Hg at study entry
CHOL	Serum cholesterol in mg/100 ml at study entry
CIG	Number of cigarettes smoked per day (estimated to the nearest five)
death	
0	Alive at follow up
1	Dead at follow up
cause	Cause of death
alive	living at follow up
chd	coronary heart disease
othercvd	other cardiovascular disease
stroke	stroke
cancer	cancer
other	other cause of death
blank cell	missing data
gender	
0	Male
1	Female

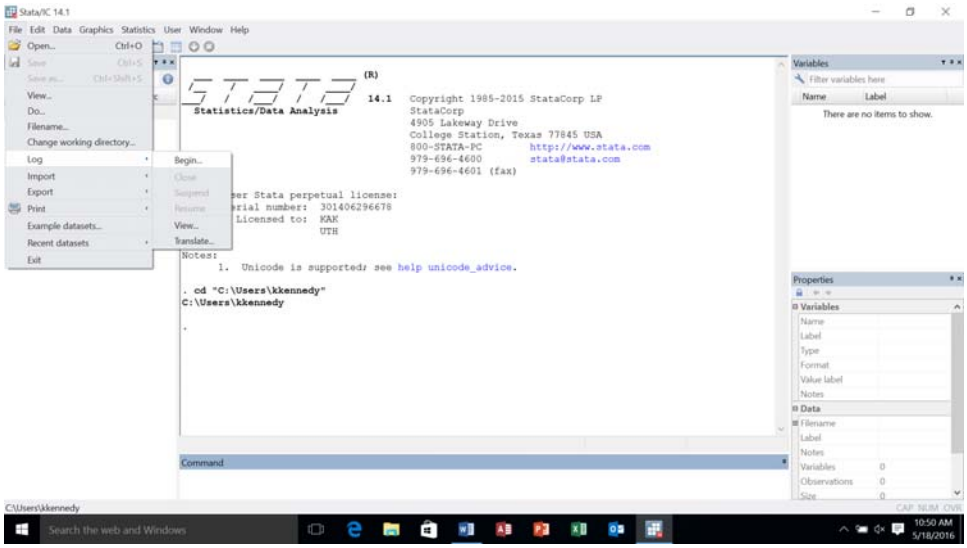
Creating a Log File

The Stata working directory is the default location that Stata will use for files and results unless it is changed. Click File—Change Working Directory and indicate the folder you want for your working directory on your computer.

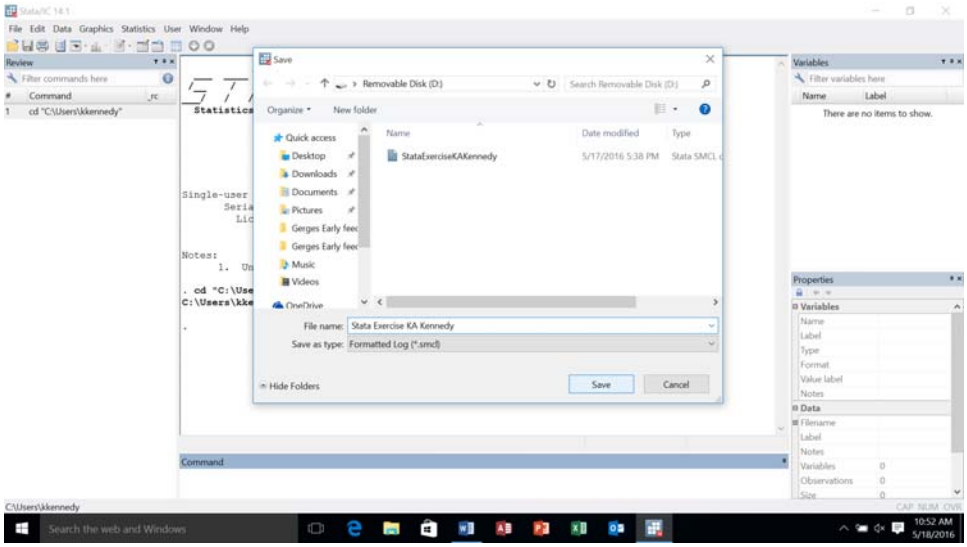
To have a record of your work for this exercise, you must save a log file. Log files can be stored in one of two formats: a text file with a .log extension or a Stata Markup and Control Language file with a .smcl extension. The *.log text files can be read and edited by any word processor program or text editor. The *.smcl files can only be read by the Stata program. *.smcl files can be translated to *.log files by clicking on File—Log—Translate.

TO GET CREDIT FOR THIS EXERCISE, YOU MUST SUBMIT A LOG FILE. NAME YOUR LOG FILE "*Stata Exercise Your Name".* YOU CAN SUBMIT EITHER AN *.smcl OR A *.log FILE.

File—Log—Begin (you can also click the Log button in the toolbar)



In the Save dialog box indicate your log's folder at the top, the log file name (File name:), and the format (Save as type:).



Click the Save button. The dialog box closes, the command you just executed is listed in the Review window and also in the Results window with some additional documentation, and log status (on) is indicated in the lower right corner of the Results window.

You can make comments (descriptive information that will not be executed by the program) in log files by typing your comment preceded by an * in the Stata Command window. You can print a log file by right clicking on the Viewer window with the open log file and selecting Print.

If you need to stop and resume this exercise, you can save the log file (it should also be automatically saved to your working directory when you exit the program). To add to your work later, re-enter Stata, click File—Log—Begin and insert the name of your saved log file in the dialog box. Then click on "Save" and select "Append to existing file" in the pop up dialog box to resume adding commands and results to your log file.

Keyboard Data Entry

Most biomedical datasets have rows containing the data from individual subjects/cases (or other analysis units) and columns containing different measurements or variables recorded on those cases. Data in this format are called wide format data. Long format is used for longitudinal and clustered data and repeated measurements on the same subject. This exercise contains data in wide format.

id	death	cause

1402	0	Alive
1403	0	Alive
1404	1	CHD
1405	1	CHD
1406	0	Alive

In Stata, data are held in a spreadsheet called the Data Editor. The Data Editor is entered by clicking Data—Data Editor—Data Editor (Edit). Use your keyboard to enter the data in rows 1-5 from the table directly into the Data Editor. (Variable names will be entered later). [Tab or the right-arrow key moves you to the right, Shift-Tab or the left-arrow key to the left, Enter or the down-arrow key down, and Shift-Enter or the up-arrow key up.] Stata automatically assigns a variable name to each column (var1, var2, etc) when data entered. Stata distinguishes text (eg, the observations for the cause variable) from numerical data by using reddish font for the former and black font for the latter.

Variable Names

When you select a column in the data table, information about the variable is shown in the Variables window to the right of the Results and Command windows. You can name or rename variables (var1, var2, etc.) here. Replace the default variable names with the appropriate names from the table above. Variable names should be short to be efficient and descriptive enough to reduce errors when manipulating the data, conducting analyses, and reading output. Stata variable names must start with a letter, they can contain letters (variable names are case sensitive), numbers, and the underscore but no spaces or other characters, and they can be up to 32 characters in length (eight or fewer is recommended).

Importing a Data File

The data in the above table was only part of the data needed for this exercise. The complete data concerning id, death, and cause for all patients have been entered into an Excel file named Data-1.xls. Before starting to enter the new dataset, clear the Data Editor by typing clear in the Command window. Import the Excel file, Data-1.xls. Click File—Import—Excel spreadsheet (*.xls, *.xlsx), in the Import Excel dialog box use the Browse button to select Data-1.xls where you stored it on your computer when you downloaded the file from our website. Click on Open. Click the Import first row as variable names button, click the OK button.

(Another simple way to import small Excel files is to copy/paste. Open the Excel file and select and copy all the cells. Paste what you have copied into the first/left cell in the first data row (below the first row with the variable names) of the Stata Data Editor. If you have variable names in the first row, select Variable Names in the dialog box that asks if you want to treat the first row as variable names).

Saving Data

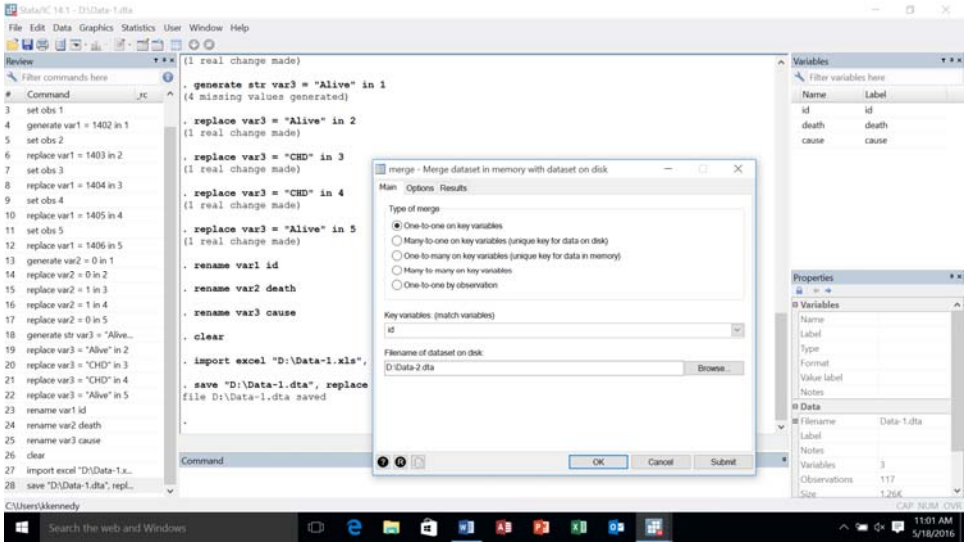
It is prudent to regularly save your work. Closing the Data Editor does not save your data. Save these data as a Stata data file (Data-1.dta):
Close the Stata Data Editor (click on X in upper right corner of the Data Editor Window).
File—Save As
The Save Stata Data File dialog box pops up.
At the top of the dialog box, locate the folder you want to store Data-1.dta if you don't want your data in your default folder.
Make sure Stata Data (*.dta) is selected in the Save as type: box.
Type Data-1 in the file name: box.
Click the Save button in the lower right corner of the dialog box.

You can open a previously generated and stored Stata data file (*.dta file) by clicking File—Open, or the Open button on the toolbar.

Merging Data Files

This exercise concerns merging two data files with different variables but the same cases in each dataset. Merge the data in Data-1.dta with the data in Data-2.dta that you downloaded to your computer from our website. One of these datasets must be currently open in Stata. In this example, Data-1.dta is currently open in Stata. The following steps merge the variables in Data-1.dta with the additional variables in Data-2.dta for the cases indicated by the key variable id.

Data—Combine datasets—Merge two datasets
The “Merge dataset in memory with dataset on disk” dialog box opens. The key variable is the variable in both files used to combine the two datasets (id).
Select One-to-one on key variables as the Type of merge.
Select id as the Key variables: (Match variables).
Browse to select Data-2.dta as the Filename of dataset on disk. Click on Open.
Click the Submit or OK button.



(You can execute commands from the pull-down windows by either clicking OK or Submit in opened dialog boxes. The latter leaves the dialog box open. It is used when you want to revise your command. OK executes the command and closes the dialog box.)

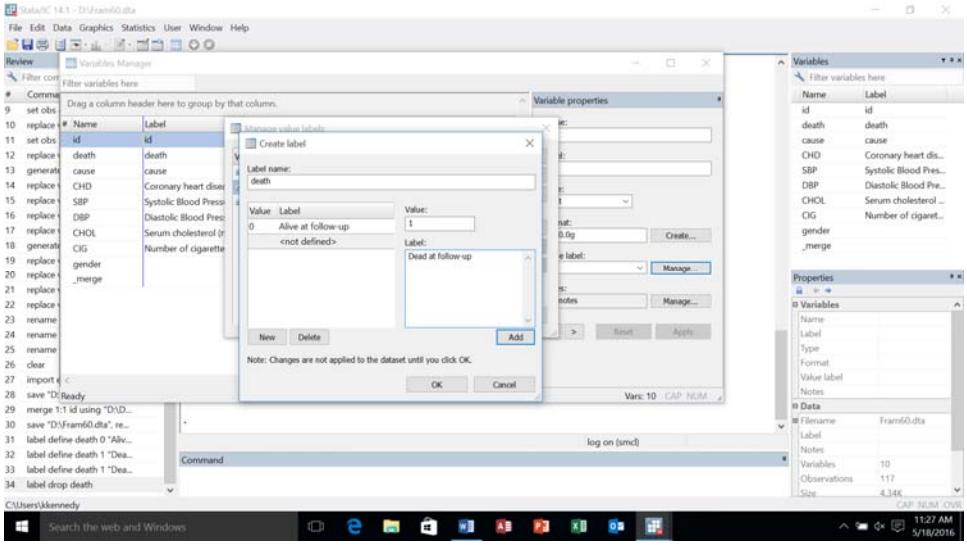
Save this merged dataset as Fram60.dta.
File—Save As
The Save Stata Data File dialog box pops up. At the top make sure you’ll save your data in the intended folder you created for this exercise.
In the Save as type: box, select Stata Data (*.dta).
In the File name: box, type Fram60.dta.
Click the Save button in the bottom right of the dialog box.

Variable Labels and Value Labels

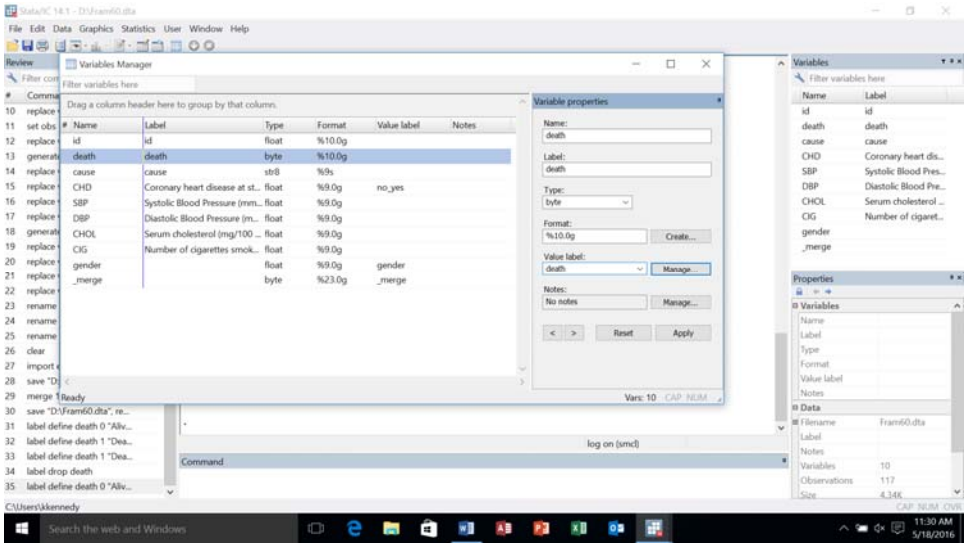
The next 2 sections are optional for this computer exercise. You will need to know this for the advanced biostatistics course and for analyzing your own data, but you won’t need this for the introductory course. You can proceed directly to the Data Exploration section (page 8) if you choose to skip this.

Variable names are often insufficient to allow the datasets to be understood easily by others (or by you if you go back to the data after a long break). Variable labels (short definitions of the variable) and value labels (how the categories are defined) provide additional information. Variable labels are used to more fully describe what the variables are and often include the units of measurement. They can be up to 80 characters in length; spaces or any characters can be used. Value labels are used when the numerical data entries have a text definition (eg, 0 and 1 for the death variable indicate Alive and Dead, respectively). Value labels are generated in two steps. First, the numeric values and the referent words are listed and saved [eg, 0 for no and 1 for yes (the same set could be used for more than 1 variable)] as Value Labels. Second, the Value Label list is assigned to the variable.

Click the pull down menu, Data— Variables Manager. In the Variables Manager dialog box, click the Manage button next to the Value Label fill-in box. In the Manage Value Labels dialog box, click the Create Label button and type a Label name for the value list you want to create (death). Then, in turn fill in each Value (number): and associated Label you want (Alive at follow-up for 0 and Dead at follow-up for 1) followed by clicking the Add button until all the values are labeled. Click the OK button and Close the Manage Value Labels dialog box (first step).



In the Variables Manager dialog box, select the death variable and type the newly created Value Label in the Value Label fill-in box and click the Apply button in order to map the named value labels to the variable (the second step).

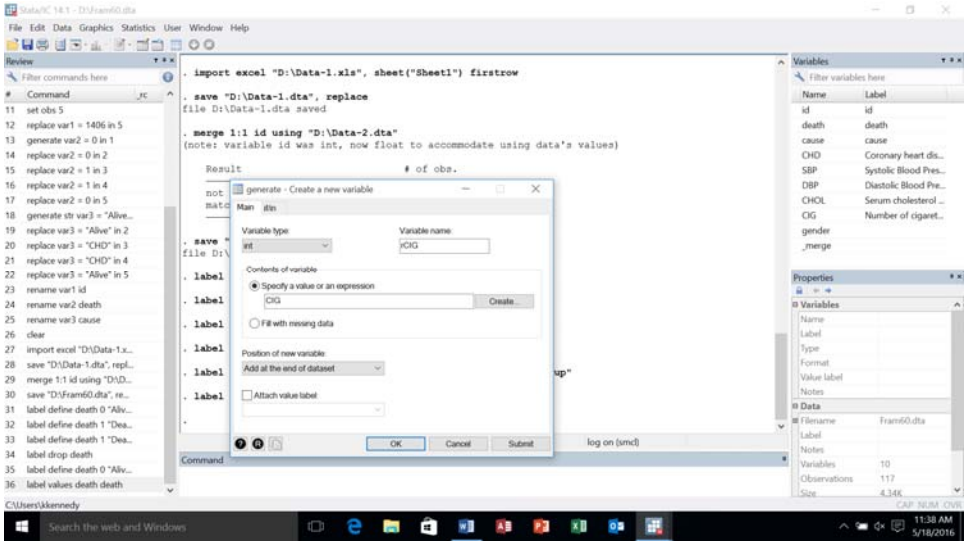


Close Variable Manager dialog box. Click on Apply to accept the changes you've made. (The Variables Manager is also used to delete variables. It can also be done directly from the spreadsheet view by selecting a column then right clicking and selecting Data-Drop Selected Data.)

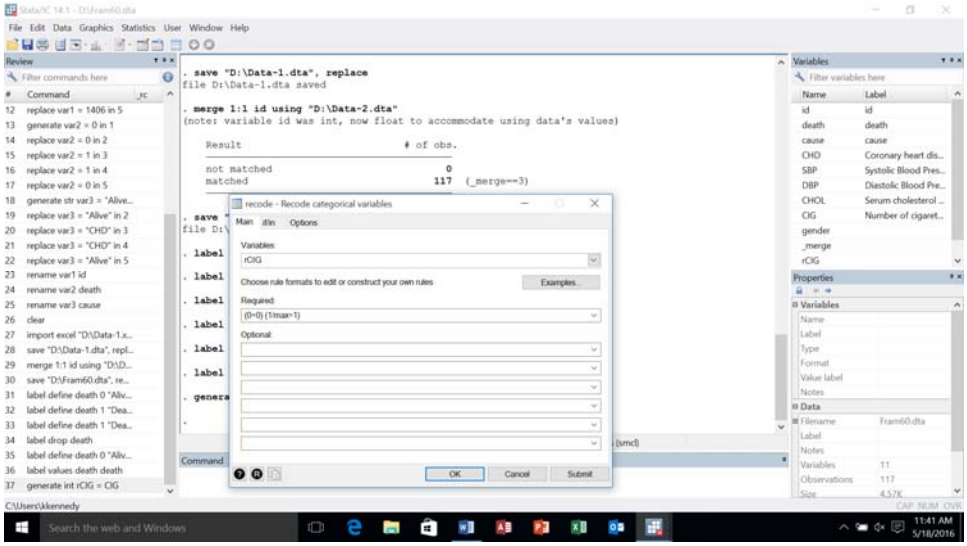
Generating New Variables and Recoding

You may need to recode variables when they are not categorized as needed for analysis. For example, CIG indicates the number of cigarettes smoked daily (rounded to the nearest 5). Recode CIG into two categories, 0 for nonsmoker and 1 for smoker, and label that new variable rCIG. We will first generate the new variable rCIG and then recode that variable.

Data—Create or change data—Create new variable
In the generate – Create a new variable dialog box type rCIG in the Variable name: box, int (for integer) as the Variable type:, CIG in the Specify a value or an expression box, and click Submit or OK.
This creates a new variable, rCIG from the data in CIG.



Data—Create or change data—Other variable-transformation commands—Recode categorical variable
In the recode – Recode categorical variables box select rCIG for the Variables:, type (0=0) (1/max=1) in the Required: box, and click Submit or OK. This stipulates that a 0 in rCIG will be recoded as 0 in rCIG and anything between 1 and the maximum value will be recoded as a 1. (Clicking the Examples button gives examples of acceptable rule formats.)



Recoding cause (of death) into two categories (1 for deaths from cardiovascular causes and 0 for cases still alive and all other causes of death) involves converting the text (string) variable to a numeric variable before combining categories. Encoding assigns numbers to a string (text) variable. This is necessary in order to perform statistical procedures on string (text) variables because Stata only performs statistics on numeric variables. This command assigns the string values as value labels to the numeric values. Stata displays them in the Data Editor as value labels (words) but computationally it considers them as numbers. Open the Data Editor and review the values in the spreadsheet for CHD and death. They appear to be words but note they are blue font rather than reddish font reserved for "true" string variable values.

The cause column is displayed in a reddish font because it is a string variable that has not been encoded. Encode this variable then use the recode command to generate the two desired codes for cardiovascular death (1) and alive or non-cardiac death (0).

Data—Create or change data—Other variable-transformation commands—Encode value labels from a string variable
Select cause as the Source string variable:
Enter cvdeath (new variable that will be encoded) under New numeric variable:
Click OK or Submit

To see the numbers Stata has assigned to the cvdeath variables, use the following command after closing the Data Editor (if still open).

Data – Data utilities – Label utilities -- List value labels
Enter cvdeath under Labels to list:
Click on Submit or OK. Stata displays the list of values for the cvdeath variable.

Data—Create or change data—Other variable-transformation commands—Recode categorical variables
In the recode – Recode categorical variables box select cvdeath for the Variables:, type (1/2=0) (3=1) (4=0) (5/6=1) in the Required: box, from the Options tab check Generate new variable, enter cvdeathyn, and click Submit or OK.
This creates a new variable (cvdeathyn) with a value of 0 for cases with 1 or 2 or 4 for cvdeath and a value of 1 for cases with 3 or 5 or 6 for cvdeath.
New value labels (No for 0, Yes for 1) would need to be created for this recoded variable because the value labels before the recode no longer apply to the recoded variable.

Create a new variable that is the difference between systolic and diastolic blood pressure (BPdiff=SBP-DBP).

Data—Create or change data—Create new variable

In the generate – Create a new variable dialog box type BPdiff in the Variable name box and SBP-DBP in the Specify a value or an expression box, float (for floating decimal point) as the Variable type:, and click Submit or OK.

When you recode your data or generate new variables you should verify that your modifications are as intended. One way to verify your variables is to list your new variables and any original variables used to calculate the new variables.

Data—Describe data—List data
In the Main tab of the list – List values of variables dialog box, select id SBP DBP BPdiff in the Variables: box and click Submit or OK.

Data Exploration

You should explore your data before analyzing to verify the accuracy of the data and to determine how the data are distributed. Erroneous or problematic data values need to be identified and corrected. The data in memory command characterizes the formatting of the data as well as other information about the variables (labels and value labels). The codebook command examines the data distributions. Review the following output to make sure that your data seems complete and the variables are correctly specified.

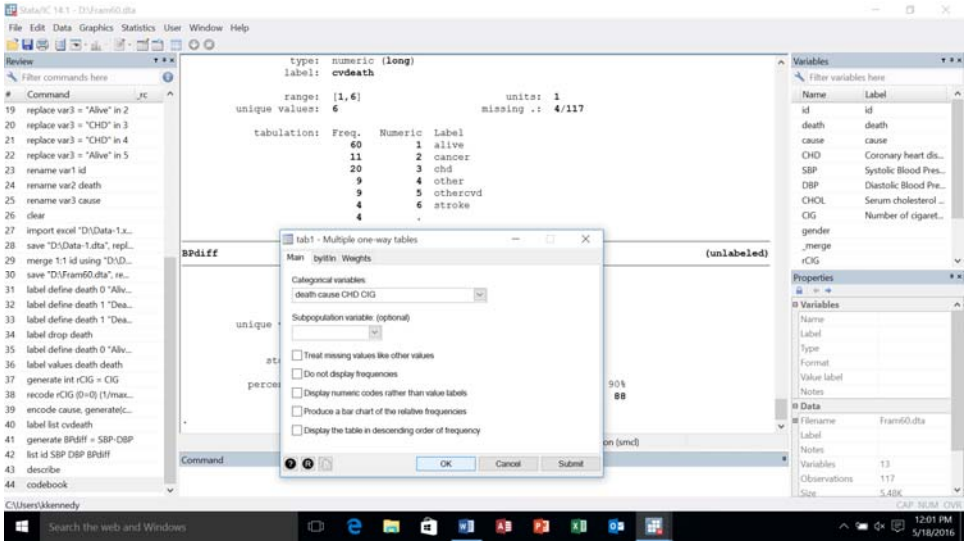
Data—Describe data—Describe data in memory or in a file
Select “In Memory” and click OK
Data—Describe data—Describe data contents (codebook)
Leave Variables: empty and click OK
Note that you may need to click on the –-more–- at the bottom left of the Results screen (or hit the Space bar) to see all of the output. You can type set more off <Enter> in the Command window to disable the more function (may already be set as the default with your installation).

a. Categorical variables

Some of the variables are categorical indicating the group/category to which a case belongs. Determine the frequencies of values for the categorical variables death, cause, CHD, and CIG. You have already generated that information by the codebook command above. (Scroll up in the Results pane to view what was generated for each of these categorical variables.) Now, you will generate frequency counts for the values for each variable.

Statistics—Summaries, tables, tests—Frequency Tables—Multiple one-way tables

Select the categorical variables (click on death, cause, CHD, and CIG) you want under Categorical variables in the dialog box. Click on Submit or OK.



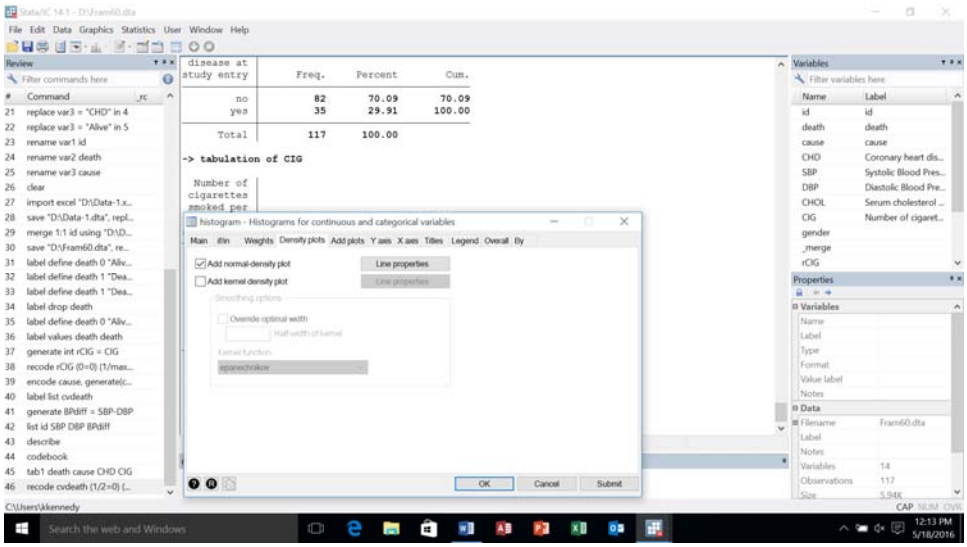
Review some of the options you can check in the Frequency Tables dialog boxes to customize the results you obtain. Review the output in the Results window. Use the window controls on the right of the Review window to examine earlier output. The Results window only stores the most recent output. You can open your log file in the Viewer window to view all your output since beginning the log file.
Window—Viewer—New Viewer
File—Open, then Browse to find your log file.

b. Continuous variables

Some of the variables are continuous (they measure something like blood pressure). Explore the distribution of values for the variable CHOL by generating a histogram.

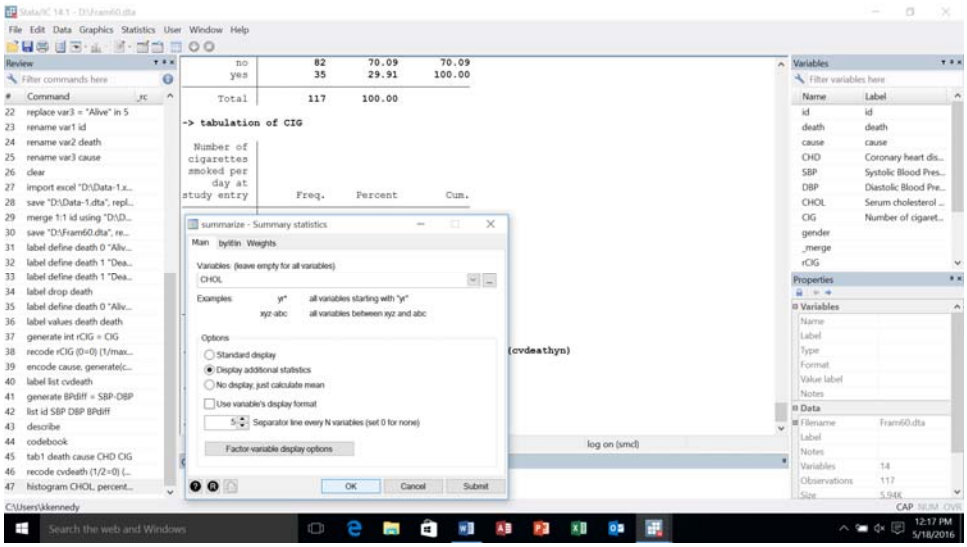
Graphics—Histogram

In the Main tab of the histogram – Histograms for continuous and categorical variables dialog box, in the Data section, select CHOL as the variable and Data are continuous, then select Percent in the Y-axis section. Include a line on your histogram indicating the expected normal distribution for a distribution of the same mean and SD as the recorded CHOL variable. In the Density plots tab of the histogram – Histograms for continuous and categorical variables dialog box select Add normal density plot. Click the Submit or OK button.



Descriptive statistics are summary measures of data that describe the data. For CHOL, calculate the mean, the median, the standard deviation, the 25th percentile (1st quartile), the 75th (3rd quartile), and the smallest (minimum) and largest (maximum) values.

Statistics—Summaries, tables, tests—Summary and descriptive statistics—Summary statistics
In the dialog box, select CHOL, Display additional statistics, and click the Submit or OK button.



Sorting

With large data sets, it can be difficult to find entries that have been identified as erroneous of missing in the data summary. In the Data Editor (click on Data—Data Editor—Data Editor (Edit)) finding extreme values can be facilitated by sorting all the values of the variable in question. Use the Data—Sort drop down menu to sort the variable name in question (eg, CIG). Note that Stata (like other database programs such as Access, but not necessarily like spreadsheet programs such as Excel), sorts the entire row of data not just the selected column. Close the Data Editor.

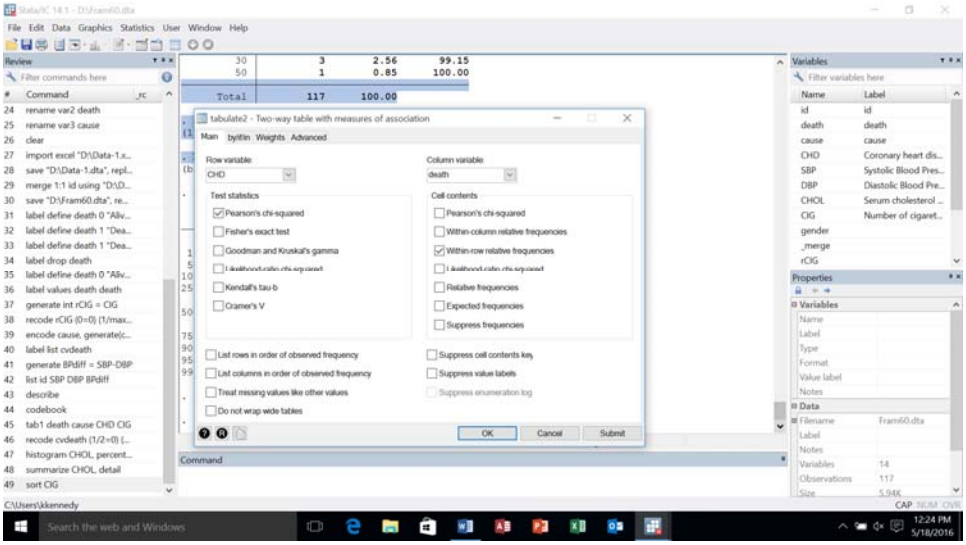
Inferential Statistics

The primary reason for using a statistical package is to compute inferential statistics (eg, chi-squared test or t-test). In this exercise, you will perform these two common statistical tests to test the following two hypotheses:

Hypothesis 1: 60-62-year-olds with CHD are more likely to die before they are 78-80 (end of follow-up) than 60-62-year-olds without CHD.

Data to test that hypothesis can be presented in a 2 X 2 table (cases with CHD as one row and cases without CHD as the other row; dead in one column and alive in the other column). Generate the 2 X 2 table with CHD as the rows and death as the columns and calculate a chi-squared test statistic (more precisely, the Pearson's chi-squared test statistic) to test Hypothesis 1. For cases with and without CHD, indicate the percentages dead and alive.

Statistics—Summaries, tables, and tests—Frequency tables—Two-way table with measures of association
In the Main tab of the tabulate2 – Two-way tables dialog box select CHD for the Row variable: and death as the Column variable:, check Pearson's chi-squared under Test statistics and Within-row relative frequencies under Cell contents, and click Submit or OK.



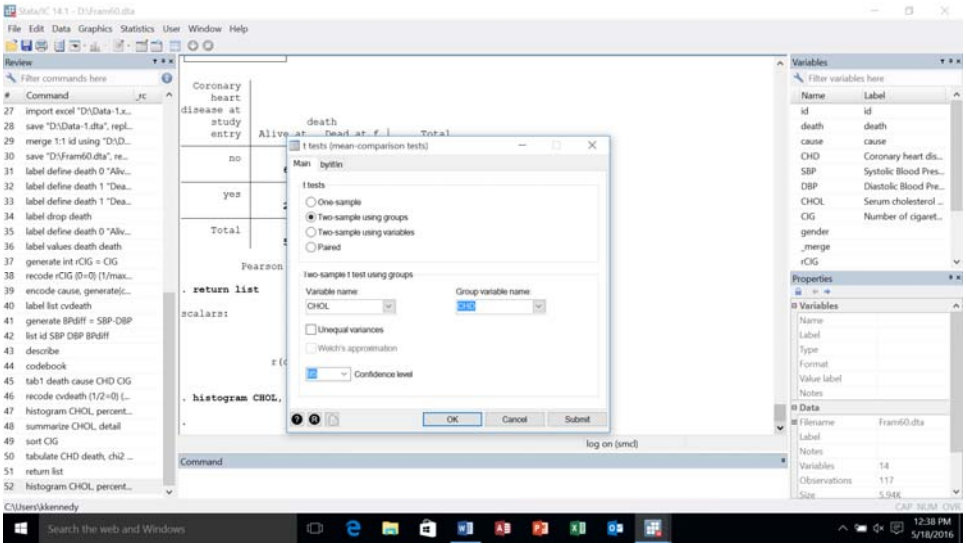
The results are listed in the Stata Results window and the log file. The chi-square is 16.1504 with 1 degree of freedom. The significance level is listed as 0.000. The Null hypothesis is rejected; 60-62-year-olds with CHD are more likely to die than 60-62-year-olds without CHD. To display the p value with greater precision, type return list in the Command window. It is displayed next to r(p) =.

Hypothesis 2: 60-62 year-olds with CHD have higher serum cholesterol levels than 60-62 year-olds without CHD.

A statistical test of this hypothesis is a t-test. The t-test assumes your data are normally distributed and the variances (dispersion of the data) are equal in the two groups. It is important to look at your data before conducting a statistical test. Before calculating a t-test, plot the distributions of CHOL values for the two groups of cases, those with CHD and those without CHD.

Graphics—Histogram
In the Main tab of the histogram – Histograms for continuous and categorical variables dialog box, in the Data section, select CHOL as the variable and Data are continuous, then select Percent in the Y-axis section (should be saved from when you did this before). In the By tab of the dialog box, check Draw subgraphs for unique values of variables, select CHD for Variables, and click Submit or OK. You should get a histogram of the distribution of CHOL with separate graphs for CHD = yes and CHD = no.

Statistics—Summaries, tables, and tests—Classical tests of hypotheses—t test (mean-comparison tests)
In the Main tab of the t test—(mean comparison tests) dialog box, select Two-sample using groups (because the CHOL data are all in one column), then select CHOL as the Variable name: and CHD as the Group variable name:, click Submit or OK.



The t-test (t= -1.3239) with 115 degrees of freedom and two-tailed significance level = 0.1882 is consistent with accepting the Null hypothesis.

Subgroup Analyses

Test Hypothesis 2 separately for women and men.

Graphics—Histogram
In the Main tab of the histogram – Histograms for continuous and categorical variables dialog box, in the Data section, select CHOL as the variable and Data are continuous, then select Percent in the Y-axis section (should be saved from when you did this before). In the By tab of the histogram – Histograms for continuous and categorical variables dialog box, check Draw subgraphs for unique values of variables, and this time select gender CHD for Variables, and click Submit or OK.

Statistics—Summaries, tables, and tests—Classical tests of hypotheses—t test (mean-comparison tests)
In the Main tab of the t-tests (mean comparison tests) dialog box, select Two-sample using groups, then select CHOL as the Variable name: and CHD as the Group variable name:. In the by/if/in tab of the t-tests—(mean comparison tests) test dialog box, check Repeat command by groups, and select gender for the Variables that define groups:, and click Submit or OK.

For gender =male, the t-test (t= -0.7962) with 60 degrees of freedom and two-tailed significance level = 0.4291 is consistent with accepting the Null hypothesis. In contrast, for gender =female, the t-test (t= -2.0226) with 53 degrees of freedom and two-tailed significance level = 0.0482 is consistent with rejecting the Null hypothesis at the 0.05 level of significance.

Congratulations! You're finished with the tasks in the exercise.

When you exit Stata, your log file will be automatically closed and saved (you can also close it manually, File—Log—Close). The data file (changes you made) will not be saved automatically; you must select Save in the dialog box when you exit the program. You can edit a *.log file using Microsoft Notepad, Word, or other programs. You can delete unintended commands and results and add comments (put an * before the comment) to make your file more readable. If you created your log file as an *.scml file, you will need to translate it (File – Log – Translate) to a *.log file before you can edit it.

Email your log file (*.log or *.scml) to Deborah.Garcia@uth.tmc.edu to receive credit for this exercise.