

Data Management, Manipulation and Exploration using Stata.

For instructional purposes, we will use .csv dataset (named BP_Example) from Prof. Cristina's handout and Daniel's book.

Data dictionary:

PATIENT = Patient ID Number

Y = Mean arterial blood pressure (mm Hg)

X₁ = Age (years)

X₂ = Weight (kg)

X₃ = Body surface area (m²)

X₄ = Duration of hypertension (years)

X₅ = Basal pulse (beats/min)

X₆ = Measure of stress

X₇ = Sex

X₈ = Diabetes Mellitus status

Data Exploration/management

- Log

Helps you to save everything that is displayed on the Stata window.

To create a log, go to: **Menu:** File>Log>Begin

To close it, go to: **Menu:** File>Log>Close or use command line by typing: log close

- Do file

A more useful way to record your successful commands.

To create a do file, go to: **Menu:** Window>Do-File Editor>New Do-file Editor

After creating a do file, go to the Command window and copy your successful commands and paste them into the do-file editor. Save the file and continue adding to it. You can add comments to it by typing `*comments`.

- Opening a dataset

There are variant ways of opening datasets in Stata depending on the dataset extension (e.g. .csv, .txt, etc).

a. **Menu:** File>Import>Text data (delimited, *.csv, ...) Select the desired file and options, and click OK.

b. **Command:** insheet using "C:\Users\user\Google Drive\Summer\Summer 2018\BIOM 559\Stata Workshop\BP_Example.csv" (note: this command does not work for Excel—you must use the import statement)

Note: Any dataset with .dta extension is a Stata dataset and you can open it directly from:

Menu: File>Open>

- View a data set

To view a dataset on Stata window, go to:

Menu: Data>Data Editor>Data Editor (Browse) or type **browse** on the command line.

- Edit a data set

To edit a dataset, go to:

Menu: Data>Data Editor>Data Editor (Edit) or type **edit** on the command line.

- Rename variables

Sometimes may like to give variables different names that are informative to you and your readers. There are many ways you can achieve this. We will show two ways of doing this. **Show these using BP_Example dataset.**

Change the variable names and labels.

- *Variable labels – Stata use this on the plotting of the datasets.*

- a. Renaming at once.

Menu (in the Data Editor or Main window): Data>Variables Manager

On the right-hand side, under **Name**, enter the name to which you want to refer to the variable and under **label**, enter the full descriptive label you wish to give variable. Then click “Apply”. Go to the next variable do the same until you are done. Close the window after you are done.

- b. Open the lock and rename directly on the “Properties” at the right side of the Main window. Click on the lock to make sure it is unlocked. Now you can change the variable name and its label. Again, make sure you are in edit mode for this to work (type “edit” into the command line). Or use command line.

Command: `rename y BP`
`label variable BP "Mean Arterial Blood Pressure (mm Hg) "`
etc...

- Describe a dataset

Knowing whether a variable in a dataset is qualitative (character) or quantitative (numeric), etc is crucial. You can easily see this from Stata outputs since it color-codes the data in spreadsheet. Numeric variables will appear black while character variables will appear red.

Menu: Data>Describe data>Describe data in memory or in a file

Command: `describe`

This describes all the variables in the data set. If you want to describe a particular variable, type its name after “describe” or select it from the drop-down menu.

- Summarize a dataset

This helps you know some basic statistical summaries of your dataset (such as mean, standard deviation, etc.).

- a. *numerical dataset:*

Menu: Statistics>Summaries, tables, and tests>Summary and descriptive statistics>Summary statistics

Command: `summarize`

To obtain additional statistics, on the menu select the button “Display additional statistics”. On the command line, after summaries, type “, detail”.

- b. *For categorical data:*

Menu: Statistics> Summaries, tables, and tests>Frequency tables>One-way table

Select your categorical variable (for this example, select sex) and click OK.

Command: `tabulate sex`

This will give the frequency (count), the percentage of the total, and the cumulative percentage.

- Missing data

If your dataset is large, it would be hard to pinpoint exactly what data is missing. However, Stata has a way for you to pinpoint what is missing.

Menu (in the Editor window): Data>Describe data>List data

Suppose you want to know which patient(s) are missing observations for variables. On the tab that opens, type “patient” into Variables. Click on the tab “by/if/in” and under “Restrict observations”, click in the box under “if: (expression)” and type missing(variable) where variable is whichever variable you think has missing values. Click OK. For example, type “missing(age)”.

Command: `list patient if missing(age)`

- Filter data

Suppose you want to view data that meet certain criteria, you can use the filter button on the editor (looks like a funnel).

Menu (in the Editor window): Click on the funnel shape icon.

In the box under “Filter by expression”, type in filtering criteria. For instance, you want to view all the patients with blood pressure measurement greater than or equal to 120. Type “BP >= 120”. Note: this does not permanently alter your data set in any way. You can simply type “browse” in the command line to bring back the entire data set or click on the “Remove filter” in the Filter dialogue box.

Command: `browse if BP>=120`

To display all the data again, type: `browse`

- Summary tables by groups

Sometimes, you want to summarize data by a grouping variable. For example, you may want to know mean blood pressure by sex, or we may want to find out who is positive for diabetes mellitus by sex.

a. Summarize a numerical variable by a group:

Menu: Statistics>Summaries, tables, and tests>Summary and descriptive statistics>Summary statistics
Enter the variable that you wish to summarize under “Variables” in the dialog box under “Main”. Go to “by/if/in” tab, check the box next to “Repeat command by groups” and enter the grouping variable (say, sex). For example, if we want to find the mean and standard deviations of the blood pressure by gender, on the “Main” tab, enter “bp” under “Variables” and on the “by/if/in” tab, enter “sex” for the grouping variable.

Command: `by sex, sort : summarize BP`

b. Summarize a categorical variable by a group:

Menu: Statistics>Summaries, tables, and tests>Frequency tables>Two-way table with measures of association

Under “Main”, select the variable that you want to summarize from the drop-down menu under “Row variable”. Select the second variable from the drop-down menu under “Column variable”. Check the box for “Relative frequencies” to display the cell percentages (otherwise you are just given the frequencies (counts)). For example, if you want to determine the number of males and females who have diabetes mellitus, under “Row variable” select “Sex” and under “Column variable”, select “dmstatus”.

Command: `tabulate sex dmstatus, cell`

Note: You can also summarize only particular groups, such as just the females, or just those with a diabetes status = 1 by clicking on the “by/if/in” tab and entering “dmstatus==1” under the “If: (expression)” statement.

- **Creating a new variable**

Sometimes, we may wish to create a new variable for some reasons.

For example, let us create a new variable called “dm2” that has “yes” or “no” instead of “1” and “0”, respectively, for dmstatus. This should be a character variable (choose “Str” for variable type), and the new name should be “DM2

Command: generate DM2=“no” if dmstatus==0
replace DM2=“yes” if dmstatus==1

- **Creating a new variable (Using mathematical operators to create new variables)**

Command: generate wtlb = weight*2.2

Graphics

- **Boxplot**

Menu: Graphics>boxplot

Under the “Main” tab, select the orientation (vertical is commonly used) and your variable of interest. Click on the “Categories” tab and check the box that says “Group 1”. Select your grouping variable and then click “OK”. For example, choose your variable of interest to be “BP”. Choose your grouping variable to be “sex”. Click “OK”. The boxplot may take a few seconds to generate.

Command: graph box BP, over(sex)

Note: You can edit the plot directly by right clicking on the graph and selecting “Start Graph Editor”. Once the editor is started, you can double click on most any part of the figure and edit it, such as adding titles, axis labels, etc.

- **Scatter Plot**

Menu: Graphics>Twoway graph (scatter, line, etc.)

Click on “Create...” A second window will open. Choose “Basic Plots” and “Scatter” (both the default choices). Under “Y Variable” choose the variable you want to appear on the Y-axis, and under “X Variable”, choose the variable you want to appear on the X axis. For example, if you are interested in whether weight and blood pressure have a linear relationship, choose “BP” for the Y variable and “weight” for the X variable. Click “Accept” and then click “OK” again on the original plot window.

Command: twoway (scatter BP weight)