Introduction to Stata
2017-18

03.
Data Description Basics

Please note I do a lot of comments!
You will see many of my commands that begin with an asterisk. I've put some of these in green (but not all) so that they are easier to see. Commands in STATA that begin with an asterisk (*) are comments. While recommended, you don't actually have to type these comments.

1. Describe Your Data - Numerical Descriptions

YOUR TURN
The following is a session that you can duplicate on your own. Tip - Commands that begin with an asterisk (*) are comments. The highlights in blue are my doing not Stata.

```
. * Import data set from the internet - Use the command use "fullrlpath" remembering quotes.
. use "http://www.pauldickman.com/survival/ivf.dta", clear

. * Reorder the variable so that they are in alphabetic order - Use the command aorder
. aorder


. * Descriptives for dicrete variable - Use the command tab1
. tab1 sex, missing

. * Descriptives for discrete variables with display of labels & missing values - numlabel, add
. numlabel, add
. tab1 sex, missing


. * Descriptives for continuous variable - Use either tabstat or summarize
. summarize bweight
. summarize bweight, detail
. tabstat bweight, stat(n, mean sd sem min q max cv) missing


. * Descriptives for TWO dicrete variables - Use the command tab2
. * tab2 rowvariable columnvariable, options
. tab2 sex hyp
. tab2 sex hyp, row
. tab2 sex hyp, row column cell exact chi2

. * Descriptives for ONE continuous variable, by groups defined by a dicrete variable
. * Must sort by the discrete variable first
. sort sex
. tabstat bweight, by(sex) col(stat) stat(n mean sd sem min q max)

. * use option LONGSTUB if you want to be reminded of the variable you are summarizing
. tabstat bweight, by(sex) col(stat) stat(n mean sd sem min q max) longstub
```

2. Describe Your Data – Graphical Descriptions

Note – For each graph, I have shown you 2 examples. The first is the simple and very basic graph. The second is the same graph with various aesthetics added (for example – titles, x and y-axis tick marks, etc)

```
. * Tell Stata what graph scheme you want to use. I like the scheme s1color
. set scheme s1color

. ******   Bar Graphs  ******
. *  BAR GRAPH, simple - Use the command histogram with the option discrete
. histogram hyp, discrete

. *  BAR GRAPH, fancy - Use the command histogram with the option discrete
. histogram hyp, discrete percent addlabels ylabel(0(20)100) xlabel(0 "Normotensive" 1
"Hypertensive") gap(50) title("Bar Graph -Hyp") subtitle("n=639") caption("hyp_barchart.png")


. ******   Dot Plots  ******

. *  DOT PLOT, simple - Use the command dotplot
. dotplot matage

. *  DOT PLOT, fancy – Use the command dotplot
. dotplot matage, center msize(vsmall) xlabel(1 "All") title("Distribution of Maternal Age")
subtitle("n=641") caption("dot_matage.png", size(vsmall))


. * DOT PLOT for more than 1 group, simple – Use the command dot plot with option over( ).
. * Must sort first.
. sort sex
. dotplot matage, over(sex)


. * DOT PLOT for more than 1 group, fancy.
. sort sex
. dotplot matage, over(sex) title("Distribution of Maternal Age") subtitle("by infant sex")
caption("dot2_matage.png", size(vsmall))


. ******  Box Plots  ******

. * BOX PLOT for more than 1 group, simple – Use the command graph box with option over( ).
. sort sex
. graph box matage, over(sex)

. * BOX PLOT for more than 1 group, fancy.
. sort sex
. graph box matage, over(sex) title("Distribution of Maternal Age") subtitle("by infant sex")
caption("box2_matage.png", size(vsmall))
```

```
.* HORIZONTAL BOX PLOT for more than 1 group, simple - Use the command graph hbox
. graph hbox matage, over(sex)


.* HORIZONTAL BOX PLOT for more than 1 group, fancy
. graph hbox matage, over(sex) title("Distribution of Maternal Age") subtitle("by infant sex")
caption("hbox2_matage.png", size(vsmall))


.  ******  Histograms  ******
. * HISTOGRAM - Preliminary:  It's a good idea to obtain min and max
. tabstat matage, stat(min max)

. * HISTOGRAM, simple - Use the command histogram
. histogram matage

. * HISTOGRAM, fancy - Use the command histogram
. histogram matage, width(5) start(20) percent ylabel(0(10)50) addlabels title("Distribution of
Maternal Age") subtitle("n=641") caption("histogram_matage.png", size(vsmall))


. ******  Scatterplots  ******

. * X-Y SCATTERPLOT -  Preliminary:  Obtain min and max of the X and Y variables
. tabstat matage gestwks, stat(min max)

.* X-Y SCATTERPLOT, simple  - Use the command graph twoway (scatter yvariable xvariable)
. graph twoway (scatter gestwks matage)

.* X-Y SCATTERPLOT, fancy
. graph twoway (scatter gestwks matage, msymbol(d) msize(vsmall)), xlabel(20(5)45)
ylabel(20(5)45) title("Scatterplot") ytitle("Weeks Gestation",size(small))
caption("scatter.png", size(vsmall))


.* X-Y SCATTERPLOT WITH OVERLAY LINEAR FIT, simple: graph twoway (lfit yvariable xvariable)
. graph twoway (scatter gestwks matage) lfit gestwks matage)

.* X-Y SCATTERPLOT WITH OVERLAY LINEAR FIT, fancy
. graph twoway (scatter gestwks matage, msymbol(d) msize(vsmall)) (lfit gestwks matage),
xlabel(20(5)45) ylabel(20(5)45) legend(off) title("Scatterplot") subtitle("with overlay linear
fit") ytitle("Weeks Gestation",size(small)) caption("lfit.png", size(vsmall))


. * X_Y SCATTERPLOT W OVERLAY FIT & 95% CI, simple: graph twoway (lfitci yvariable xvariable)
. graph twoway (scatter gestwks matage) (lfitci gestwks matage)


. * X_Y SCATTERPLOT WITH OVERLAY FIT AND 95% CI, fancy
. graph twoway (scatter gestwks matage, msymbol(d) msize(vsmall)) (lfitci gestwks matage),
xlabel(20(5)45) ylabel(20(5)45) legend(off) title("Scatterplot") subtitle("with overlay linear
fit and 95% CI") ytitle("Weeks Gestation",size(small)) caption("lfitci.png", size(vsmall))
```