

Stata Lab 2

Remember: If you are not using Stata on your own machine, first you need to log in to the Virtual Computing Lab (VCL) at <https://vcl.unc.edu>, choose Shibboleth (UNC-Chapel Hill) and proceed to login, click on New Reservation, choose 'Stata 12, WinXP (vmware)' from the pull-down list of environments, and click Create Reservation, click connect (and then get RDP file if needed).

Additionally, all datasets are on Sakai under Resources->Labs->Week of August 27.

- 1.) Open a log file and name it.
- 2.) Open the data file `unicef.dta`. The percentages of low birth weight infants in various countries around the world are contained in this dataset. The measurements themselves are saved under the variable name `lowbwt`. The variables `life60` and `life92` give the life expectancy for someone born in 1960 and 1992, respectively.
- 3.) Use the data browser to take a look at the data.
- 4.) To get a better look at the data enter **describe** in the command window. To only look at particular variables you can type, for example, **d lowbwt**
 - a.) How many variables are in this data set?
 - b.) Which variables nominal, ordinal, or continuous?
 - c.) How many observations are contained in the data set?
- 5.) Construct a boxplot and then a histogram of the variable `lowbwt`. The commands for this are from the first lab. Note: if you would like to see both graphs at once, name them. For example type:
graph box lowbwt, name(box1, replace)
The first graph will remain when you type the command for the second graph. If the name for the histogram is `hist1`, you can see the graphs side-by-side if you type:
graph combine box1 hist1, cols(2)
 - a.) Do the data appear to be skewed? If so, are they skewed right or left?
 - b.) Do the data contain any outlying observations of percentage of low birth weight? If so, to which countries do they belong? To find these countries it will be helpful to sort the data:
gsort lowbwt

6.) Calculate the mean, median and standard deviation of the percentage of low birth weight.

Try the following commands:

mean lowbwt

codebook lowbwt

summarize lowbwt, detail

7.) Create a scatterplot of lowbwt vs. life92:

graph twoway scatter lowbwt life92

What can you say about the relationship between low birth weight percentage and life expectancy?

8.) Create a new variable called *change* which is the change in life expectancy from 1960 to 1992.

Remember the **generate** command from the first lab. Which country had the highest increase in life expectancy? You can sort *change* in descending order by typing:

gsort -change

*Now we will be moving on to a different data set. Clear the unicef data from Stata by typing **clear** in the Command window.

Eight individuals experienced an unexplained episode of vitamin D intoxication that required hospitalization; it was thought that these unusual occurrences might be the result of excessive supplementation of dairy milk. Blood levels of calcium and albumin – a type of protein – for each subject at the time of hospital admission are provided below.

1.) Open the data editor and input the following data into Stata:

Calcium (mmol/l)	Albumin (g/l)
2.92	43
3.84	42
2.37	42
2.99	40
2.67	42
3.17	38
3.74	34
3.44	42

2.) Find the following summary statistics for the recorded calcium levels:

- Mean
- Median
- Standard deviation
- Range

3.) For the given albumin levels, compute the:

- Mean
- Median
- Standard deviation
- Range

4.) For *healthy* individuals, the normal range of calcium values is 2.12 to 2.74 mmol/l, while the range of albumin levels is 32 to 55 g/l. Do you believe that patients suffering from vitamin D intoxication have normal blood levels of calcium and albumin?

***Now we will examine a third and final data set.**

The data set *lowbwt.dta* contains information recorded for a sample of 100 low birth weight infants – those weighting less than 1500 grams – born in two teaching hospitals in Boston, Massachusetts. Measurements of systolic blood pressure are saved under the variable name *sbp*. The dichotomous random variable *grmhem* indicates whether or not the child had germinal matrix hemorrhage (GMH). Don't forget to **clear** data from the previous problem!

1.) What is the mean systolic blood pressure of this sample of 100 babies?

2.) Construct a pair of box plots for the systolic blood pressure measurements – one for babies without GMH and one for babies with GMH. Use the command:

graph box sbp, over(grmhem)

3.) Compute the mean and *standard error* of the systolic blood pressure measurements for the two groups. Use the command:

mean sbp, over(grmhem)

To get the *standard deviation* for each group one would type:

summarize sbp if grmhem==1

summarize sbp if grmhem==0

a.) Which group has the larger mean? The larger standard deviation?

4.) How many babies had germinal matrix hemorrhage in this data set? Use the command:

tabulate grmhem

5.) How many male babies had germinal matrix hemorrhage? How many female babies? Use the command:

tabulate grmhem sex

Bonus: Estimate the probability that a female baby has germinal matrix hemorrhage.

6.) When finished type: **log close**