Intro to Stata

The main goal with these two exercises it to demonstrate your basic familiarity with the Stata user interface. The basic idea is for you to issue a few simple commands in Stata, collect the code and results, and assemble them in a Word document similar to what you might turn in for an assignment.

Mathematically, we will be examining different transformations of data, and the effect of these transformations on the mean and standard deviation. These will be important later on in this course, so you will be revisiting this topic later.

Problem 1. The Price of a New Car

(a) Open the data set "autos.dta", found in the folder "Y:\Stata".

Find the mean and standard deviation of the "price" variable, using the **`summarize`** command. Make a histogram of this variable, using the **`histogram`** command (scale the y-axis with percents).

This tells us about the mean price and the variability of prices in 1978.

Suppose we are interested in the cost of each car plus the cost of title, registration, and delivery. Add $500 to the price of each car (**`generate`** a new variable), recalculate the mean and standard deviation, and redraw the histogram.

How has the mean changed? How about the standard deviation? Can you express this as a formula?

Looking at the histogram, you will see that while the x-axis has shifted, the distribution of the data has the same shape as before. This kind of change of variable is called a "translation."

(b) Now suppose we are interested in what these prices would look like in 2013 dollars. Inflate the price of each car by 350% (again using **`generate`**). Again recalculate the mean and standard deviation, and draw a new histogram.

How has the mean changed this time? The standard deviation? You will see that the shape/distribution of the data is still the same, although the numbers along the x-axis have changed. Does "rescaling" change the ratio of the mean to the standard deviation?

(c) Combine both "translation" and "rescaling". Suppose we want 2013 prices, and the cost of title, registration, and delivery is still $500. How does this affect the mean and standard deviation?

This is a "linear transformation" ("linear" in the sense of linear algebra), and it has the very nice property that the shape of our data distribution remain the same.

(d) Finally, look at a non-linear transformation. Consider the fuel efficiency of these cars as measured in miles per gallon, the "mpg" variable.

Calculate the mean and standard deviation, and draw a histogram of the data.

Now suppose we would rather analyze fuel efficiency in terms of gallons per mile. Generate a new variable that is the reciprocal of the original variable. Recalculate the mean and standard deviation, and draw a new histogram.

Does this look like a linear transformation?

Assemble all your results together in a Word document. Extra credit: If you panel all four graphs together in a two by two grid (with appropriate captions), it makes the mathematical point very clear.

Problem 2. Ages and Wages

(a) Open the data set "nlsw88.dta", found in the "Y:\Stata" folder. These data are from a national longitudinal survey of women that began in 1967. They were re-interviewed in 1988, and some of that data is here.

Calculate the mean and standard deviation of the "age" and "wage" variables, and draw a histogram of each.

A statistic that describes the relationship between age and wage is the correlation coefficient, which you will learn about in detail later. For now, just calculate it with the command

```
correlate age wage
```

(b) A very useful data transformation is to "normalize" your data, something else you will examine in detail.

Calculate two new variables "zage" and "zwage" by subtracting the mean of each variable from its column, and dividing by the standard deviation, so

$$\text{zage} = \frac{\text{age} - \text{mean(age)}}{\text{stdev(age)}}$$

$$\text{zwage} = \frac{\text{wage} - \text{mean(wage)}}{\text{stdev(wage)}}$$

These are linear transformations. What is the mean of each variable? The standard deviation? Draw two new histograms to show that the distributions have retained their shape.

Finally calculate the correlation coefficient between zage and zwage.

An important feature of linear transformations is that they not only preserve the distribution of each variable, but also preserve the relationship between variables.

As before, assemble your statistical results and graphs into a single Word document. This can be the same document you used for Problem 1.