 January 4, 2011

Problem Set 1

This problem set will be due Friday, January 21 by noon. It can be dropped off in my mailbox in the economics department mailroom. No late problem sets will be accepted. You may work in groups but everyone must turn in their own problem set. Note that this does not mean turning in multiple copies of the same problem set. Each group member must produce their own Excel results and write up their own solutions. For Excel output, please only turn in the final results and graphs you are asked for. Do not print out and turn in the complete datasets (use the 'Set Print Area' feature in Excel).

1. **Working With Cross-sectional Univariate Data**

   For this question, you will be working with life expectancy data for the year 2006 from the World Health Organization. The datafile (world-health.csv) can be found on Smartsite in the data folder. Note that links to definitions of the variables are provided at the bottom of the spreadsheet.

   (a) Create a histogram for female life expectancy at birth with bins having a width of 5 years. Absolute frequency should be on the vertical axis. Does the distribution of life expectancy appear to be symmetric, left-skewed or right-skewed?

   (b) Calculate two measures of central tendency and two measures of dispersion for female life expectancy at birth. Calculate these same descriptive statistics for female healthy life expectancy at birth.

   (c) Create a new variable that is defined as the difference between female life expectency and female healthy life expectency. Calculate the mean and standard deviation of this new variable. In one sentence, explain what this new variable in measuring.

   (d) How does the mean of this new variable compare to the difference between the mean female life expectency and the mean female healthy life expectency? How does the variance of this new variable compare to the difference between the variance of female life expectancy and the variance of female healthy life expectancy?

   (e) Let's think about generalizing your results from part (c). Suppose you have two variables x and y and you create a new variable z that is equal to $x - y$. Use the definition of the sample mean and the summation rules from class to show that:

   $$\overline{z} = \overline{x} - \overline{y}. \tag{1}$$

   Will the same thing be true for the variance of z? In other words, is it possible to show that:

   $$\sigma_z^2 = \sigma_x^2 - \sigma_y^2 \text{ ?} \tag{2}$$

2. **Working with Time-series Univariate Data**

For this question, use the annual GDP data for the United States from 1900 to 2000 available on Smartsite. GDP stands for gross domestic product and is a measure of the total output of the economy. The data file is in the data folder and the filename is gdp-1900-2000.csv. The variables are the following:

YEAR: the year of the observation, ranging from 1900 to 2000

NOMINAL: the nominal GDP for the United States in that year in billions of dollars

REAL: the real GDP for the United States in that year in billions of dollars

(a) Nominal GDP is measured in billions of each year's current dollars. This means that nominal GDP changes due to both changes in output and inflation. Real GDP is measured in a billions of year 2000 dollars so that differences in real GDP correspond to just differences in output. Do you think the standard deviation in nominal GDP will be larger or smaller than the standard deviation of real GDP? Explain your answer and then calculate both standard deviations to see whether you were right.

(b) Suppose that instead of measuring nominal GDP in billions of dollars, we measured it in trillions of dollars. We could do this by dividing GDP measured in billions of dollars by 1,000:

$$GDP_{trillions} = \frac{1}{1000} GDP_{billions}.$$

Use the definition of the sample standard deviation given in class, the summation rules from class and the standard deviation for $GDP_{billions}$ you calculated in part (a) to calculate the standard deviation of $GDP_{trillions}$. (Note: You don't need to calculate $GDP_{trillions}$ to answer this question.)

(c) We are often interested in the growth rate of GDP. Create a new variable that gives the annual growth rate of real GDP over the past year. This can be calculated by taking the difference between the current year's real GDP and the previous year's GDP, dividing by the previous year's real GDP and multiplying by 100. For example, the growth rate corresponding to the 1910 observation would be calculated as:

$$g_{1910} = 100 \cdot \frac{GDP_{1910} - GDP_{1909}}{GDP_{1909}} \tag{3}$$

(d) Create a line chart for the growth rate of real GDP from 1901 to 2000, with year on the horizontal axis and growth rate on the vertical axis. Does it appear that the growth rate of real GDP has become more volatile or less volatile over time?

(e) Economists often use a mathematical shortcut for calculating growth rates. For small changes in a variable, the percent change is approximately equal to the difference in natural logs. Using this approximation, we could calculate the growth rate for the 1910 observation as:

$$g_{1910} = 100 \cdot (ln(GDP_{1910}) - ln(GDP_{1909})) \tag{4}$$

Construct a new variable that calculates the growth rate of real GDP using this approach and another variable that shows the magnitude (absolute value) of the difference between the growth rate calculated in part (c) and the growth rate calculated in this part for each year.

(f) Create a line chart showing both the growth rate of real GDP (use either one of your growth rates) and the difference between the two measures of the growth rate for 1901 to 2000 with year on the horizontal axis. Does the log approximation look like a good approximation of the growth rate? Does the approximation work better when growth rates are large or when they are small?