McGill University
Department of Economics
Economics 257D: Economic Statistics

Assignment 1: Descriptive Statistics

J. W. Galbraith ©

Exercise 1

Obtain data on the civilian unemployment rate and the *percentage change in* US real gross domestic product (GDP) from the FRED (Federal Reserve Economic Data) data section of the web site of the Federal Reserve Bank of St. Louis: `https://fred.stlouisfed.org`. If you scroll down on the main page, each of the series will appear (as of the time of writing) as one of the featured series ('at a glance'), so that you can just click on the name of the data series (in blue) and you should see a graph of the series. Note that the particular transformation you need is the one that comes up automatically (again, as of the time of writing). You can then download the data by clicking on the download button on the right side of your screen. This will transfer a file to your computer in the format you choose. While I normally prefer to keep copies of small data sets such as these in a human-readable ASCII text format, you might prefer to download them simply as .xls files as inputs to Excel, Matlab etc.

We can label these two data series as U and Δy.

Finally, obtain data on the Industrial Production Index (call this I) from the same source. You will note that this series is trending upward: we will not be able to analyze this in the same way. See 1e below.

If you download the files in human-readable ASCII text, you will have to edit the files to eliminate explanatory information and dates before processing with Excel or Matlab. You can use a plain-text editor (e.g. Notepad, OxEdit, etc) to do this. If you download as an excel file, you may be able to process the file directly in your software.

In your assignment, indicate the day on which you downloaded the data, since the data sets may be updated. **For U and Δy only:**

1. Compute the sample values of the mean, variance, standard error, coefficients of skewness and kurtosis of each random variable.

2. Compute the medians, noting that the series may have an even number of observations depending on when you download.

3. Produce histograms of the data with a bin width of 0.3, to start with. Comment on the apparent appropriateness, or otherwise, of this bin width for these data sets and sample size. Describe what would happen if you were to use bin widths of 0.1 or 1.00. Switch to another bin width for either or both of the series, if you think that this would be more revealing of the actual distribution of the data.

4. Compare the coefficients of skewness and kurtosis with the values that you would expect in a Normal distribution: 0 and 3 respectively. Do these appear Normal? Of course, we haven't yet studied this distribution or learned how to do a formal test, so just make a rough judgment.

5. **Now consider the third series,** I. Because this series has an upward trend, it does not make sense to talk of the mean or other measures calculated above as being approximately constant, as we might be willing to do for the first two series. It is nonetheless possible to compute sample measures, *even though they are not estimating anything well defined*. This illustrates the importance of doing some theoretical study of probability and distributions so that we can understand what is going on before we start calculating things. Meanwhile, compute the sample mean on the first half and the latter half of your data points for I : you will notice the you get substantially different numbers. Now, compute the percentage change in the series as $\frac{I_t - I_{t-1}}{I_{t-1}}$, where $t = 1, 2, \ldots N$ is the index set for this time series. Notice that when you compute the proportionate change you will lose one observation, because you can't do it for the first observation (you would need observation zero, which by definition you don't have). For this transformation, again compute the sample mean on the first half and the latter half of your sample. The numbers should now be closer to each other. We will discuss this in class after the assignment is handed in; for now just compute some numbers as a basis for our discussion.

Exercise 2

Consider the following random variables derived from U which you obtained above (we'll just use U here). $S = U+2$; $Z = 2U$; $W = 4U+3$; $V = W+2S$. In each case, give an estimate of the mean and variance of the new random variable (these are estimates because the mean and variance obtained in the first part of the question are estimates; if those were population values, you could compute population values for the new random variables). If you prefer you can just assign a name to the mean and variance of U and give your answer as a function of that variable, eg. define the mean of U as μ, variance of U as v_u, give answers as functions of μ and v_u.

If you find this non-transparent or have any doubts about your answers, you can actually construct the new variables and do the computations again from scratch.

Exercise 3

Consider two investments which have each produced returns over a period of seven years. Investment A produced a return of 20% in year one, zero in year two, 20% in year three, zero in year four, 20% in year five, zero in year six, and 12% in year seven. Investment B produced a return of 10% in each year.

1. Compute the arithmetic and geometric mean returns for each investment.

2. Beginning with \$1000, work out what the value of each of the investments would have been after seven years, and determine which was the better investment.

3. The two mean-return calculations, arithmetic and geometric, rank the investments differently. Which one of the two correctly reproduces the ranking by total final value? Why does the other mean calculation get it 'wrong'?

Exercise 4 (Optional)

If you are familiar with software that allows you to do this (e.g. Matlab, Python, R, etc.) generate a sample of 200 pseudo-random numbers (x_i) from some standard distribution that your software can reproduce, such as the

Uniform or Normal, with values mostly in a small interval around 0 (eg within +/- 3 or within 0 to 1, or some such interval).

Alternatively, just work with the unemployment rate data above, and modify the series as described below.

Next compute $(n^{-1}) \sum_{i=1}^{200} (x_i - \overline{X})^k$, for exponents $k = 2, 4, 6, 8$.

Now consider the following two experiments:

1. multiply all values in the original series by 2, and recompute the results (of course you can figure this out without doing it, as in Q2).

2. replace just one single value in the original series by the number 20 (use 50 if you are using unemployment rate data), and recompute the results.

Which experiment produces the greater change in the statistics, for each value of the exponent k? Interpret the result.