MONT 105N – Analyzing Environmental Data
Project on Variability in Sampling
March 9, 2020

Introduction

The descriptive statistics we have introduced are important tools, but they are not the end of the story. One basic goal of statistics is to analyze the information that is contained in a data set and to make inferences from that information about the population the measurements came from. The data sets used in environmental questions might come from sampling some larger population of organisms. Even when that is not the case, we can use the idea of sampling from a larger population of all possible measurements as a way to conceptualize how the data was generated.

To understand some of the statistical topics we will need for the next steps of our work with the carbon footprint audit data, it will be good to have some more intuitive understanding of the process of taking samples (random or otherwise) from a population, and the variability that is built into the process of computing statistics that are produced by that sampling process. This may seem counterintuitive at first—after all, there is just one mean or SD (for instance) associated to any particular sample from a larger population. There is no variation or uncertainty in dealing with any one data set. But the idea is that those sample means and SDs and other statistics produced from *different* sample data sets *can vary a lot* and we need to have a good working understanding of several aspects of that sort of variation.

We will also practice with some additional features of spreadsheets that may be useful for your work. Unlike some of the other chapter projects, this project will not focus on an environmental data set, but instead will use synthetic data produced in different ways. Most of your work for this will be contained in a single spreadsheet file.

Constructing a First "Population" and Sampling

Open a new spreadsheet and enter the formula `=RAND()` in cell `A1`. This function computes a random value from a certain distribution on the interval 0 to 1. The outputs of `RAND` are called *random numbers*, but they are produced using a definite algorithm, as is true for essentially everything done with computer software. So a better name would be *"pseudo-random" numbers*. They are only random in the sense that they have no apparent patterns. Copy and paste that into all the cells in the block extending from column `A` through column `Z` and row `1` through row `100` (2600 cells in all). You now have a random sample of size 2600 from that uniform distribution.

Because of a "quirk" in the way spreadsheets containing this `RAND()` function in formulas are handled, you will see that if you now enter a formula in another cell of the spreadsheet and calculate that other value, all of the numbers in this block will be recomputed too. To turn that feature off (it's annoying and it will mean your results are constantly changing),

- highlight the whole block
- CTRL-V to copy, then

- Edit/PasteSpecial, and
- Paste Values Only (or CTRL-Shift-V).

If you examine the contents of the cells in the block now, you should see just numbers, not the formulas that generated them, and that's what we want. This means that new random numbers will not be generatedeach time the spreadsheet is updated.

First Calculations

Now perform the following calculations:

(A) Compute the average (mean), the SD, and the five-number summary of the whole population of 2600 numbers.

(B) Next compute the data needed for a *frequency histogram* showing the distribution of this population of 2600 numbers. (Look back at Chapter 6 in the text if you need a refresher on this.) Since all the numbers are between 0 and 1 and we have a relatively large data set with $N = 2600$, we can use 10 equal bins with upper boundaries at $0.1, 0.2, \ldots, 1.0$. Constructing the histogram will be done as follows:

- Enter the upper bin boundaries $0.1, 0.2, \ldots, 0.9, 1.0$ into cells in one column in the spreadsheet at least three rows below the block of 2600 cells above (say in column `A` and rows `103` to `112`).
- Then in the parallel column in cell `103`, enter `=FREQUENCY(A1:Z100,A103:A111)`.
- When press ENTER/RETURN to execute that, the frequency table will generated in column `B`.

Then you can generate the histogram plot by making a bar chart with the frequency values. Generate a frequency histogram using the computed frequencies with the bins as above. This should be nearly (but not exactly) even across the range 0 to 1.0.

(C) Now, we consider sampling from this "population." Notice that the block `A1:J20` can be thought of as either:

- 100 rows (horizontal) with 26 cells in each, or as
- 26 columns (vertical) of 100 cells each.

We will think of those as two different groups of samples from the population consisting of all 2600 of the numbers—one group of 26 samples of size 100 from the columns, and as second group of 100 samples of size 26 from the rows. (These are not really random samples, of course, since they are constructed in this systematic way. However the whole block was computed by processes that produce random-looking results, so it will do no harm to think of them as random.)

(1) Compute the averages of each of the 100 rows and put those averages in the cells `AA1:AA100` (you'll need to insert another column to the right of column `Z` to do this).

(2) Find the mean and SD of the 20 row averages from part (1). Generate a frequency histogram for those 100 values using bins of width 0.05 rather than 0.1. (You will need to decide on appropriate bin boundaries by examining the values you get. You want to set the bin boundaries to include all the data, but you also don't want lots of bins with 0 counts.)

(3) Compute the averages of each of the 26 columns and put those averages in the cells `A102:Z102` (i.e. in the cells in row 102 under the block).

(4) Find the mean and SD of the 26 column averages from step 3. Since we only have 26 of these, making a frequency histogram is not too meaningful, so you do not need to do that.

Constructing a Second Population and Sampling

Open a second tab (sheet) in your spreadsheet and enter the "*black magic*" formula

```
=SQRT(-2*LN(1-RAND()))*COS(RAND()*2*PI())
```

in cell `A1`. (This might be tricky. If you get an error message, check the parentheses carefully and make sure they are exactly as above.) Copy and paste that into all the cells in the block extending from column `A` through column `Z` and row `1` through row `100` (2600 cells in all). Show 3 decimal places in all of the numbers. Repeat parts A to C above for this new data set, except that you will need to decide on appropriate bin boundaries for the histograms. (Don't worry about exactly what the formula means. The only thing you need to know for now is that it's designed to produce random values that are distributed differently from the ones you used in the first section.)

Further Questions

(D) How was the population histogram for the second population different from that of the first population?

(E) What can you say about the SD's of the original population, the SD's of the averages of the samples of size 26, and the SD's of the averages of the samples of size 100? In particular, was one of these consistently the largest and one consistently the smallest? If the pattern was not entirely consistent, was there a general trend?

(F) If we redid these computations with any fixed rule for computing the population values, would it always be true that the average of the row means is the same as the population mean? Similarly for the average of the column means? (Hint: algebra.)

(G) What would truly random samples of sizes 26 or 100 from these populations look like? Suggest a way to generate such samples.

(H) If we had 26 truly random samples of size 100 from either of these populations, would the average of the sample means be the same as the population mean? Explain your answer.

Assignment

Submit your spreadsheet and your answers to questions A - H in a separate document.