MONT 105N – Analyzing Environmental Data
Chapter 9 Project
February 6 and 8, 2019

Introduction

The descriptive statistics we have introduced in this chapter 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 look at starting in the next chapter, 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 on larger assignments for this and other courses. 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 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 uniform distribution on the interval 0 to 1.[1] Copy and paste that into all the
cells in the block extending from column `A` through column `J` and row `1` through row `20` (200 cells
in all). Show 3 decimal places in all of the numbers. You now have a random sample of size 200
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

---

[1]Technical note: We will say more about what this means in Chapter 10. 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 and if you compute a large collection of them, their distribution will be approximately uniform
over the interval from 0 to 1.

off (it's annoying and it will mean your results are constantly changing), highlight the whole block and 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 generated each time the spreadsheet is updated.

First Calculations

Now perform the following calculations:

(A) Compute the average (mean) and the SD of the 200 numbers. Compute the five-number summary. (Optional: If you want to experiment, try producing a "candlestick plot" for this in Google Sheets.)

(B) Next compute the data needed for a *frequency histogram* showing the distribution of this population of 200 numbers. Look back at Chapter 6 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 $N = 200$, we can use 10 equal bins with upper boundaries at $0.1, 0.2, \ldots, 1.0$. Constructing the histogram could be done as we discussed in the Chapter 6 project, *but* that doesn't quite give us enough control over the number of bins, etc. There is another method for counting how many numbers are in each bin. Enter the first nine upper bin boundaries $0.1, 0.2, \ldots, 0.9$ into cells in one column in the spreadsheet at least three rows below the block of 200 cells above (say in column `A` and rows `31` to `39`). Then in the parallel column in cell `B31`, enter `=FREQUENCY(A1:J20,A31:A39)`. When you 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.

(C) Generate a frequency histogram using the computed frequencies with the bins as above.

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

- 20 rows (horizontal) with 10 cells in each, or as
- 10 columns (vertical) of 20 cells each.

We will think of those as two different groups of samples from the population consisting of all 200 of the numbers—one group of 10 samples of size 20 from the columns, and as second group of 20 samples of size 10 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 20 rows and put those averages in the cells `L1:L20` (i.e. in the cells in column `L` parallel to the block)

(2) Find the mean and SD of the 20 row averages from part (1). Generate a frequency histogram for those 20 values using 5 bins. (You will need to decide on appropriate bin boundaries by examining the values you get.)

2

(3) Compute the averages of each of the 10 columns and put those averages in the cells `A22:J22` (i.e. in the cells in row 22 under the block).

(4) Find the mean and SD of the 10 column averages from step 3. Since we only have 10 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

$$\texttt{=SQRT(-2*LN(1-RAND()))*COS(RAND()*2*PI())}$$

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 `J` and row `1` through row `20` (200 cells in all). Show 3 decimal places in all of the numbers. Repeat parts A to D 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

(E) Note that `RAND()` appears in both formulas you used to generate both of the population data sets above. Again, this is a built-in Excel formula that generates random (more precisely "pseudo-random") numbers in the interval from 0 to 1. Looking at your population histogram from the first data set, what can you say about the distribution of those numbers? Do they seem to be reasonably uniformly distributed over the interval? What about the second data set where you used the *black magic* formula? How is that one different? How do the two population box plots compare?

(F) What about the distributions of the row and column averages in both cases? What did the distributions look like there? Did it depend which data set they came from?

(G) What can you say about the SD's of the original population, the SD's of the averages of the samples of size 20, and the SD's of the averages of the samples of size 10? 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?

(H) 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.)

(I) What would truly random samples of sizes 10 or 20 from these populations look like? Suggest a way to generate such samples.

(J) If we had 20 truly random samples of size 10 from either of these populations, would the average of the sample means be the same as the population mean? (This is the same question as H above, but for a different way of producing the samples.) Explain your answer.

Assignment

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