Building Frequency Distributions

Did I mention that understanding frequency distributions will be critical to developing an understanding of the analytical techniques that we are going to cover? What better way to understand frequency distributions than to build a few?

The Excel workbook you will use for the remainder of this exercise can be downloaded by clicking HERE.

The workbook contains 3 worksheets, as indicated by the tabs at the bottom. The first of these ("Oncho") contains abundances (worms/fish) of *Onchocleidus flieri*, a gill parasite of fliers (*Centrarchus macropterus*) caught in the Okefenokee Swamp.

We will start with this set of data because it is relatively small. Our goal is to rearrange the data, such that each abundance is listed only once, and the number of times that abundance was observed is indicated in the cell next to that abundance. With a small data set such as this, it would be relatively easy to simply do the counts, i.e., no fish with 0 abundance, 2 fish with an abundance of 1, 2 fish with an abundance of 2, etc., but this would be quite cumbersome with larger datasets. Fortunately, Excel simplifies this task, requiring only that we determine an appropriate class size.

As a first step, it will be helpful to determine the range of our observations. In cell A22 type:

=min(a2:a20)

As we learned before, it generally is easier to highlight the cells being referenced than to type in the range. As you might have guessed from the name of the function, the smallest value in that series will appear in cell A22. Again, this is not terribly necessary with such a small set of data, but doing it here will allow you to validate what the function does, so that you can use it when the list of observations spans several pages. Here is another hint for working with large data sets: you can highlight an entire row or column, regardless of length, using the "Shift" key to anchor the first cell, and the "End" key followed by an arrow to move to the last filled cell. Let's do this for our next command. In cell A23, type the following (but do not hit enter or one of the arrow keys):

=max(

Now, click on cell A2 to select it. Then, while holding down the "Shift" key, hit "End" on your keyboard, and then hit the down arrow (you do not need to hold the "End" key down). This will highlight all of the values in the column. Key in the right parenthesis and hit enter to complete the formula. You also can use "Page Down" while holding the "Shift" key to highlight large blocks of values. It may not seem like a huge timesaver now, but you will thank me for it later.

We now know (if it wasn't obvious before) that our abundances range from a minimum value of 1 to a maximum value of 18. When displaying a frequency distribution, we generally want to have class sizes that are as small as possible, but still produce a graph without too many ups and downs. We will start with class sizes of one. Starting in cell C2, enter a value of 1 , and fill the cells below it with a step value of 1, and a final value of 18, using the "Fill" function that you learned last week. There should already be labels in C1 for the abundance ("number") and in D1 for the frequency, but if there are not, you should probably add some.

To get Excel to put the data in column A into a frequency distribution using the classes you assigned in column C, we will use the "FREQUENCY" function, which takes the form:

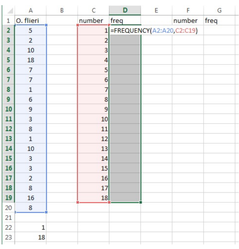
=FREQUENCY(*data*, *bins*)

The "*data*" argument is filled by highlighting all of the observations. Then enter a comma, and highlight the classes you filled into column C for the "*bins*" argument. These are referred to as "bins", because they are where Excel will place the observations. Finish the formula with the right parenthesis and hit "Enter". Your completed formula for cell D2 should look like this:

=FREQUENCY(A2:A20,C2:C19)

Click on cell D2, and hit "F2" to see the formula, and the cells being referenced. Before performing the next step, click on an empty cell to avoid accidentally editing the formula in cell D2.

Now comes the odd bit. Highlight the cells in column D that are adjacent to the "bins", including cell D2, and then hit the "F2" key:



Now, hit "Ctrl"+"Shift"+"Enter", holding down all 3 keys simultaneously. If you have done this correctly, and are kind to animals and small children, column D should fill in with the appropriate counts (frequencies). Because the data set is small, this will not be difficult for you to verify. If the numbers in column D continually increase, and sum to a number that is obviously greater than the actual sample size (the number of rows in column A), then you have done something wrong (most likely you dragged the formula down).

If you are using a ~~boat anchor~~ Mac computer, the last time that I checked, the equivalent procedure (after completing the formula) was "to click and drag" from the cell with the formula to select all the cells to the right of the "bins", click at the end of the formula in the formula bar, and press "Command" and "Enter" on the keyboard. If that doesn't work, you will have to hunt on Google...

We can see that if we were to graph these frequencies (which go on the Y-axis), the graph would be far from smooth. Thus, it would be prudent to rebuild the frequency distribution using larger class sizes. Starting with a value of 2 in cell F2, fill in the values to 18 using a step value of 2. When building the distribution, Excel will count all observations that are the value of the bin value and less, so observations less than or equal to 2 will be placed in the "2" bin, and observations greater than two, and equal to or less than 4 will be counted for the "4" bin, and so on. Build this distribution with the "FREQUENCY" function, and then build another starting the classes (bins) in cell I2, with a starting value of 3, and a step value of 3. Graph both of these distributions following the guidelines we learned last week, including moving the graph to its own worksheet, saving the first as "Chart1", and the second as "Chart2".

Question 6: Which of the 2 graphs would be the best (i.e., smoothest) presentation of these data, and why?

The spreadsheet labelled "Physa" contains another set of data. You may want to save your spreadsheet at this point if you haven't already done so (as always, save it as "*yourlastname*ex2"). These data are shell lengths for the pulmonate snail *Physa pomilia* collected from a single locality in October of 2012 in the West Pond at Brick Pond Park in North Augusta. Use the "MIN" and "MAX" functions in Excel to determine the range of the observations, and use the "FILL" function to build your classes (bins). The step parameter in the "FILL" function will work with fractions, so start small (perhaps 0.1 mm classes), and then increase the class size until the graph of the frequency distribution looks smooth. Remember that you **DO NOT** drag-copy the frequency formula...you have to use "F2" followed by "CTRL+SHIFT+ENTER". Once you settle on an appropriate (or at least appealing) class size, graph those data following the guidelines we already have used, and save it in its own worksheet as "Chart3".

Now comes the fun part! In the worksheet labelled "RAND" (have you saved your worksheet?) are 500 observations. These represent integers ranging between 1 and 10, drawn randomly using the software package R (the program used can be viewed here, for what it is worth). Remember that "random" means that each integer between 1 and 10 should be equally likely to be drawn for every observation. We can examine whether this is the case by building a frequency distribution of these data. BUT WAIT! THERE'S MORE! We are going to compare that distribution to a frequency distribution generated by running the same process in Excel. First, we need to recall the formula for generating random integers between 1 and 10:

=INT(RAND()*9+0.5)+1

OK. Maybe "recall" isn't the right word, because I just gave the formula to you. Type that formula into cell B2. Dragging the formula to copy it to 500 cells will not take too much effort, but there is an easier way. Click on cell B2 and copy the formula using "Ctrl+c". Now click on cell A2, and then hit the "End" key, and then the down arrow (do NOT hold down the shift key). This should take you to cell A501. Now, use the arrow to move over to cell B501, hold down the "Shift" key, hit the "End" button, and then the up arrow. This should highlight all of the cells in column B that are adjacent to a value in column A. Hit "Ctrl+v" to paste the formula into all of those cells. You should now have 500 random integers ranging from 1 to 10 drawn by Excel. Don't be alarmed if these values change as you build your distribution. Remember that Excel will redraw these numbers every time that you perform a function.

For this exercise, our class sizes are pre-determined: all integers from 1 to 10. All that remains is to build frequency distributions for the Excel and R draws, and compare the two. The best way to do this is to make a single graph depicting both distributions (again, following the same guidelines we established earlier). Save the graph as "Chart4".

Question 7: Which of the 2 software packages appears to do a better job of generating random numbers? What is the basis for your conclusion, i.e., what is the expected distribution, and which method comes the closest to meeting this expectation?

Save your Word document and Excel workbook as *yourlastname*ex2, and submit via Blackboard.

Tune in next week...same Bat-time, same Bat-website!

Week 2 Objectives

Be able to categorize independent and dependent variables as continuous, discrete, or nominal variables

Understand the distinction between fixed and random variables

Understand the importance of replication, randomization, and control in experimental design

Know how to construct a frequency in Excel and understand how to interpret a frequency distribution

Understand and recognize pseudoreplication