Environmental Data Laboratory
Professor Kirchner

Laboratory 2: Exploring the Central Limit Theorem

The Central Limit Theorem has far-reaching implications for almost every aspect of data analysis, but the formal mathematical proof of the theorem is sometimes hard to grasp. Here, rather than proving the central limit theorem, you will demonstrate it. Remember that a mere demonstration does not carry the logical force of a formal proof, because it only generates results for specific examples rather than for general cases. Hopefully, however, these demonstrations will give you a better intuitive grasp for how the Central Limit Theorem works. By demonstrating rather than proving the theorem, we're deliberately trading generality for intelligibility.

Loosely put, the Central Limit Theorem says that if you take measurements on randomly selected samples and average them, the mean of your measurements will have a more normal distribution than the individual measurements themselves. As you average over larger and larger numbers of measurements, the distribution of means will be more and more "normal", or bell-shaped. This will happen no matter how badly behaved (how skewed, how flat or spiky, how bimodal, etc.) the distribution of the individual measurements is. The more non-normal the individual measurements are, however, the more measurements you will need to average, to obtain a roughly normal distribution of the means.

In the second part of this exercise, you will explore the practical implications of the Central Limit Theorem. As the preceding paragraph suggests, this theorem is not merely a theoretical curiosity; it is also important in practical day-to-day work. For example, the theorem indicates that you can accurately measure the average of a property (say, the average concentration of an air pollutant over time at a given location, or the spatially-averaged concentration of a contaminant in groundwater), even if those concentrations are very unevenly distributed, as long as: 1) your samples are representative, and 2) you have enough of them. Here, you will estimate how many samples are required to get a reasonable estimate of the average streamflow leaving a watershed.

This lab will also give you lots of experience--maybe more than you want--with using the formula editor in JMP (remember last time, when you used the formula editor to transform flow measurements to logarithms?). The instructions will be as explicit as possible, but if you're still confused, you may want to consult with chapter 3 in JMP Start Statistics, or consult with a classmate, or consult with the TA.

Remember, too, that help is always available online, either through the "help" function in JMP, or by selecting the question mark from the "tools" menu, and using it to click on anything you want explained.

As in last week's exercise, you should label a sheet of paper with your name, "EPS 120, Lab #2", and today's date. Keep this handy to jot down your musings about each question in the following exercise.

Part A: Thought experiments with synthetic distributions

In this part, you'll construct your own "thought experiments" to illustrate how sampling affects the distribution of averages, where the individual measurements are not normally distributed. You'll do this by creating a set of measurements with a known distribution--one that is not at all normal. You'll then sample randomly from this "parent" distribution, and calculate the average of your random sample. By repeating this process over and over, you'll construct a distribution of the averages, which you can compare to the distribution of the individual measurements. Clear? I didn't think so...but follow along and you'll get the hang of it.

Launch JMP and select File >> New >> Data Table (or just pick the "New Data Table" option from the launch box). You will be presented with a data table that has one column, and no rows. How helpful, right? But don't fret. Select "add rows" from the "rows" menu, and at the prompt, add 200 rows. Now you have a one-column, 200-row table full of black dots (that's how JMP shows missing data). Not much more helpful, I know...but be patient.

Step 1: Create a parent distribution

Our first step will be to create a "parent" distribution in column 1 that we can then work with. We could work with real data, but it's perhaps easier if we start with synthetic numbers that have known properties. Let's start with a *uniform distribution*, one with values that are evenly distributed between zero and one. (Of course, they won't be *exactly* evenly distributed, since they're random numbers).

Select your new empty column (by clicking on its header, "column 1"), and then select Cols >> Column Info. In the resulting dialog box, replace the "Column Name" with something appropriate (in this case, say, 'parent distr'). Then press on "New Property" and pull down "Formula". If the formula window doesn't appear automatically, click "Edit Formula". This is how you open a formula window in order to create a formula for a column of a JMP data table. Complicated, I know, but that's what you have to do... and you'll need to do it a fair bit in this lab.

Now it's time to enter our formula. On top of the function page, you will see two lists. The right-hand list contains categories for built-in functions. Click "Random" from the right-hand list. This brings up a selection of random number generating functions. Click "Random Uniform". This inserts uniformly-distributed random numbers into your column. Your formula should look like:

"Random Uniform()"

Then close the formula calculator windows by clicking the two "OK" buttons. Note that the missing values disappear, and are replaced by 200 random numbers. Simple, no?

Now, use "Distribution" from the "Analyze" menu to look at the histogram of this distribution. Note that it does not have the classic bell shape of a normal distribution.

> *Very important note to remember, for this lab and for labs throughout this course:* it may take *time* for JMP to recalculate rows whenever you insert a new formula, add rows, or, most importantly, change an existing formula. If you change a formula, JMP must recalculate that column, and all other columns that depend on it. *You should wait until the recalculations are finished before you do anything else with JMP*. How do you tell when JMP is finished recalculating? Look for "evaluations done" in the lower left corner of the JMP window.

Step 2: Calculate averages from random samples of the parent distribution.

Now, you've just seen what the distribution of the individual measurements looks like. The question is, if you took pairs of measurements and averaged them, how would the averages be distributed? Let's now simulate this. What we want to do is to calculate a whole column of averages, where each cell in that column is the average of two points *chosen at random* from the parent distribution. Here's how to do it.

First, create a new formula column called, say, "average of 2". Then, do the following to build the formula (this may be tricky or tedious... but hang on). First, from the function window, select "Statistical" (you may need to scroll the list to do this), then select "Summation" from the sub-list that appears. Then insert "parent distr" from the left-hand variable list into the big empty box marked "body". As it stands, the formula would calculate the sum of all the measurements in your "parent distr" column, which isn't what you want. Now, change the "NRow()" at the top of the summation to "2" (click or double-click on "NRow()", type "2", and press return or enter). What you'll see at this stage is:

$$\sum_{i=1}^{2} parent\ distr$$

This still isn't quite what you want, for two reasons. First, it's not an average, it's just a sum. Second, it's simply the sum of the first two values, not two randomly chosen values. To get randomly chosen values, here's what you need to do. Select the box enclosing 'parent distr' in the formula. Then from the list of functions (the right-hand list), select "Row" and then "Subscript" from the sub-list that appears. Now, highlight the subscript box by clicking on it

(if it is not highlighted already), and from the list of functions select "Random" (you will probably need to scroll down to see this), and then select "Random Integer" from the sub-list that appears. Next, highlight the box inside the parentheses of the "RandomInteger()" function, and from the list of functions select "Row" and then "NRow" from the sub-list.

Finally, you need to divide by two. So, select the whole function (by clicking on the outline box around the whole thing), click "÷", and enter the value "2" below the line. Now you should see:

$$\frac{\sum_{i=1}^{2} parent\ distr_{RandomInteger(NRow())}}{2}$$

Now, this is just what you want. What it says is, "Pick a random integer between 1 and the number of rows, then go get the value of "parent distr" that is stored on that row. Do this twice, sum the results, and divide by two."

Now, draw the histograms of the "parent" distribution and the two-point averages. Note that although the parent distribution is roughly evenly distributed, the two-point averages are not. In the two-point averages, there are relatively few values like 0.1 or 1.0, and relatively more values between, say, 0.3 and 0.8.

Question 1: *Why?* Why are the extreme values under-represented, and the middle values more widely represented? After all, what you're looking at is the average of samples from a parent distribution that is *virtually flat*. •1. Take a minute to think about this, and jot a brief answer on your paper.

It's important to be clear about what your column "average of 2" represents. It is a set of 200 different means, each of which is calculated from two randomly chosen measurements from your parent distribution. Thus, the distribution of "average of 2" represents the likely distribution of two-point averages that you could get from pairs of measurements.

Note that in the resulting report window, the red pull-down triangles have many handy features. For example, you can select "Distributions >> Stack" to put the histograms in the more familiar horizontal orientation. Or you can select "[distribution name] >> Fit Distribution >> Normal" to superimpose a normal curve that has the same mean and standard deviation as your histogram, so you can compare your histogram to a normal distribution with the same center and spread as your data (or try other distributions as well). Third--and this is important--you can select "Uniform Scaling" from the pull-down triangle next to "Distributions" to put all of your histograms on the same scale. If you don't use uniform scaling, each histogram will be automatically rescaled according to the range of its data; this shows the shape of the distribution more clearly, but makes it hard to compare the width of two different histograms.

Finally, remember that using the "grabber hand" tool and moving it up/down and left/right changes the resolution in your histogram (the width of the bins), and varies the alignment of the bin boundaries.

Step 3: Extend the analysis to include averages of more measurements.

Your "average of 2" column is based on the average of just two data points from the parent distribution. Remember that the central limit theorem says that as you average over more and more data points, the distribution of your averages should become more and more normal. Let's see whether this is the case (or, more accurately, *how much* the case it is).

Make another column called "average of 10", and give it a formula just like the one you created above... but replace the 2's with 10's, like this:

$$\frac{\sum_{i=1}^{10} parent\ distr_{RandomInteger(NRow())}}{10}$$

Handy hint: You can save yourself a lot of work by *copying the formula* from your previous column. Note that copying the formula is different from copying the values in the column. Instead, you open the calculator window, select the formula (so that the outermost outline box is highlighted), and copy it. Then you open the calculator window for the new column, and paste it there. Then you'll need to change the 2's to 10's, and you're done.

Remember to give JMP enough time to recalculate your columns.

How does the shape of the distribution of "average of 10" differ from "average of 2"? •2: On your paper, sketch all three distributions (the parent distribution, the 2-point average, and the 10-point average) on the same axes to show how they compare. When you do this, remember that *the areas under the three different curves need to be the same* (since what you are trying to sketch are probability density functions, and the total probability--the area under the curve--doesn't change from one case to the next). That means that narrower distributions need to be proportionally taller. •3: Are the means of the distributions substantially different? Why or why not? •4: Are the standard deviations different? Why or why not?

One implication of the central limit theorem is that as more and more measurements are included in an average, its standard deviation shrinks as $1/m^{0.5}$, where m is the number of measurements in the average. (Note that the parent distribution itself can be considered a distribution of "averages" for m=1...that is, for single points). •5: Do your averages seem to fit this pattern? In other words, does increasing the number of points in the average from 2 to 10 decrease the standard deviation roughly according to the prediction of the Central Limit Theorem: s.d.(mean)=s.d.(parent data)/$m^{0.5}$?

Step 5: Extend the analysis to less well-behaved parent distributions

What if the parent distribution isn't as symmetrical--and generally well-behaved--as the uniform distribution used above? Let's investigate that possibility by repeating the analysis outlined above for another distribution. First, save your work so far, so you can go back to it. To do so, save your JMP data file as an appropriately named file on the desktop, or on a floppy if you remembered to bring one.

Now, to redo the analysis above with a different distribution, you don't need to reconstruct the whole spreadsheet; all you need to do is change the parent distribution. To do this, double-click on the heading of the parent distribution column (or click once and select "Cols" >> "Column Info". Then click "Edit Formula".

Now, what appears before you is your familiar formula, "Random Uniform()". Highlight this, then click on "x^y", then change the "2" to "3", yielding:

$$\text{Random Uniform}()^3$$

Then close the window, and notice that a new set of numbers has replaced your previous ones. (And remember to *wait* for JMP to update the other columns in the data table.)

Now, redraw the histograms of the parent distribution, the 2-point average, and the 10-point average. Use the grabber hand to divide the parent distribution more finely, to better reveal how skewed it is. Remember that you can use the "uniform scaling" option to better compare the spread of the three distributions.

•6: Again, sketch all three distributions (the parent distribution, the 2-point average, and the 10-point average) on the same axes to show how they compare. •7: Does the distribution for the 2-point average look acceptably

"normal" to you? Does the 10-point average? Why do you think this might be? •8: Again, does the standard deviation seem to obey the $1/m^{0.5}$ rule (see above for details) as well as it did in the other examples? •9: Do you think that averaging over even greater numbers of measurements might make the distribution more "normal"?

Part B: Can you estimate the average flow from a watershed?

Now, let's leave the nice clean world of theory behind and dive into the less tidy realm of real-world data. This part of the exercise concerns a very practical problem: can you reliably estimate the average flow rate of a stream (say, from a series of roughly weekly measurements)? There are many reasons one might want to know the average streamflow. One might want to know, for example, how much water will be available for use downstream. Or, perhaps you are studying the physiology of the forest, and you want to know how much water the trees are transpiring; you might try to find out by comparing the flow of water leaving the forest in streamflow with the flow of water coming in via rainfall. Or, perhaps you are looking for evidence of global warming, and you want to see whether streamflow is changing (possibly indicating either a change in precipitation, or a change in evapotranspiration).

While this may seem like a simple problem, it is not at all as trivial as it first appears. Remember that streams go through long periods of very low flow, with occasional brief flood flows that are so high that they can account for a large fraction of the total flow, even though they account for a very small fraction of the year. So how many streamflow measurements does it take to determine the average streamflow to an acceptable level of precision (where what's "acceptable" obviously depends on the task at hand)? Note that this problem is virtually identical to many other problems in environmental science. Does average air quality meet legal standards? What is the average, or total, flow of pollutants into the Bay from various sources? In each case, you are faced with the problem of accurately defining the average of a series of measurements that may vary wildly from time to time and place to place.

Step 1: Peek at the problem at hand.

Open the JMP data file "Kaarvatn flow data". This is a series of flow measurements taken on a stream in the western fjord country of Norway. To get some idea of what the time series looks like, you might want to plot flow against time (choose "fit y by x" from the "analysis" menu, then use "time" as the x-variable and "flow" as the y-variable). For another look at the flow data, plot a histogram of the flow measurements, and use the "grabber hand" to make the bars narrow enough that you can see the shape of the distribution clearly. Pretty skewed, eh?

Step 2: Average randomized samples of streamflow measurements.

As with the earlier, synthetic data, construct a formula that averages 10 random measurements. You can either construct the following formula, or copy the formula from your earlier file (again, copy the *formula*--from the calculator window--don't copy the data values themselves) and edit it. The result you want is:

$$\frac{\sum_{i=1}^{10} Flow(L/s)_{RandomInteger(NRow())}}{10}$$

Now, draw the histograms of the raw flow measurements and the 10-point averages. Note the differences in the shapes of these distributions (you don't need to write anything down yet); note also that the standard deviation of the distribution of 10-point averages is still large compared to its mean.

Step 3: Estimate distributions for averages of larger samples.

Let's say we need to know the average flow to within fifteen percent or so. How many measurements would need to be included in the average, to make its standard deviation smaller than, say, 200 L/s? Try this by a series of experiments. Create three more columns, for 20-point averages, 50-point averages, and 100-point averages. You can save yourself the trouble of recreating the complicated formula above by opening the formula calculator for that column, selecting the whole formula, copying it (see the "edit" menu), and then pasting it into the formula calculator for the new column. You will then need to change the values of "10" to the appropriate values, both in the summation and in the denominator.

If the 100-point averages calculate too slowly, then you can approximate them using a formula that computes 10-point averages of data values that are randomly selected from the column of 10-point averages you've already calculated...that is, the same formula as before, but using the column of 10-point averages as its source data, rather than the raw "Flow (L/s)" measurements. This works because the average of 10 ten-point averages is the average of 100 points. It's not as thoroughly randomized as the average of 100 randomly selected individual points, but it's pretty close.

•10: On your paper, sketch the parent distribution and the distributions of two different averages, to illustrate the effect of changes in the number of samples being averaged. •11: Write down a table where you compare the standard deviation for each sample size with the standard deviation you would expect if the $(1/m)^{0.5}$ rule held exactly. Does the rule seem to hold reasonably accurately? •12: Roughly how many samples do you need in your average to make its standard deviation about 200 L/s?

Step 4: Extrapolate to sample sizes needed for high-precision averages.

If you were trying to see the effect of climate or land-use changes, you might be looking for changes in the mean flow of just several percent. To see these sorts of effects clearly, you would need an average flow that was known to within roughly a percent. How many samples would you need to include in your average to make it reproducible to within one percent, or roughly 15 L/s? (Remember that you can think of the reproducibility of an average as being the standard deviation of replicate averages from each other... where of course those averages are estimated on different groups of samples, all drawn from the same population.) Don't try to do this by extending the numerical experiment you've conducted before, summing over ever-more measurements, because you might well exhaust the data set.

Instead, assume that the standard deviation of the average is indeed inversely proportional to the square root of m, and calculate the sample size you would need to be reasonably confident that you knew the mean to within roughly 1 percent. •13: Write the answer, plus a brief explanation of how you did it. •14: If these samples were taken once each week, how many years would you need to measure the average precisely enough that you could be confident that you knew it to within 1 percent? •15: What does this imply about your ability to detect the effects of climate change or land use changes?

That's it. I hope this exercise has left you with at least two clear impressions. First, the central limit theorem is your friend; it makes means look normal (even when they are the means of samples from truly nasty parent distributions), and it makes their standard deviations shrink. But second, not even the central limit theorem can work miracles. When the variability in your parent data is very high, there may be no alternative to sampling quite extensively if you want to know the average to very high precision.

Finally, one closing note. Everything you've done in this lab assumes that your measurements are *representative* of the whole parent distribution, and that they are taken *randomly*, or without bias. For example, if you only sample the stream during low flows (if you can't get to the stream when it is in flood) then obviously you will never approach the true mean, no matter how many measurements you take. Your estimates of the mean may be *precise*, but they will not be *accurate* unless your measurements are representative.