Environmental Data Laboratory
Professor Kirchner

Laboratory 1: Introduction to JMP and qualitative analysis of distributions

This lab gives you a chance to become familiar with some of the basics of using JMP. You will work through some simple introductory exercises provided by the makers of JMP. You will then do a short qualitative assessment of water quality in a Norwegian watershed.

To log on, use the class user ID and password. These will allow you to log on to any of the lab computers, either during lab periods or during drop-in hours, throughout the semester. Consult with the front desk at the lab for the current drop-in schedule. This user ID is for use only by students enrolled in this course. You will want to get in the habit of bringing a floppy disk or a memory stick, in case you want to save any files you create (you won't need to save any files from this exercise).

In general, you should feel free to work with others, and discuss your work freely. Each student should hand in his/her individual written work (answers to questions, etc.), but the results--and thoughts about those results--can be derived collaboratively. Places where you need to respond are numbered and marked by "•". You need only provide a numbered list of your responses.

Step 1: Log on to a station, double-click on the server, double-click on the "classes" folder, and double click on the "EPS 120" folder. Everything pertaining to this class will be nested somewhere within the EPS 120 folder. Everything pertaining to the first laboratory exercise, for example, will be in the "Lab 1-Introductory Lab" folder.

Step 2: Read through the following pages of your text, "JMP Start Statistics", trying out each of the demonstrations in them.

pp. 8-17: A simple introduction to analysis using JMP IN

Note that you can launch JMP IN either by double-clicking on the JMP IN icon itself, or by double-clicking on any JMP data file (which will both launch JMP IN and open the data file).

Note also how a data table and all its graphical reports are linked to one another. For example, highlighting one or more rows in the data table will simultaneously highlight the corresponding parts of a frequency distribution (histogram). Likewise, if two histograms are shown simultaneously (either in the same report window or different windows), highlighting one or more bars of one histogram will highlight the sections of the other histograms that correspond to the selected data.

pp. 27-36: Entering data and constructing simple graphs

Along the way, demo the help system both by using the "Help" menu, and by using the question-mark tool in the toolbar (just to the right of the 'arrow' cursor tool). Pick up the question mark tool, place it over an item on a data sheet or analysis platform, and click to see information related to that item.

pp. 57-62: An introduction to the formula editor

pp. 75-79: Another example with the formula editor

Step 3: Included in the folder for this lab are water quality data files from measurements on two Norwegian watersheds: a mountainous watershed near the western coast, called Kaarvatn, and a forested watershed in southern Norway called Langtjern. These data were collected over the past 10-20 years by the Norwegian State Pollution Control Authority and the Norwegian Institute for Water Research, which has kindly made the data available to us. Open one or the other data file and follow along through the exercises below. On a separate sheet, answer the indicated questions. At the top of the sheet, put "EPS 120L, Lab #1", your name, and the name of the watershed you've selected.

The water quality variable that we'll be interested in here is alkalinity, denoted Alk (the unit of measure is microequivalents per liter, for the chemically curious). Low alkalinities can indicate acidic conditions, and may make waters uninhabitable for some organisms. Here, we'll try to describe the alkalinity of these watersheds, and we'll look at how the distribution of alkalinity varies with different times of year, and different streamflow conditions.

First, use "Distribution" under the "Analyze" menu to compile a distribution of alkalinities from your chosen watershed. Observe how the visual appearance of the distribution changes when you change the width of the bins (by moving the grabber hand sideways), or the boundaries of the bins (by moving the grabber hand up and down).

•1. Describe, in a sentence or two, the shape of the alkalinity distribution. In a couple of sentences, describe in your own words how the visual appearance of the distribution changes as you change bin width and bin boundaries.

Next, recall from lecture and/or reading that the central tendency of data is often described in three different ways: 1) the arithmetic mean, 2) the median--the 50th percentile, with equal numbers of data points on either side, and 3) the mode--the peak of the distribution. Let's see how <u>efficient</u> these different measures of central tendency are...that is, how well they work on relatively small samples.

> •2a. First, from the alkalinity distribution you just drew, write down the mean, the median, and your best estimate of the mode (you'll have to eyeball the mode from the histogram). Write each of these down.
>
> Second, randomly select twenty rows from various parts of the the data table. To do this, select "Subset" from the "Tables" menu. Then click "random sample" under "rows options" and insert a sample size of 20. Construct a distribution of alkalinities for this randomly chosen subset. •2b. Again, write down the mean, the median, and the mode for this subset.
>
> Randomly select another set of twenty rows from the original data table, and repeat the analysis outlined in the previous paragraph. (Be careful not to randomly select your 20 rows from your previous subset of 20, which would give you the same subset back again!) •2c. Again, write down the mean, the median, and the mode for this second random subset.
>
> Do this a third time: randomly select 20 rows, and •2d write down the mean, the median, and the mode for this third random subset.

•3. Ponder the following questions, and write a brief response. Of the three measures, the mean, the median, and the mode, do any seem to be efficient estimators of central tendencies in the data? That is, do the estimates for your three small samples seem to agree with the values derived from the bigger sample, or don't they? If you were faced with the problem of assessing the alkalinity of many Norwegian watersheds, and you could only take a few samples at each watershed, are there any measures would you recommend using, and any you would avoid?

Now, let's look at whether low-alkalinity periods tend to happen during particular times of year or particular flow conditions. For these purposes, "low alkalinity" means less than zero at Langtjern, and less than 10 at Kaarvatn. Draw distributions for month, flow, and alk, using the full data set, not one of your small subsets.

Whoops! Notice that all the high flow data are invisible, because the flow distribution is so skewed that a single bar, near zero, dominates the histogram. Notice that some much higher flows occur, but that they are rare. To make the flow data more visible, we'll need to <u>transform</u> the flow data in a way that spreads them out more evenly. It turns out that a logarithmic transformation will work reasonably well for these data. Create a new column called "log (flow)", and in the dialog box press ""New Property" and pull down "Formula". In the formula window click "transcendental" to look at the possible transcendental functions (of which the common log is one). Select the common log ("Log10"), then click on the empty box in the common log formula (if it's not already highlighted), then select the flow variable to fill that box, so the formula reads "log10(flow(L/s))". Then close the window.

Now then, draw distributions for month, flow, log(flow), and alk. Using the arrow tool, as before, shift-click to select the range of low alkalinities from the alk distribution (by highlighting the lower bars of the histogram). Now look at the highlighted parts of the other distributions (notice how you can barely see these on the flow distribution, but how they are relatively clear in the log(flow) distribution--this is one advantage to transforming data so that it is not all bunched up in one place). Remember that each unit of log(flow) represents a ten-fold change in flow. Ponder the following question, and respond in no more than a few sentences.

•4: Are there certain times of year, or flow conditions, when low alkalinities tend to be more common? If you were concerned about possible biological damage because of low alkalinity, what flow conditions or times of year would you target for more intensive study?

That's it! You're finished with the first lab. We hope you've found it illuminating. Put away your windows, quit JMP, log out (click the logout icon on the desktop), and hand in your page of responses. We'll see you next week.