**A Guide to Data Analysis in Excel**

Alpine Hydrology Workshop

**March 14, 2015**

**Learning goals:**

* I am able to access my data in Excel.
* I know what the difference between mean and median is, how to calculate them, and know when to use median or mean.
* I know how to assess the variability of a data set using standard deviation.
* I know how to create a time series graph in Excel.
* I know how to compare two data sets using scatterplots and linear regression.
* I know how to formulate questions and hypothesis about the data, and which figures and analyses to use to answer these questions.

***Graphical Analysis***

Graphs serve two main purposes, to provide insight into analysis of data and illustrate important concepts when presenting the data to others (Helsel and Hirsch, 2002). Once a dataset is obtained, the next step is finding visual ways to represent the data that may support the “story” of the data, or inspire the idea for the storyline itself.

There are several methods of graphing in Excel. Each graph will emphasize a different property of the data set, and so it is important to try several different methods during analysis, and choose the right one for presentation.

**Statistical Methods in Water Resources** (an electronic version is included in the course CD)**,** written by Helsel and Hirsch is an excellent resource for approaching data analysis. Though it is aimed for hydrology, Chapters 1, 2, and 16 provide a readable introduction to statistics and graphical presentation for all types of data.

Today, we will be working with the SNOTEL data from Scotch Creek, for water years 2011 and 2014, which you downloaded in the last activity. Open the 2011 and 2014 documents you saved from the last activity (if you didn’t get that far, open the documents entitled ScotchCreek\_SNOTEL\_2011WY.xlsx and ScotchCreek\_SNOTEL\_2014WY.xlsx on the course CD).

Clean up your spreadsheet with these steps:

* Title columns with clear header titles. The data descriptors used by SNOTEL:

WTEQ: Snow water equivalent (SWE), inches

PREC: Daily precipitation accumulation, inches

TOBS: Observed air temperature, °C

TMAX: Maximum recorded daily air temperature, °C

TMIN: Minimum recorded daily air temperature, °C

TAVG: Average daily air temperature, °C

SNWD: Snow depth, inches

* Delete unnecessary columns (the site number and time columns). Also, make sure to delete the last row of data (the second 9/30/2011).
* Open the 2014 snow year and copy and paste it into the 2011 workbook as a new worksheet (right click on the worksheet tab on the bottom and select “Move or copy…”)
* Clean up and format the 20124data in a similar way, and save the document under a new name. (ScotchCreek\_SNOTEL\_2011\_2014.xlsx on the course disc)
* SnoTel uses -99.9 for any data errors. Delete any cell values of -99.9.
* Note: When doing data analysis/management, save your work often. In case you want to return to an earlier version of the data, save versions of the workbook under a different name (using numbers or dates to identify, e.g. Scotch\_Analysis\_1; Scotch\_Analysis\_2, etc.). It’s a lot of files, but makes your work easier in the long run!

***Mean vs. Median***

Both mean and median are a method of identifying the central tendency of the data, and it is important to know the difference in order to represent that data correctly. The mean, which is denoted as “average” in Excel, is the sum of all data values *Xi* divided by the sample size *n*.

The median, or the 50th percentile, is the central value of the distribution when the data are ranked in order of magnitude (Helsel and Hirsch, 2002).

Both values are easily computed in excel using the commands **=AVERAGE(**values**)** and **=MEDIAN(**values**).**

The main difference between a median and a mean is the effect of outliers on the value. Helsel and Hirsch (2002) refer to the mean as the “balance point” of a data set. If there is an anomalous very large (or small) value in the data set in comparison to the others, the mean will be strongly influenced, whereas the median will not be.



In other words; for a symmetric distribution of data, the mean is equal to the median. If the data are not symmetric, as in figure 1.3a, then the mean is pushed towards outlier values more than the median.

Hence, the median is more robust, and more resistant to being affected by a single observation.

In many environmental data sets, the median may represent a better estimator of the “middle” value in a dataset with several outliers (which are likely to occur in the real world!).

Using the Excel commands, calculate the mean (remember in excel, mean is =AVERAGE(values) and median is =MEDIAN(values)) and median snow water equivalent for water years 2011 and 2012. To find the maximum snow water equivalent, use the command =MAX(values).

**Questions:**

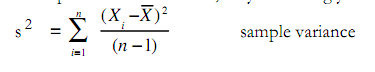
Which year had a higher snow water equivalent?

What other parameters are you curious about?

***Variability***

Assessing the variability of a data set can be as important as finding the central tendency. Variability is a measure of the spread and can be defined using several statistics. The most commonly used are standard deviation and quartiles.

Standard deviation is calculated as the square root of the variance. Variance is a measure of how far values lie from the mean.





The command for standard deviation in Excel is **=STDEV(**values**).**

Take a minute to calculate the standard deviation and interquartile range of snow water equivalent (SWE) for the water years 2011 and 2012 at Niwot.

**Questions:**

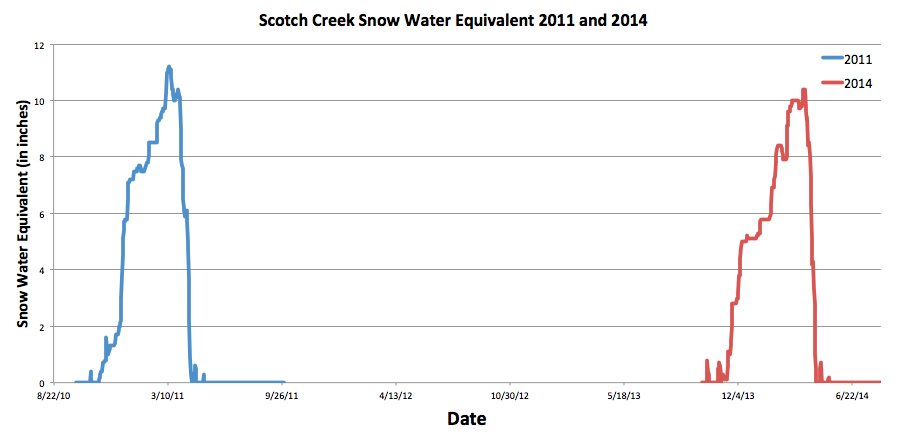
Which year was had higher variability?

Is there a relationship between variability and the median of SWE?

***Times Series Graphs***

A time series graph is the simplest way to present a set of data collected over a period of time. This is an excellent place to start to look at trends or stories in the data.

* Select the Date and SWE (WTEQ) data for 2011. Under the Charts tab, select Scatter, Straight Lined Scatter.
* Right click on the graph, and select Move chart. Select New Sheet, and name the sheet “Time Series”. A new worksheet will appear with the graph enlarged.
* To add 2012 to the time series, right click on the graph and select Select Data…
* First, select Series 1 and name it 2011 WY. Next, add a second series and add the 2012 data by selecting the appropriate cells with you cursor in the 2012 worksheet.
* Spend a few minutes adding a title, axes titles, formatting the legend, and making the graph look presentable. Use the Chart Layout tab.



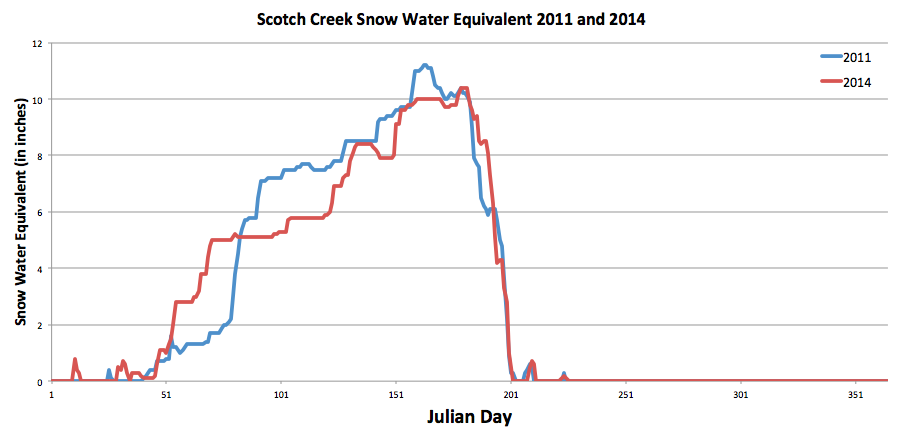
**Questions:**

Take a step back. What is this graphing showing?

What story does it tell? Is there a better way to represent this story?

* To overlay the two years, add a column in each of your data sheets named Julian Day. In the first couple of cells, add the numbers 1;2;3;4. Select the first 4 cells with numbers, and click on the tiny box in the bottom right. This should fill in the rest of the cells in the column. Check at the bottom of the column to see that the numbers go up to 365.
* Create a new graph, using Julian Day on the x axis. This way, both years will be aligned. An efficient way to do this is to select the Time Series worksheet by right clicking on the bottom tab that says Time Series, and copy the worksheet. Click

Select Data and change your selected x values. You will most likely have to change the range of the x axis, by right clicking on the axis and selecting Format Axis.

****

**Questions:**

Now what does your graph tell you? What story can you tell about this data?

What other questions do you have now that you have this graph?

Do you have the necessary data to answer those questions?

What graphs can you make next?

*Integrating Data: Finding relationships between variables*

Climate data provides an excellent opportunity to examine relationships between continuous variables, especially in instances where several parameters are measured at one site (in this case precipitation, temperature, SWE).

First, formulate a question that relates the SWE data to temperature and/or precipitation, e.g. What is the relationship between SWE and temperature? Is there a relationship between SWE and precipitation? Is there a relationship only in some seasons and not others?

**Questions:**

Your research question:

Your hypotheses:

After coming up with some guiding questions and hypothesis, begin to examine the temperature and precipitation data using similar methods to the ones you used with SWE; find the mean, median, standard deviation, and create a time series. Plot a time series of temperature and/or precipitation on the same graph as SWE.

**Questions:**

Do you see anything interesting? What figure can you create next to show those relationships?

Note: If you are plotting two different variables on one chart ( i.e. SWE and temperature), it may be necessary to put temperature on a secondary axis. To do this, right click on the temperature line and select “Format Data Series...”. Select secondary axis under the axis tab.

*Simple Linear Regression*

Linear regression, or Ordinary Least Squares analysis is used to describe the covariation between some variable of interest and one or more other variables. It can be used to describe the relationship between two variables or estimate or predict the values of one variable based on another variable with more available data. Today we will just use it to describe the relationship between two variables.

To begin create a scatterplot with your independent variable on the x-axis, and the dependent variable on the y axis. Use your guiding question to decide which is which, e.g. if your question is “does precipitation predict SWE?” precipitation would be independent(x), and SWE would be dependent (y).

**Questions:**

Your independent variable (x):\_\_\_\_\_\_\_\_\_\_\_\_

Your dependent variable(y):\_\_\_\_\_\_\_\_\_\_\_\_

Is there a clear relationship between the variables?

If not, can you isolate a season when there is a clear relationship between the variables?

To fit a linear regression line, right click on the points, and select “Add Trendline…” Select Linear under Trend/Regression type. Click on the “Options” line below “Type” in the box on the left hand side of the box. Check the boxes for “Display equation on chart” and “Display R-squared value on chart”. A line should appear on your chart, with an equation and R2 number displayed as well.

The higher your R2 is, the stronger the correlation. Remember, correlation does not equal causation, be careful about the language that you use.

**Questions:**

What are your final conclusions about your analysis?

Were you able to answer your research question? Why or why not?

**110**

What other data would be helpful to answer your question? Where could you access this data? (hint: look at the Downloading Data worksheet for ideas)

Note: Figures/equations taken from:

Helsel, D.R. and R. M. Hirsch, 2002. Statistical Methods in Water Resources Techniques of Water Resources Investigations, Book 4, chapter A3. U.S. Geological Survey. 522 pages.