Distribution Workbooks

These workbooks work with Microsoft Excel Version 2010 or later, with file extension .xlsx. The user should be familiar with Excel, including the ability to select and sort data and change scales, fonts, and colors on graphs. These workbooks create Percentile Plots, Box Plots, Histograms, Kernel Density, and Probability Plots. These standard statistical graphs appear in almost every issue of AAAS Science and Springer Nature. They were first developed in the age of main-frame computers. The statistical routines and plotting procedures were first documented by Chambers et al. (1983), and are standard plots in most applied statistics texts like Montgomery and Runger (2011).

Each workbook accommodates 5000 points (12,000 Kb file size). If you need to handle more than 5000 points, contact me at johndunham76@gmail.com and tell me how many points you need.

You can download a range of different workbooks that match several formats that you might use. The workbooks make calculations automatically, but you as the User still have to define the scale range for the plots. To make setting these scales easier, I produced workbooks pre-loaded with data similar to what you may collect in your research.

A workbook formatted for Isotopes is pre-scaled to handle negative and positive numbers. A workbook formatted for measurements is scaled to process numbers like lengths, areas, weights, and flow rates; essentially, any measurement scaled from zero to infinity. A workbook formatted for percentage is scaled for proportions measured between Zero and One, such as porosity ranging from 0 to 100% or fractions (ratios) ranging from 0 to 1.0.

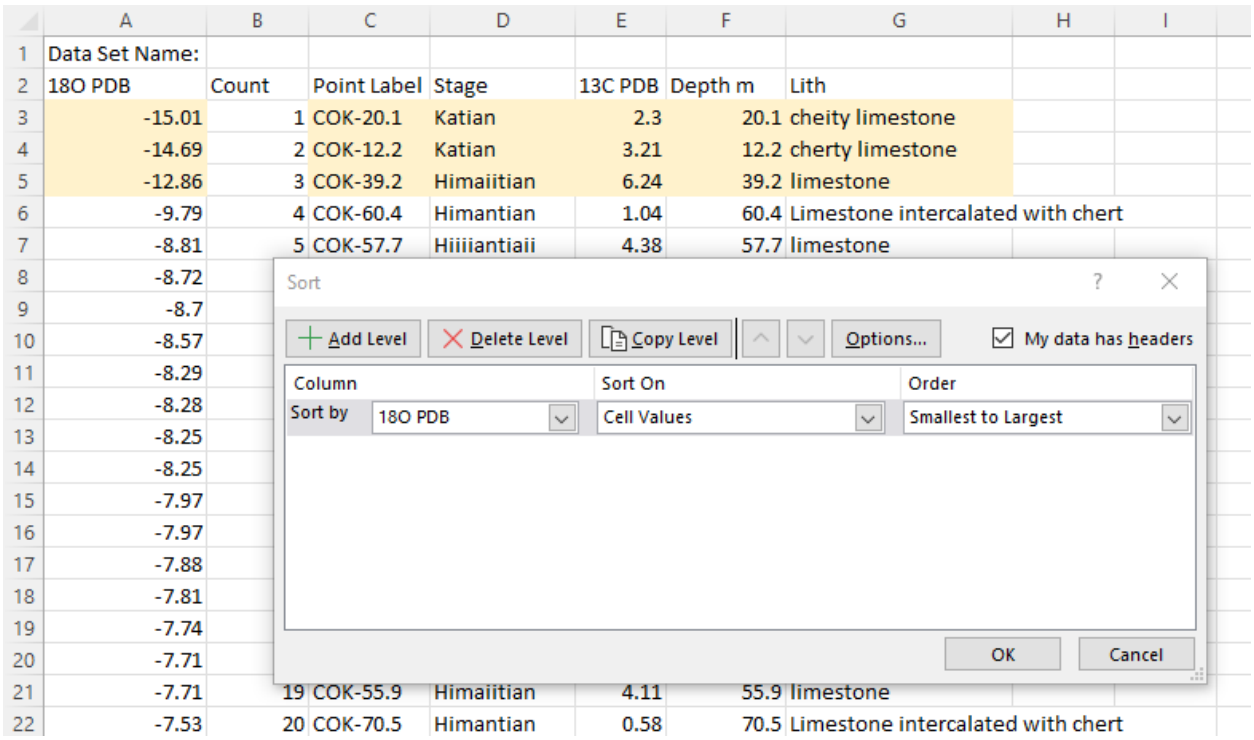
Once the user has assembled a spreadsheet of raw data, the user will copy and paste the data into Column A of the Raw_Data worksheet. At this point, the data automatically transfer to all the other worksheets in the workbook. Next step is to verify that all data are correct in Raw_Data Column A, and then Sort the data from Smallest value at the top to Largest value at the bottom of the column.

Users will then go to each worksheet tab and set the X and Y axis scales for each graph. Cells that require User Entry are highlighted in Yellow Color Fill. Be sure that entries are correct in the yellow cells.

The first example discussed below is an **Isotope Scale Workbook** that examines del 18O measurements of Upper Ordovician limestones in Nevada. Open the Github Mountjoy3 Folder, and select 1_Isotope_Scale_Distribution_Plot.xlsx.

The processing steps for all the workbooks are the same. Step one is to go to the Raw_Data worksheet tab. Enter your data into Column A. Double check your entry to ensure that there are no “extra” or left-over values from a previous data set in cells below the data you just entered. Enter the Data Name in Cell A2; this will automatically appear on successive worksheets.

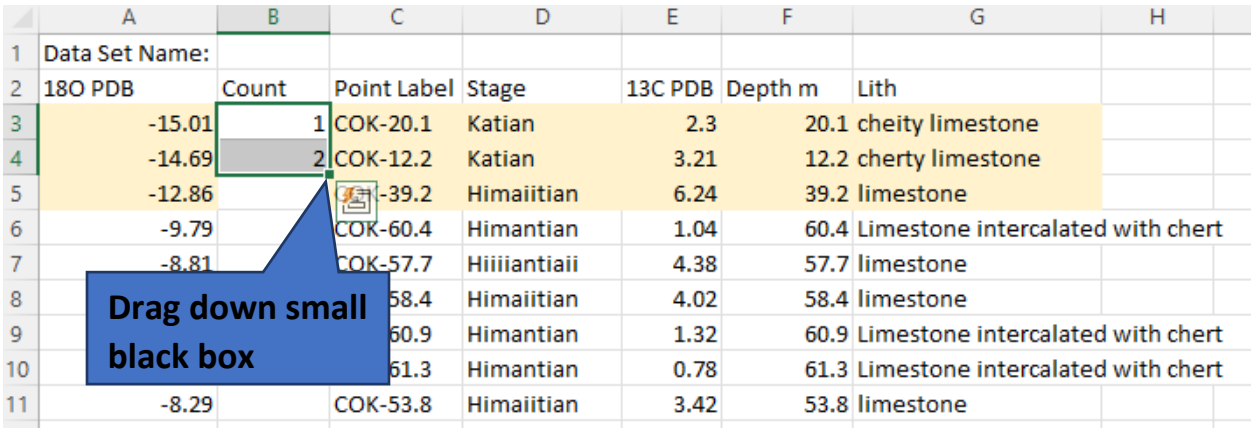
Once you have confirmed that your data is correct in Column A, go to Sort & Filter, and Sort Column A from Smallest to Largest.



	A	B	C	D	E	F	G
1	Data Set Name:						
2	18O PDB	Count	Point Label	Stage	13C PDB	Depth m	Lith
3	-15.01	1	COK-20.1	Katian	2.3	20.1	cheity limestone
4	-14.69	2	COK-12.2	Katian	3.21	12.2	cherty limestone
5	-12.86	3	COK-39.2	Himaiitian	6.24	39.2	limestone
6	-9.79	4	COK-60.4	Himantian	1.04	60.4	Limestone intercalated with chert
7	-8.81	5	COK-57.7	Hiiiiantiaii	4.38	57.7	limestone
8	-8.72						
9	-8.7						
10	-8.57						
11	-8.29						
12	-8.28						
13	-8.25						
14	-8.25						
15	-7.97						
16	-7.97						
17	-7.88						
18	-7.81						
19	-7.74						
20	-7.71						
21	-7.71	19	COK-55.9	Himaiitian	4.11	55.9	limestone
22	-7.53	20	COK-70.5	Himantian	0.58	70.5	Limestone intercalated with chert

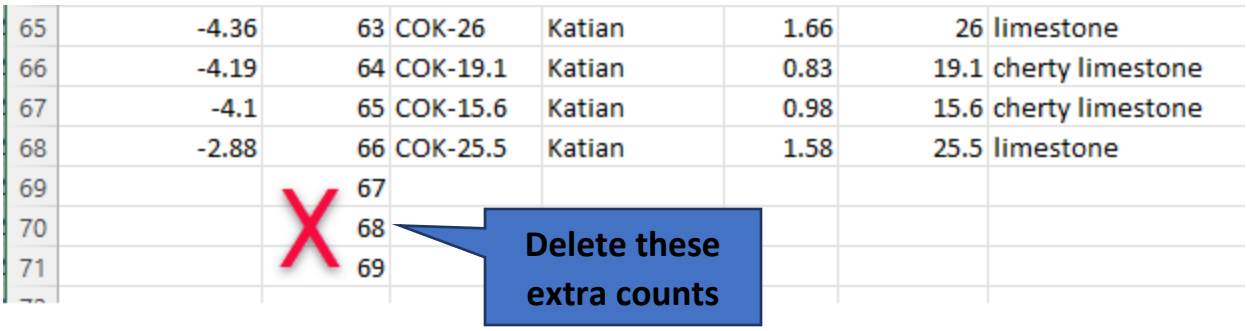
Next, go to Column B and set the Count number to match the number of values in column A. You have to do this manually. A correct count number is necessary to trigger the rest of the automatic calculations in the worksheets. Make sure that every cell in column A has a corresponding count number in column b, and make sure that no extra count numbers are present in column B below the last value in column A.

An easy way to fill the Count column is to enter 1 and 2 in the first two rows, then select the cells for 1 and 2, and look for a small black square at the lower right corner of the selection box; then, drag the black box down to fill all the rest of the cells in the Count column.



	A	B	C	D	E	F	G
1	Data Set Name:						
2	18O PDB	Count	Point Label	Stage	13C PDB	Depth m	Lith
3	-15.01	1	COK-20.1	Katian	2.3	20.1	cheity limestone
4	-14.69	2	COK-12.2	Katian	3.21	12.2	cherty limestone
5	-12.86		-39.2	Himaiitian	6.24	39.2	limestone
6	-9.79		COK-60.4	Himantian	1.04	60.4	Limestone intercalated with chert
7	-8.81		COK-57.7	Hiiiiantiaii	4.38	57.7	limestone
8			58.4	Himaiitian	4.02	58.4	limestone
9			60.9	Himantian	1.32	60.9	Limestone intercalated with chert
10			61.3	Himantian	0.78	61.3	Limestone intercalated with chert
11	-8.29		COK-53.8	Himaiitian	3.42	53.8	limestone

Make sure to delete any extra numbers in the Count column below the last data entry in Column A.



Confirm that your entries in Columns A and B are correct, and then go to the Perc_and_Box Worksheet tab. The data you entered in the Raw_Data sheet appear in Columns A and B on this sheet. Column C automatically calculates the Percentile assigned to the sorted data values.

Percentiles indicate the percentage of values that fall below each particular data point, i.e., the rank of each score relative to all other scores.

The formula used in these workbooks to calculate the Percentile is the "Benard" formula: (i-.3)/(n+.4). Different software packages use different formulas to calculate percentiles. I use Benard because it matches one of the more popular packages. If you compare your output to someone else's graph and find that the percentile points are slightly different from yours, the likely reason is using a different method to calculate percentiles. The difference is very slight among all methods.

There is no problem with duplicate values since each duplicate has a different count number, and scoring is simply a function of the count number and the total number of data points.

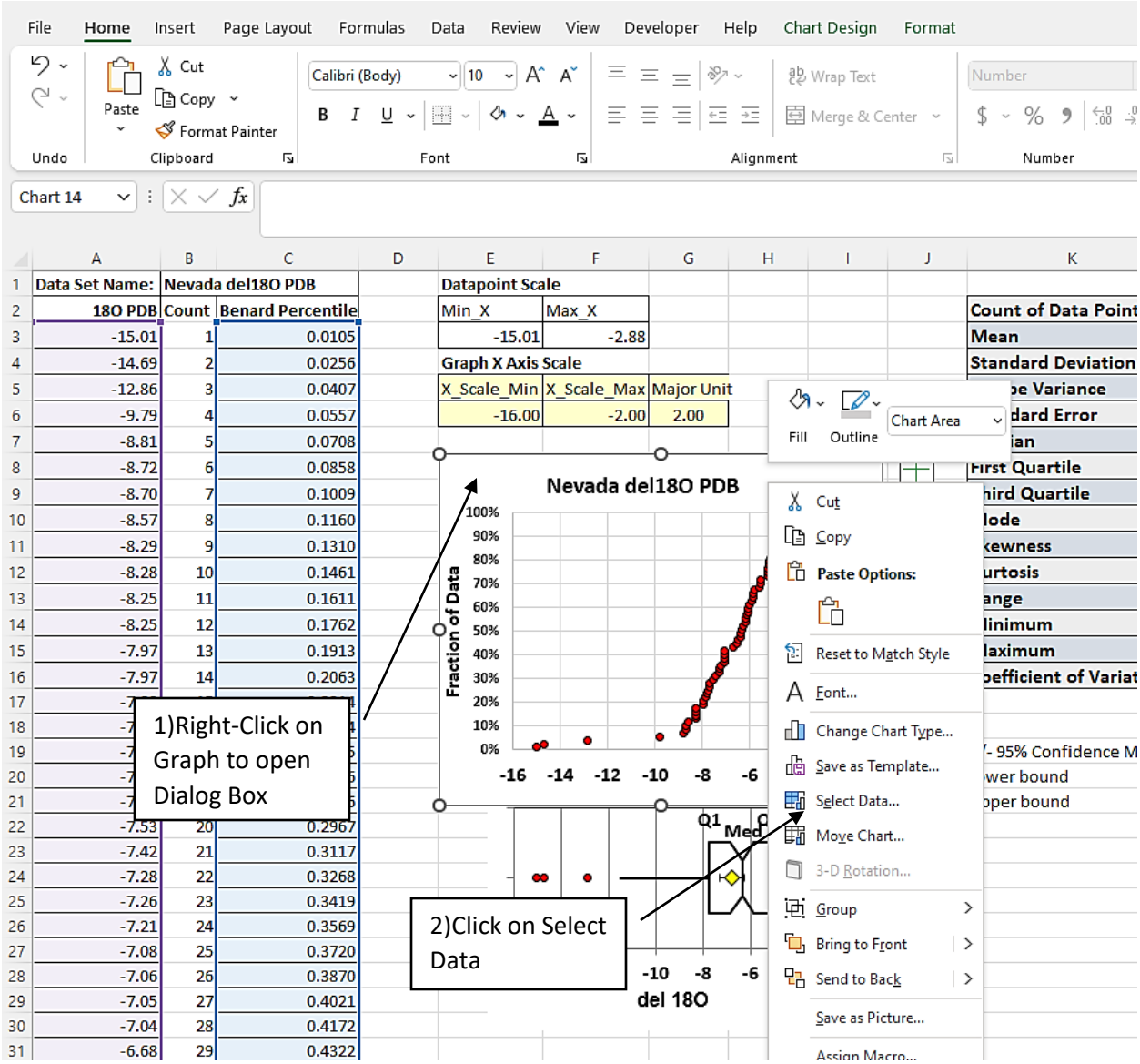
Specific percentiles have unique names: the 50th Percentile is the Median, the First Quartile, Q1, is the value of the 25th Percentile, while the Third Quartile, Q3, is the 75th Percentile.

Percentile Score Methods

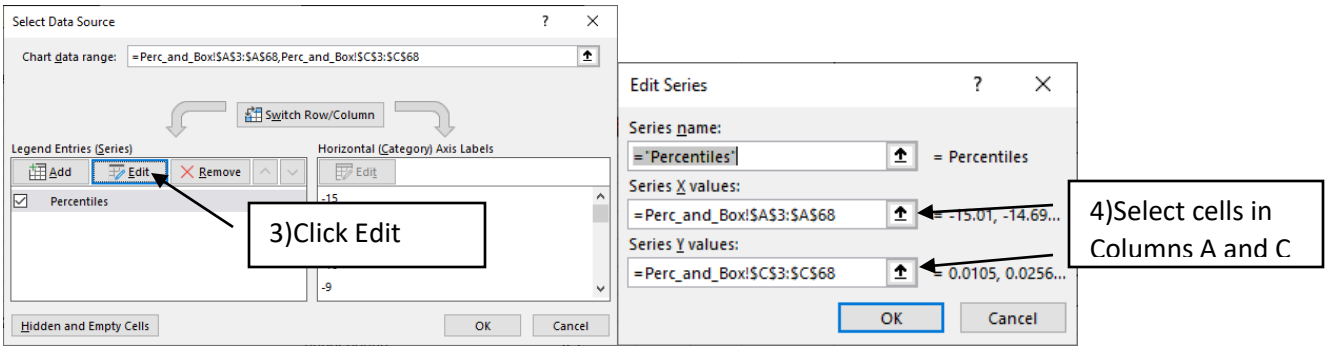
Input data is sorted from smallest to largest, and the count number of the sorted data is scored using one of the methods listed in this table. (i) is the count number and (n) is the total number

Method	Percentile Score based on (i) and (n)
Blom	(i-0.375)/(n+0.25)
Benard	(i-0.3)/(n+0.4)
Hazen	(i-0.5)/n
Van der Waerden	i/(n+1)
Kaplan-Meiner	i/n

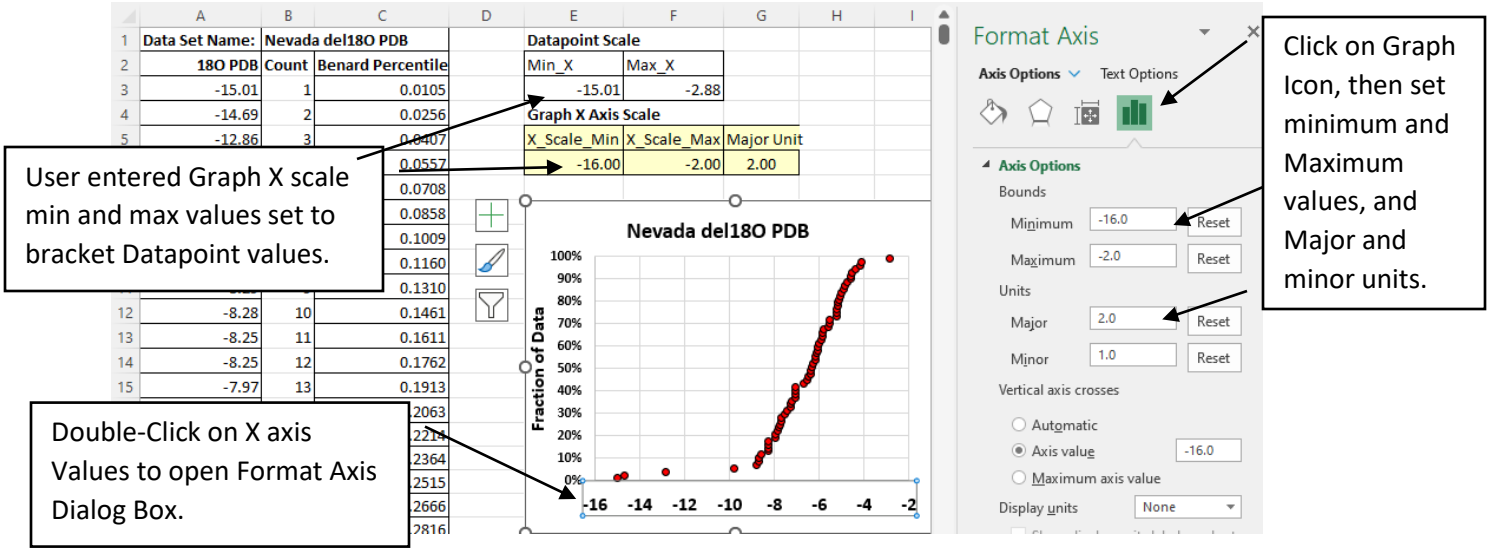
The next step on the Perc_and_Box sheet is to select the ranges and scales for each graph. Right-Click on each graph to open the Select Data dialog and select the X value and Y value ranges.



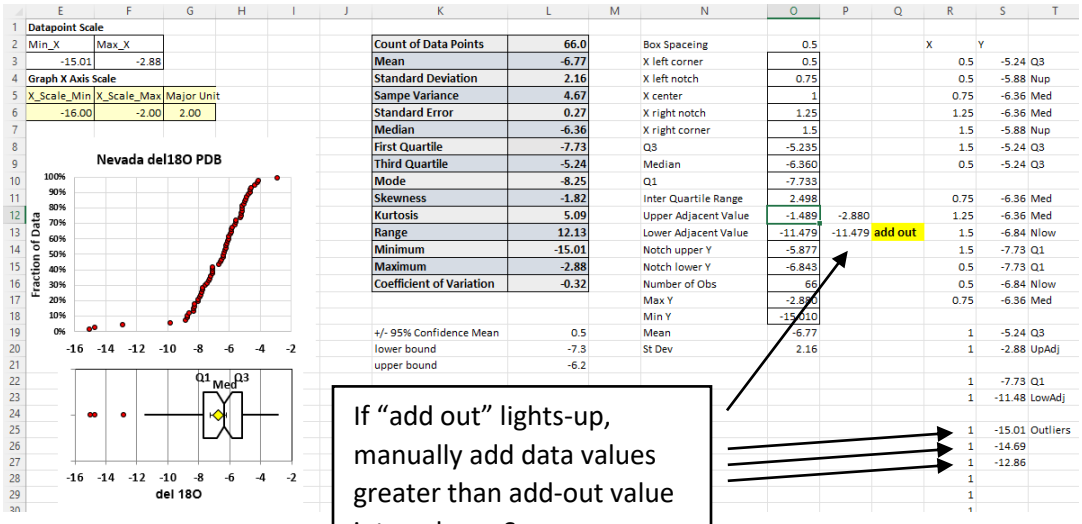
Select Data Box opens; there is only one data series to select, which is Percentiles, click the Edit Box and select cells in column A for X values and Column C for Y values.



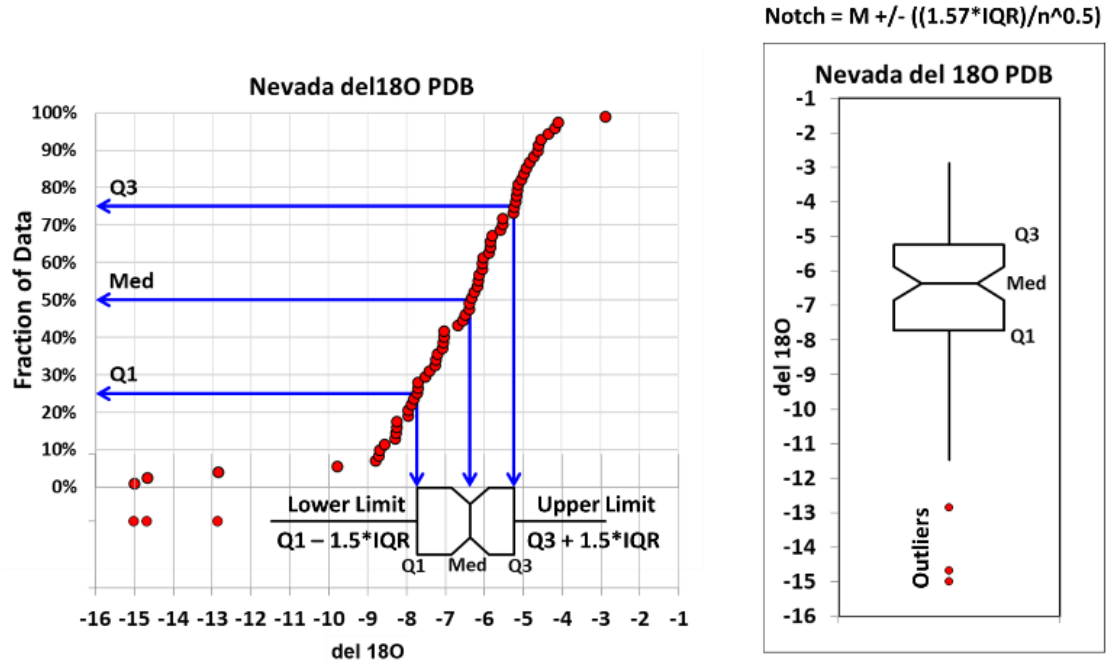
Ensure there are no empty cells above or below the selected X and Y data Ranges. The next step is to set the X-axis and Y-axis Scales. Yellow-filled cells in Cells E6, F6, and G6 are used to help you set the Axis Scales. Yellow cells are the ones that require User Input. In this example, Cells E3 and F3 automatically display the minimum and maximum X-axis data values. Your job as a user is to decide how you want to set the scale of the X axis on your graph. In this example, I decided to enter a Minimum X-axis scale value of -16.0 into Cell E6 and a Maximum X-axis scale value of -2.0 into Cell F6 since these values are one whole digit above and below the data max and min values. You could decide to use other numbers if, for example you were trying to match the scales of other graphs. Put whatever you want in cells E6 and F6, as long as these values capture the entire range of your data.



You have now completed all entries for the Percentile Plot. The next step is to format the Box Plot. Click on the Box Plot and set the X scale to the same numbers you used for the Percentile Plot. There is no need to set the data range for the Box Plot; these are automatic. One detail is to check for "Outliers". The "Whiskers" on the box plot are the Upper and Lower Adjacent Limits, equal to 1.5 times the interquartile range (Q3-Q1 distance) above Q3 and below Q1. Outliers are points that plot beyond the extent of the whisker limits. If Cell Q13 lights up Yellow and says "add out ", you must manually enter all values greater than Adjacent Values into Column S, as shown below.



There are several graphs on this page. Once you confirm the correct ranges and scales for these graphs, you can format them to fit your purpose. The box-plot notch shows the 95% confidence interval for the location of the population median.

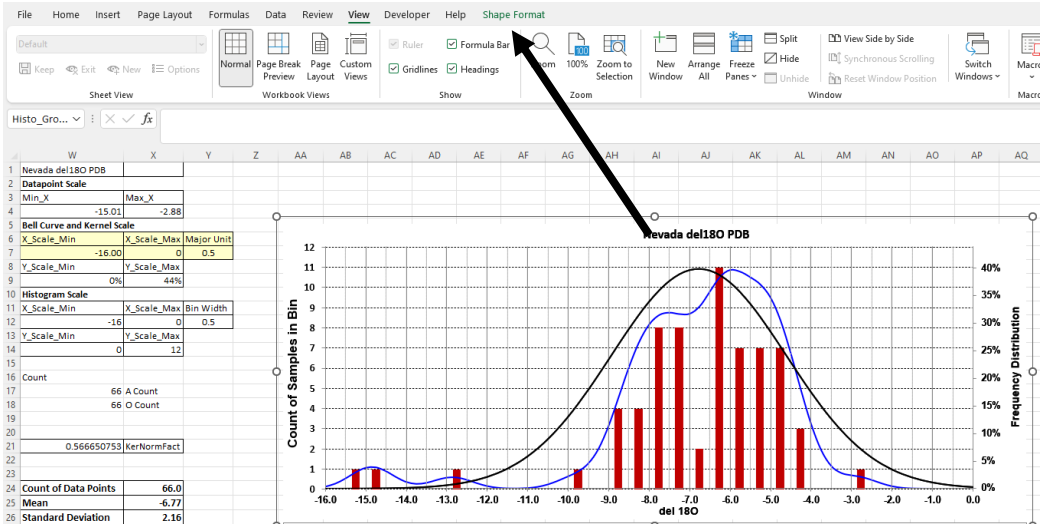


Histogram Tab

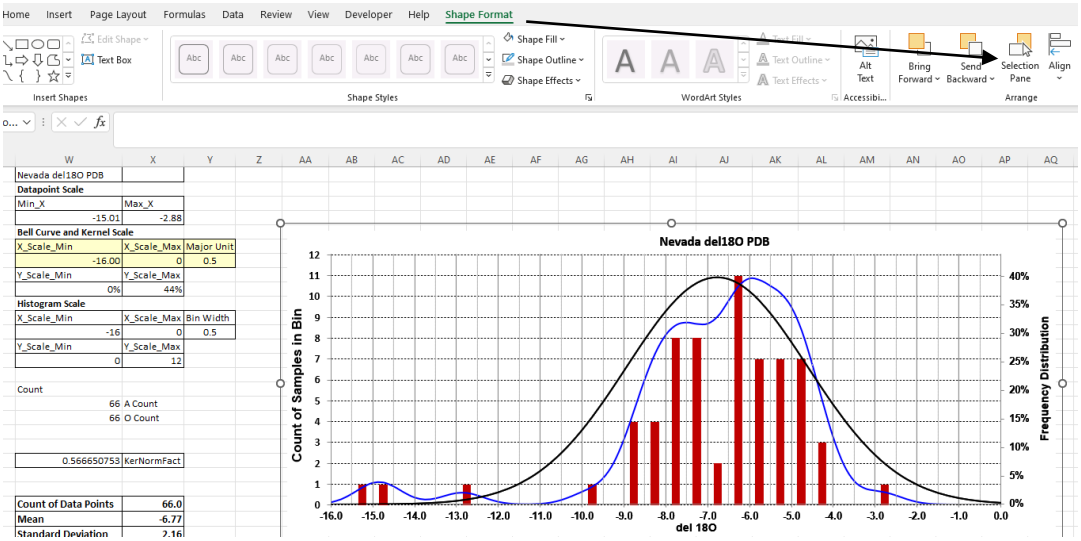
The next step is to go to the Histogram worksheet tab. Raw data is automatically entered into Columns A, B, and C, which trigger a series of calculations in columns D through K. Do not change any entries into these cells.

The next step for the histogram is to set scales for each of the different plots on the page. This is a bit complicated.

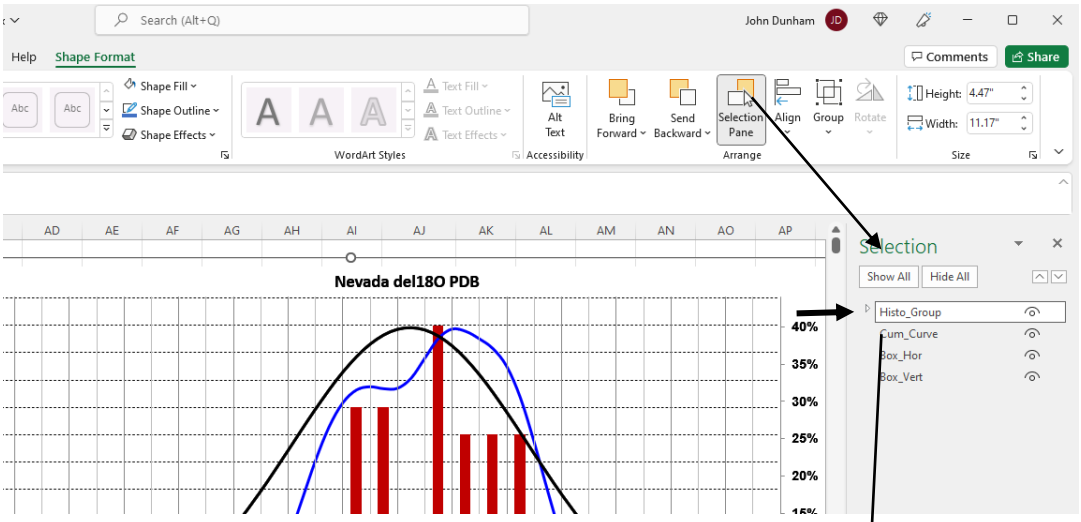
First, click on the graph in column AA. An underline appears under Shape Format at the top right.



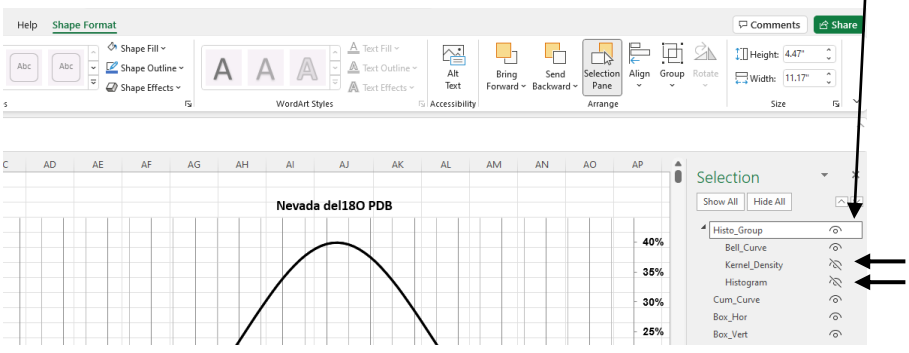
Click on the underlined Shape Format words, and a new set of icons appears on the top right. Locate “Selection Pane” in the middle of the Arrange Group.



Click on Selection pane, and a “Selection” list appears.



The first item on the list is “Histo_Group”. Click on a small triangle to the left of the words “Histo_Group”, and a drop-down list appears with Kernel_Density, Bell_Curve, and Histogram; each has an “eyeball” icon on the right side.



The Histo_Group consists of three overlain charts. The task is to scale each of these charts, which requires several steps that start with turning off the Kernel_Density chart and Histogram chart by clicking on the eyeball icon; this turns off these graphs, leaving only the Bell Curve.

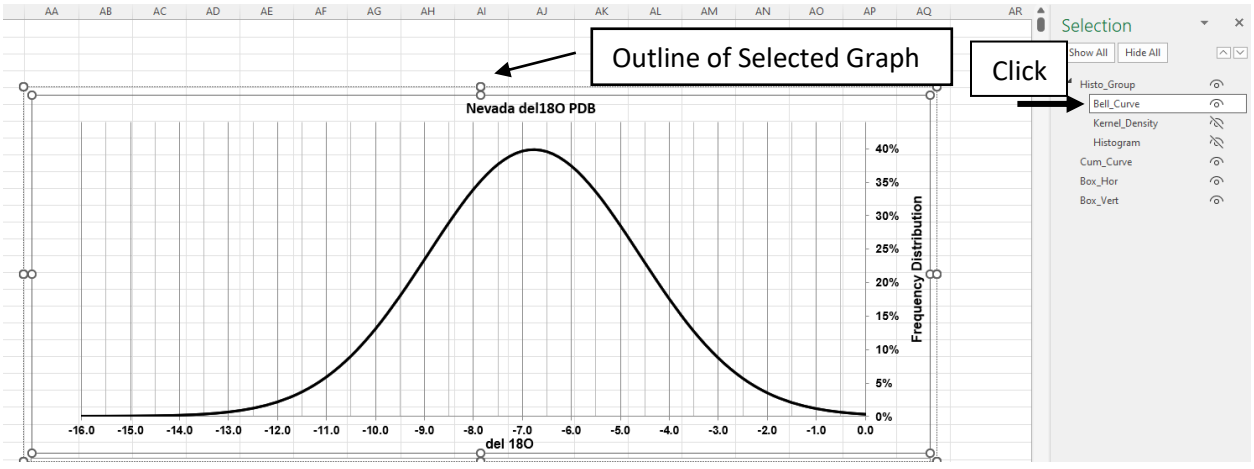
The picture above shows slashes through the eyeballs indicating they are turned off, and only the Bell Curve remains.

Worksheet columns W, X, and Y, contain a table for setting the X and Y axis scales for these plots. The Datapoint Scale Table shows your raw data's minimum and maximum X values. The Bell Curve and Kernel Scale set the X scale minimum value to the first whole number less than the Raw Data Minimum. I set the X scale maximum value to zero, but it could have been set as low as -2.0, the first whole number above the maximum Raw Data Maximum. Your choice of scales might be different, especially if you wish to compare a set of histograms with the same scales. The point is that you can enter any X minimum and X maximum values that you want into this table. Once set, you will use these same values to set the X-axis scales of all three of the overlain charts.

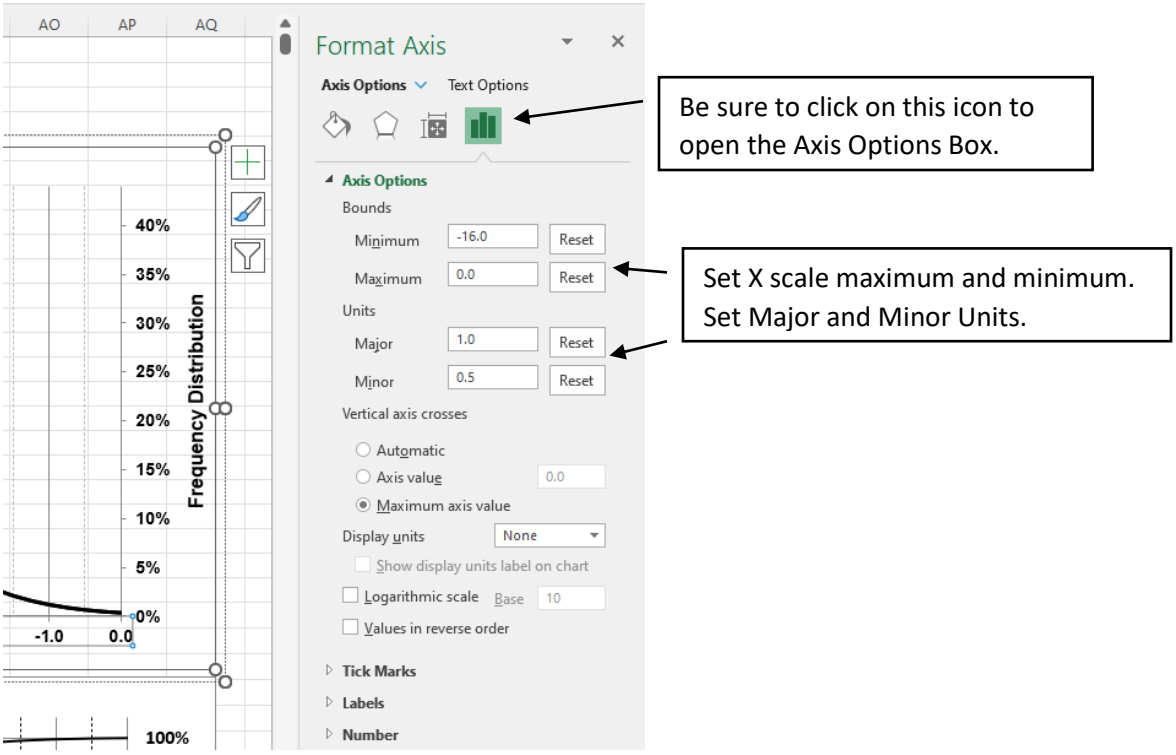
The spreadsheet sets the Bell Curve Y axis maximum by calculating the maximum percentage height of the Bell Curve, and then adding 10% to the max percentage so that there is a small space between the top of the Bell and the top of the graph.

W	X	Y
Nevada del18O PDB		
Datapoint Scale		
Min_X	Max_X	
-15.01	-2.88	
Bell Curve and Kernel Scale		
X_Scale_Min	X_Scale_Max	Major Unit
-16.00	0	0.5
Y_Scale_Min	Y_Scale_Max	
0%	44%	

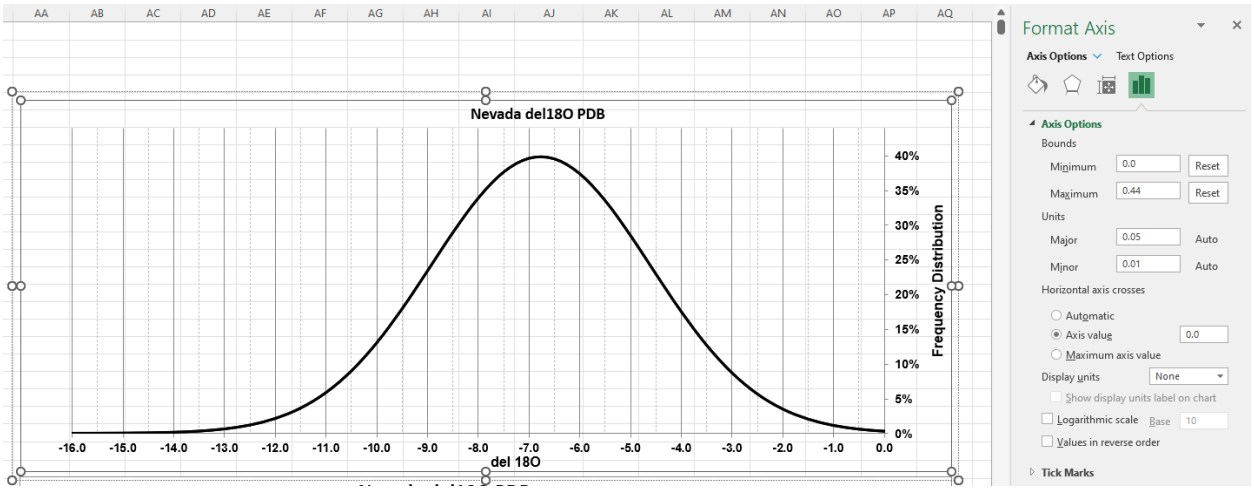
Next, you set the X and Y scales on the Bell Curve. On the Selection Pane, click on the Bell_Curve name; a selection box appears surrounding the Bell Curve graph.



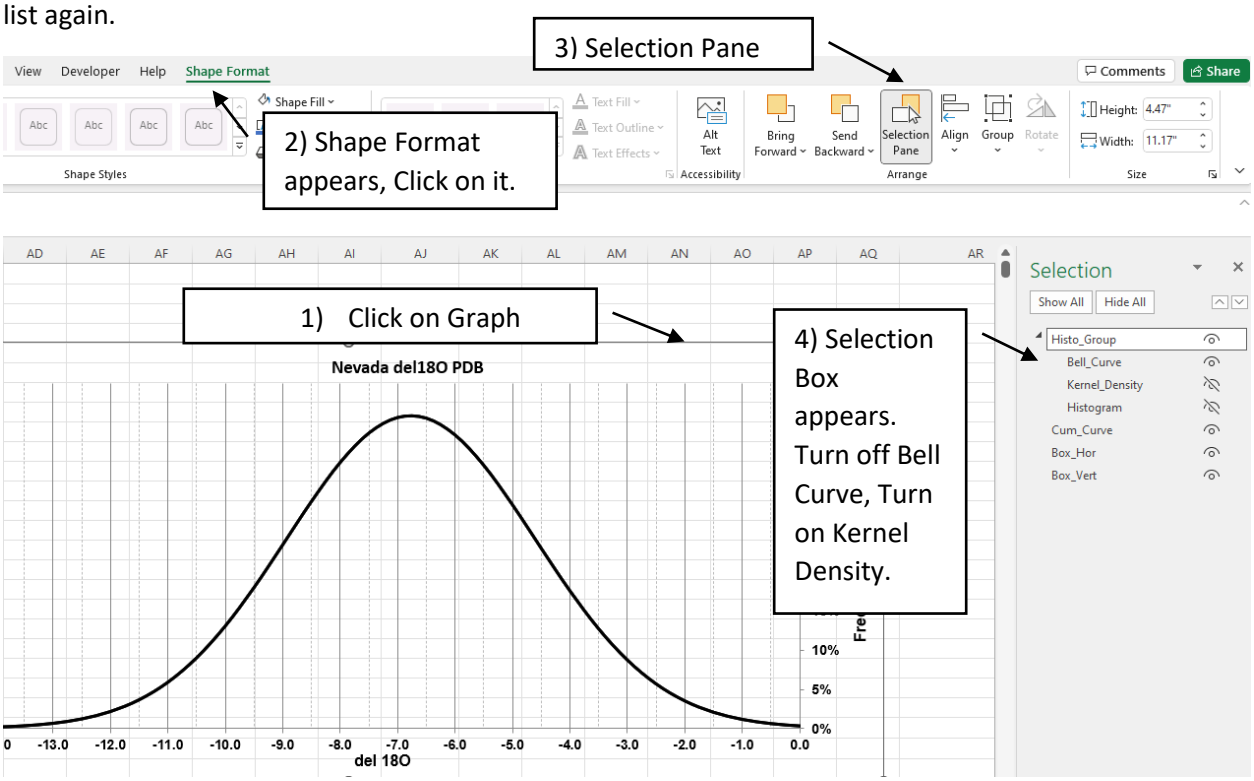
Next, double-click on the X-axis numbers on the graph; a set scale dialog box appears. If you don't see the Bounds Minimum and Maximum boxes, click on the Bar-Graph Icon shown below. Then, the Bounds will appear. As noted above, I set the minimum to -16 and the maximum to 0.0. Note that I set the Major unit to 1.0 because the axis numbers are close to overposting if I set the units to 0.5. Depending on the size and scale of your graph, you could set the Major unit to any interval that looks good on your graph. Note that the Minor Units are set to 0.5; this posts dashed minor gridlines at 0.5 unit intervals between the major gridlines. You are now finished setting the Bell Curve X-axis scale.



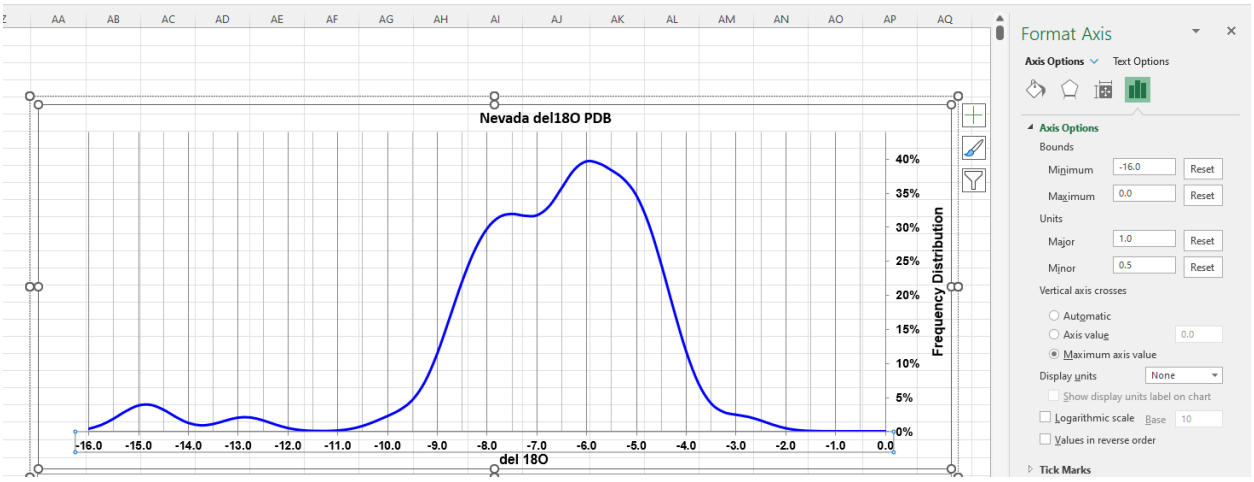
Next, double-click on the Y-axis numbers, the Y-axis format box opens. Set minimum to 0 and maximum to 0.44, major unit to 0.05, and minor unit to 0.01. This creates the final version of the Bell Curve.



Next, we set the Kernel Density Curve Scale. Click on the Bell Curve Graph, then, go back up to click on Shape Format, then Selection Pane, and this re-opens the Selection Box, where we see the Histo_Group list again.



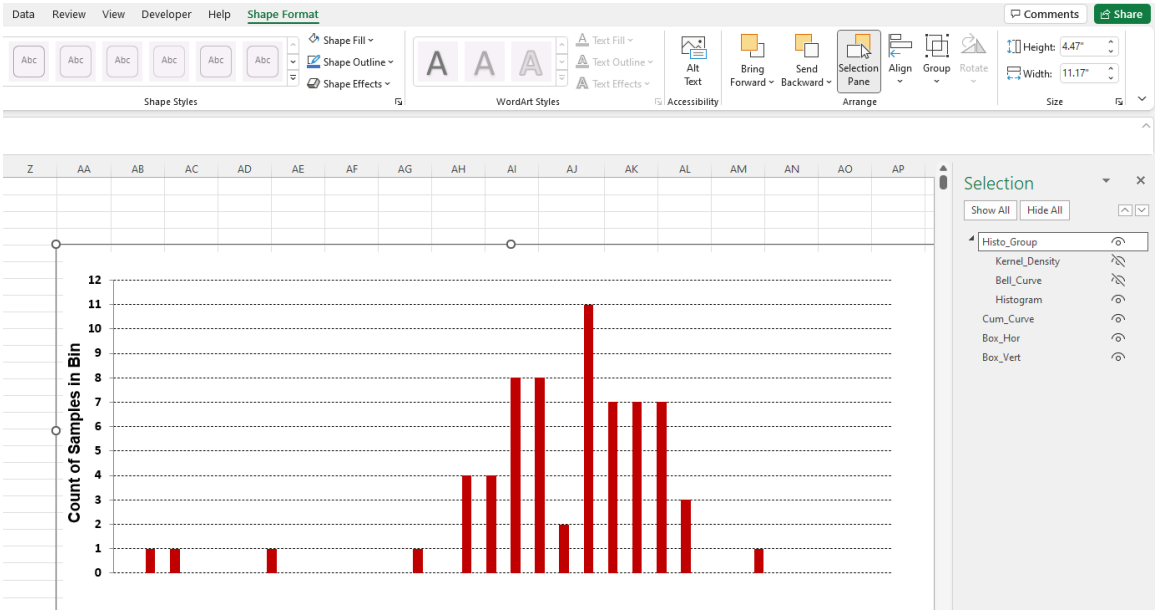
Turn off the Bell curve, turn on the Kernel Density curve. Set the X and Y axes to the same values that you used to set the Bell Curve scales.



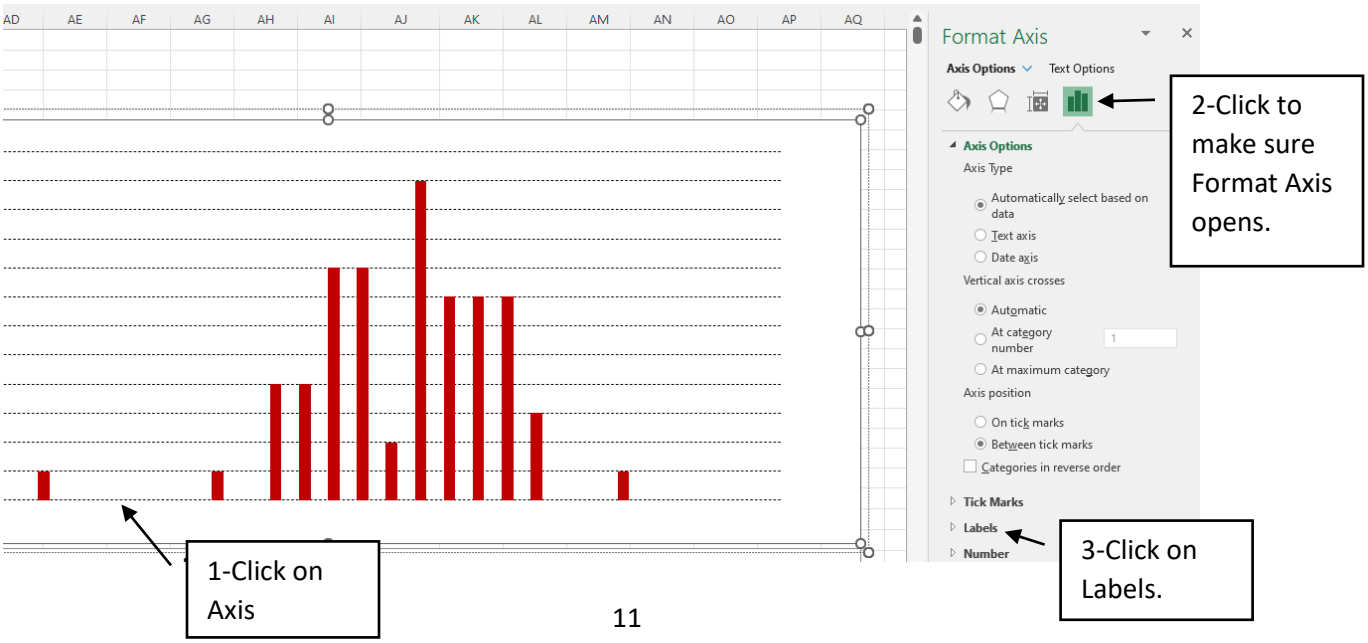
The Kernel Density Curve differs from the Bell Curve in that the Kernel looks at every data point, while the Bell Curve is based only on the mean and standard deviation of the Raw Data set; it doesn't see the individual data points. The name Kernel refers to the "Seed" used to make the smoothing function. In this example, the seed is a Gaussian Kernel, which creates a miniature bell curve around each data point and then stacks up or integrates the curves at each data point. A critical factor that influences the shape of both the Kernel Density Curve and the histogram is the choice of Bin Width. As you will see in the

Histogram Graph, each bar corresponds to the number of data points that lie within the boundaries of the bin. In this case, I have set the Bin Width to 0.5 units. For example, the height of the histogram bar and the height of the Kernel Density Curve correspond to the number of data points that lie between -8.0 and -7.5 per mil. The point is that the spreadsheet sets both the Kernel Density Curve and the Histogram to the same bin width; in this case, 0.5 units.

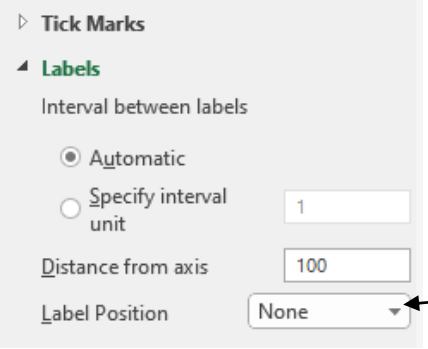
Finally, we set the scale for the histogram. Following the same steps as before, click on the graph, click on Shape Format, click on Selection Pane, turn off the Kernel Curve, and turn on the histogram.



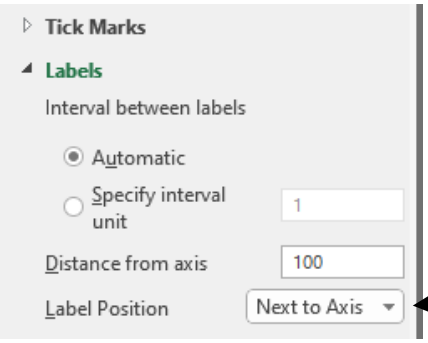
Note that there are no labels on the histogram X-axis; this is because I turned them off so they would not over-post when combined with the Bell and Kernel graphs. We will turn the histogram labels back on so we can see what we're doing. Double click on the bottom line of the graph, which is the X axis. When the Format Axis box appears, be sure to click on the Bar Graph Icon to ensure that the Format Axis box is displayed.



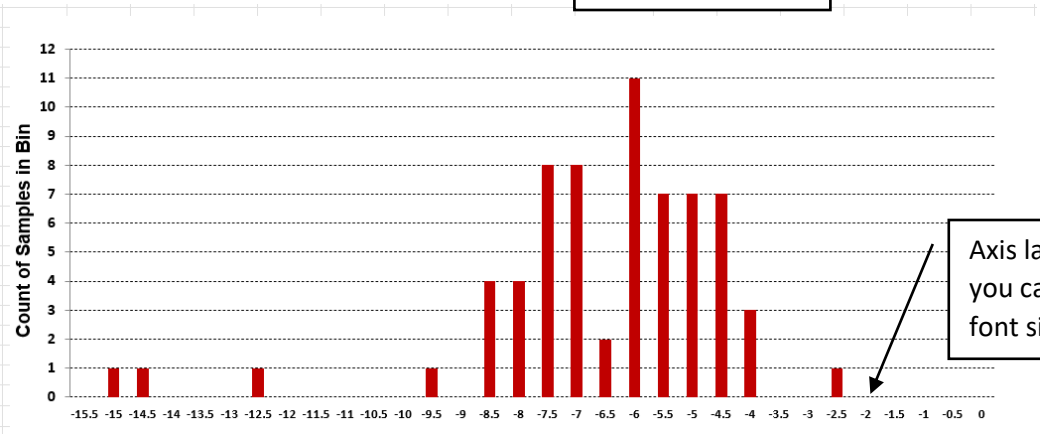
Click on the word “Labels”, which opens a dialog; the last line is the important one which reads “Label Position”; note that it is presently set to “None”. Click on None, and change to “Next to Axis”. Numbers appear on the X axis. Of course, leave the labels on if you intend to show the histogram by itself.



Click on None



Change to “Next to Axis”



Axis labels appear; you can increase the font size if desired.

The spreadsheet automatically counts the number of values within each bin based on the numbers you entered into Columns W, X, and Y.

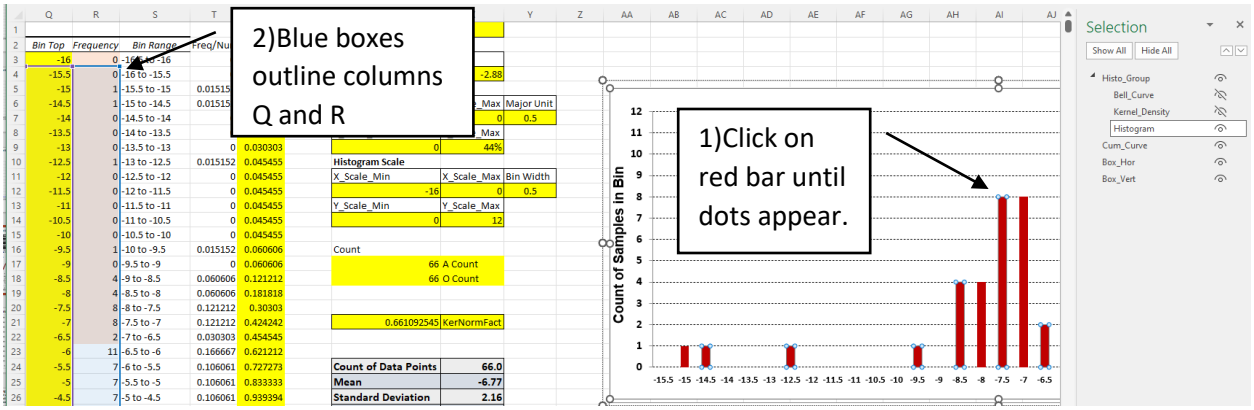
W	X	Y
Nevada del18O PDB		
Datapoint Scale		
Min_X	Max_X	
-15.01	-2.88	
Bell Curve and Kernel Scale		
X_Scale_Min	X_Scale_Max	Major Unit
-16.00	0	0.5
Y_Scale_Min	Y_Scale_Max	
0%	44%	
Histogram Scale		
X_Scale_Min	X_Scale_Max	Bin Width
-16	0	0.5
Y_Scale_Min	Y_Scale_Max	
0	12	

Histogram X Scale automatically set from -16 to 0.0 based on Bell Curve scale, while Bin Width is a User Choice; here I used 0.5.

Histogram Y Scale set from 0 to 12, because the maximum bar height is 11. Add 1 to the maximum to give some space above graph peak.

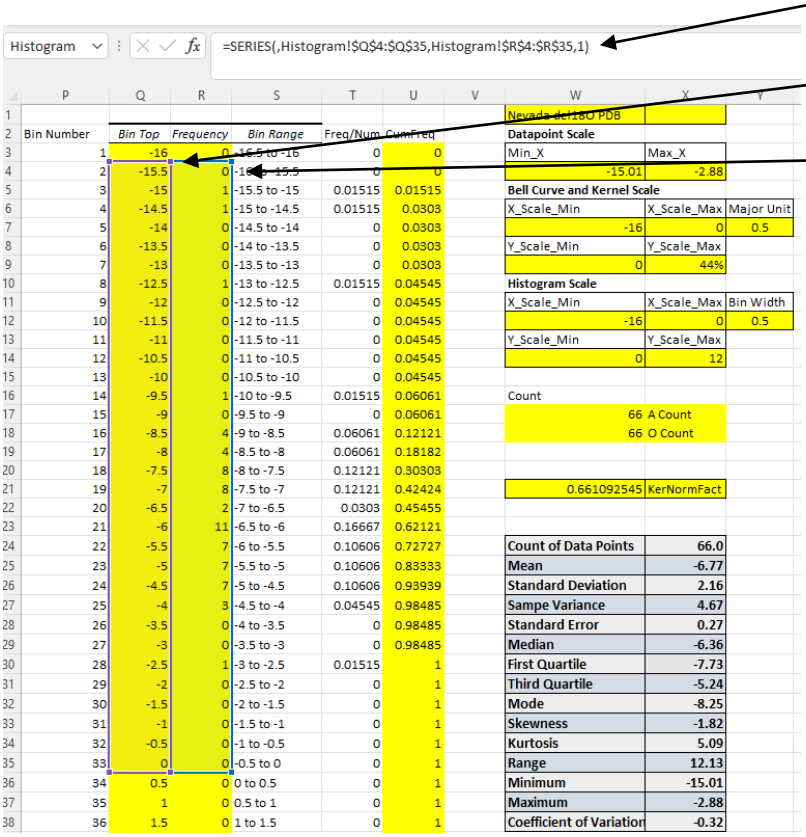
Bin-Width is an important consideration. If it is too narrow, the graph is too “spikey,”, if too broad, it is overly smoothed and obscures detail in the actual data distribution. In this case, I could have set the bin width to 0.25 units, 0.5 units, or 1.0 units. I chose 0.5 because it wasn’t too spikey while still showing detail.

Setting the histogram scale follows different procedures than the other graphs. First, click on any of the red bars on the histogram; keep clicking until a series of dots appear around the bars; this indicates the selection of the data series for the bars. Next, look at Columns Q and R of the spreadsheet; you will see blue outlines surrounding numbers in these columns.



The first number in column Q is the Minimum X value you put in Cell W7 for the Bell curve. The rest of the numbers in column Q increase in steps of 0.5 units, which is the Bin Width that you put in Cell Y12.

Each of the numbers in Column R corresponds to the number of samples that lie within the Bin Range shown in Column S. For example, 1 sample lies within the range 15.5 to -15, while 11 samples lie within the range -6.5 to 6.0 PDB. The spreadsheet does the counting. Now, you will adjust the histogram to capture the range of values between the minimum (-16) and maximum (0) values shown on the Bell and Kernel curves.



Bin Number	Bin Top	Frequency	Bin Range	Freq/Num	CumFreq
1	-16	0	-16.5 to -16	0	0
2	-15.5	0	-16 to -15.5	0	0
3	-15	1	-15.5 to -15	0.01515	0.01515
4	-14.5	1	-15 to -14.5	0.01515	0.0303
5	-14	0	-14.5 to -14	0	0.0303
6	-13.5	0	-14 to -13.5	0	0.0303
7	-13	0	-13.5 to -13	0	0.0303
8	-12.5	1	-13 to -12.5	0.01515	0.04545
9	-12	0	-12.5 to -12	0	0.04545
10	-11.5	0	-12 to -11.5	0	0.04545
11	-11	0	-11.5 to -11	0	0.04545
12	-10.5	0	-11 to -10.5	0	0.04545
13	-10	0	-10.5 to -10	0	0.04545
14	-9.5	1	-10 to -9.5	0.01515	0.06061
15	-9	0	-9.5 to -9	0	0.06061
16	-8.5	4	-9 to -8.5	0.06061	0.12121
17	-8	4	-8.5 to -8	0.06061	0.18182
18	-7.5	8	-8 to -7.5	0.12121	0.30303
19	-7	8	-7.5 to -7	0.12121	0.42424
20	-6.5	2	-7 to -6.5	0.0303	0.45455
21	-6	11	-6.5 to -6	0.16667	0.62121
22	-5.5	7	-6 to -5.5	0.10606	0.72727
23	-5	7	-5.5 to -5	0.10606	0.83333
24	-4.5	7	-5 to -4.5	0.10606	0.93939
25	-4	3	-4.5 to -4	0.04545	0.98485
26	-3.5	0	-4 to -3.5	0	0.98485
27	-3	0	-3.5 to -3	0	0.98485
28	-2.5	1	-3 to -2.5	0.01515	1
29	-2	0	-2.5 to -2	0	1
30	-1.5	0	-2 to -1.5	0	1
31	-1	0	-1.5 to -1	0	1
32	-0.5	0	-1 to -0.5	0	1
33	0	0	-0.5 to 0	0	1
34	0.5	0	0 to 0.5	0	1
35	1	0	0.5 to 1	0	1
36	1.5	0	1 to 1.5	0	1

Nevada del18O PDB	
Datapoint Scale	
Min_X	Max_X
-15.01	-2.88
Bell Curve and Kernel Scale	
X_Scale_Min	X_Scale_Max	Major Unit
-16	0	0.5
Y_Scale_Min	Y_Scale_Max
0	44%
Histogram Scale	
X_Scale_Min	X_Scale_Max	Bin Width
-16	0	0.5
Y_Scale_Min	Y_Scale_Max
0	12
Count	
66	A Count
66	O Count
0.661092545	KerNormFact

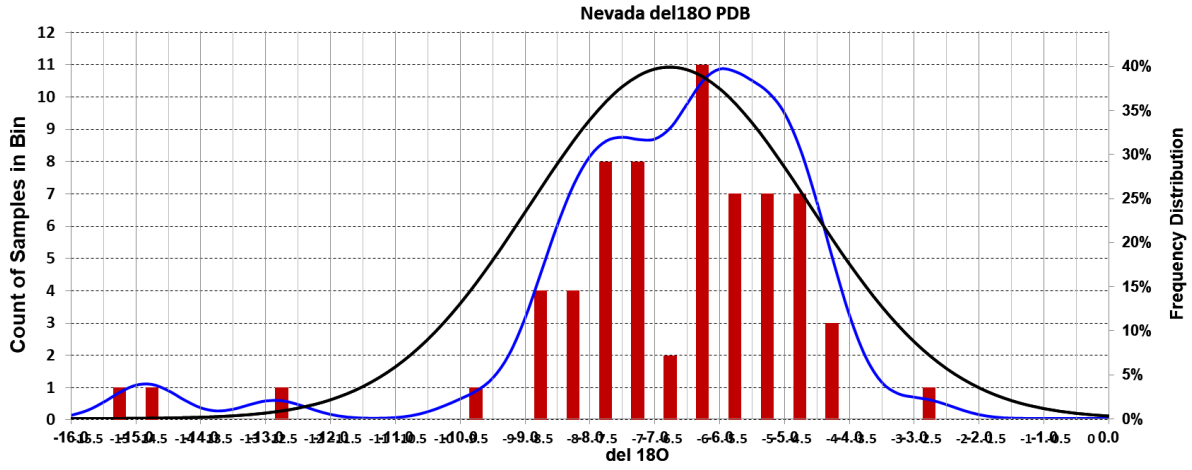
Count of Data Points	66.0
Mean	-6.77
Standard Deviation	2.16
Sampe Variance	4.67
Standard Error	0.27
Median	-6.36
First Quartile	-7.73
Third Quartile	-5.24
Mode	-8.25
Skewness	-1.82
Kurtosis	5.09
Range	12.13
Minimum	-15.01
Maximum	-2.88
Coefficient of Variation	-0.32

Click on Red Bar, and the Histogram data series appears.

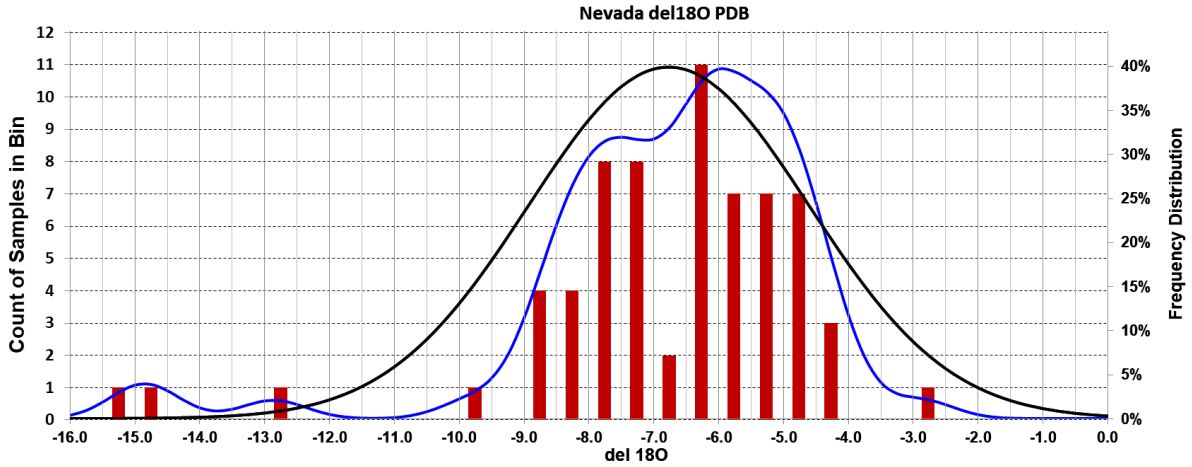
The first cells in the data range start at Q4 and R4. This corresponds to a bin range extending from a base at -16 and a top at -15.5. We did this because the base at -16 matches the minimum value of -16 on the Bell and Kernel charts.

Next we determine the last cells in the data range. The minimum value set for the Bell and Kernel charts was 0.0. Therefore, we want to have the last cell extend to a bin top of 0.0. That matches cells Q35 and R35. This is how you match your histogram scale with your Bell and Kernel data when you use your own data. Adjust the data range to match your other curves.

At this point, the Bell, Kernel, and Histogram charts all have the same scale. We go to the Selection Pane and turn on all three graphs. They look like this:

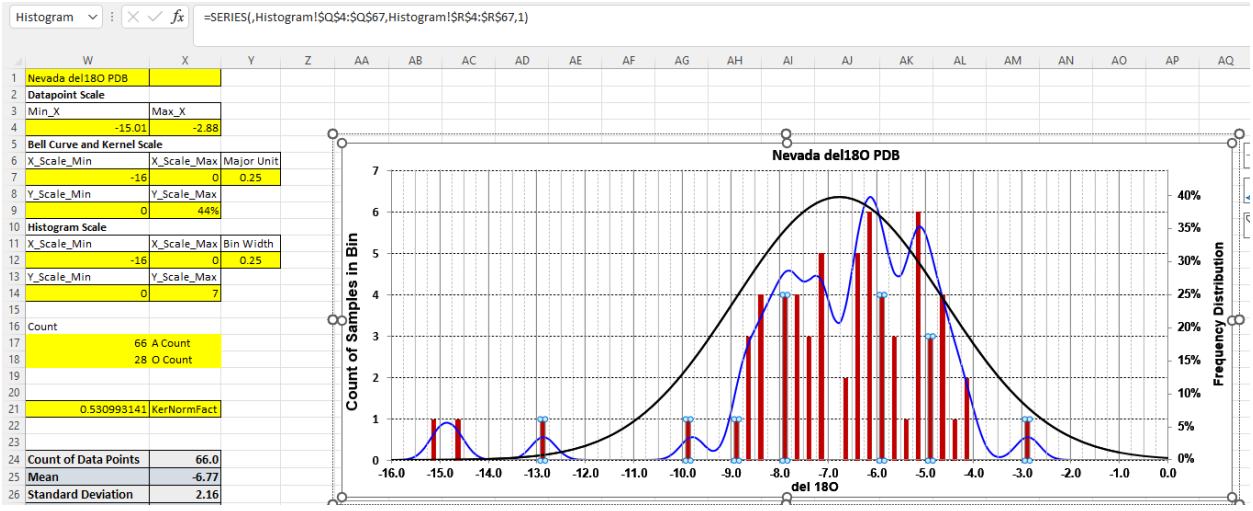


But note how the X-axis numbers are overposting. This is why it is necessary to turn off the labels on the histogram chart. We return to the selection pane, turn off the Bell and Kernel charts, and then double-click on the X axis of the histogram. The Format Axis box appears on the right; we go down to click on Labels and set Label Position to None. Turn the other charts back on, and we get this:



The dashed lines match the 0.5 unit wide histogram bin width. You could set the major unit to 0.5, but the numbers would overpost at this size. However, if you made the graph wider, that would solve the overpost problem. You can format these graphs in any way you want. You can turn each graph on or off at your discretion to show them separately or stacked in different ways.

But what if you wanted to use a different bin width than 0.5? No problem; just change the histogram bin width in cell Y12 to 0.25 (or anything else you want to try). Then, make sure to re-scale the histogram using the same methods described above. In the example below, clicking on the red bar shows the selected data range. Cells Q4 and R4 start the range at -16.0, while cells Q67 and R67 end the range at 0.0. Naturally, this is a broader range than when the bin width was set to 0.5. Also, the new bin width changes the height of the peak red bar; before, it was 11; now, it is 6. Following the same methods described above, the Y axis maximum scale is set to 6 + 1 extra equals 7. The result is shown below:



Note that this is “spikier” than the plot with bin width set at 0.5. There is always a balance between making a plot that shows the essential character of the data set without oversimplifying or overcomplicating the story. In this case, I chose the 0.5 bin width, which tells the story without too much spikiness.

This type of chart is used in scientific publications. The plot on the left shows a NASA study of the velocity of micro-meteoroids that might strike satellites like the Webb telescope; the mean is 21.4 km/sec = 77,000 kilometers per hour!! The plot on the right shows histograms of detrital zircon ages that are fitted with kernel density curves.

Examples of Histograms and Kernel Density Plots used in scientific publications:

Micrometeoroid Velocity: NASA, 2015

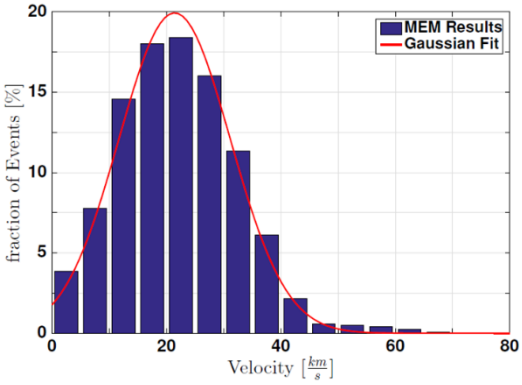
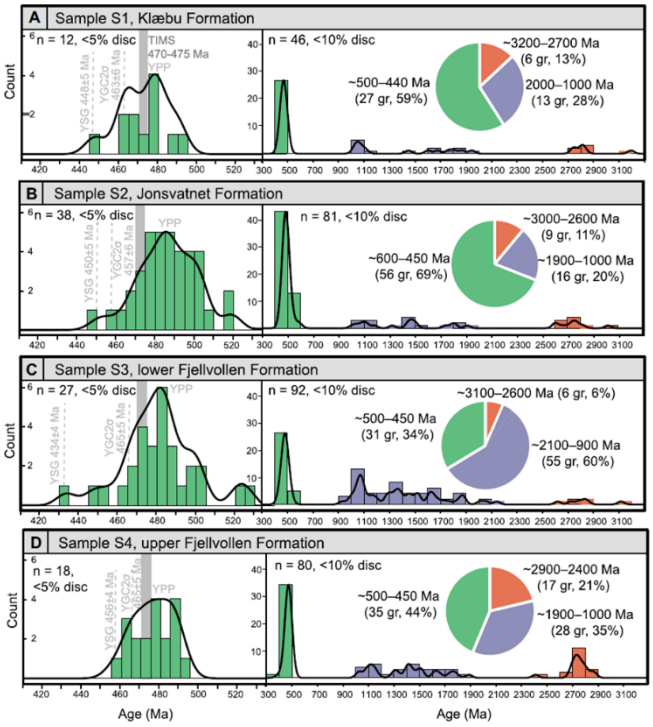


Fig. 3. Probability distribution of impact velocities for LPF micrometeoroid collisions during science operations. Histogram in blue are estimates from the NASA Meteoroid Engineering Model and a representative ephemeris for LPF. The fit in red is a best-fit normal distribution.

Detrital Zircon Ages: GSA Bull., 2022



Violin Plot Tab

The next tab is the Violin Plot, which combines the Box Plot and Kernel Density Plot with all the data points scattered within the box to prevent overposting. The scales are easier to set on this plot. The picture below explains the data series that make up the Violin Plot.

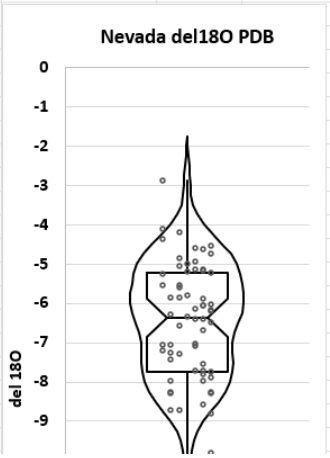
Columns A and B are labeled “Jitter Data Points”. Jitter is a standard method that randomly scatters data points in the horizontal direction, so they do not over-post one atop another. If you don’t like the scatter pattern, you can press the F9 key to change the scatter to a pattern that you like.

Column F automatically displays the Raw Data values that you entered earlier.

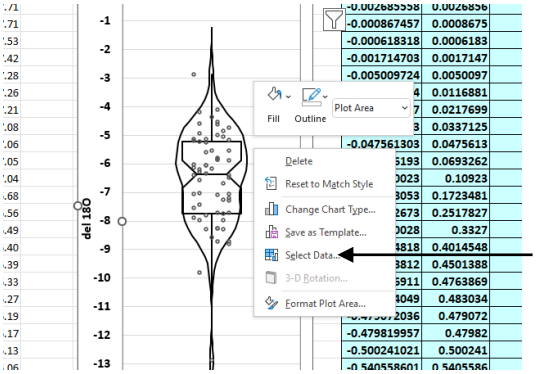
Columns H and I show the Minimum and Maximum scale values that you used to make the histogram, and the Bin Width used for the histogram.

Columns K, L, and M are the values used to make the Violin Plot. These are calculated from all the rest of the columns in the worksheet. You do not need to make any entries in any of these columns; the entries and calculations are all automatic.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Jitter Midline:	0				Data Name					Violin	Violin	Violin
2	-3	3			Label	18O PDB		Chart Parameters			Plot Left	Plot Right	Plot X
3	-3	-0.30000	Data Parameters			-15.01		n	66		-0.0063	0.0063	-16
4	-2	-0.20000	Mean	-6.77		-14.69		Chart min from histo.	-16		-0.0161	0.0161	-15.75
5	-2	-0.20000	StDev	2.16		-12.86		Chart max from Histo.	0		-0.0325	0.0325	-15.5
6	3	0.30000	Q3	-5.22		-9.79					-0.0522	0.0522	-15.25
7	3	0.30000	Q1	-7.76		-8.81		Bin Width from Histo	0.5		-0.0668	0.0668	-15
8	-1	-0.10000	IQR/1.34	1.89		-8.72		increment spacing	0.25		-0.0684	0.0684	-14.75
9	-2	-0.20000				-8.70		Kernel	Gausian		-0.0560	0.0560	-14.5
10	2	0.20000	Min	-15.01		-8.57					-0.0372	0.0372	-14.25
11	3	0.30000	Max	-2.88		-8.29		Data Max	-2.88		-0.0216	0.0216	-14
12	-2	-0.20000	n	66		-8.28		Left+Right Width	0.7	<Goal Seek to 0.7	-0.0153	0.0153	-13.75
13	3	0.30000				-8.25		Width Factor	3.0299035	<Change i13	-0.0187	0.0187	-13.5
14	-2	-0.20000				-8.25					-0.0277	0.0277	-13.25
15	-2	-0.20000				-7.97					-0.0354	0.0354	-13
16	2	0.20000				-7.97					-0.0358	0.0358	-12.75
17	3	0.30000				-7.88					-0.0283	0.0283	-12.5
18	2	0.20000				-7.81					-0.0174	0.0174	-12.25
19	3	0.30000				-7.74					-0.0083	0.0083	-12
20	1	0.10000				-7.71					-0.0031	0.0031	-11.75
21	2	0.20000				-7.71					-0.0010	0.0010	-11.5
22	2	0.20000				-7.53					-0.0007	0.0007	-11.25
23	-2	-0.20000				-7.42					-0.0020	0.0020	-11
24	-1	-0.10000				-7.28					-0.0058	0.0058	-10.75
25	-2	-0.20000				-7.26					-0.0136	0.0136	-10.5
26	-3	-0.30000				-7.21					-0.0254	0.0254	-10.25
27	1	0.10000				-7.08					-0.0393	0.0393	-10
28	-2	-0.20000				-7.06					-0.0555	0.0555	-9.75
29	-3	-0.30000				-7.05					-0.0809	0.0809	-9.5
30	1	0.10000				-7.04					-0.1274	0.1274	-9.25
31	2	0.20000				-6.68					-0.2011	0.2011	-9
32	-1	-0.10000				-6.56					-0.2937	0.2937	-8.75
33	3	0.30000				-6.49					-0.3882	0.3882	-8.5



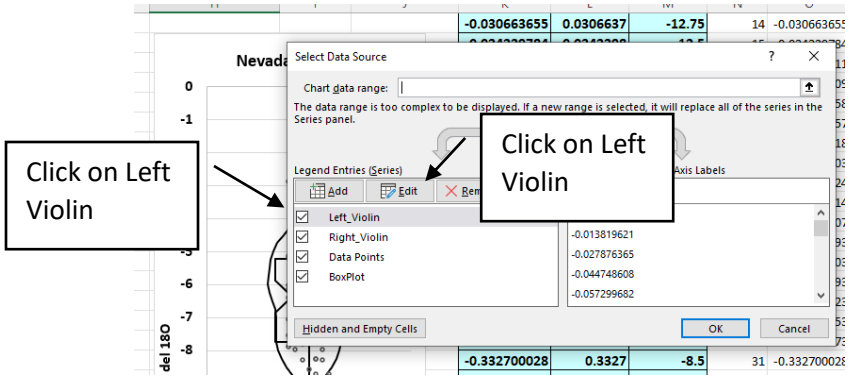
To set the scales for the Violin Plot, click on the plot, then right-click until a box opens that says “Select Data”.



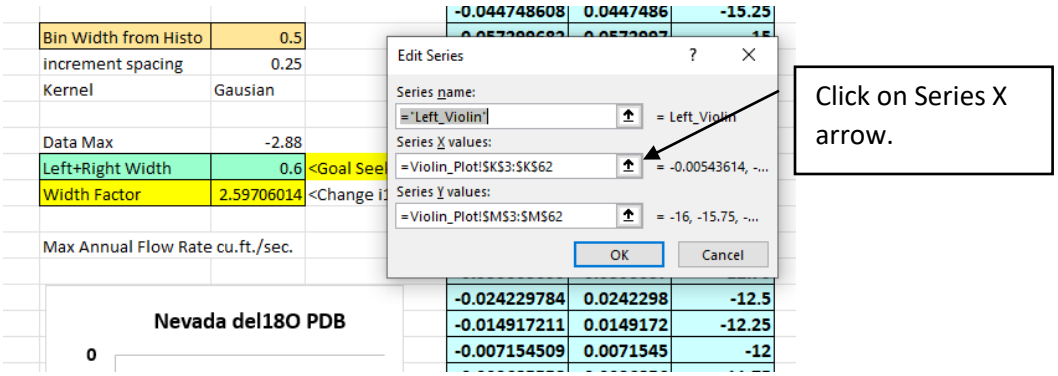
Right click on graph to open this box, then, click on Select Data. The picture below shows the Select Data box.

The select data box opens to show four different data series that make up the Violin Plot. You only have to set the Left and Right Violin data ranges; the ranges for Data Points and Box Plots are selected automatically.

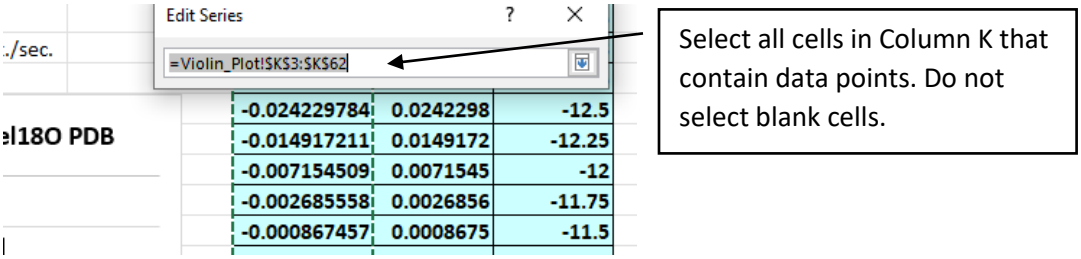
First, click on Left Violin, then Click on Edit.



The Edit Series Box opens:

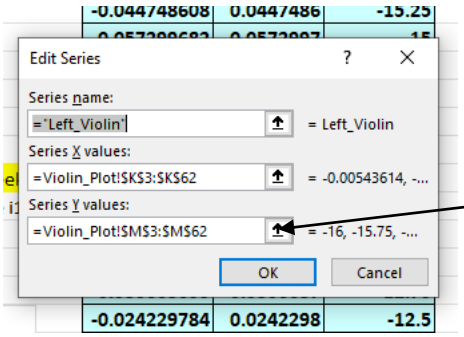


Click on the arrow to the right of Series X values, the Edit Series box opens:



The most important thing here is that the first cell starts at K3, while the last cell stops at the last data value in Column K. In this example, K62 is the last cell with data. Your raw data may have fewer or more data points than this example, in which case, the data range for the Left Violin will stop above or below cell K62. If you select any blank cells in the data range, the plot will not work. Be sure to select all cells that have data in Column K, and be sure not to select any blank cells.

Next, click on the arrow next to Series Y values: select all cells in Column M that contain data points.

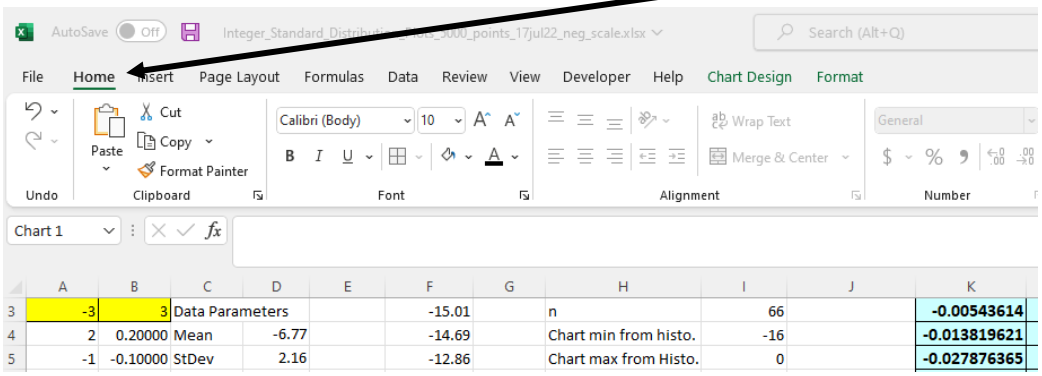


Select all cells in Column M that contain data points. Do not select blank cells.

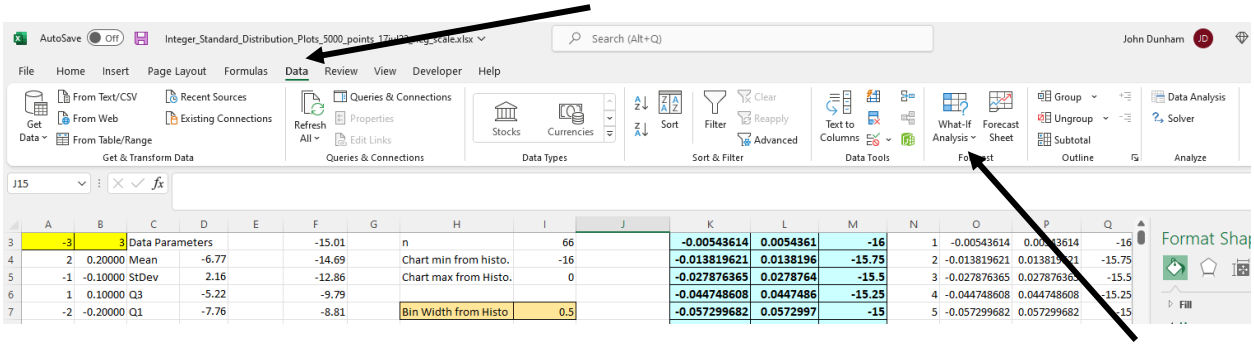
Next, repeat this process for Right Violin, which has X data in Column L, and Y data in Column M. Do not select any blank cells in these columns.

The final task is to define the width of the violins, so that they are close to the maximum width of the Box plot. We use "Goal Seek" to set this width.

The Goal Seek routine is hard to find if you don't know where to look. Note on the picture below that several names are listed side by side at the top of the worksheet; it is likely that you are located on the "Home" Screen, which is indicated by the line beneath the word Home.

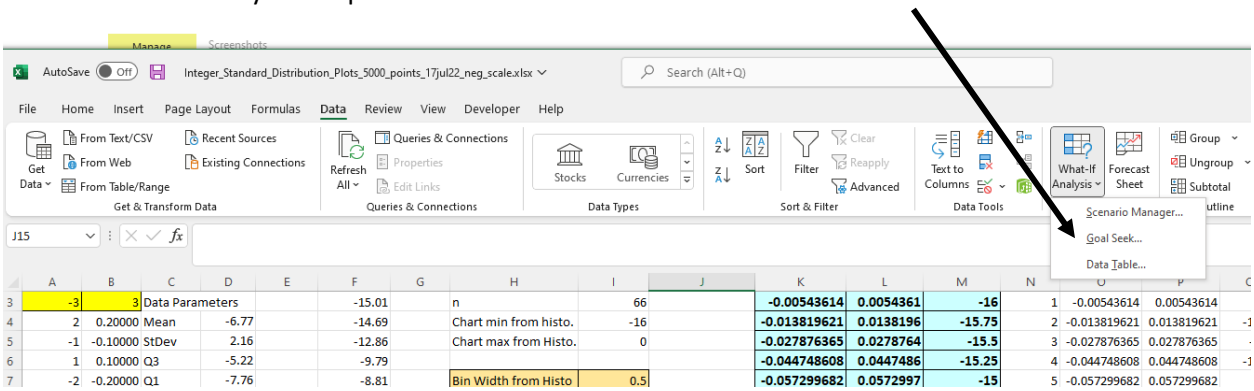


Goal Seek is located on the Data Screen, click on data, and an underline will appear under data.

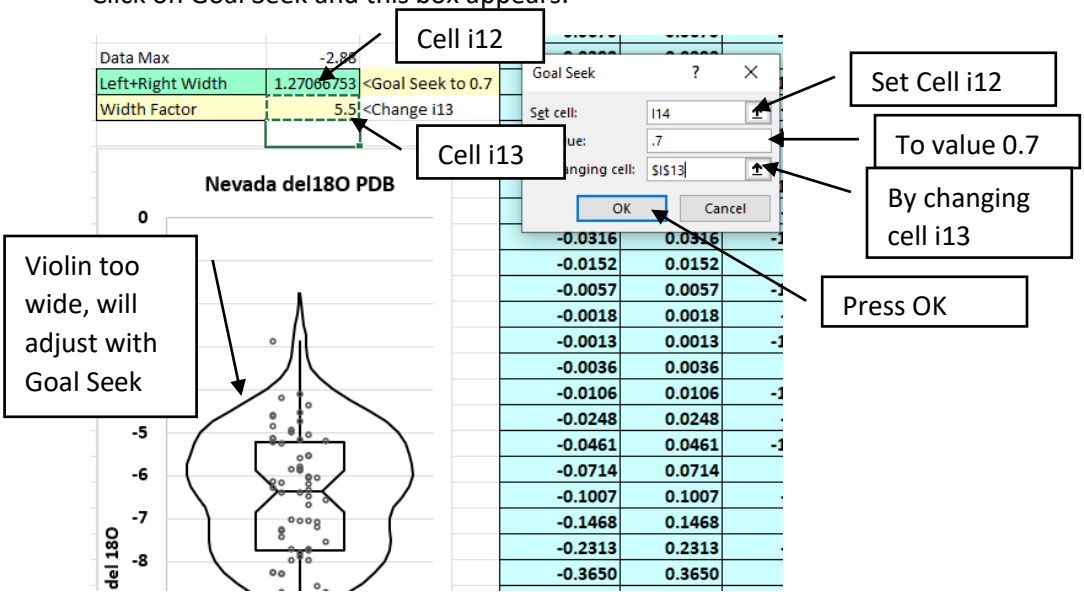


Once on the Data Screen, look at the series of names and icons beneath the top line. Note "What if Analysis" on the right side of the screen.

Click on What If Analysis to open this box. Note Goal Seek on the second line.



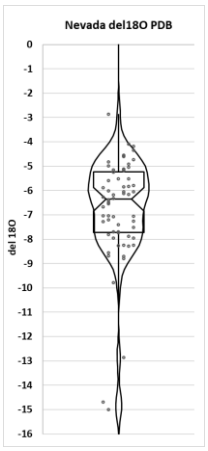
Click on Goal Seek and this box appears:



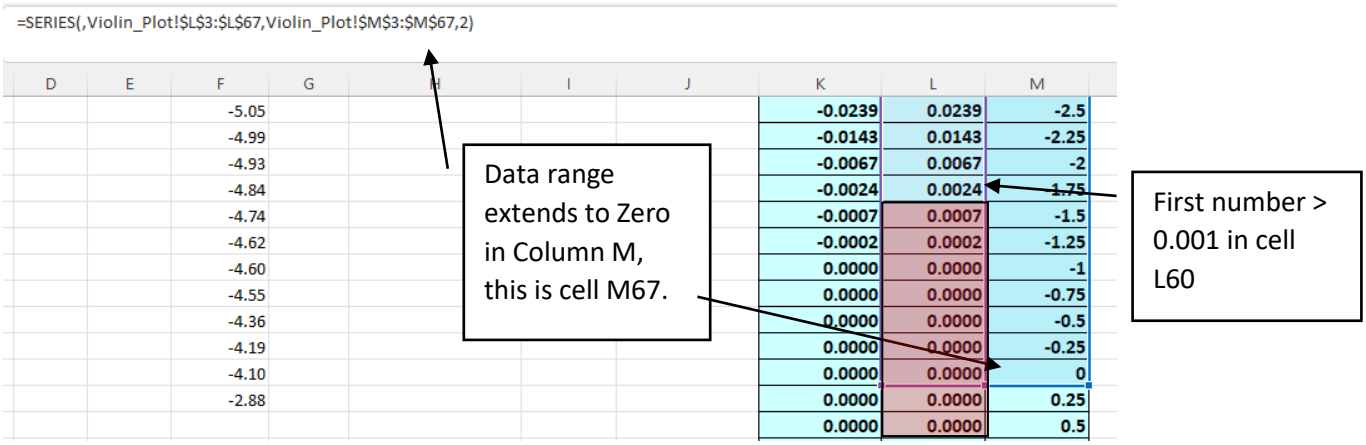
Goal Seek changes the Width Factor so that the Violin Width matches the Box Plot Width. Of course if you want a wider plot, you just manually increase the width factor until you get something you like.

Data Max	-2.88	
Left+Right Width	1.27066753	<Goal Seek to 0.7
Width Factor	5.5	<Change i13

The plot below has a narrower width based on Goal Seek. However, note the long straight line between -2 and Zero; this is the region where the left and right kernel density frequency curves decrease to zero percent.



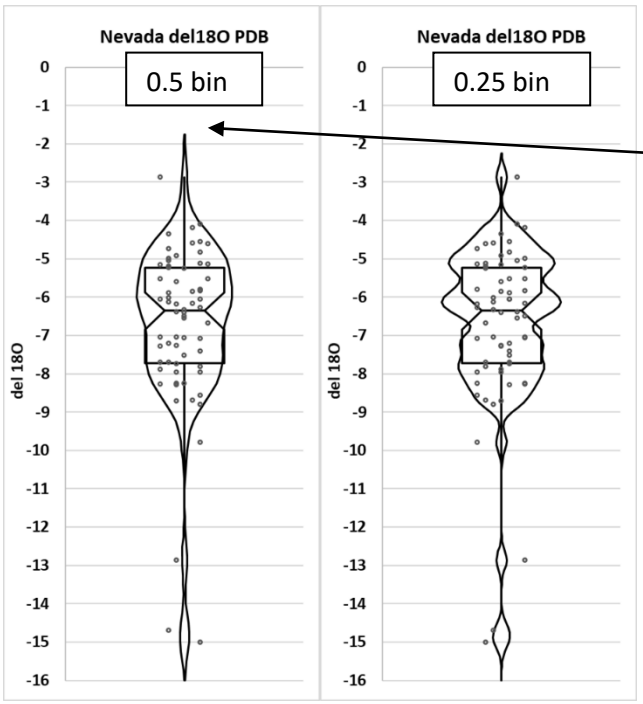
You should “trim” the left and right violin curves to prevent a long straight Zero line at the top and bottom of the curves. In this example, the base of the data range is at Zero in Column M.



D	E	F	G	H	I	J	K	L	M
		-5.05					-0.0239	0.0239	-2.5
		-4.99					-0.0143	0.0143	-2.25
		-4.93					-0.0067	0.0067	-2
		-4.84					-0.0024	0.0024	-1.75
		-4.74					-0.0007	0.0007	-1.5
		-4.62					-0.0002	0.0002	-1.25
		-4.60					0.0000	0.0000	-1
		-4.55					0.0000	0.0000	-0.75
		-4.36					0.0000	0.0000	-0.5
		-4.19					0.0000	0.0000	-0.25
		-4.10					0.0000	0.0000	0
		-2.88					0.0000	0.0000	0.25
							0.0000	0.0000	0.5

Click on the right curve and the Right_Violin_Plot data range will appear in the formula line. Scroll down to look at the numbers in Column L, which become very small as Column M approaches Zero. Counting up from Zero in Column M, look for the first number in column L that is greater than 0.001. I used conditional formatting to fill red-color whenever values are less than 0.001. In this example, the first number greater than 0.001 is in cell L60 where the number is 0.0024. In the formula bar, change the data range from L67 to L60, and change M67 to M60. This trims off the long zero tail. Now, click on the Left_Violin and change its range to stop at cells K60 and M60.

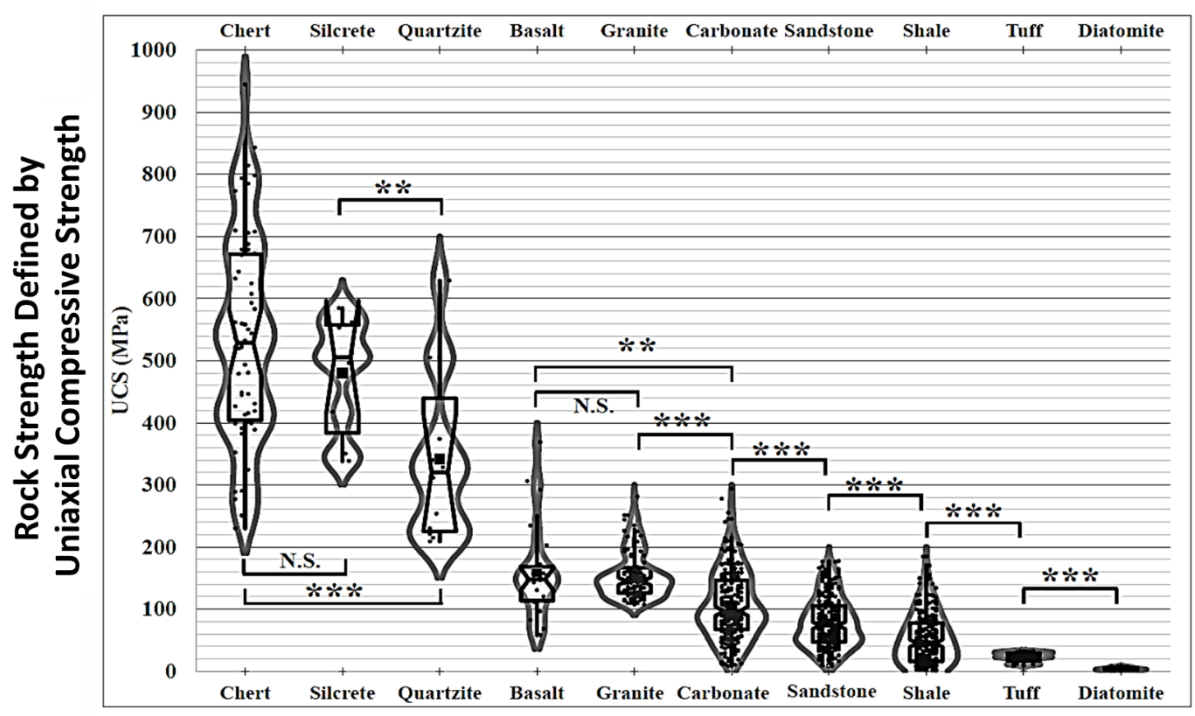
If you did decide to use the 0.25 unit bin width for the histogram, you will find that your Violin Plot changes to a 0.25 bin width and is spikier than the 0.5 bin width violin.



No need to show a long zero tail. Just trim the tail as described above.

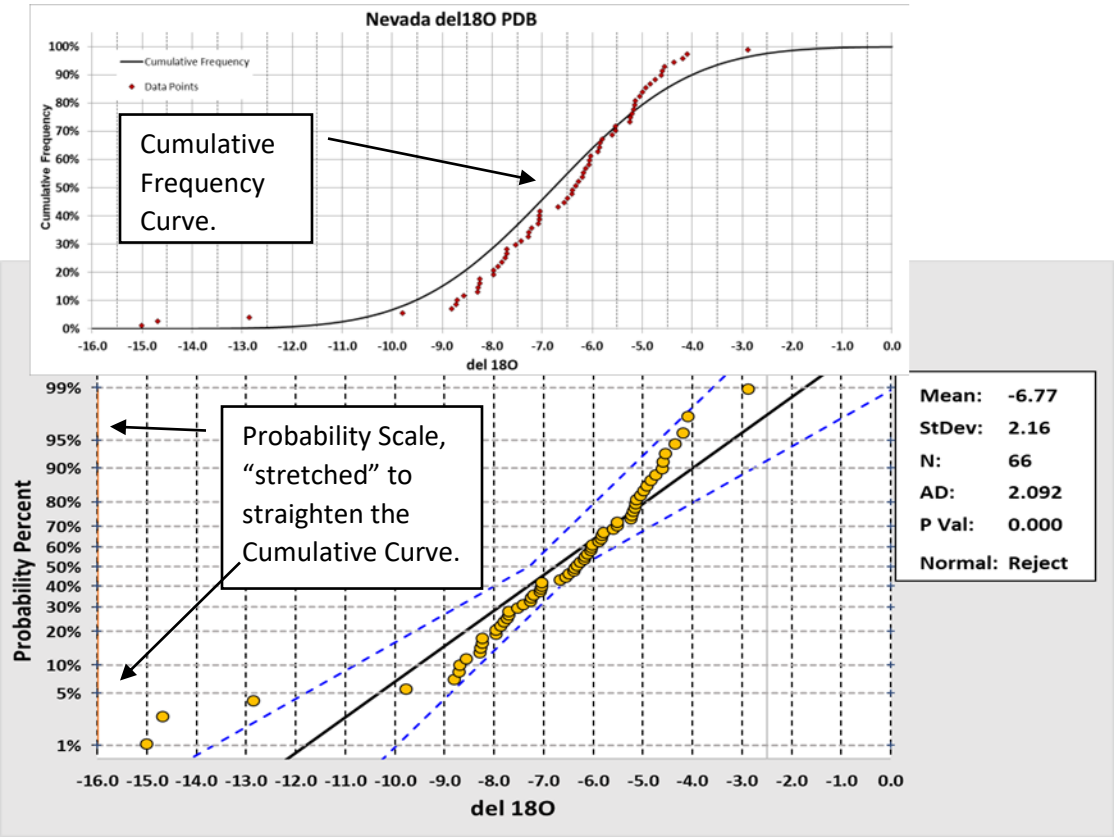
In my opinion, I would stick with the 0.5 bin width.

You have completed scaling the Violin Plot. These are especially useful for side by side comparisons as shown by this example: *** indicates means are different at 99% confidence limit; ** means different at 95% confidence, N.S. indicates no significant difference between mean values. The mean difference test is done with the Excel Data Analysis Toolpak, t-Test: Two_Sample Assuming Unequal Variances.



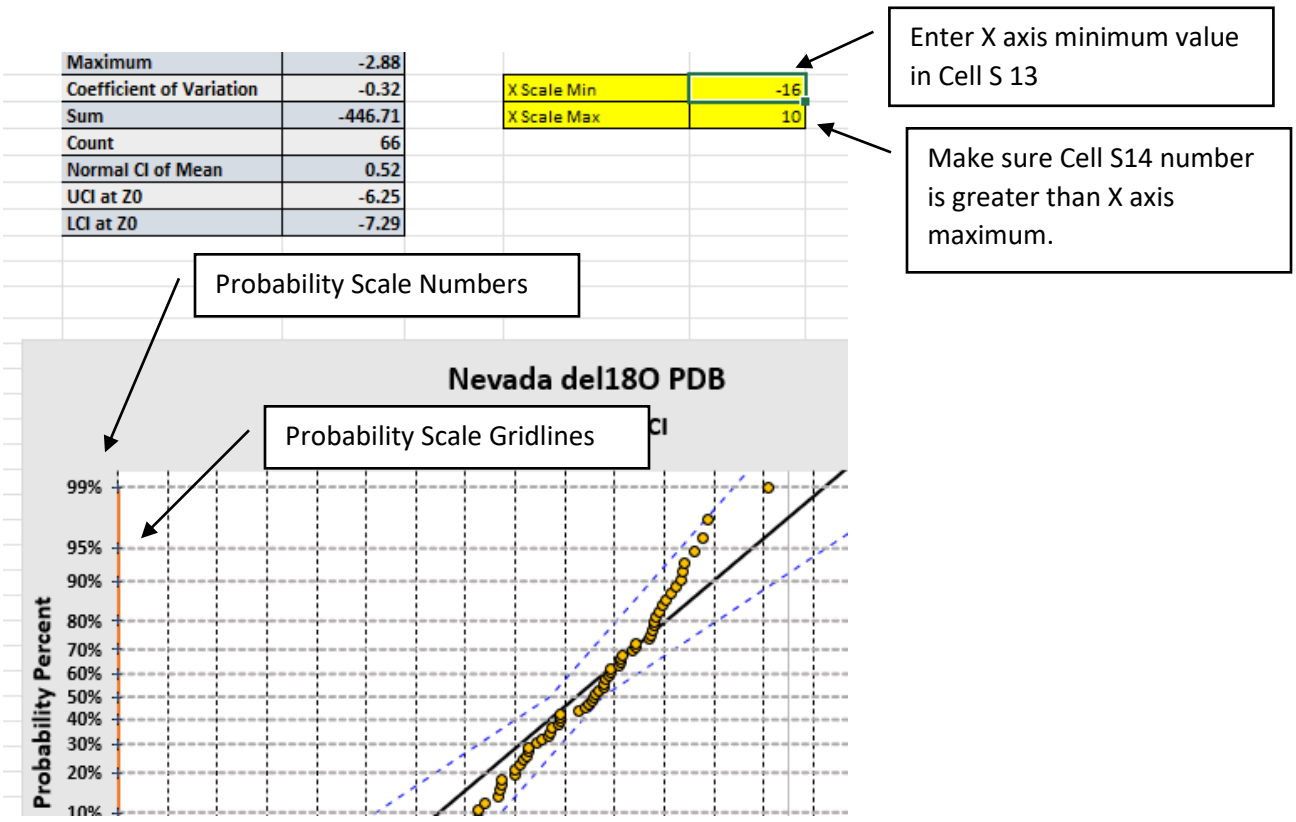
Probability Plot Tab

Raw data is automatically entered into Column A, and all the rest of the calculations are automatic. The Y-axis Probability Scale is stretched at the tails such that a Cumulative distribution curve is bent into a straight line.



The black line shows the trend of a Normal (Gaussian) distribution with the Mean and Standard Deviation of the Raw Data set. The blue dashed lines show the 95% confidence intervals for the actual location of the population probability line. Note that the outliers plot outside the confidence intervals, which are wide apart, thus indicating considerable uncertainty in the location of the population line. The summary box has an entry labeled AD; this is the Andreson-Darling Statistic, which tests whether the data come from a normally distributed population. The P-Value is calculated from the AD statistic; if P is less than 0.05, there is a 95% chance that this data set does not come from a normal distribution. The P-Value is very small in this case, and the Normal Distribution hypothesis is rejected. Commercial software packages also include AD tests on probability plots.

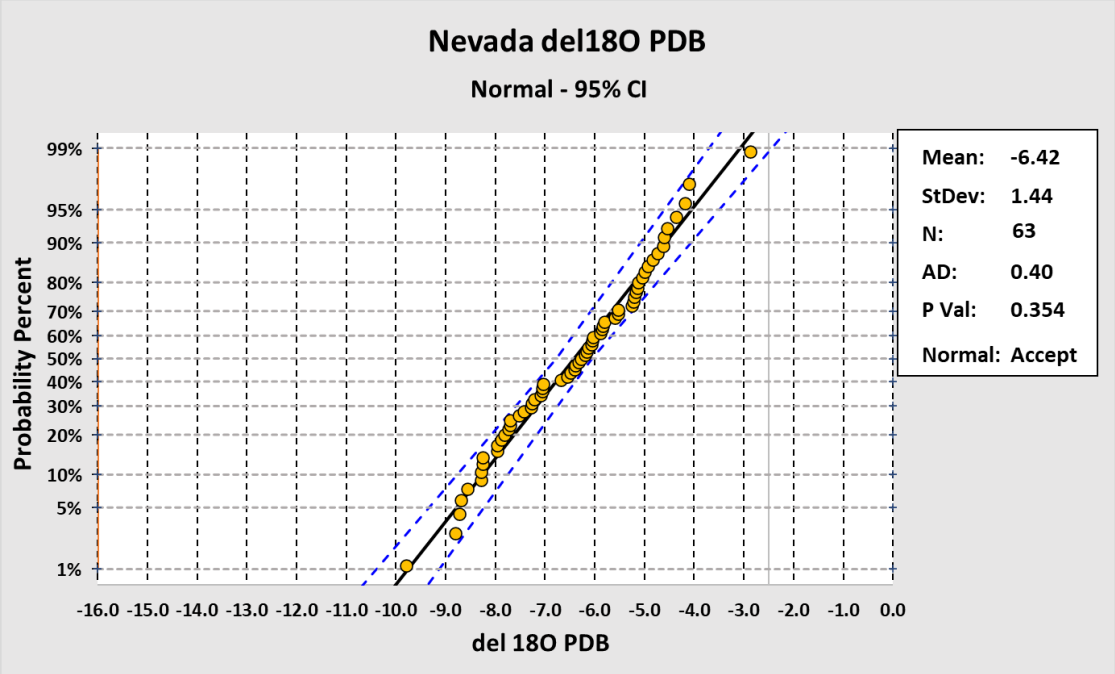
You have to set the X-axis scale. Click on it, and ensure it matches the ranges you used for the plots on the Histogram tab. Also, you have to enter X minimum value into cell S13, ensuring that the Probability Percent Scale Numbers post correctly on the left Y axis of the plot. If you don't see these numbers, double-check to ensure that the X minimum is correctly entered into cell S13. Also, make sure that the number in cell S14 is greater than the X axis maximum value in order to get gridlines to display properly.



Truncated_Prob_Plot Tab

I suspect that these data fail the normality test because of the presence of outliers that come from a different population than the rest of the data. Therefore I will see what happens if I delete the three outliers and replot the truncated data. Whenever outliers are present, it is essential to try and determine what factors might influence the extreme values. Possibilities include mislabeling of samples, errors in analysis, typographical errors in data tables, or valid measurements that simply come from a different population than the rest of the samples. In the case of these Nevada carbonates, I can confirm that the outliers come from samples that contained thin veins of white sparry calcite that filled tectonic fractures in the limestones. The remaining samples were microcrystalline calcite in lime-mudstone to wackestone textured rock that did not contain tectonic fractures with calcite fracture-fills. I interpret the fracture fills as late diagenetic products that formed during burial at higher temperatures and different conditions than the microcrystalline matrix calcite in the non-fractured lime mudstones.

The data in column A on this worksheet are not automatically entered from the Raw Data sheet. Rather, I simply copied cells A6 to A68 on the Raw Data tab, and pasted them into Column A on the Truncated_Prob sheet. This gets rid of the 3 outliers. Whenever you paste data into a worksheet, make sure there are no “left-over” cells in the column where you just pasted your data. Next, you have to manually set the count number in Column B. In this case, the count stops at number 63, corresponding to 18O value -2.88 per mil. Make sure there are no extra numbers in the Count Column. The result is shown below:

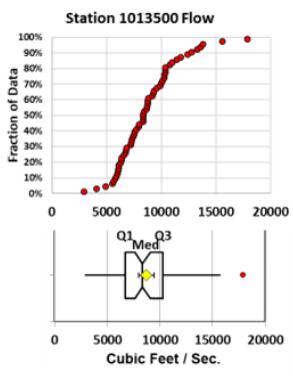


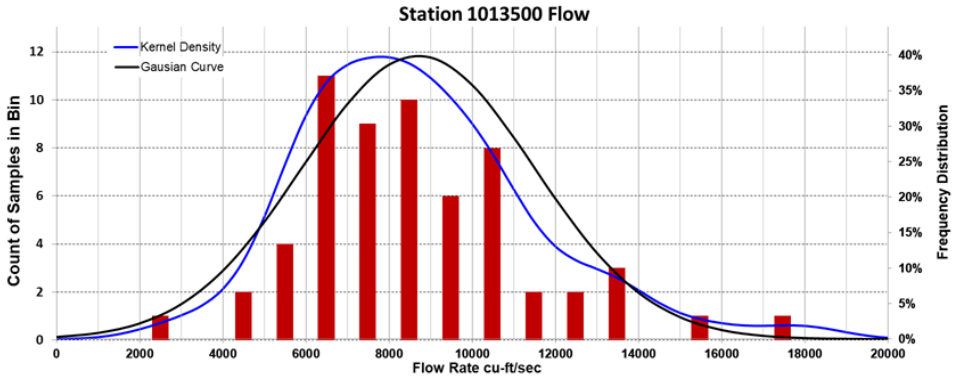
Getting rid of the outliers makes a big difference. The data points lie within the 95% confidence intervals for the Normal Distribution trend line, and the data pass the Anderson Darling test for normality. There is no mathmatical reason why any geologic data set should follow a normal distribution, but, if the samples come from the same population, then they often are.

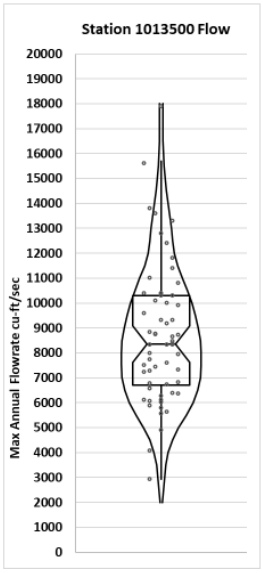
The presence of outliers that can't be explained by measurement errors or typos are probably points that come from a different population. Always try to explain the outliers. Never delete data points just to fit your model. If you do delete or truncate values in a data set, be sure that you report it, and can justify your action.

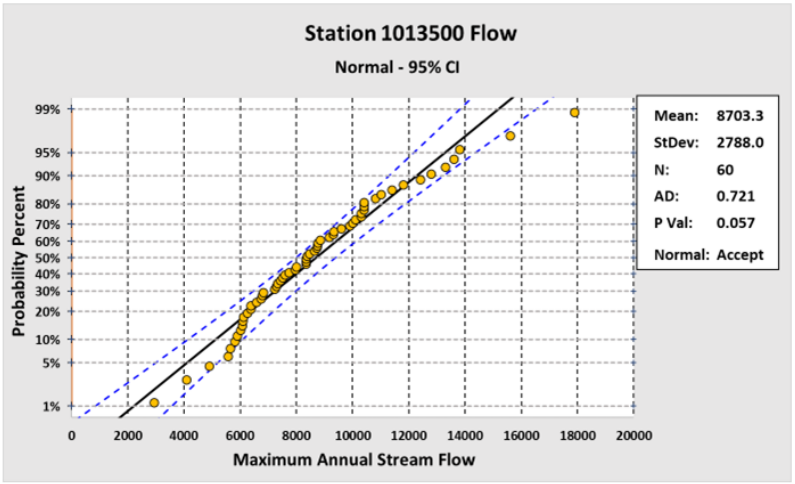
Measurement Scale Workbook

Open the Github Mountjoy3 folder and select 2_Measurement_Scale_Distribution_Plot.xlsx. The data pre-loaded into the measurement workbook are maximum annual flow rates for each of the past 60 years measured at USGS stream gauge number 1013500, located in northern Maine, USA (Farmer et al., 2019). The data range extends from 2,940 cubic feet/second to 17,500 cubic feet/second. Given this range, I scaled each plot from Zero to 20,000 cu-ft/sec. This is likely to capture the range of measurement data that you might collect in your research. You will load your own data into the Raw Data worksheet tab, and then change the chart titles and data names from stream flow to whatever it is that you're going to plot on your graph. The procedures for scaling the graphs are the same as described for the Isotope Scale Graph. Here is the set of graphs for the stream flow example:







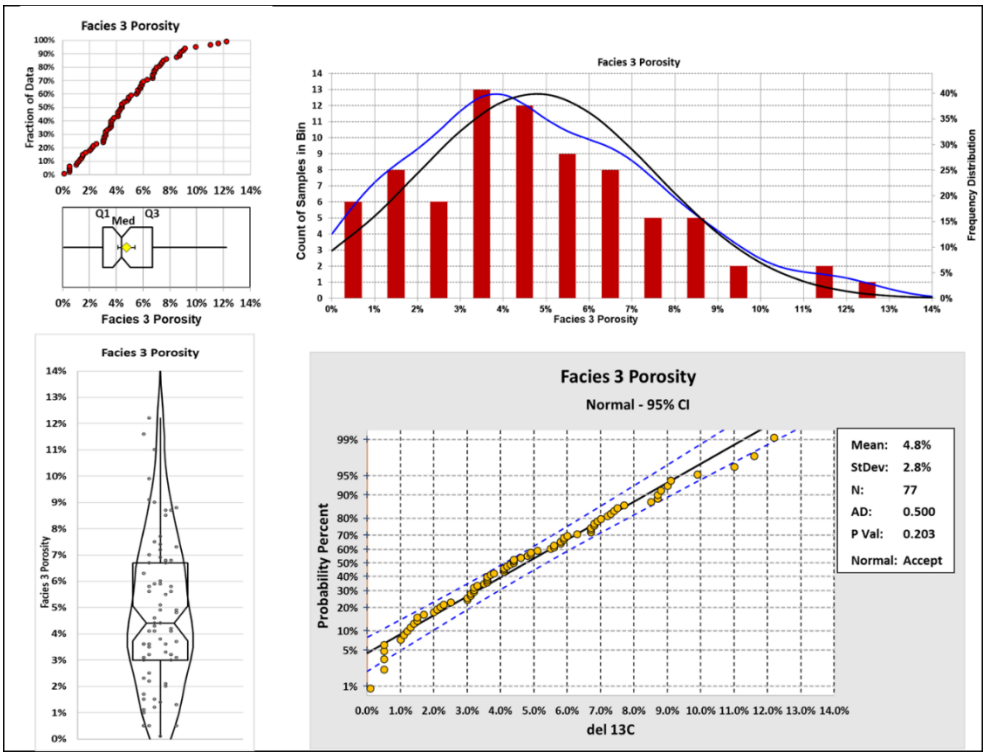


Percent Scale Workbook

Open the Github Mountjoy3 folder and select 3_Percent_Scale_Distribution_Plot.xlsx. This example comes from Middle Devonian Slave Point Formation dolostones with moldic porosity developed from leaching of fossil fragments. Facies 3, described by Dunham and Watts (2017), is Amphipora rudstone to floatstone, consisting of Amphipora fragments in a matrix of microcrystalline dolomite, which replaced an original lime-mud matrix during early burial diagenesis; in contrast, the Amphipora fragments were not replaced by dolomite and retained calcium carbonate mineralogy. During a later phase of burial diagenesis, the calcium carbonate Amphipora fragments were dissolved, while the microcrystalline dolomite was unaffected. The result is a rock with open moldic pores in the shapes of original Amphipora fragments encased in a matrix of impermeable microcrystalline dolostone. Whole-core analysis accurately measures the volume of Amphipora moldic pores and documents that the rock is impermeable if the molds are not touching one another.

Data entry and processing for the Percent Scale Workbook are the same as for the Isotope Scale example. It all starts with manually entering your porosity data into column A of the Raw_Data worksheet and manually entering count numbers into column B of the Raw_Data sheet. Make sure no “left-over” numbers are present in column A below your data. You will then progress to through the

remaining worksheet tabs, where your task will be to set the scales for each plot. The result should look like this:



Cosmetic Adjustments to Graphs

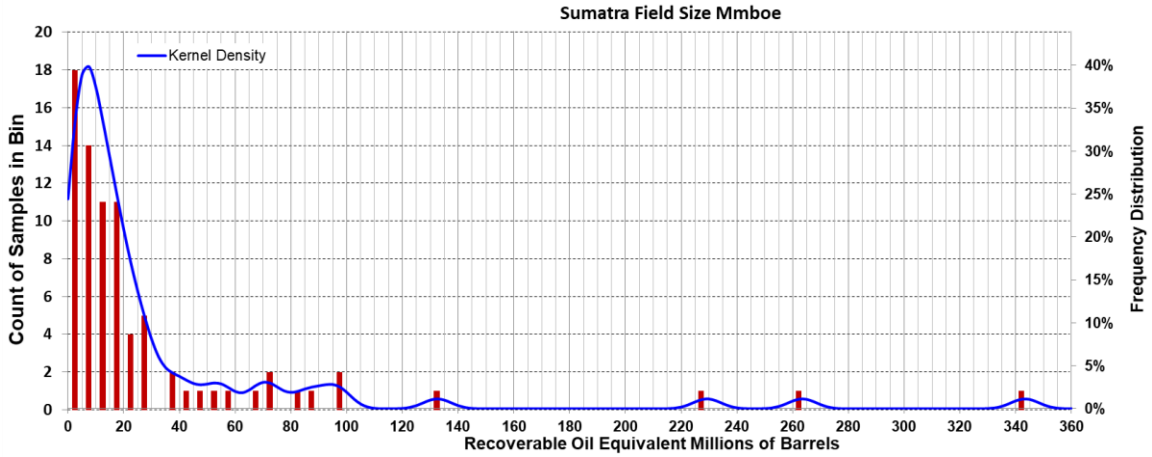
The more familiar you become with Excel, the easier it is to change fonts and colors on these graphs. If you don't like the gray color that I use to fill the Chart Area of these plots, you can change it. Click anywhere on the gray area of a graph; keep clicking until a box outlines the gray area. Then click on Format on the right side at the top of the workbook in the same list as File and Home. When you click on Format, a box will open at the top left just under the File and Home line; this box has a drop-down menu. Click on the drop-down and select Chart Area. After clicking on the Chart area, click on Format Selection (just under the Chart Area drop-down). Format Chart Area box opens on the right side of the workbook. Click on the "bucket" icon on the top left, and then click on Fill. Select Solid fill, then click on the Color drop-down menu where a color pallet will open. The default colors shown are bad choices, for the most part, so click on "more colors" below the color pallet; another box will open with the Custom tab highlighted. In the case of my graphs, I've filled the Chart Area with light gray color with Red Green Blue (RGB) code 230, 230, 230. This gives the light gray color. If you don't like the light gray, you can click on the Standard tab on the top left, this opens a more complete color pallet. You could select White in the center of the pallet if you don't want any color, or you could select any other of the lighter colors to fill the Chart Area without obscuring the black lettering. Or, you could go back to Custom and try different RGB codes to get something you like. This all seems complicated at first, but once you get some experience with Excel, it gets easier.

Lognormal Scale Workbooks

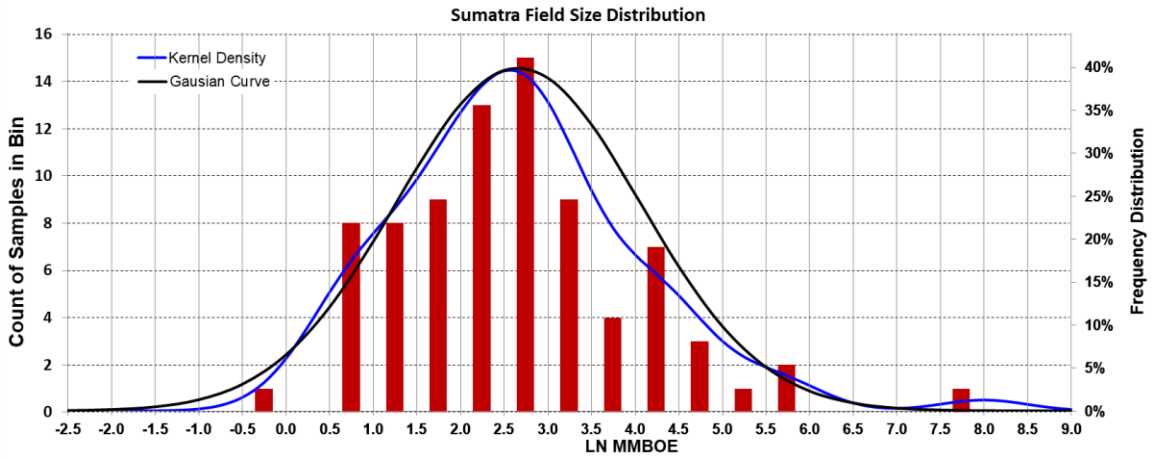
Permeability and Field Size Distribution within a basin are variables that tend to show a lognormal distribution, meaning that the natural logarithm of each value is normally distributed. Normally distributed parameters like porosity tend to be functions of a single property, such as the volume of void space in the case of porosity. On the other hand, when more than one property is responsible for changes in a variable, the result may be a lognormal distribution. In the case of permeability, several factors are in control, including porosity, pore size, and degree of interconnectedness of the pore system. Similarly, field size within a basin is controlled by area, net pay, recovery factor, and several other variables.

Two example lognormal workbooks are included here, one showing permeability distribution in a Devonian Vuggy Carbonate rock, and another showing distribution of recoverable oil reserves from fields in the Sumatra Basin of Indonesia.

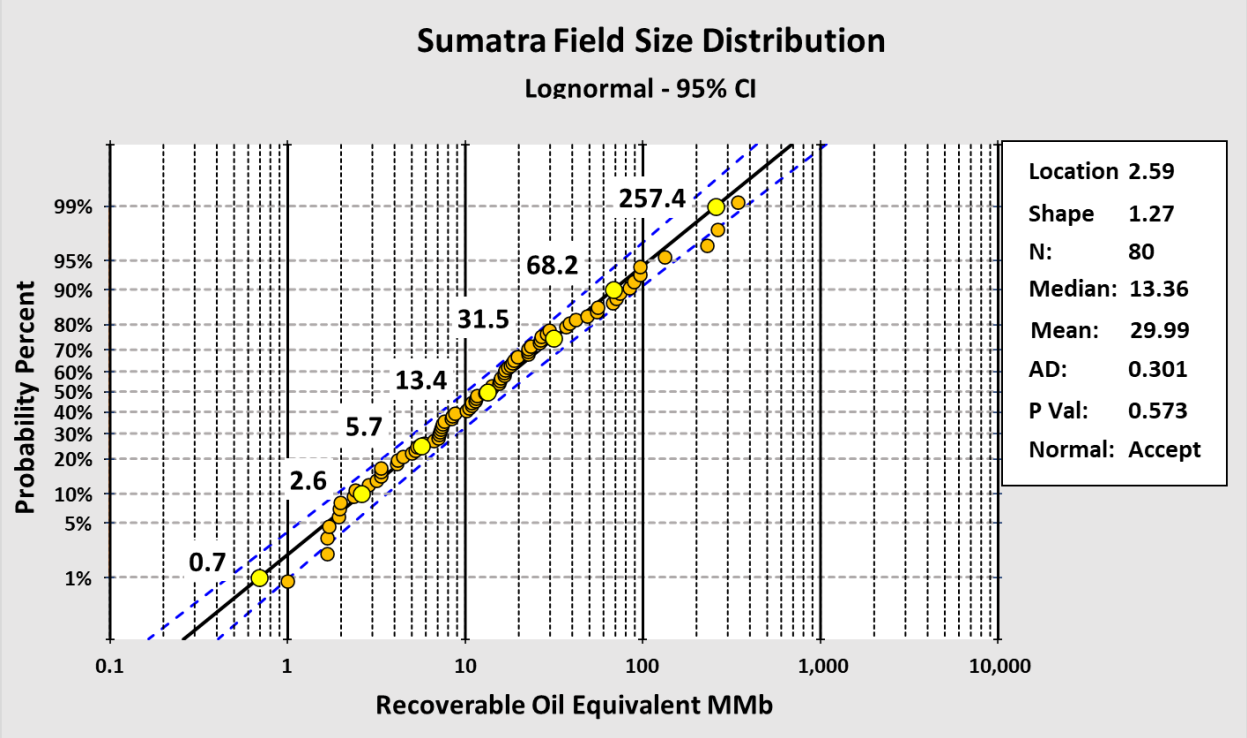
The plot below shows the size distribution of 80 oil and gas fields that produce from lacustrine-deltaic sediments in the Sumatra Basin of Indonesia. This is a very typical field size distribution with a large peak at low values but with a long tail skewed toward higher numbers. Clearly, these values are not following a normal distribution.



The next plot shows the distribution of the natural logarithms of the field-size values:



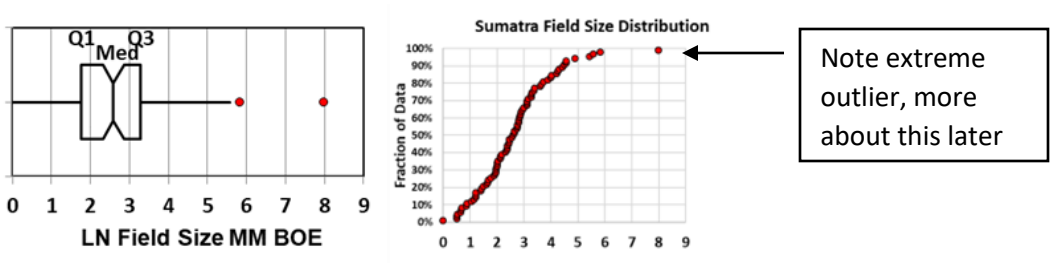
The close match between the black Gausian curve and the blue Kernel Density Curve and the Histogram indicate that the logarithms are normally distributed. Lognormal probability plots are commonly used as a reality check on recoverable-reserve estimates for exploration prospects within a basin. The plot below documents the Sumatra field size distribution:



It indicates that 10% of the fields are smaller than 2.6 million barrels, 50% are smaller than 13.4 million barrels, and 90% are smaller than 68.2 million barrels. If someone presents an exploration prospect with a reserve estimate of 100 million barrels, you would be correct to ask what factors account for the fact that 100 million barrels is larger than 95% of any field discovered in this basin. The plot is not saying that 100 million is impossible; it's just saying that extraordinary size should be justified by extraordinary evidence.

Open the Github Mountjoy3 folder and select 4_Field_Size_Lognormal_Distribtion_Plot.xlsx. You will process the Sumatra Field Size workbook just the same as the standard distribution workbooks. Enter your data into Column B of the Raw_Data Sheet, and be sure to add or delete count numbers in Column A to match the length of Column B. Then, Column C automatically calculates the natural logs of the numbers in Column B.

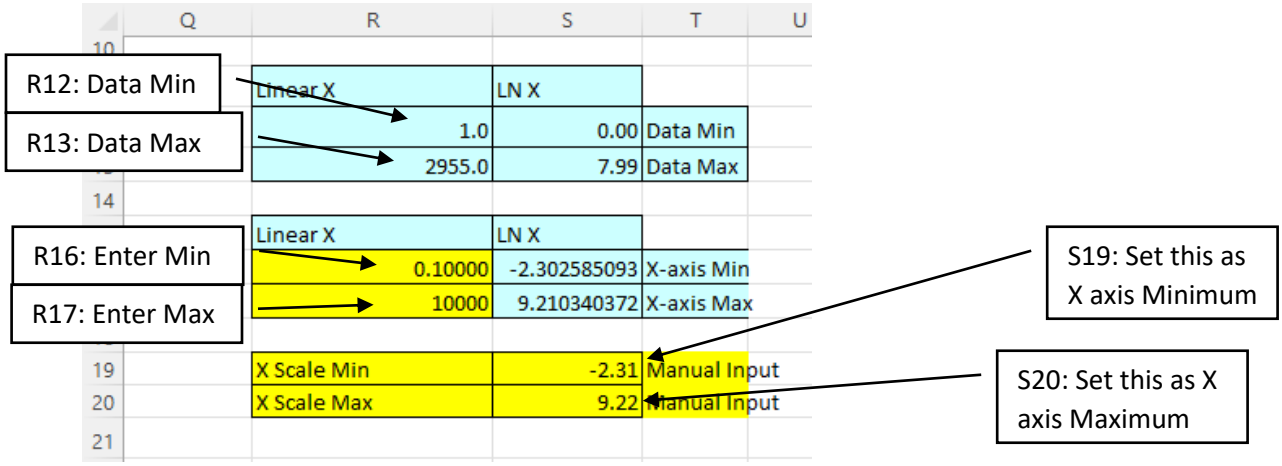
The Percentile and Box Sheet shows the distribution of the natural log values. Be sure that the scales are set to capture the minimum and maximum values in the distribution.



On the Histogram sheet, make sure to set the minimum and maximum scales to capture the data range, and to set the bin width to a reasonable scale; in this case I used a bin width of 0.5.

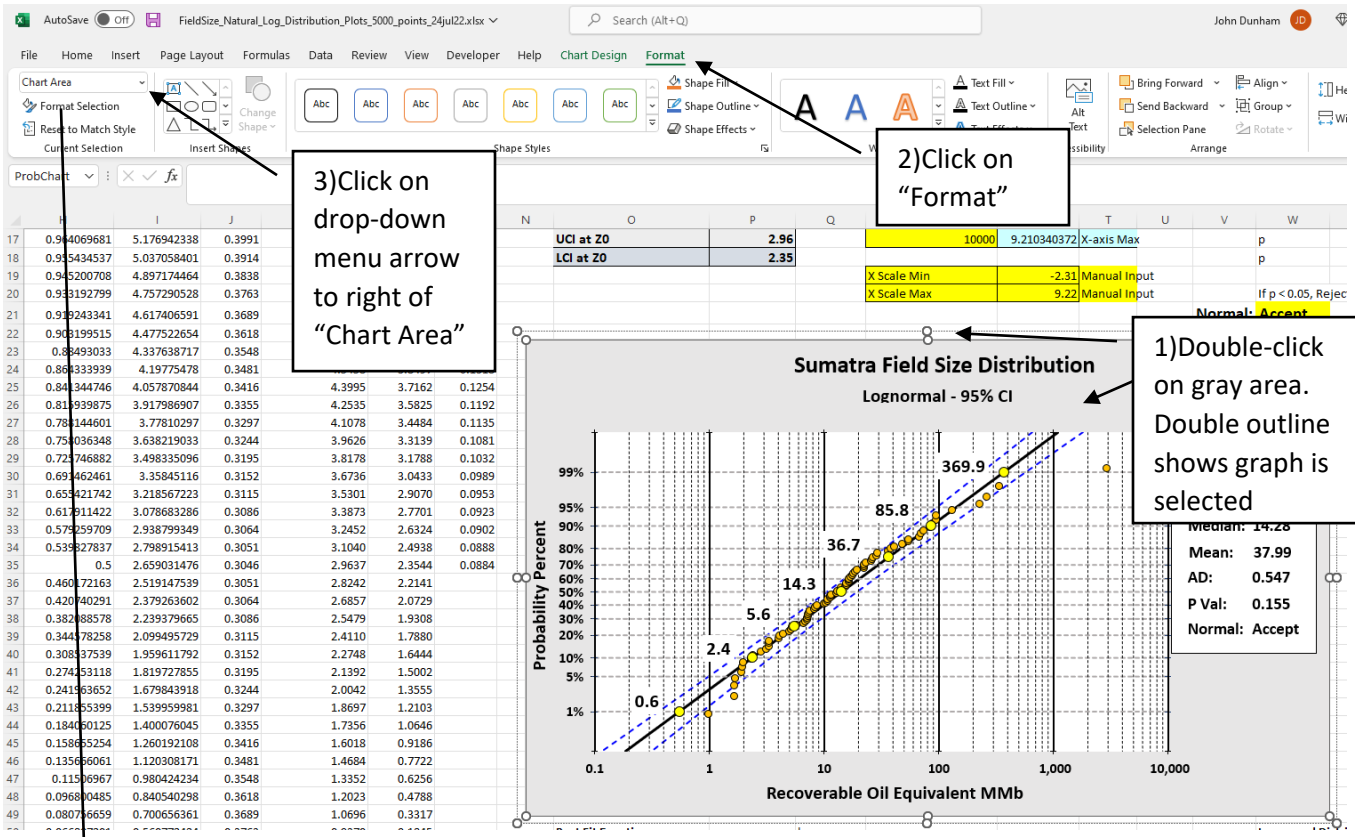
The Violin plot shows the normal distribution of the logs of field size.

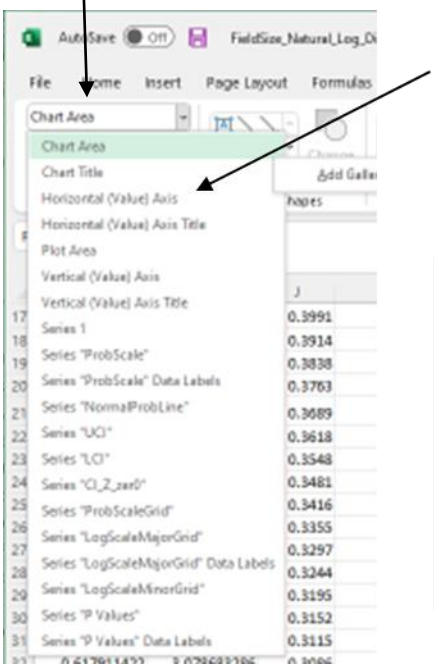
The scales for the lognormal probability plot are set as described below. The X axis of the lognormal plot translates the logarithms into real numbers, and allows direct reading of field size numbers for various percentiles within the distribution. However, the minimum and maximum numbers used to set the X axis scale range are the log numbers. To set the X axis scale, first, look at the data minimum and maximum data values in cells R12 and R13; in this Sumatra case, the data minimum and maximum are 1 and 2955. Then, look at the yellow-colored cells R16 and R17. These are the target minimum and maximum scale values. Our graph has a log scale, the major grid lines increase by 10 times from one to the nest. We want our scale to start at one-major-grid line less than our data minimum, and to finish at one-major-grid line greater than our maximum data value. In this case, the data minimum at 1.0 means that we want to start the scale at 0.1; the data maximum is 2955, so we want to set the scale maximum value to 10,000, which is the first major gridline above 2955. Type 0.1 into cell R16 and 10,000 into cell R17. The worksheet automatically displays the desired log-scale X axis minimum and maximum values that you will enter into the Probability Plot.

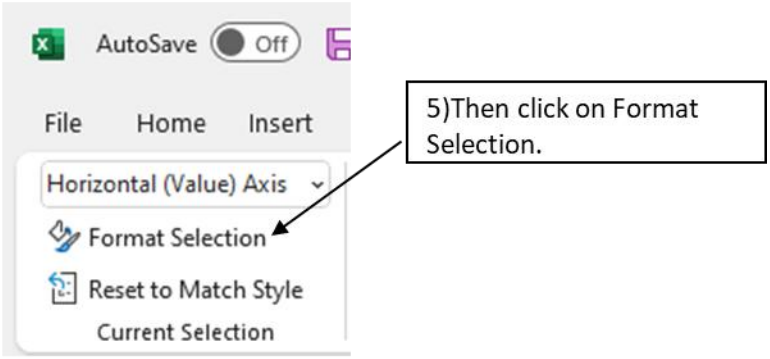


Now that we have the numbers, we open the Horizontal Axis dialog box to set the scales.

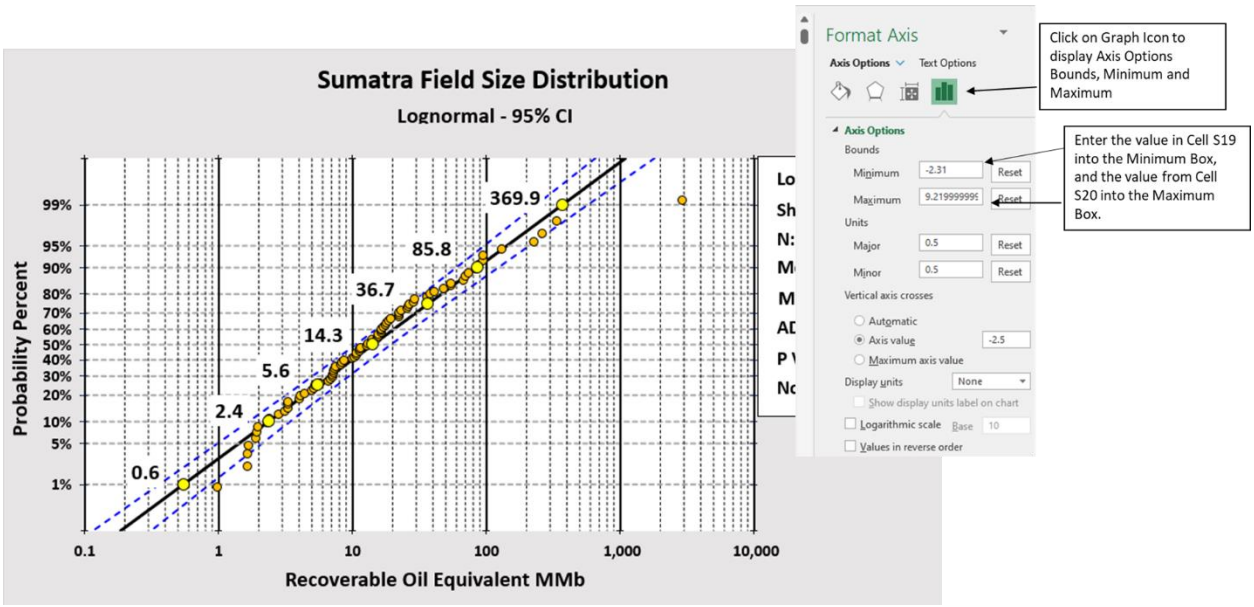
Click on the gray-area of graph two times to select it; the first click selects the group, the second click selects the graph itself. Note the double-outline on the picture below, which shows that the graph is selected. Then, look at the top right line on the sheet, and click on the word “Format”. At this point, a box opens on the left side of the sheet, the words “Chart Area” appear in a box. Click on the drop-down menu arrow to right of Chart Area; a long list appears.







When you click on Format Selection, The Format Axis Dialog Box opens. Click on the Chart Icon to display Minimum and Maximum Axis Values. Enter the value from Cell S19 into the Minimum Box, and enter the value in Cell S20 into the Maximum Box. Then, click on the graph to close the format axis box. The graph should look like this:



The properties of a lognormal distribution are different from a normal distribution. You will recall that the normal distribution is defined by the mean and standard deviation of the data set in question. In contrast, a lognormal distribution is defined by two parameters called Location and Shape, where Location is the Mean of the logarithms and Shape is the Standard Deviation of the logarithms. Important properties of the lognormal distribution are functions of Location and Shape. For example, the Mean of a lognormal distribution is not the mean of the real numbers.

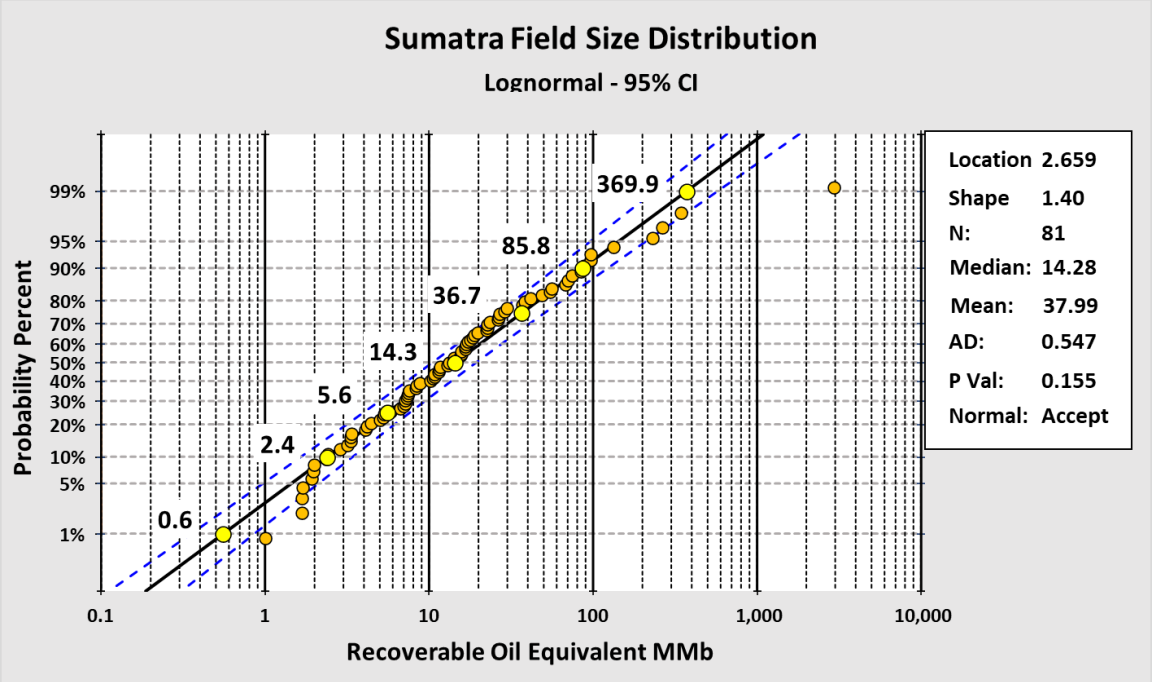
Properties of a Lognormal Distribution:

Location Parameter is the Mean (m) of the Natural Logs of the data values. In this example, Location = 2.66.

Shape Parameter is the Standard Deviation (s) of the Natural Logs of the data values. In this example, Shape = 1.40. Mean = (exp(1)^(2.66+(.5*(1.4^2))) = 37.99 Mmboe. Median = exp(1)^2.66 = 14.3 Mmboe.

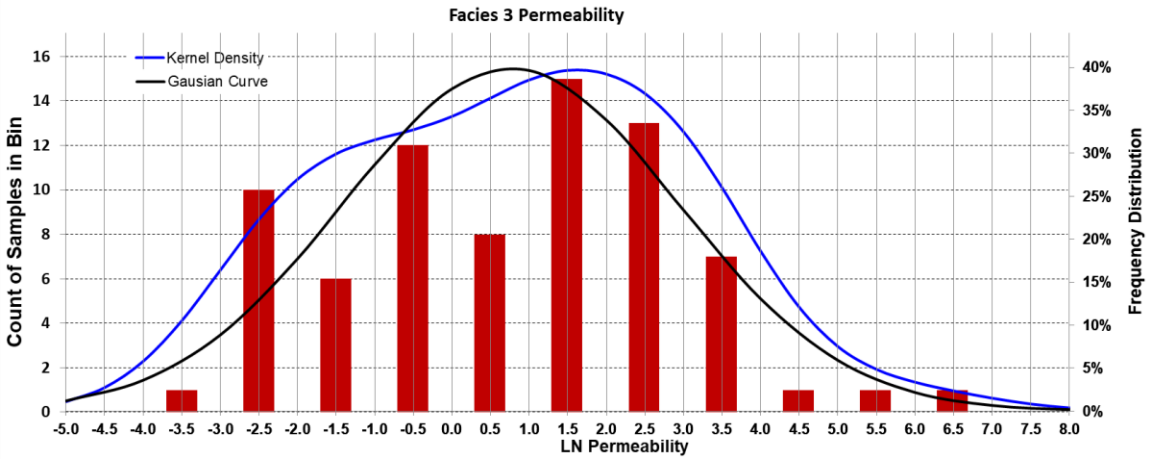
The chart below displays key parameters about the lognormal distribution and reports in this case that the data pass the Anderson Darling Test for a Lognormal Distribution.

The field sizes are expressed in units of Barrels of Oil Equivalent, since both oil and gas are produced from the basin. One barrel of oil is equivalent to 5,800 standard cubic feet of gas, so barrels of oil equivalent (BOE) include both recoverable oil and gas from these Sumatra fields.



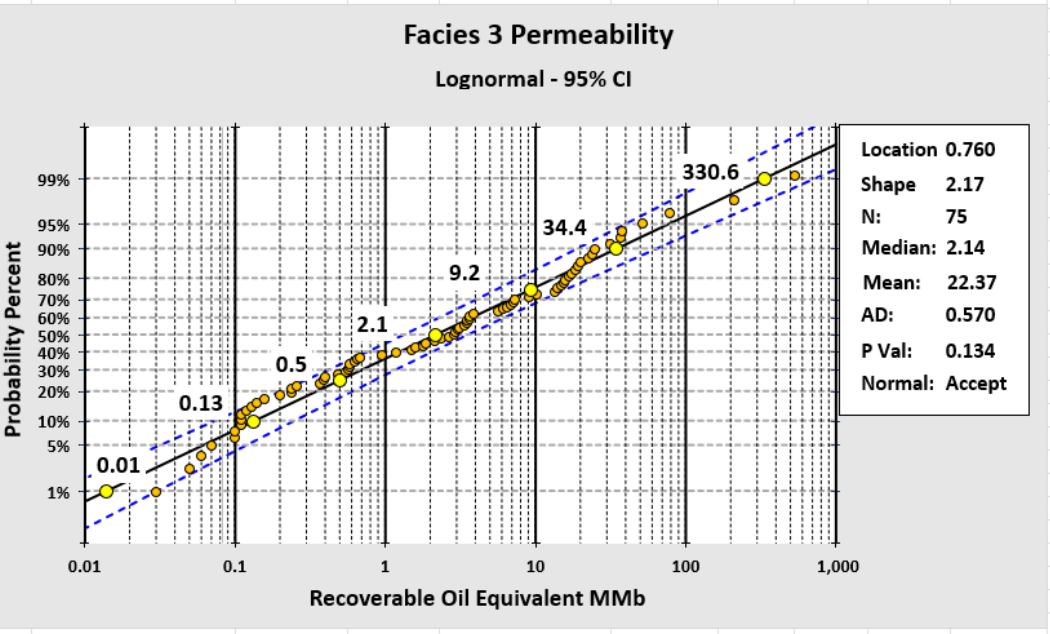
Note the single point plotting at nearly 3 billion barrels oil equivalent. This is clearly different from anything else in the population of Sumatra Basin fields. The outlier is Arun Field, which is a large carbonate buildup that produces from interparticle and vuggy porosity in coral-algal boundstone. Arun is the only significant carbonate reservoir in the region; all the rest of the fields produce from lacustrine-deltaic sandstones. The earlier Sumatra Field Size plot shown above came from the Truncated_Prob_Plot worksheet, where Arun was deleted from the rest of the distribution. If you compare the percentile values between the two plots, you will see that using the Truncated Plot gives more appropriate values for recovery from lacustrine-deltaic reservoir sands.

The second lognormal distribution workbook plots permeability measurements from Devonian vuggy dolostone from northern Alberta. Open the Github Mountjoy3 folder and select 5_Permeability_Lognormal_Distribtion_Plot.xlsx.



The distribution is “spikier” than the field size plot, but the bell curve seems to approximate the Kernel Density and Histogram plots.

The Lognormal Probability Plot lists various key percentiles of the distribution; 10% of the permeability values are less than 0.13 mD, 50% are less than 2.1 mD, and 90% are less than 34.4 mD. These numbers would be important input for modeling oil and gas recovery from this type of reservoir. Also included are tables that list additional percentiles and values.

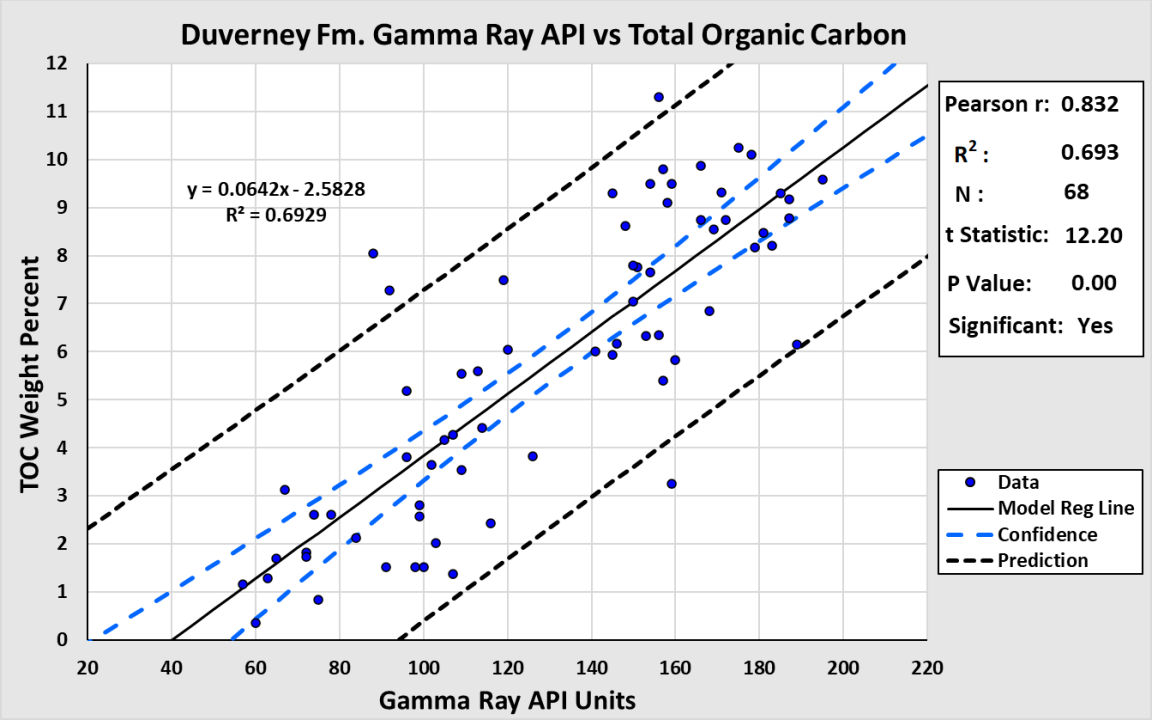


	Best Fit Equation	a	b					Lognormal Distribution Properties	
		2.166976097	0.759897	Enter Prob %	Get X Value	Z Prob	LN x	Location	0.76
Enter X	LN x	Get Prob %	Z Prob	0.1%	0.003	-3.09023	-5.9366	Shape	2.17
1	0	36%	-0.350672	1%	0.01	-2.32635	-4.2812	Mean	22.37
10	2	76%	0.7119082	10%	0.13	-1.28155	-2.0172	Median	2.14
100	5	96%	1.7744881	25%	0.5	-0.67449	-0.7017	Mode	0.02
				50%	2.1	0	0.7599	Variance	54295.09
				75%	9.2	0.67449	2.2215	St.Dev.	233.01
				90%	34.4	1.28155	3.53699	Skewness	1161.18
				99%	330.6	2.32635	5.80104		
				99.9%	1730.8	3.09023	7.45636		

This completes the explanations for the Normal and Lognormal distribution worksheets. If you have any problems, send me a note at johndunham76@gmail.com and I’ll get the sheets to work for you. Next, is a discussion of Linear Regression Workbooks.

Linear Regression Workbooks

X-Y Scatterplots are among the most useful statistical graphs since they may reveal relationships among different parameters. The most common X Y cross plots that you will see will show data points and a single best-fit line through the points, along with the equation for the line and an R-squared number. This type of display omits additional key information. More sophisticated commercial software packages include not only the r and R-squared numbers, but also t-tests to determine the significance of the correlation, and graphics that highlight the uncertainty of the model, such as the examples below. Open the Github Mountjoy3 folder and select 6_Linear_Regression_Linear_X_Linear_Y_Scales.xlsx.

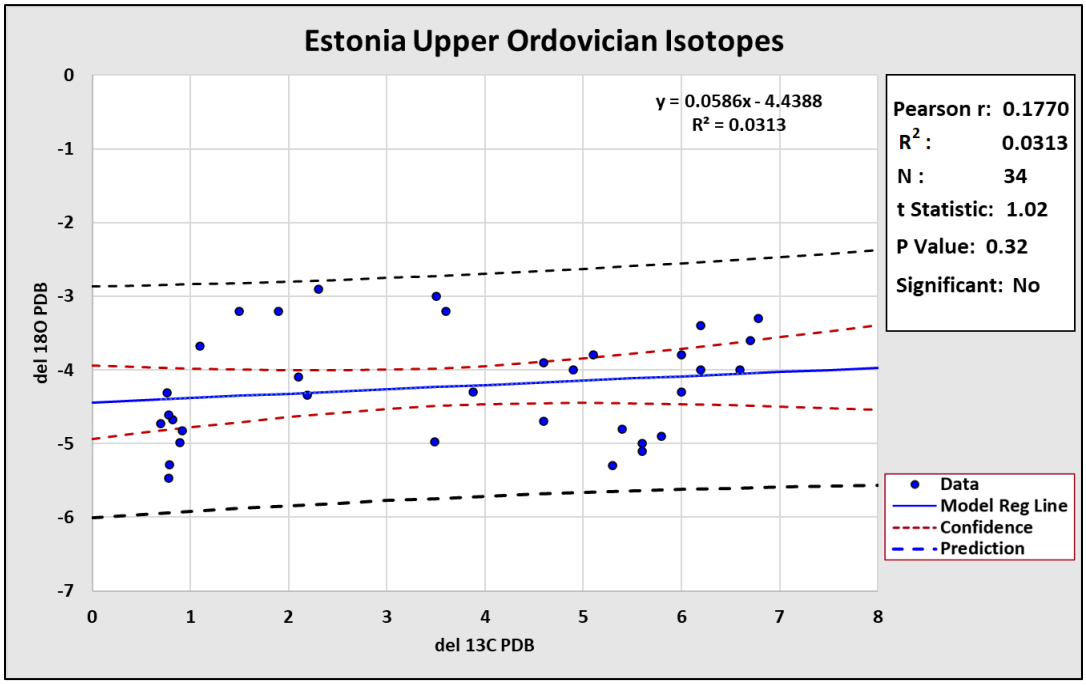


The inner lines bracketing the regression line define the 95% confidence interval for the location of the line; the actual line could be anywhere within this interval. The broadly spaced lines are 95% prediction intervals. These are here to give you a reality check on the confidence of your predictions. There is a tendency among some users to use the regression equation to make firm predictions of Y based on X, but this is definitely wrong. The confidence intervals bracketing the regression line simply define the boundaries within which the true correlation line could be present; they do not give a plus/minus number to attach to your estimate. The prediction lines give the 95% confidence plus/minus interval that you should assign to any prediction of Y based on X.

Note that the Pearson correlation coefficient (r) indicates a positive correlation, and the t Statistic confirms a significant correlation. The R^2 value indicates that this model can explain 69.3% of the variation of Y with respect to X, but that leaves 30.7% of the variation in TOC that can't be explained by Gamma Ray. Factors that might account for the unexplained variation include borehole conditions such as "washouts", borehole-rugosity, mudcake thickness, or any other condition that would cause the GR log to measure lower natural radioactivity than in a better quality borehole. Other factors that might involve TOC include mislabeled samples, mismatch between sample depths and gamma-ray-log depths, and errors in sample analysis. It is not surprising that 30% of the variation in TOC is not explained by our model; however, the model is useful as long as plus/minus values are assigned to estimates. Certainly, the model demonstrates that "hot-gamma" layers are indications of good source rock intervals.

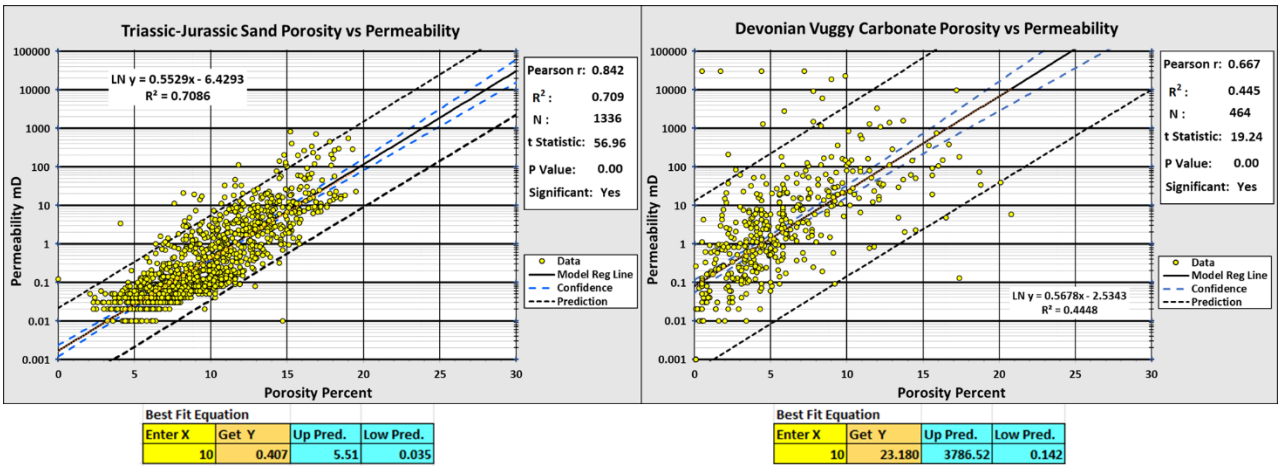
The only user input required to use these regression workbooks is to enter your X data into Column A and your Y data into Column B. In order to keep track of things, you can enter Sample Names and Numbers into Column C. Just make sure that each X value is paired with the appropriate Y value from the same sample.

The next example below shows that there is no significant correlation between these values. Obviously, if a correlation is not significant, prediction of Y is based on X is not possible. This is a good example that explains the P Value, which is the probability that this model is no different from an "intercept-only" regression line; "intercept-only" is a flat line paralleling the X axis and anchored at the Y intercept. The t Statistic tests the hypothesis that the fitted regression line is no different from an intercept-only line and calculates a P Value. If we want a 95% confident estimate, we say that if P is less than .05, then the intercept-only hypothesis is false, and the correlation really is truly significant. In the case shown below, the P value is much greater than 0.05, indicating that this line is no different from "intercept-only" and clearly this correlation is not significant.



Linear regression plots include several scaling options including Linear X Linear Y like the plots shown above, as well as Linear X Log Y, Log X Linear Y, and Log X Log Y.

Linear X Log Y is the standard scale for Porosity – Permeability cross plots. Open the Github Mountjoy3 folder and select 7_Linear_Regression_Linear_X_Por_Log_Y_Perm_Scales.xlsx. The regression equation calculates the natural log of permeability (LN y) based on linear-scale porosity (x). Permeability equals e (base of natural logs) raised to the power of LN y; in Excel this is =exp(1)^LN y. The spreadsheet makes the calculation for you: enter porosity, get permeability and prediction intervals.



Best Fit Equation

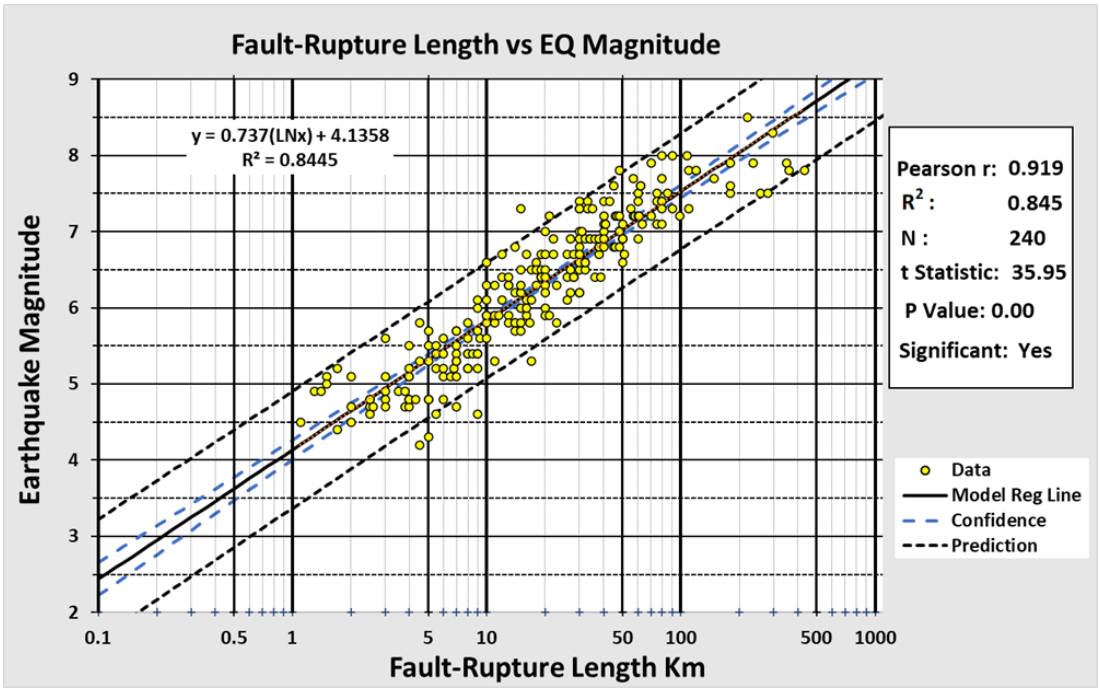
Enter X	Get Y	Up Pred.	Low Pred.
10	0.407	5.51	0.035

Best Fit Equation

Enter X	Get Y	Up Pred.	Low Pred.
10	23.180	3786.52	0.142

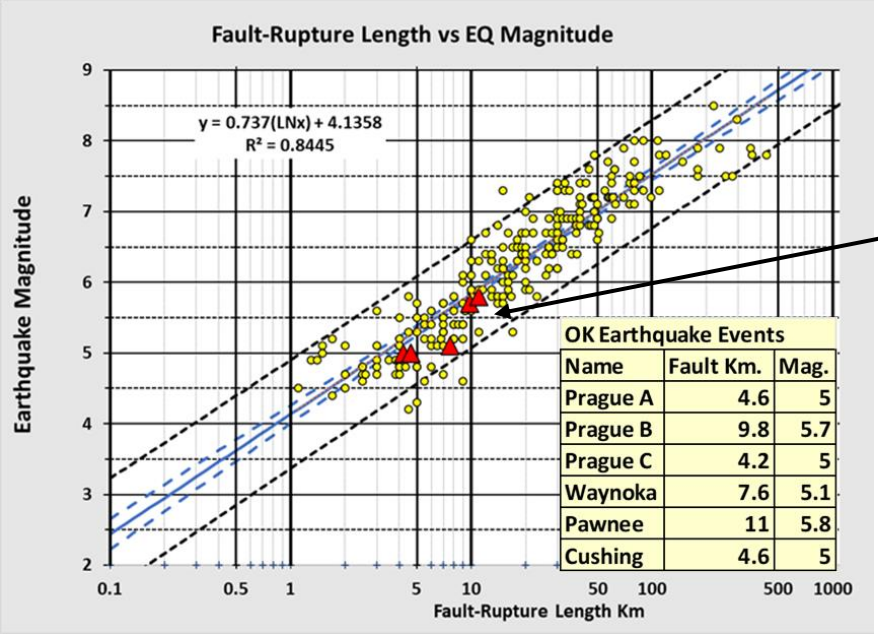
As you might expect, there is much more variation in vuggy carbonate than in sandstone. Both correlations are significant.

Log X Linear Y is a plot that you might use in some cases. Open the Github Mountjoy3 folder and select 8_Linear_Regression_Log_X_Linear_Y_Scales.xlsx. The example below indicates that earthquake magnitude is a significant function of fault-rupture length. Rupture length is plotted at a log scale on the X-axis, and earthquake-moment magnitude is plotted at a linear scale on the Y axis.



There is a significant correlation between magnitude and fault-rupture length. It's possible to tell an interesting story with this graph. Earthquakes are rare in the midcontinent of North America, but beginning in 2008, the frequency has risen by over 10 times. In Oklahoma, this increase coincided with injection of wastewater into Basement faults. Most quakes were small, but several have equaled or exceeded 5.0 magnitude, with these larger quakes resulting in damage to buildings and injury to people.

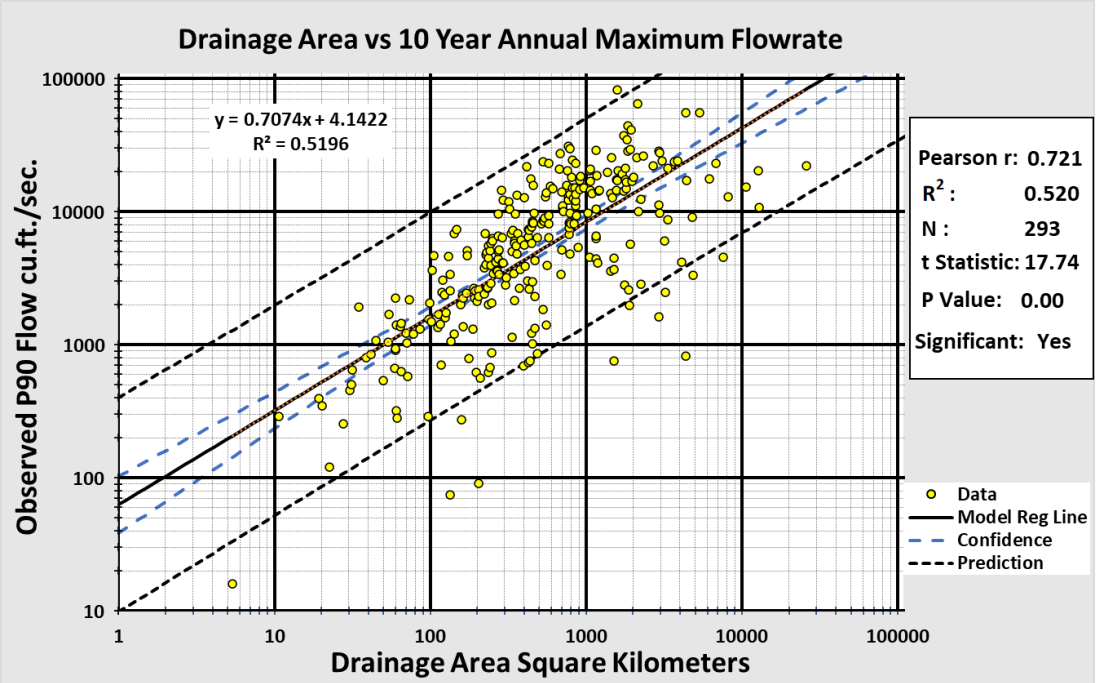
When these large induced quakes are added to the plot, they fall well within the prediction interval of the correlation.



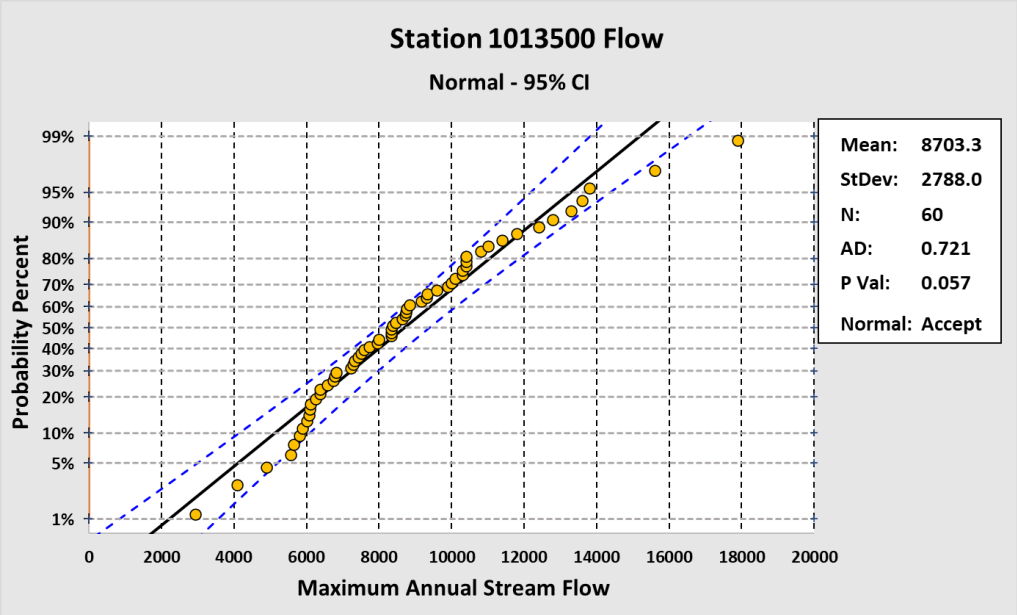
Six quakes => 5.0 magnitude plot within prediction intervals.

The clear interpretation is that if you inject fluid into a basement fault that exceeds 4 km in length, there is a strong probability that you will trigger a damaging earthquake. That is a model you can use.

Log X and Log Y scales: Open the Github Mountjoy3 folder and select 9_Linear_Regression_Log_X_Log_Y_Scales.xlsx. In this example, Stream Flow (Y) measured at several gauge locations is plotted as a function of basin drainage area (X). One would expect that the larger the drainage area, the higher the flow rate would be at the gauge location.



The Y values represent 10 year flood events; these are calculated by making probability plots ofsixty years of maximum annual flow rate data for each gauge station, and then picking the P90 value, which is the flowrate expected to be exceeded only 10% of the time, meaning once in 10 years.

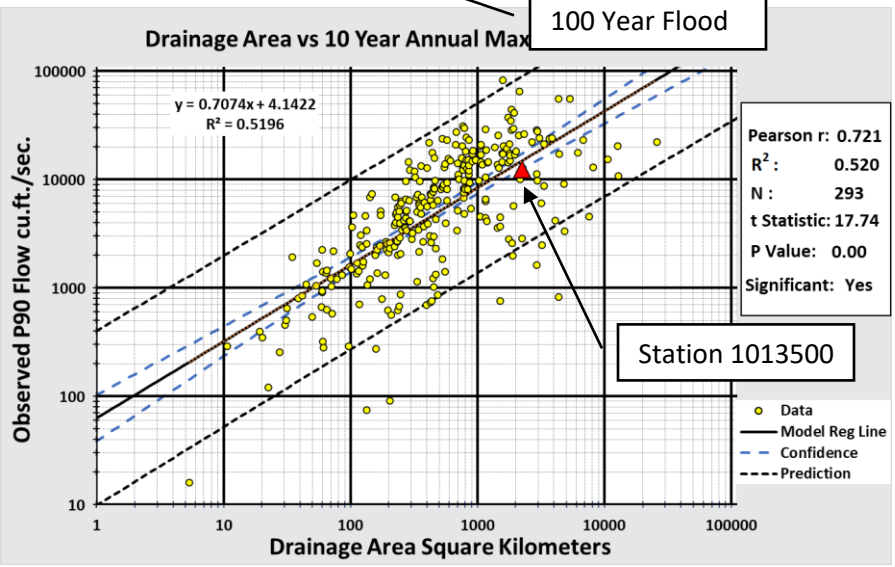


Enter Prob %	Get X Value
0.1%	88
1%	2217
10%	5130
25%	6823
50%	8703
75%	10584
90%	12276
99%	15189
99.9%	17319

This is how USGS estimates 10 and 100 year flood events.

10 Year Flood

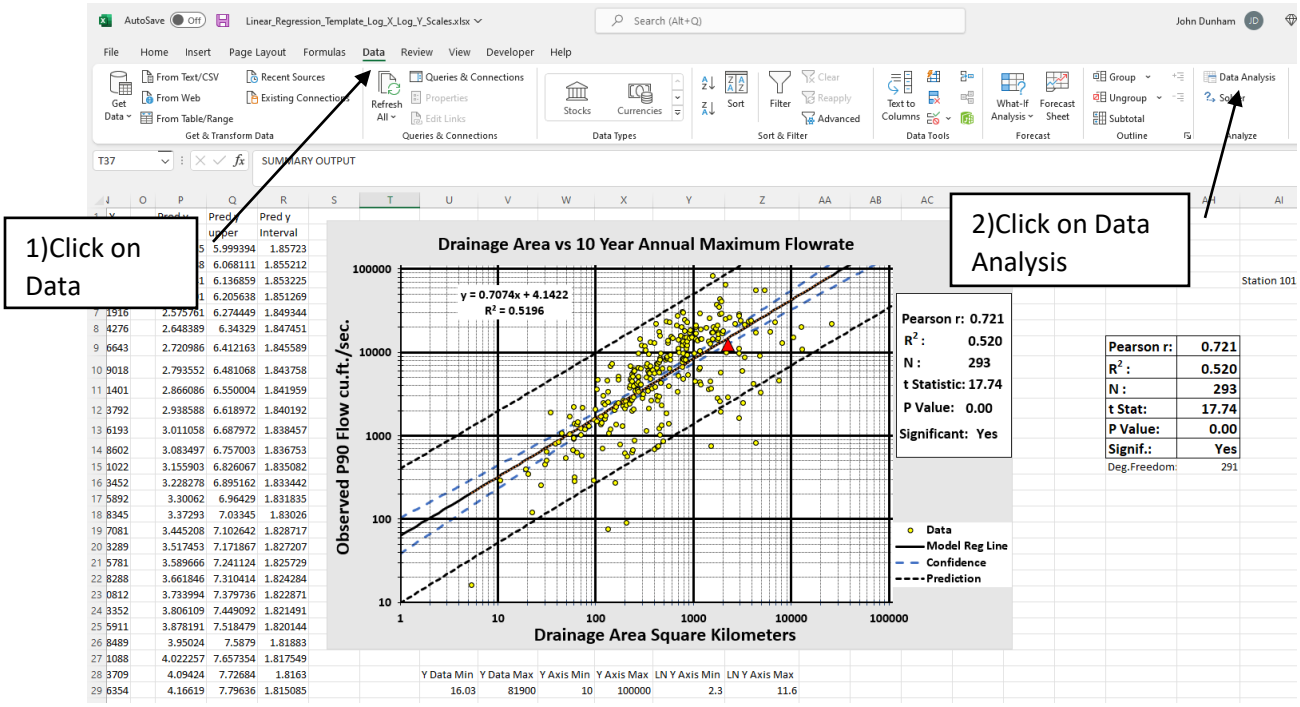
100 Year Flood



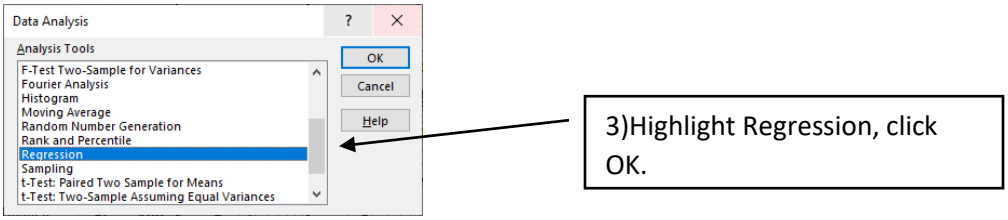
The USGS made the Drainage Area vs 10 Year Annual Maximum Flowrate graph by picking the P90 flow for 293 different gagues. The correlation is significant but the R^2 number indicates that this model only explains about 52% of the observed flow rate. Other factors must be involved. This gets to the topic of using Excel **Multiple Regression** techniques.

Open the Github Mountjoy3 folder and select 10_Multiple_Regresion_Log_X_Log_Y_Scales.xlsx workbook. I use the same Log_X_Log_Y_scales in this example, but you can use this same method for plots with different scales. To use Excel Multiple Regression, you have to make sure that the **"Data Analysis"** ToolPak is active in your version of Excel.

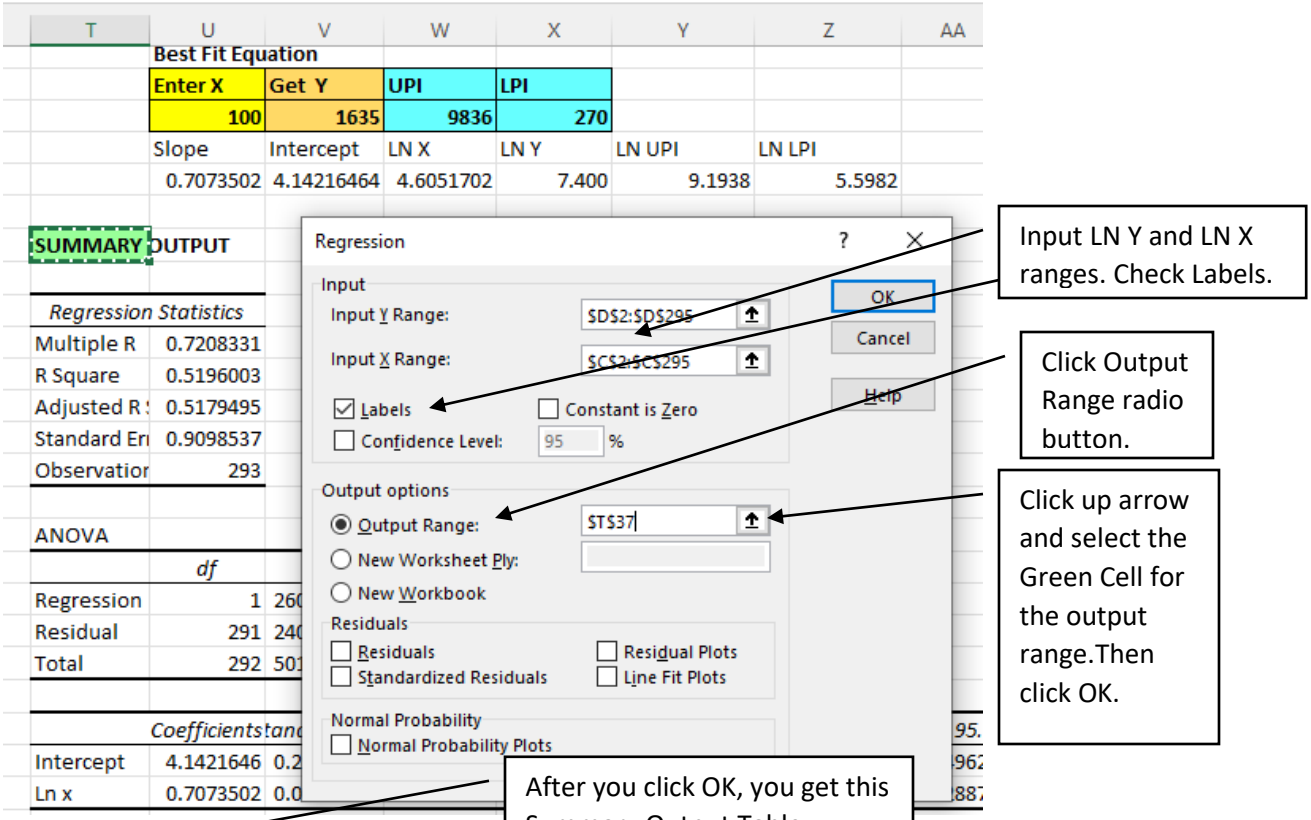
Open Excel, then click on **File** at upper left, look at the very bottom of the left side, and click **Options**, then on the lower left side, click the **Add-Ins** category. In the **Add-Ins** box, look for a drop-down box that says **Manage**: and select **Excel Add-ins** in the drop-down menu and click the Go Button, check the **Analysis ToolPak** check box, and then click **OK**. If you are prompted that the Analysis TookPak is not currently installed on your computer, click **Yes** to install it. Analysis ToolPak contains many useful routines. We will test the Regression Routine in Data Analysis Toolpak using the Log X Log Y workbook.



First, click on data, then click on Data Analysis at upper right corner. This opens a long list of Analysis Tools. Scroll down to locate regression, highlight regression, and click OK.



After clicking OK, a Regression Dialog Box opens, asking for input Y and input X data. In this example, where we are ploting Log Y versus Log X, we input LN Y values and LN X values into these boxes as shown below. Include the top line labels in the input and check the Labels box. Then, hit the Output Range radio button, click on the up-arrow box, and select cell T37 on the worksheet. I put a GREEN color fill in this box to help you locate it. Always make sure that you select the box with the green fill. Click OK, a message will appear saying that you will overwrite existing data, click OK, because this won't be a problem as long as you selected the Green colored cell.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.721
R Square	0.520
Adjusted R Square	0.518
Standard Error	0.910
Observations	293

Pearson r: 0.721
R^2: 0.520
N: 293
t Statistic: 17.74
P Value: 0.00
Significant: Yes

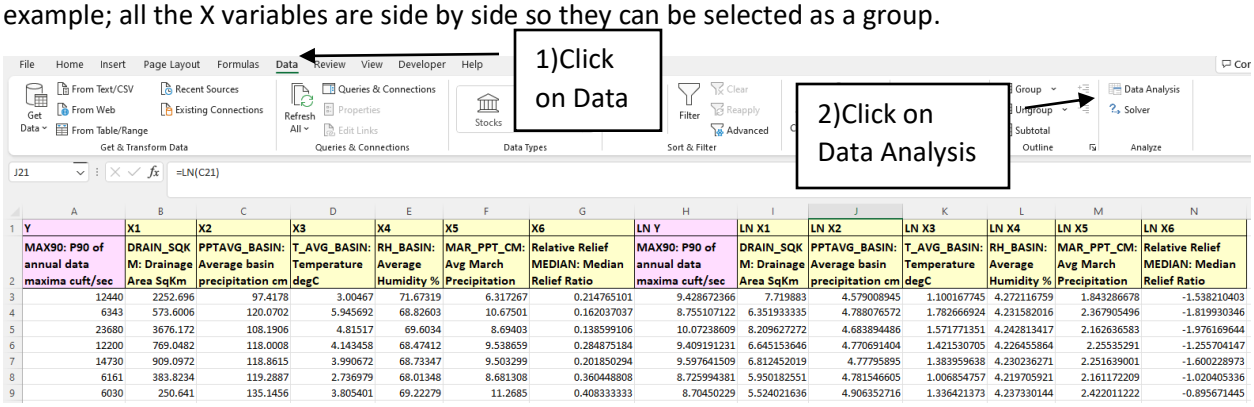
ANOVA

	df	SS	MS	F	Significance F
Regression	1	260.5570455	260.55705	314.74564	3.04521E-48
Residual	291	240.8996059	0.8278337		
Total	292	501.4566513			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	4.142164645	0.250255331	16.551754	7.999E-44	3.649624727	4.634704563	3.649624727	4.63470456
Ln x	0.70735024	0.039870772	17.741072	3.045E-48	0.628878598	0.785821883	0.628878598	0.78582188

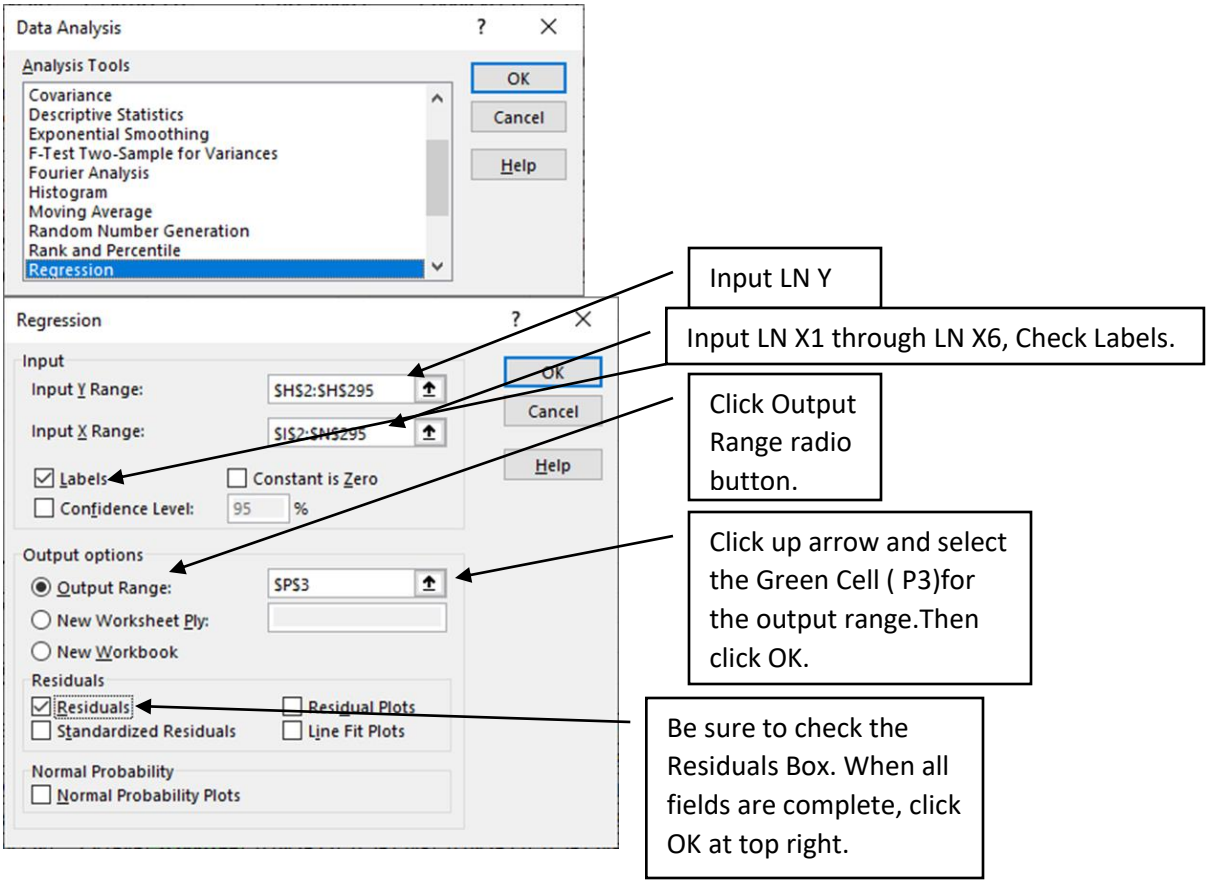
The regression summary output table gives the same numbers calculated by the worksheet.

However, when you try this with multiple regression, the complex matrix algebra calculations can't be done on a spreadsheet; you have to use Excel's data analysis routine. Multiple regression is appropriate in this case because our R^2 number using only drainage area only explains about half the variation in flow rate. The USGS cataloged other variables that might also contribute to flow rate; these include measurements at each station of annual precipitation, average temperature, relative humidity, March precipitation, and basin relief ratio. Each of these might combine with basin area to affect maximum annual flow rate. Excel's multiple regression analysis allows use of multiple X variables as shown by this example; all the X variables are side by side so they can be selected as a group.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Y	X1	X2	X3	X4	X5	X6	LN Y	LN X1	LN X2	LN X3	LN X4	LN X5	LN X6
2	MAX90: P90 of annual data maxima cuft/sec	DRAIN_SQK M: Drainage Area SqKm	PPTAVG_BASIN: Average basin precipitation cm	T_AVG_BASIN: Temperature degC	RH_BASIN: Average Humidity %	MAR_PPT_CM: Avg March Precipitation	Relative Relief MEDIAN: Median Relief Ratio	MAX90: P90 of annual data maxima cuft/sec	DRAIN_SQK M: Drainage Area SqKm	PPTAVG_BASIN: Average basin precipitation cm	T_AVG_BASIN: Temperature degC	RH_BASIN: Average Humidity %	MAR_PPT_CM: Avg March Precipitation	Relative Relief MEDIAN: Median Relief Ratio
3	12440	2252.696	97.4178	3.00467	71.67319	6.317267	0.214765101	9.428672366	7.719883	4.579008945	1.100167745	4.272116759	1.843286678	-1.538210403
4	6343	573.6006	120.0702	5.945692	68.82603	10.67501	0.162037037	8.755107122	6.351933335	4.788076572	1.782666924	4.231582016	2.367905496	-1.819930346
5	23680	3676.172	108.1906	4.81517	69.6034	8.69403	0.138599106	10.07238609	8.209627272	4.683894486	1.571771351	4.242813417	2.162636583	-1.976169644
6	12200	769.0482	118.0008	4.143458	68.47412	9.538659	0.284875184	9.409191231	6.645153646	4.770691404	1.421530705	4.226455864	2.25535291	-1.255704147
7	14730	909.0972	118.8615	3.990672	68.73347	9.503299	0.201850294	9.597641509	6.812452019	4.77795895	1.383959638	4.230236271	2.251639001	-1.600228973
8	6161	383.8234	119.2887	2.736979	68.01348	8.681308	0.360448808	8.725994381	5.950182551	4.781546605	1.006854757	4.219705921	2.161172209	-1.020405336
9	6030	250.641	135.1456	3.805401	69.22279	11.2685	0.408333333	8.70450229	5.524021636	4.906352716	1.336421373	4.237330144	2.422011222	-0.895671445

As in the previous example, the Y value is the 10 Year Maximum Flow Rate (P90) value for each station. However, instead of using one X term, the basin drainage area, the multiple regression model will use six X terms highlighted in yellow above. Click on Data on the top line, then click on Data Analysis on the upper right. In the Data Analysis window, scroll down to select Regression.



This routine uses all six X terms to model the P90 Flow (Y). The Summary Output table shows significant improvement in the correlation.

P	Q	R	S	T	U	V	W	X
6 X Terms								
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.897							
R Square	0.805							
Adjusted R Square	0.801							
Standard Error	0.585							
Observations	293							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regression	6	403.4974691	67.24958	196.3408	2.61E-98			
Residual	286	97.95918229	0.342515					
Total	292	501.4566513						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-5.67793263	1.545611565	-3.67358	**0.000285**	-8.72015	-2.63572	-8.72015	-2.63572
DRAIN_SQKM	0.833754403	0.028626519	29.12525	**1.48E-87**	0.777409	0.8901	0.777409	0.8901
PPTAVG_BASIN_CM	1.258018664	0.276733997	4.545949	**8.08E-06**	0.713325	1.802712	0.713325	1.802712
T_AVG_BASIN	0.172691256	0.045244328	3.81686	**0.000166**	0.083637	0.261745	0.083637	0.261745
RH_BASIN	0.563871957	0.523641696	1.076828	**0.282465**	-0.46681	1.594552	-0.46681	1.594552
MAR_PPT_CM	0.220703865	0.169567217	1.301572	**0.19411**	-0.11305	0.554462	-0.11305	0.554462
Relief_Basin	0.095870365	0.089450029	1.071776	**0.284724**	-0.08019	0.271934	-0.08019	0.271934

Using 6 X terms improves R^2 from 0.52 to 0.805; a major improvement

However, a check of the P-values for each of the X terms shows that Drainage Area, Average Precipitation, and Average Basin Temperature are significant at 95% confidence level, while Relative Humidity, March Precipitation Level, and Basin Relief are not significant at 95% confidence level. P-Value is significant at a 95% confidence level if it is less than 0.05. Therefore we could remove the insignificant terms and re-run the multiple regression using only the three significant X terms. As you can see, this is a very useful feature of Excel.

The way to display the improvement of multiple regression versus single-X-term regression is to make a plot with LN Observed P90 Stream Flow on the X axis, and our 6-Term modeled LN P90 Stream Flow on the Y axis. This is shown on the worksheet tab labeled Multi_X_Obs_vs_Model. The observed P90 stream flow is from the USGS dataset, and LN P90 observed is shown in Column L above. The 6-X-Term modeled P90 stream flow is shown in the Residual Output section of the Summary Output on worksheet tab Multi_X_Reg; recall that we checked the box that said show Residuals. The predicted LN max 90 shown in Multi_X_Reg Column Q was calculated from the multiple regression formula:

Model Y = Intercept + (X1*coefficient X1) + (X2*coefficient X2) + (Xn * coefficient Xn). The intercept and the six X coefficients are shown in the above table, and are used to calculate the residual output table shown here:

RESIDUAL OUTPUT	6 X Terms	
Observation	*Predicted LN max90*	*Residuals*
1	9.377300415	0.051371951
2	8.683559378	0.071547744
3	10.01098776	0.061398329
4	8.870158457	0.539032773

Model Y in output table on worksheet Multi_X_Reg. Copy all numbers in this column to worksheet Multi_X_Obs_vs_Model worksheet Column B.

The predicted LN max90 is the model Y calculated from the intercept and coefficients in the equation above; the Residuals equal the Observed P90 – Predicted P90; in this example, Observed LN P90 was 9.428672366 and Modeled LN P90 was 9.377300415 giving a residual of 0.051371951.

You'll recall that when we crossplotted Y = Observed p90 Flow against X = Basin Area, we got an r^2 of 0.51. Now, we are going to crossplot Y = Modeled P90 Flow against X = Observed P90 Flow, and we get this plot. Go to worksheet tab Multi_X_Obs_vs_Model. Here the observed LN max P90 is automatically entered into Column A, but we will have to manually copy the Modeled LN P90 from the Multi_X_Reg worksheet to column B of the Multi_X_Obs_vs_Model worksheet.

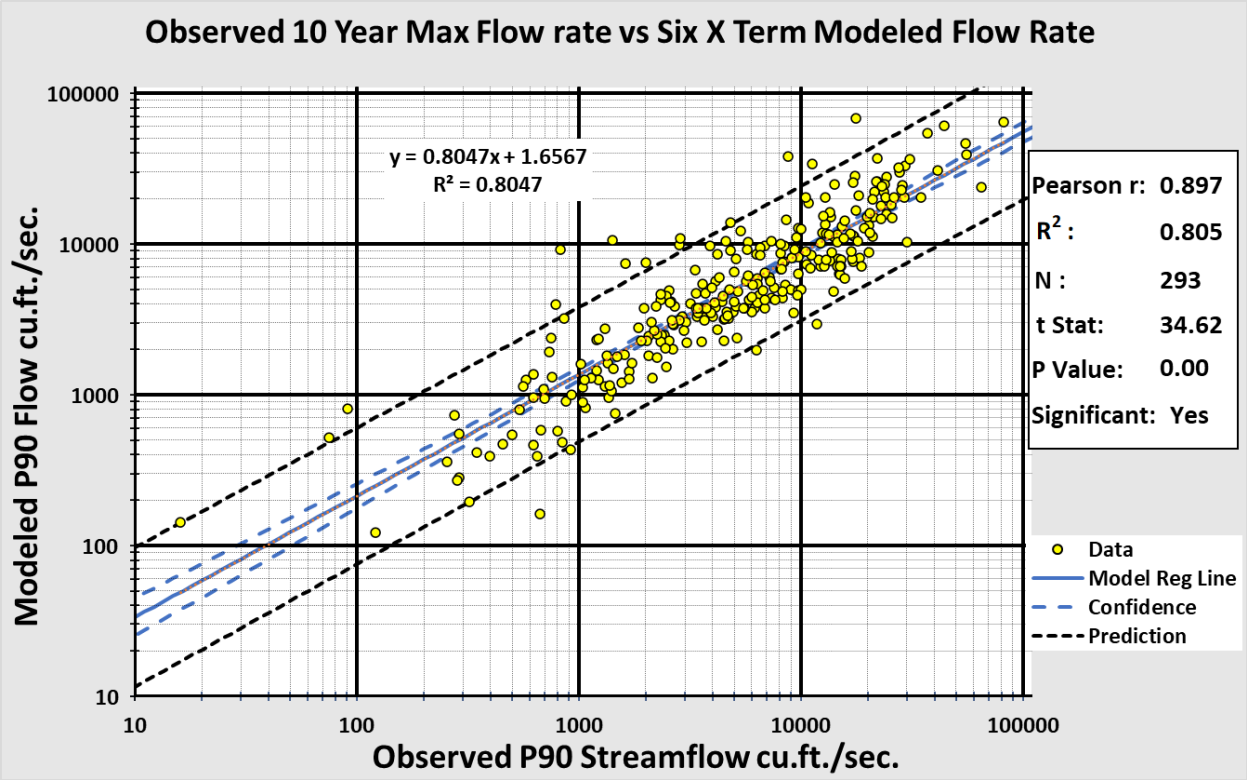
	A	B	C	D	E	F	G	H	I
1	Dataset Name:	6 X term regression							
2	Observed LN max90: P	y= Model LN p90 Flow			Slope	0.804651			
3	9.428672366	9.377300415			Intercept	1.65669			
4	8.755107122	8.683559378		% Conf.	alpha	0.05			
5	10.07238609	10.01098776		Count	n	293			
6	9.409191231	8.870158457		Avg X	mean(x)	8.480659		0.3	1.898085594
7	9.597641509	8.980580997		Sxx	devsq(x)	501.4567		0.4	1.978550669
8	8.725994381	8.230734416		Sxy	steyx	0.520451		0.5	2.059015743
9	8.70450229	8.168806834		df = n-2	t-crit	1.96815		0.6	2.139480818
10	7.881182202	7.602820729						0.7	2.219945893
11	6.216805682	6.293157122		Model X 100 Step Interval:		0.093		0.8	2.300410967
12	7.789454566	7.808648784						0.9	2.380876042
13	7.111512116	7.131919913						1	2.461341117
14	7.094234846	7.280209733						1.1	2.541806191
15	7.846589975	7.727359017						1.2	2.622271266
16	8.291295852	8.102803053						1.3	2.702736341
17	7.630461262	7.800686039							
18	6.290457411	6.67725088							
19	8.506536611	8.166228216							
20	8.666130304	8.351164266							
21	6.947937069	7.020682857						1.8	3.105061714
22	5.666772649	5.644044467						1.9	3.185526789
23	8.390949465	8.332712147						2	3.265991864
24	7.210079628	7.382889112						2.1	3.346456938
25	9.603057907	8.754890685						2.2	3.426922013

Paste Predicted LN max90 output into Multi_X_Obs_vs_Model worksheet Column B.

Multi_X_Obs_vs_Model worksheet tab.

LN-Scale_Template | Multi_X_Reg | Multi_X_Obs_vs_Model

The result is this graph:



Our 6-term model explains over 80% of the observed stream flow. We used this on a Log X Log Y plot, but you can use multiple regression on any type of linear or log scale plot. Just make sure that you use LN X and LN Y values when one or the other or both of those values are plotted on a log scale. Whenever a log scale is used, the graphs will automatically scale the axes to show natural log numbers translated back into whole numbers.

That completes the "user manual" for this set of statistical graphs. The more familiar you become with Excel, the easier it will be to use these graphs. Commercial software does all of this with the click of a button. On the other hand, you have much more flexibility in selecting fonts, symbols, colors, and sizes in Excel plots than you have with commercial software. Let me know if you have any questions or problems: John Dunham, johndunham76@yahoo.com. We'll get it done!

References

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A., 1983, Graphical Methods for Data Analysis: Duxbury Press, Boston, MA, 395 p.

Dunham, J.B., and Watts, N., 2017, Moldic-pore distribution, basement paleotopography, and oil production from a Devonian dolostone reservoir, Peace River Arch, Western Canada: SEPM Special Publication 109, p. 87-105.

Farmer, W.H., Kiang, J.E., Feaster, T.D., and Eng, K., 2019, Regionalization of Surface-Water Statistics using Multiple Linear Regression, U.S. Geological Survey Techniques and Methods book 4, Chapter A12, https://doi.org/10.3133/tm4A12.

Montgomery, D.C., and Runger, G.C., 2011, Applied Statistics and Probablity for Engineers, Fifth Edition: John Wiley and Sons, Hoboken, NJ, 768 p.