The Visual Display of Quantitative Geologic Data

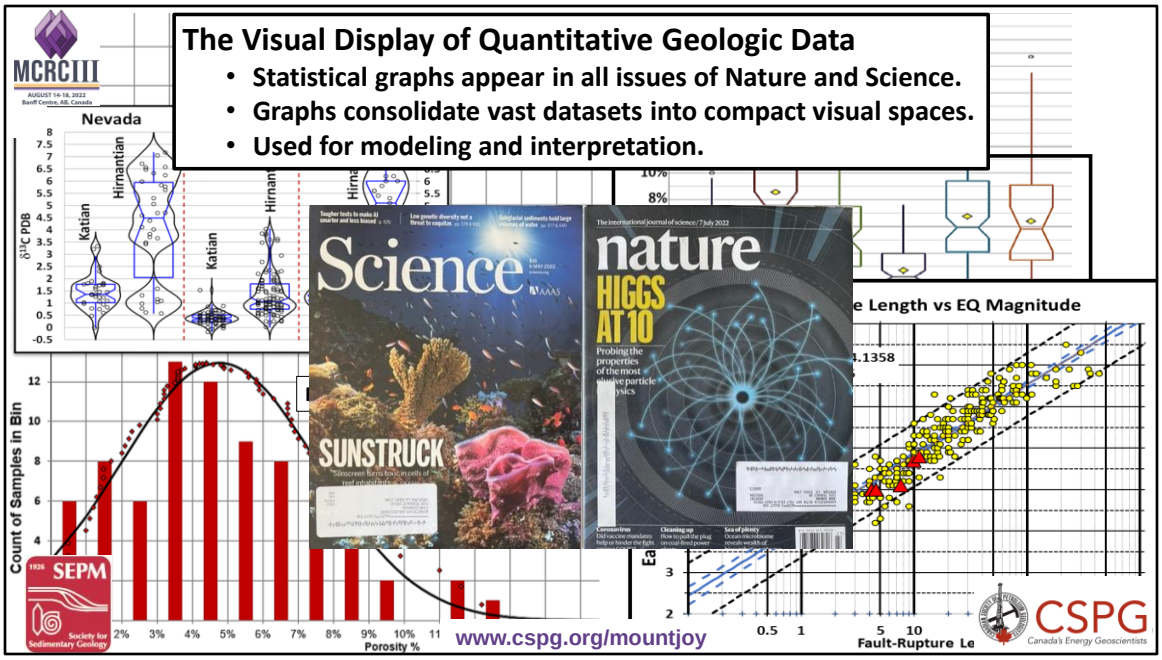
John B. Dunham | Union Oil Company of California (Retired)

SEPM Society for Sedimentary Geology

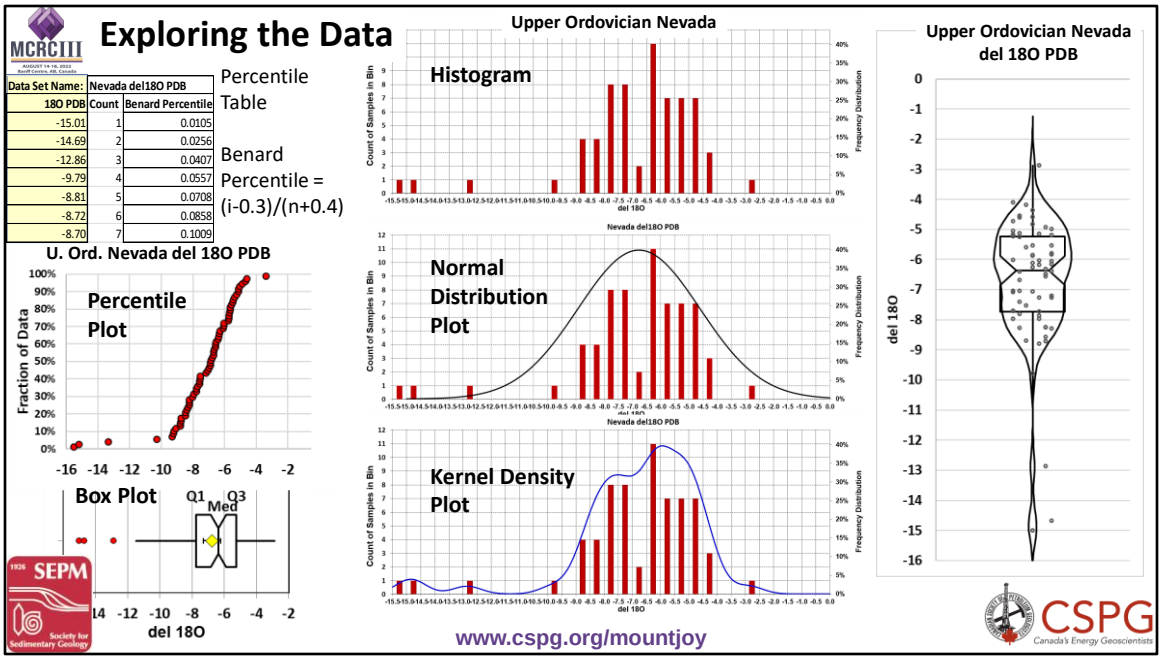
www.cspg.org/mountjoy

CSPG Canada's Energy Geoscientists

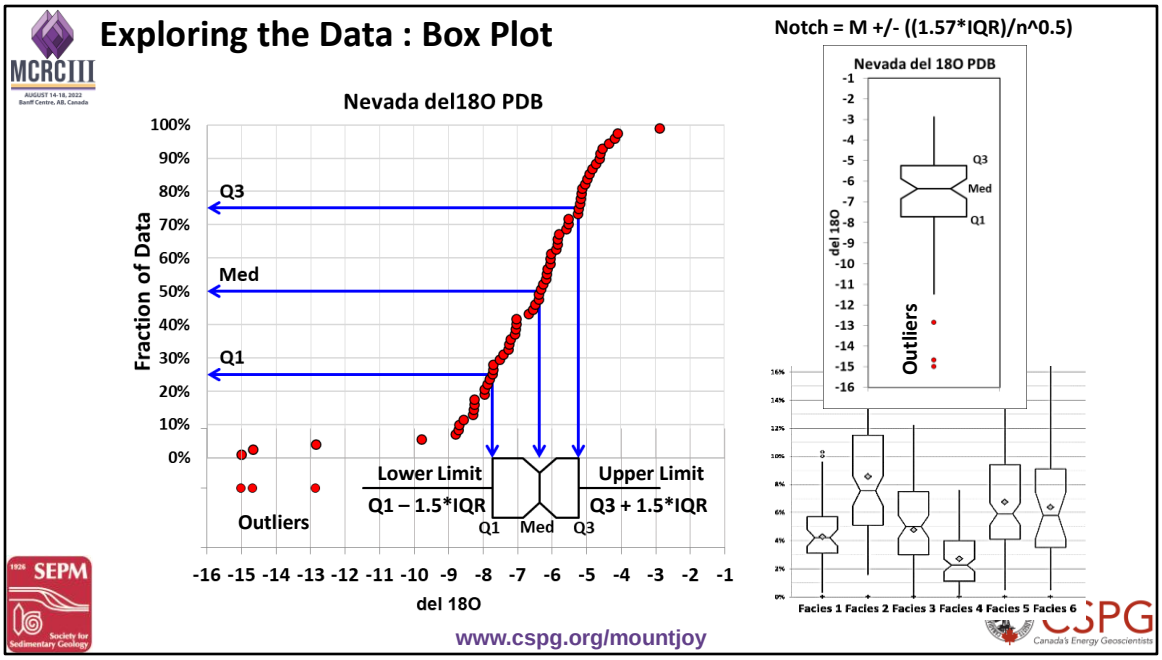
This is a rapid-fire review of Microsoft Excel workbooks that you can use to make your own publication-quality statistical graphs.



Statistical graphs appear in prestigious journals. They convert tables of numbers into portraits that reveal the structure of data. They are not just pretty pictures, they help you to explore your data, to look for patterns or problems, and communicate your interpretations to other researchers.

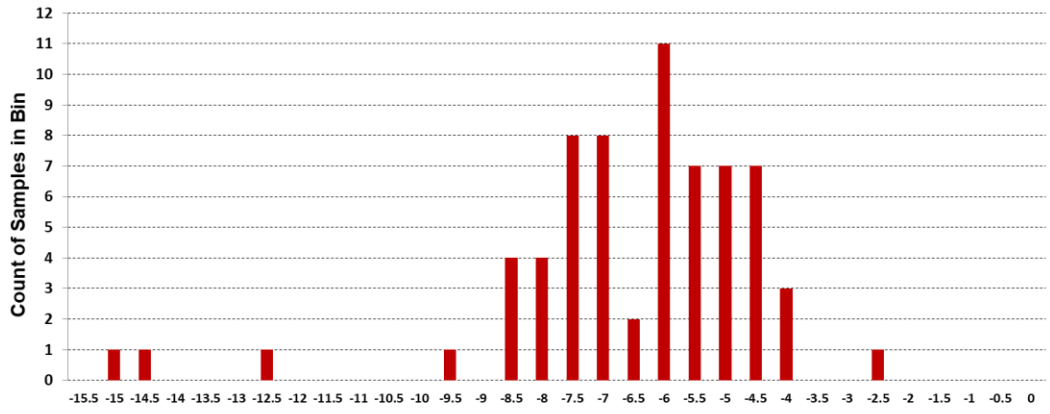


If your organization has a site license for commercial software, you should use it. But if you can't afford a commercial package, you can use this set of free templates. Enter your data into a column, and Excel will automatically create this set of graphs. In this case, showing the distribution of oxygen isotope values within Ordovician limestones.



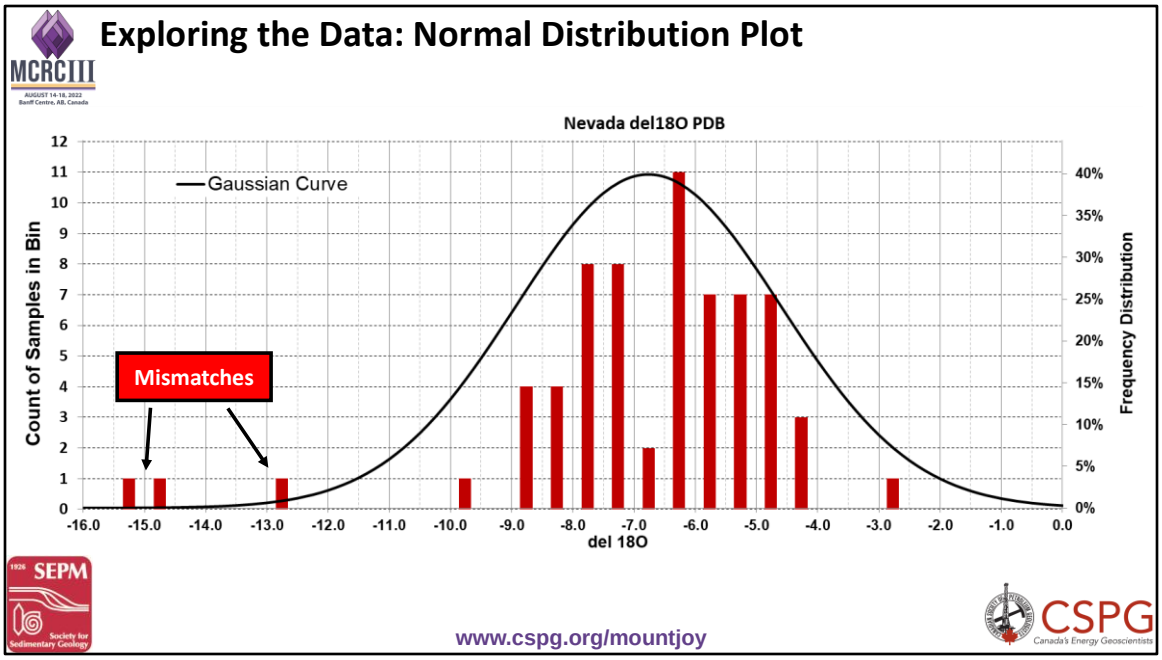
Box plots give a summary display of the distribution, showing first and third quartiles and the median, along with a formal definition of outliers. The notch shows the 95% confidence interval for location of the population median.

Exploring the Data: Histogram



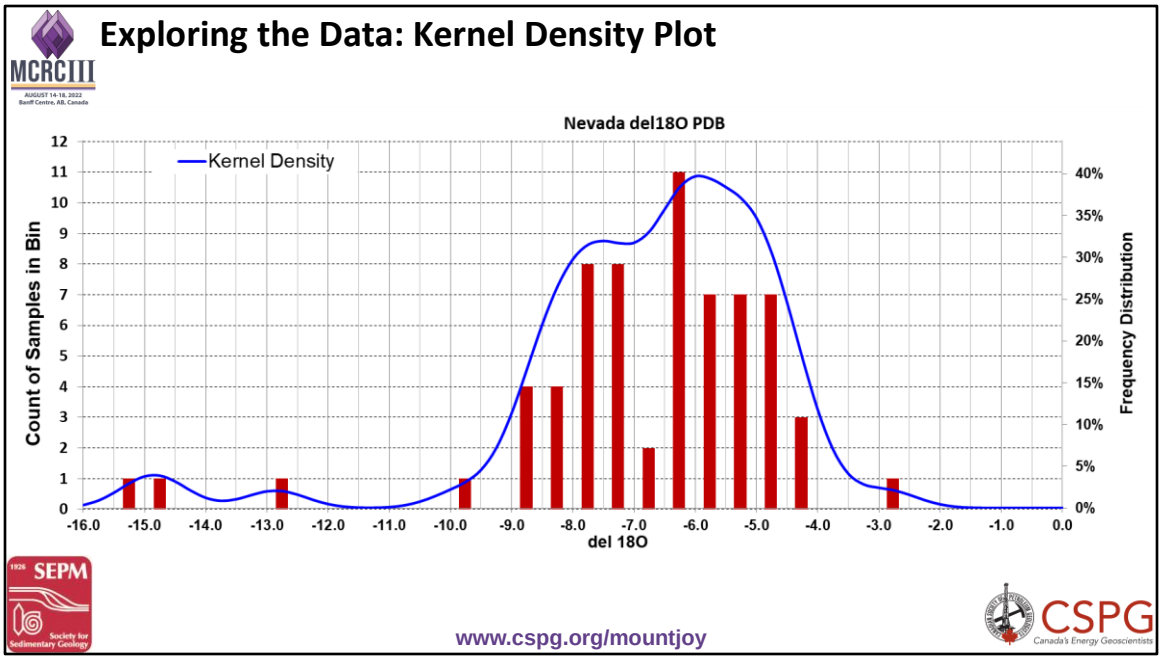
www.cspg.org/mountjoy

Histograms are familiar ways to summarize a data distribution, where the relative height of the bars represents the density of observations within intervals.



The normal distribution plot is the familiar Gaussian bell curve which is calculated solely from the mean and standard deviation of the sample data set. Here the bell and histogram don't match up, implying that the data may not correspond to a Gaussian distribution.

Exploring the Data: Kernel Density Plot



In contrast, a Kernel Density plot uses every data point. The curve show areas of higher and lower data density. The kernel refers to the smoothing function used to make the estimate, a Gaussian Kernel in this case, where bell curves are placed at each data point and then integrated to make the overall blue curve.

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

Micrometeoroid Velocity: NASA, 2015

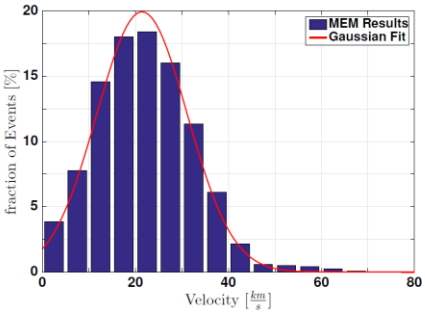
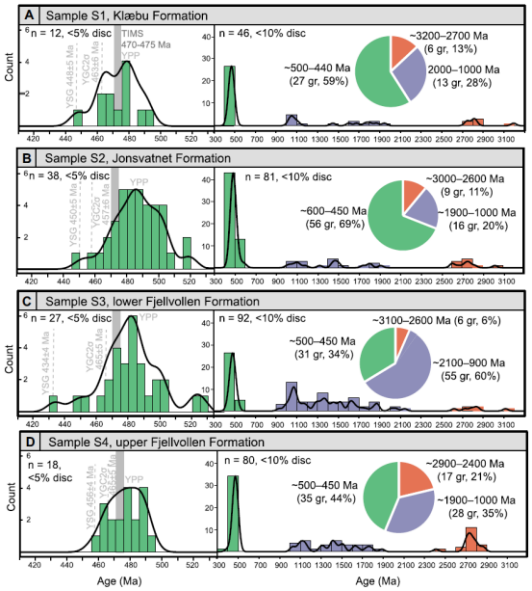
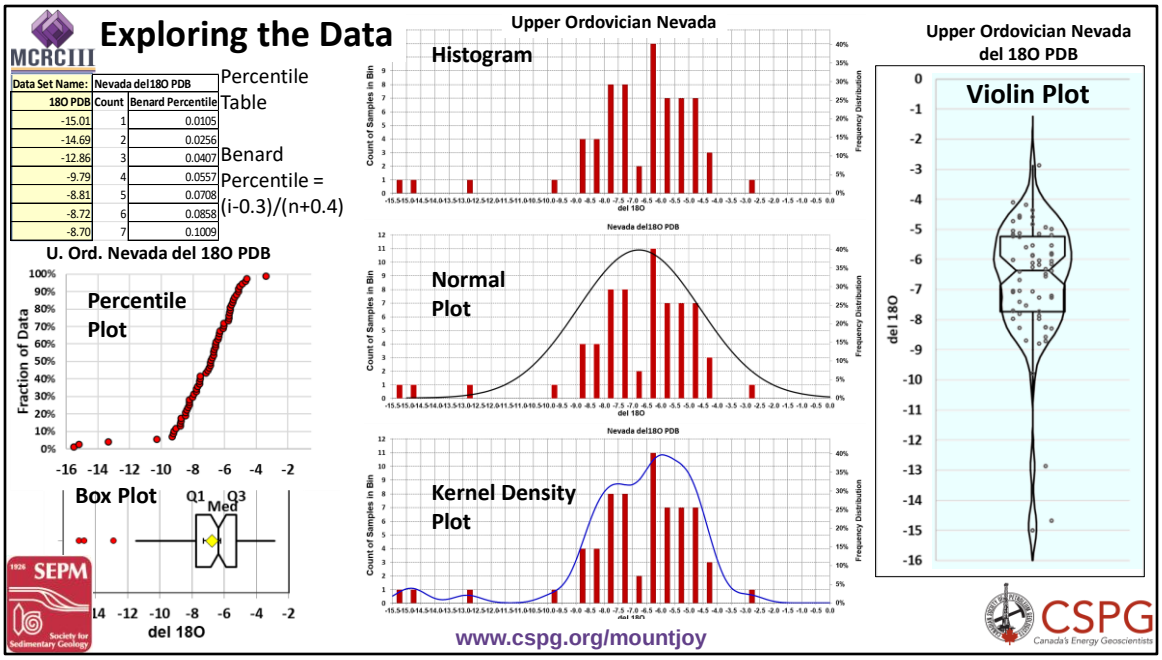


Fig. 3. Probability distribution of impact velocities for LPF micrometeoroid collisions during science operations. Histogram in blue are estimates from the NASA Meteoroid Engineering Model and a representative ephemeris for LPF. The fit in red is a best-fit normal distribution.

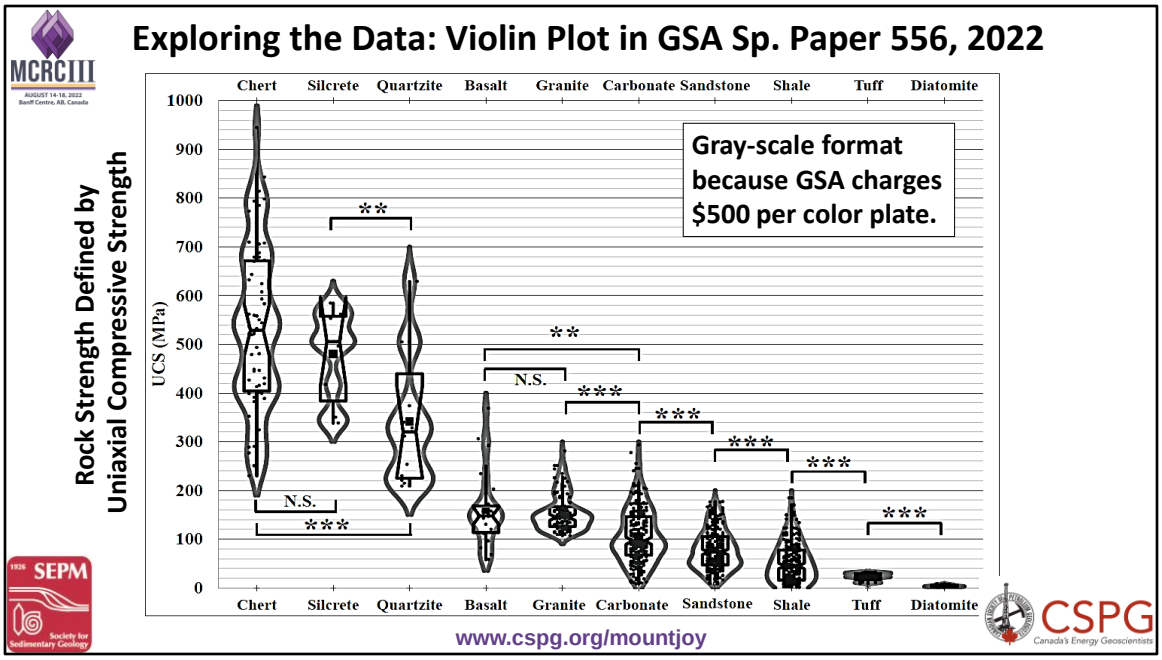
Detrital Zircon Ages: GSA Bull., 2022



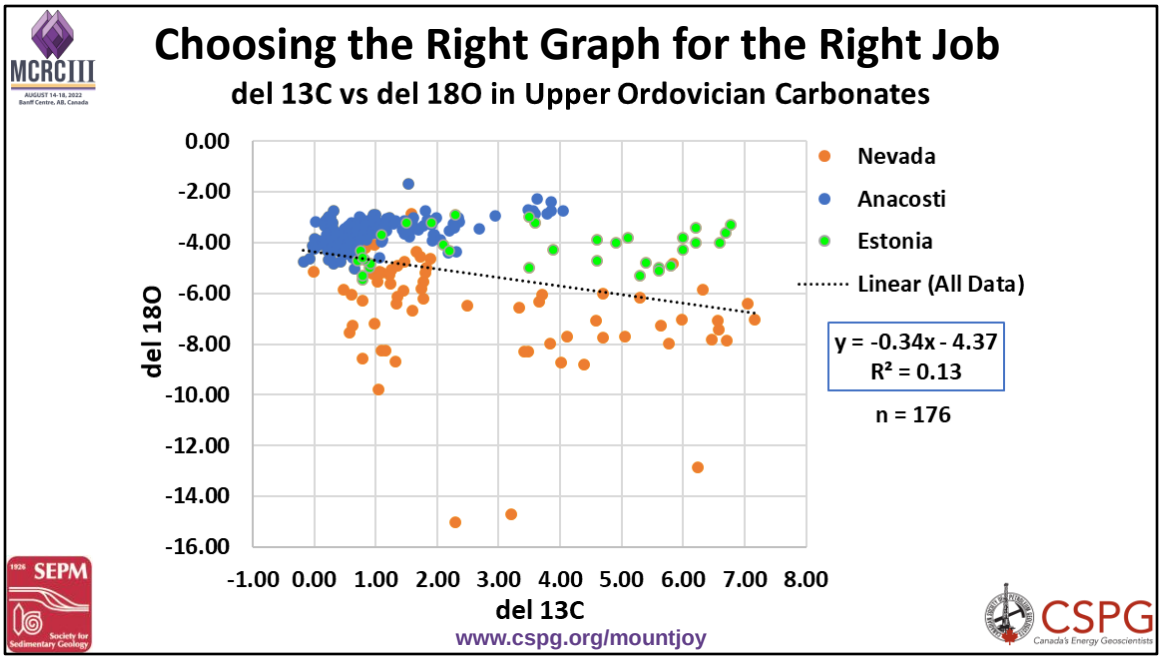
Histograms, bell curves, and Kernel Density plots appear in many scientific publications.



One of the most powerful statistical graphs is the Violin Plot. It combines the Kernel Density and Box Plots with all the data points, and provides as much information in one small space as all the other plots combined.



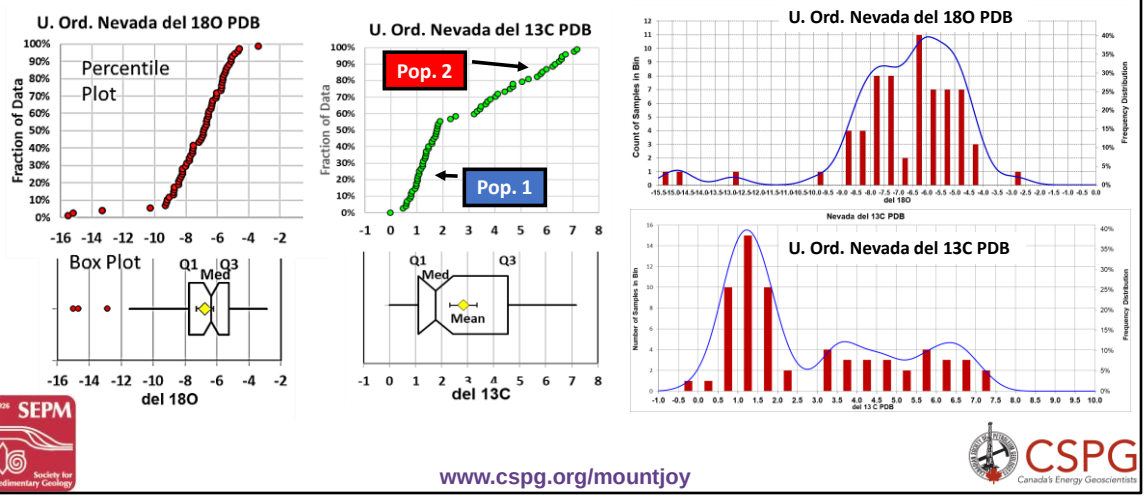
Violin plots are great for comparing data sets. This plot consolidates over a thousand observations into a comparison of rock strength among different rock populations.



Some graphs are better than others in certain situations. This XY scatterplot shows carbon and oxygen isotope data from 176 samples of Upper Ordovician limestone. It's a shotgun pattern. It records each isotope number, but there is no correlation, and that's the only obvious story you can tell from this graph.

Telling a Story with Statistical Graphs

- First visual look at the data: Are there typos, measurement errors, sample mixups?
- If valid, are the outliers part of a different population from the rest of the samples?
- Symmetry of the distribution: Normal? Skewed? Why?

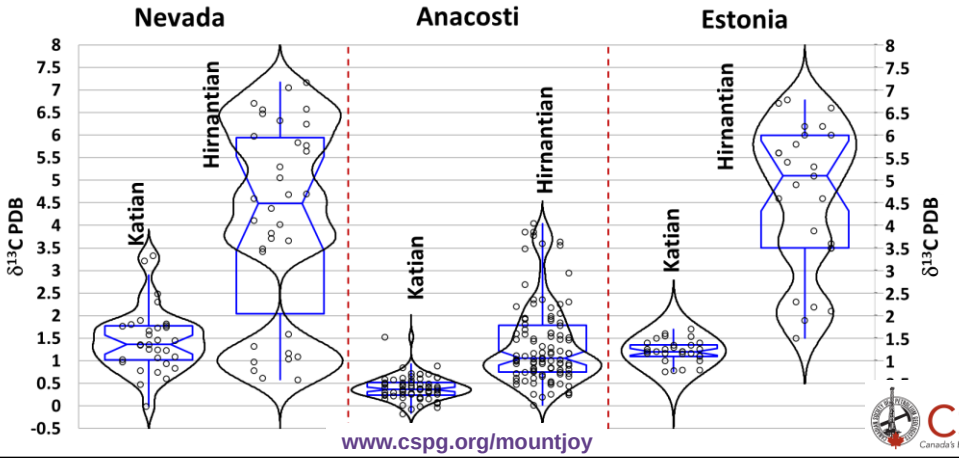


www.cspg.org/mountjoy

You can tell a more interesting story by using different graphs. This example uses Percentile, Box, and Kernal Density plots to indicate two different populations of carbon isotope values in Upper Ordovician carbonates.

Violin Plots are well suited for comparisons among different groups.

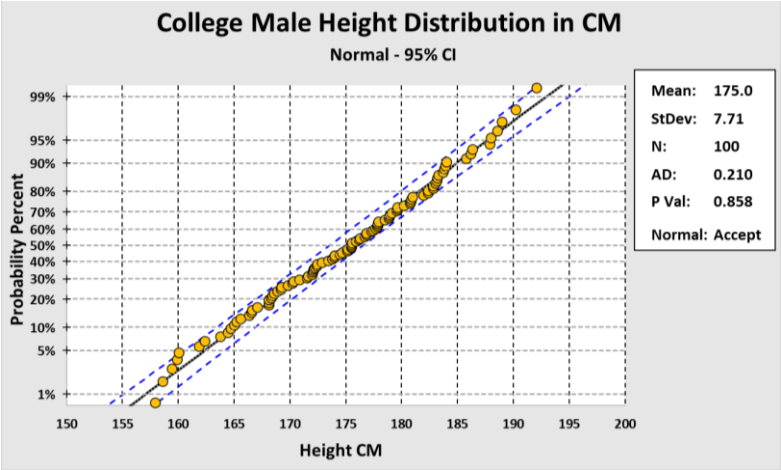
$\delta^{13}C$ numbers from Western North America, Eastern North America, and Central Europe record an "excursion" to higher $\delta^{13}C$ values from Katian to Hirnantian stages, interpreted as a shift from greenhouse to icehouse conditions, coincident with a global extinction event.
(Ahm et al.,2017, http://dx.doi.org/10.1016/j.epsl.2016.09.049)



Violin plots show an "excursion" of carbon isotope values from Katian to Hirnantian stages at widely separated outcrops. The published story is that the excursion records a global shift from greenhouse to icehouse conditions, coincident with a Late Ordovician global extinction event.

Probability Plots: Modified Cumulative Distribution Plots

- Useful for comparing samples to populations.
- Useful for comparing populations.
- Useful for detecting errors in calculation or experiment design.

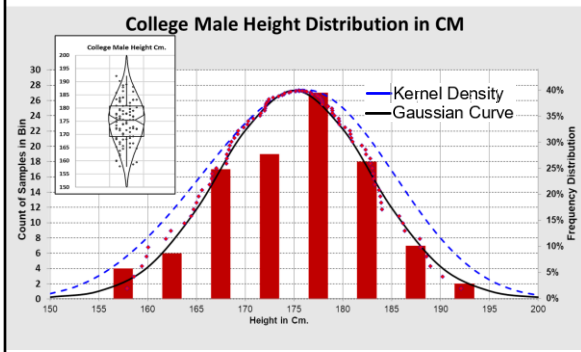


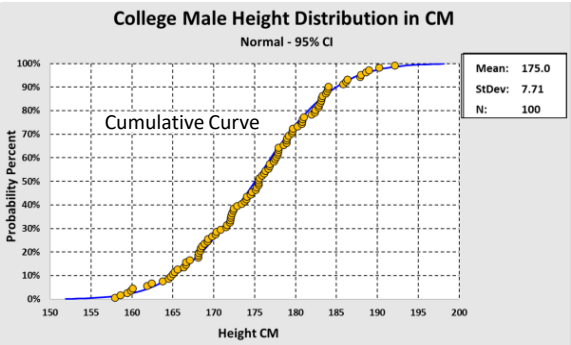
www.cspg.org/mountjoy

Probability plots are modified cumulative distribution curves that are useful for comparing populations, as well as for finding errors in your data.

Probability Plots: Modified Cumulative Distribution Plots

- Height distribution of 100 male college students.
- Bell Curve transformed to Cumulative Curve.
- Next step is to transform Cumulative Curve into Probability Plot.



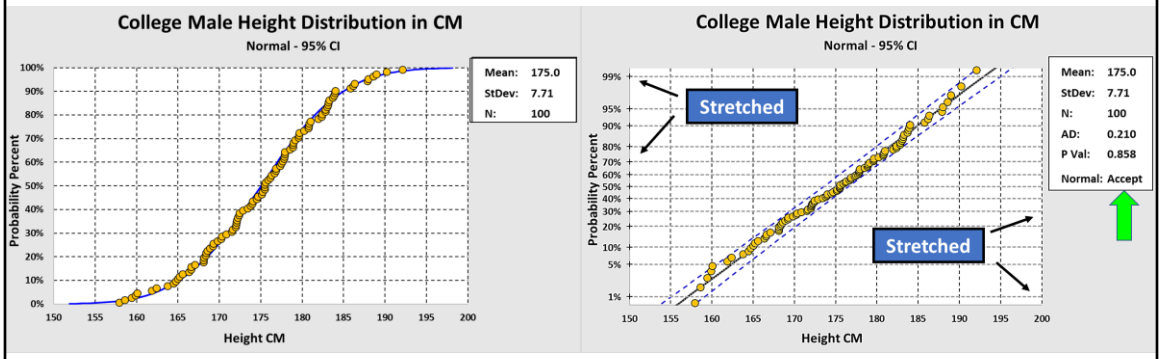


www.cspg.org/mountjoy

In this example, the heights of 100 male college students are plotted as a histogram, along with a hypothetical gaussian curve and an actual kernel density curve. A violin plot summarizes the data. The bell curve is replotted as a cumulative curve. The bell and cumulative curves are theoretical constructions based on the mean and standard deviation of the data set, and not on location of the individual data points. The fact that the observed orange points plot close to the theoretical curve suggests that these data may come from a normal distribution.

Probability Plots: Modified Cumulative Distribution Plots

- Cumulative Curve transformed into Probability Plot.
- Confidence Bands constrain location of the normal-fit line.
- Anderson-Darling Statistic tests hypothesis of a Normal Distribution fit.

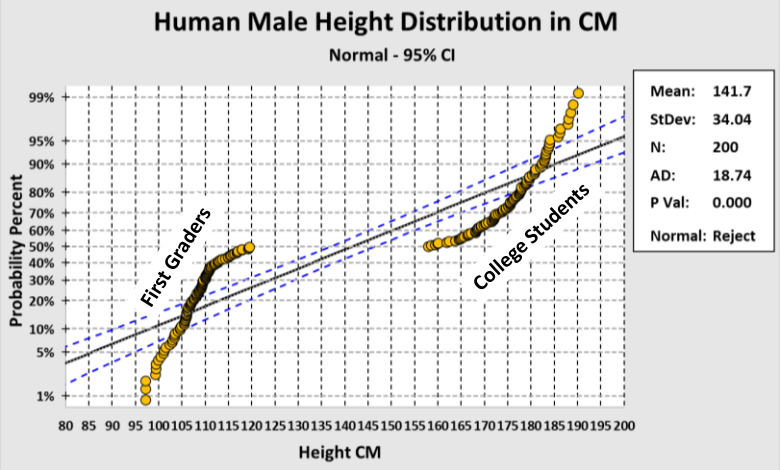


www.cspg.org/mountjoy

The cumulative curve is converted to a probability plot by stretching the Y-axis at the tails. In addition we add confidence bands that show the uncertainty in location of the model probability line. In this case, the bands are narrow, constraining the location of the line. Another difference is the addition of the Anderson Darling Test for Normality. The close correspondence of the actual orange data points to the calculated normal-fit line suggest a normal distribution. But by adding the AD test, it is possible to confirm whether or not the data come from a normal distribution. The AD test is found in all commercial software packages. The output from this template will match the output of any commercial package.

Probability Plots:

- There is no mathematical reason to assume that data follow a normal distribution.
- However, if samples are drawn from the same population, they often are.
- Outliers might be typos, measurement errors, or a mix of different populations.

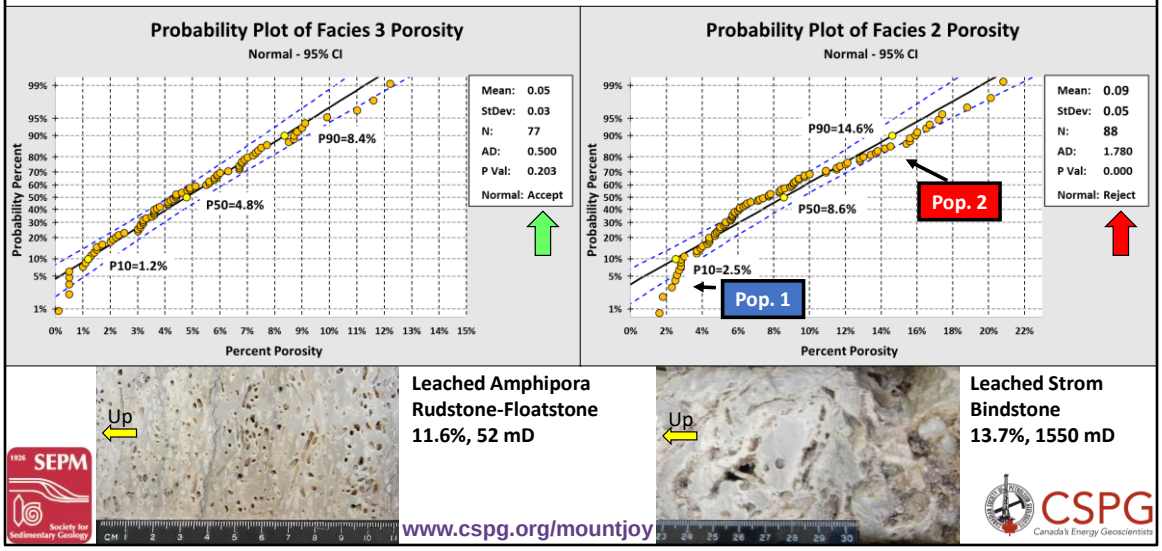


www.cspg.org/mountjoy

There is no mathematical reason to assume that a data set should follow a normal distribution. But if the samples are drawn from the same population, the data often are normal. This means that probability plots are useful for detecting outliers that could be measurement errors, or mixes of different populations. In this example, imagine telling a robot to go out and find a data set of human male height distribution. The AI follows orders and brings back a table of 200 measurements. If you didn't take a closer look, you might just use these numbers, but if you made a probability plot, you'd see that this is a mix of two different populations. Closer investigation would reveal that the AI managed to grab one data set of first grade students and another data set of college males. Clearly, any inference you might make based on this normal-fit line would be meaningless.

Probability Plots and Test for Normality

- Porosity within leached Amphipora reflects uniform size and composition.
- Porosity within leached Strom bindstone reflects variable leaching intensity.



Leached Amphipora
Rudstone-Floatstone
11.6%, 52 mD

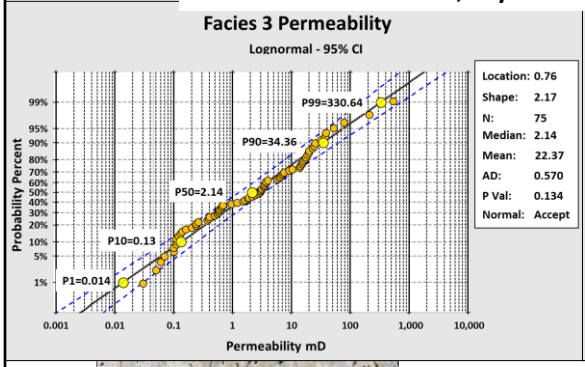
Leached Strom
Bindstone
13.7%, 1550 mD

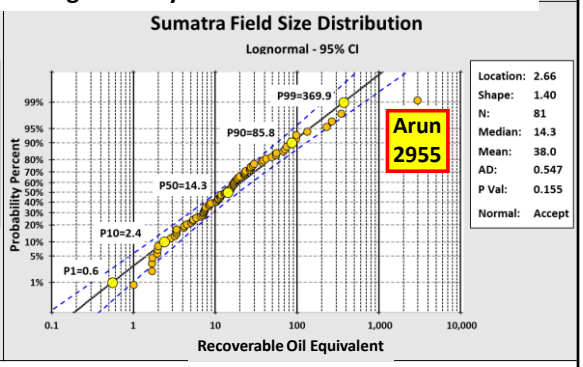
www.cspg.org/mountjoy

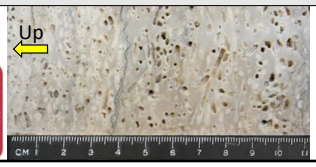
Probability plots show both passing and failure of normality tests. In this case, Amphipora of uniform size and shape were leached from a dolostone, creating moldic pores. Whole-core analysis of Amphipora Floatstone facies show a normal distribution of porosity. In contrast, intensity of leaching of stromatoporoid bindstone facies varies from partial to complete, resulting in a mix of populations and failure to pass the test for normality. One explanation is that Amphipora are uniform in size and shape and leaching intensity, while binding stromatoporoids are more variable in size and shape and resultant leaching intensity. A follow-up study could try to explain the low and high porosity bindstone populations.

Log Probability Plots

- The natural log of the variable is normally distributed.
- Permeability and Field-Size within a basin tend to be lognormally distributed.
- Arun is a different; anyone want to guess why?







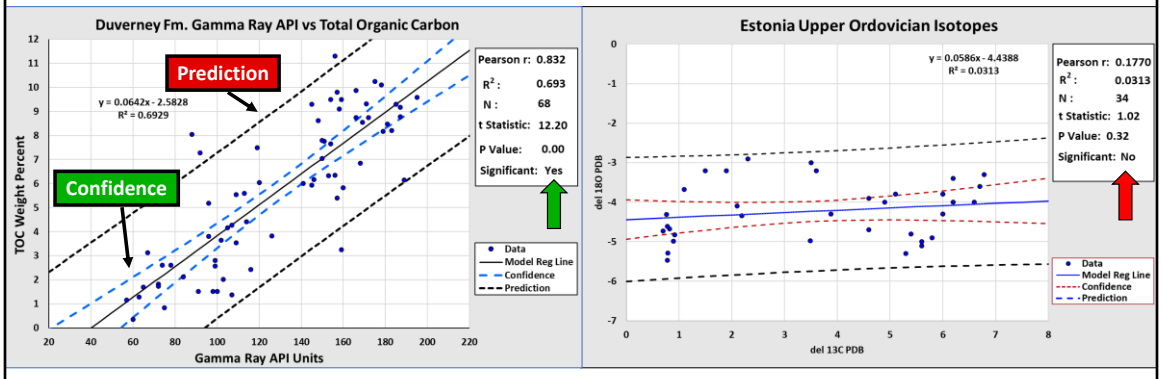
Leached Amphipora
Rudstone-Floatstone
11.6%, 52 mD

www.cspg.org/mountjoy

Log probability plots test whether the distribution of the natural log of a parameter is normally distributed. Permeability and field-size distribution within a basin tend to be log-normally distributed. Permeability in leached Amphipora floatatone is lognormally distributed. Size of 80 fields in the Sumatra Basin of Indonesia is lognormally distributed; but Arun, the 81st field is a clear outlier. Why is it different? It's a carbonate buildup; all the rest are lacustrine fluvio-deltaic sand reservoirs.

Modeling the Data with Linear Regression

- First, describe and understand the sample distribution.
- Then, design a model: estimate value of Y based on observed value of X.
- Linear Regression Plots with Linear and Log Scales.

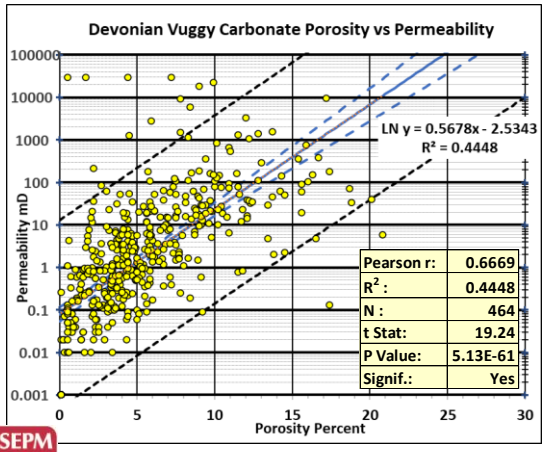


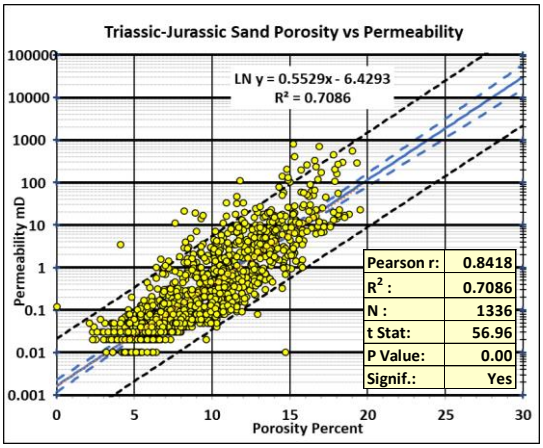
www.cspg.org/mountjoy

X-Y scatter plots are useful if you can use them to tell a story, but often these plots lack key information. Commercial packages include t-tests to assess the significance of the correlation, and graphics that bracket the uncertainty of the model. My template adds these details, showing the 95% confidence interval for location of the regresssion line, as well as 95% prediction intervals. These are here to give you a reality check on your predictions. The graph on the left shows a significant correlation while the plot on the right shows no correlation.

Modeling the Data with Linear Regression

- Linear X Axis Scale, Log Y Axis Scale.
- Standard porosity vs permeability cross plot.
- There is a significant relation, but prediction interval is very broad.



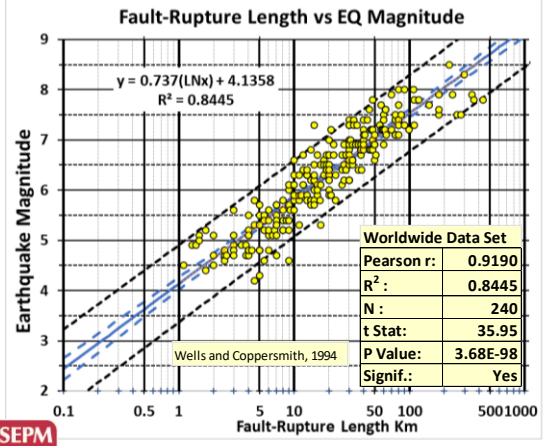


www.cspg.org/mountjoy

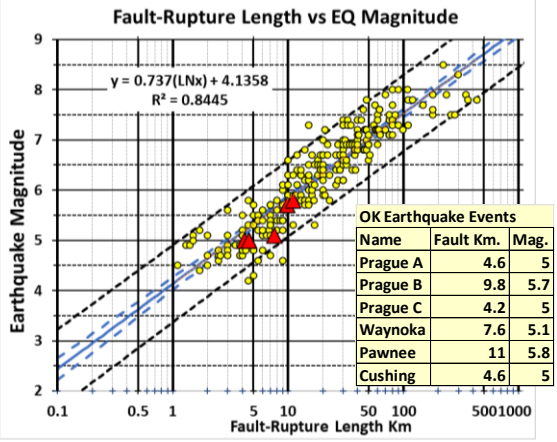
This example shows Porosity on a linear scale and Permeability on a log scale, with much more scatter in the vuggy carbonate reservoir than the sandstone. Both correlations are significant, but the prediction limits are 4 orders of magnitude wide in the carbonate; not very useful.

Telling a Story with Linear Regression

- Log X Axis Scale, Linear Y Axis Scale.
- Earthquake Magnitude as a Function of Fault-Rupture Length.
- Injection into faults of length greater than 4 km will produce damaging earthquakes.



Worldwide Data Set	
Pearson r:	0.9190
R^2 :	0.8445
N :	240
t Stat:	35.95
P Value:	3.68E-98
Signif.:	Yes



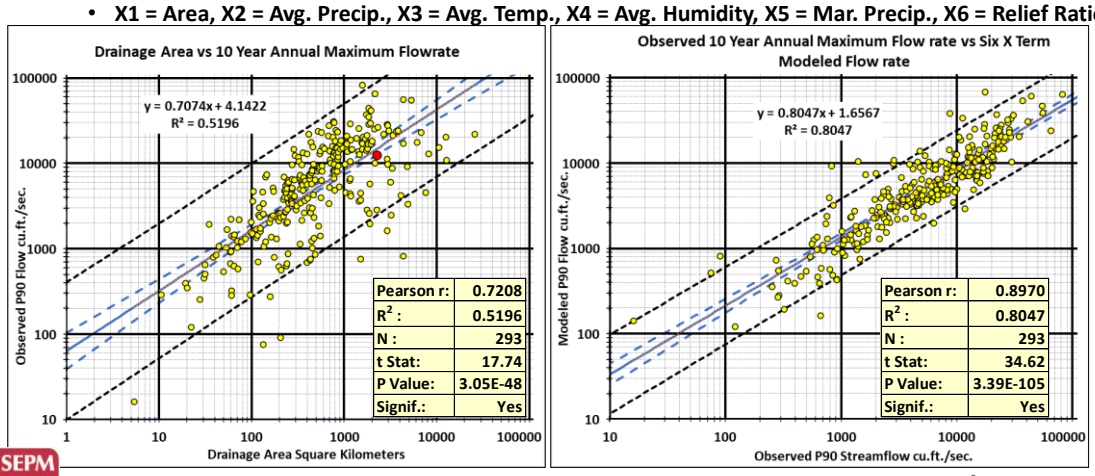
OK Earthquake Events		
Name	Fault Km.	Mag.
Prague A	4.6	5
Prague B	9.8	5.7
Prague C	4.2	5
Waynoka	7.6	5.1
Pawnee	11	5.8
Cushing	4.6	5

www.cspg.org/mountjoy

Telling a story with linear regression; the graph on the left shows a significant correlation between fault-rupture length and earthquake moment-magnitude. The red triangles in graph on the right are magnitudes of 6 additional earthquakes that were induced by injection of waste water into basement faults. The red triangles plot well within the prediction interval of the correlation. The 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.

Multiple Linear Regression

- Estimate of Y improves with additional X variables.
- Y = Intercept + (X1*mX1) + (X2*mX2) + (Xi*mXi) …
- X1 = Area, X2 = Avg. Precip., X3 = Avg. Temp., X4 = Avg. Humidity, X5 = Mar. Precip., X6 = Relief Ratio



www.cspg.org/mountjoy

Multiple linear regression uses additional x parameters for modeling. In this example, stream-gauge data from the USGS is used to model 10-year flood events based on basin-drainage area. One would expect that the larger the area, the greater the flow. The correlation is significant, but the r-squared number says that basin area explains only about 52% of observed stream flow; other factors must account for the rest of the variation. Multiple regression adds parameters like annual precipitation, average temperature, humidity, March precipitation, and basin-relief ratio, to see if one or more of these parameters might improve the correlation. The graph on the right shows significant improvement by using these parameters, which now explain more than 80% of stream flow.

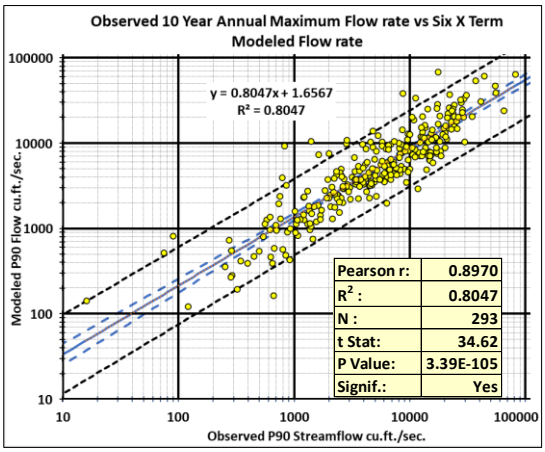
Multiple Linear Regression

- Did all six X terms really add value?
- Excel Data Analysis Output Table reveals the significance of each term.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-5.67793263	1.545611565	-3.67358	**0.000285**
DRAIN_SQKM	0.833754403	0.028626519	29.12525	**1.48E-87**
PPTAVG_BASIN_CM	1.258018664	0.276733997	4.545949	**8.08E-06**
T_AVG_BASIN	0.172691256	0.045244328	3.81686	**0.000166**
RH_BASIN	0.563871957	0.523641696	1.076828	**0.282465**
MAR_PPT_CM	0.220703865	0.169567217	1.301572	**0.19411**
Relief_Basin	0.095870365	0.089450029	1.071776	**0.284724**

Drainage Area, Average Precipitation, and Average Temperature are significant.

Relative Humidity, March Precipitation, and Basin Relief are not significant.

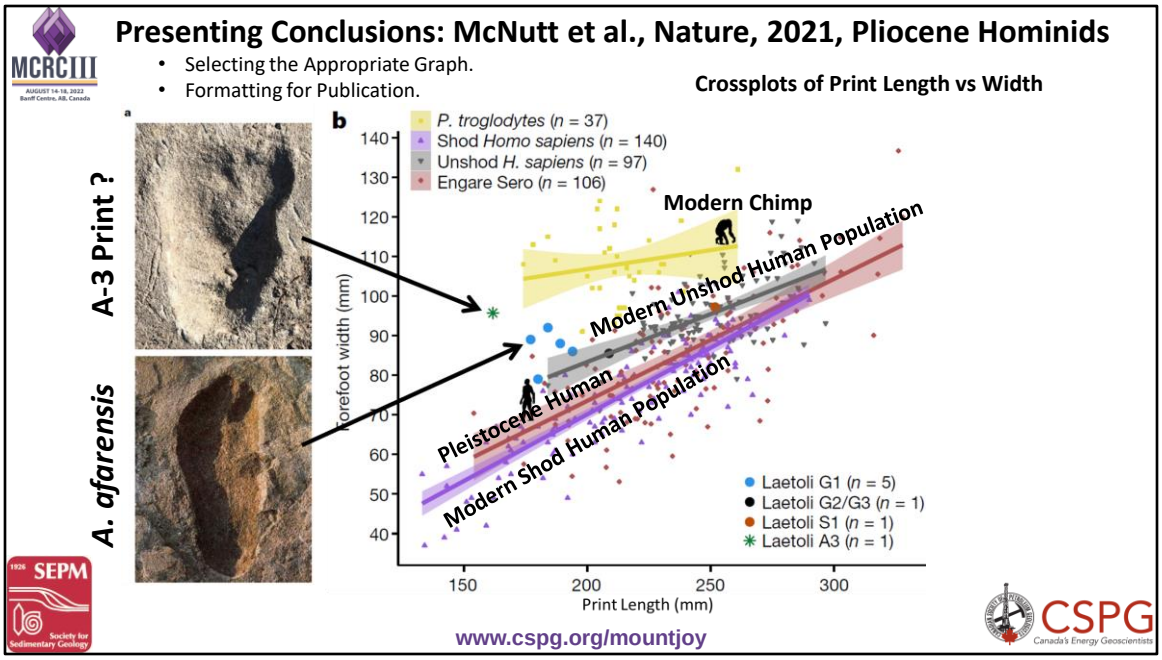


www.cspg.org/mountjoy

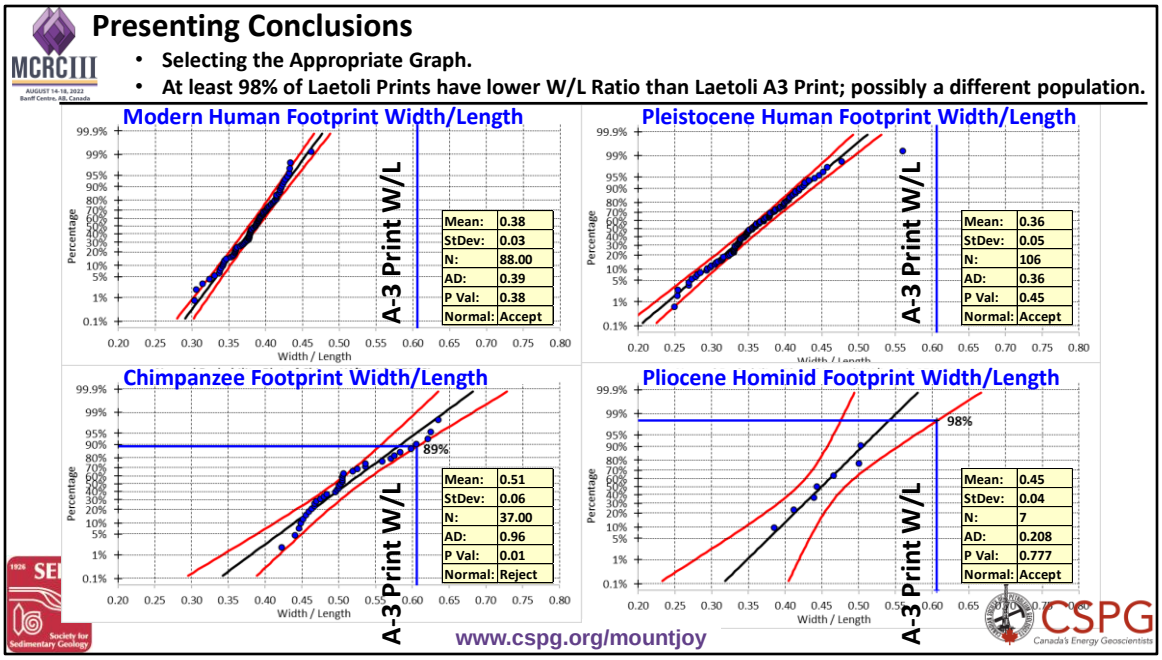
But did we really need to use all six of these X terms? The Excel Data-Analysis Output Table shows that drainage area, average precipitation, and average basin temperature are significant to the correlation, while relative humidity, March precipitation events, and basin relief are not adding significantly to the correlation. You could leave these out of the model.

Presenting Conclusions: McNutt et al., Nature, 2021, Pliocene Hominids

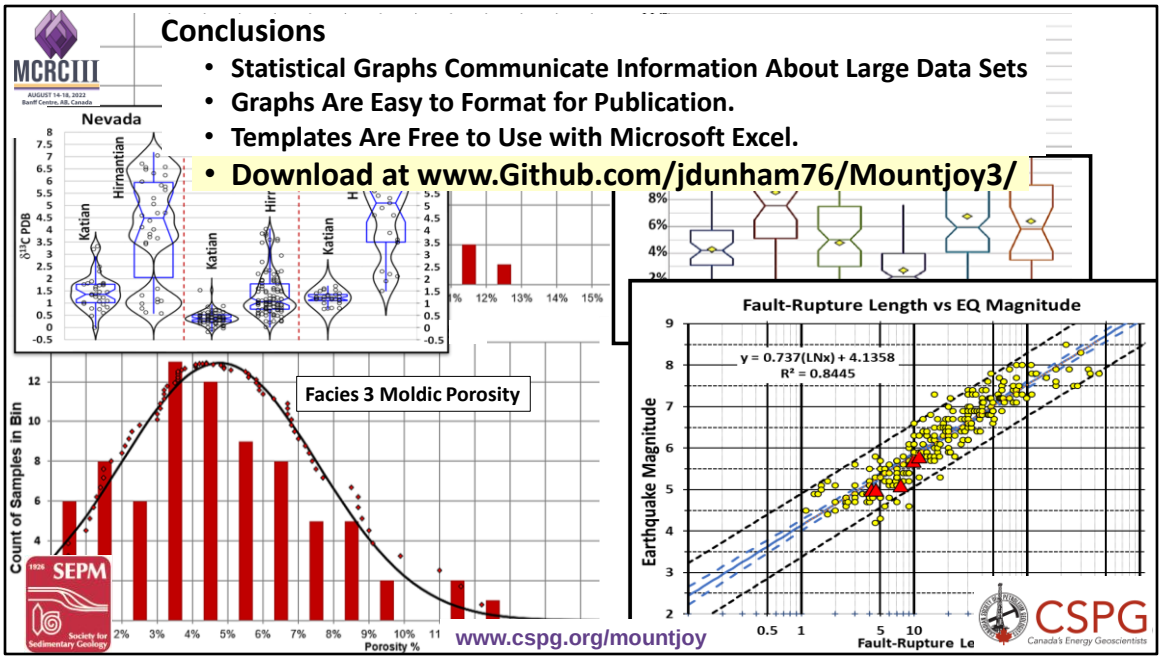
- Selecting the Appropriate Graph.
- Formatting for Publication.



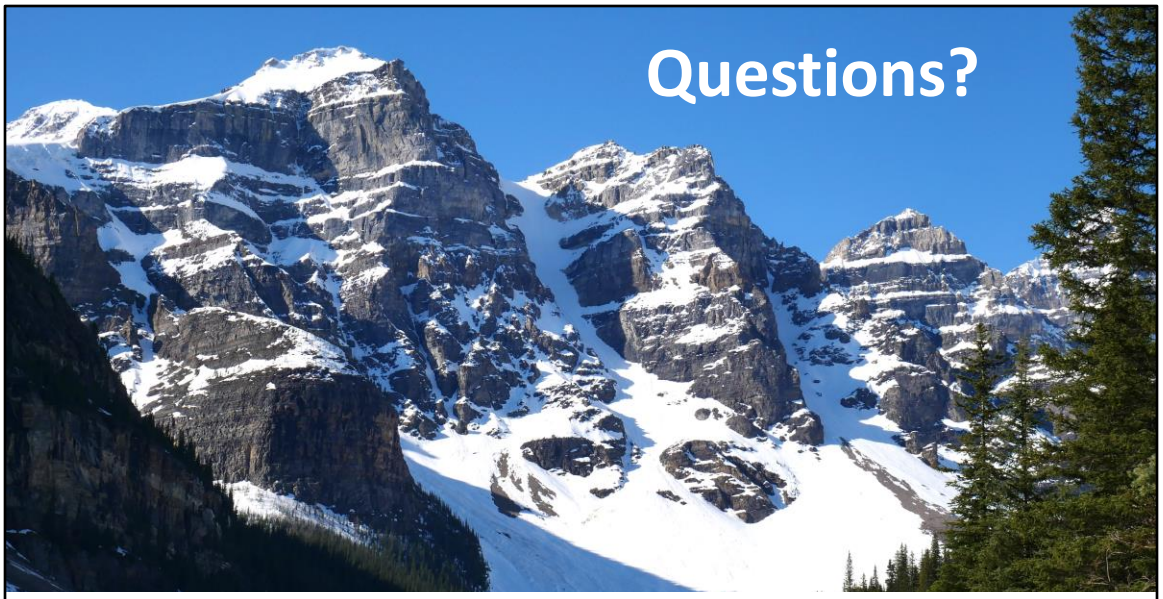
Choosing the right graph for the right job is important. This example was published in Nature. 8 Pliocene hominid footprints were discovered at a single locality in east Africa. The authors made a graph that plots print length versus print width, relative to modern and Pleistocene human populations and modern chimpanzees. The author's point was that the Pliocene prints appeared to be statistically different from both the human and chimp prints, but more significantly, that one of the Pliocene prints was very different from the other 7 prints.



I took the same data and created a width divided by length parameter, and made probability plots to test the conclusion that the A-3 Pliocene print was different from the other 7 Pliocene prints. The A-3 print is more than 99.9% different from modern or Pleistocene humans. On the other hand, the A3 print intersects the lower confidence line of the Chimp plot at the 89th percentile. The A-3 print intersects the lower confidence line of the other Pliocene prints at the 98th percentile. A3 could be part of the same population, but if so, it lies at the upper extreme of the Pliocene distribution. The authors concluded that there was a reasonable chance A-3 was made by a different hominid, other than Australopithecus afarensis, one with a wider foot more like that of a chimp. I sent my plots to the authors, and received a reply that they were pleased to see their data visualized in a different way that still supported their conclusions. These graphs make the additional point that more footprint data would tighten up the confidence interval of the Pliocene hominid plot and improve the model. This was exactly Professor McNutt's point. She needs more funding to find more prints. Nature agreed, and published her paper.

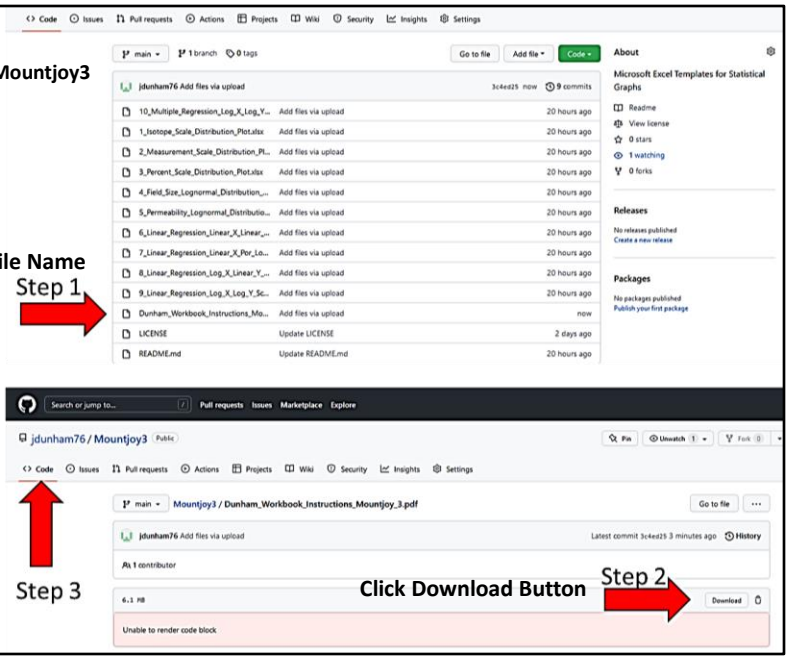


Statistical graphs help people to visualize large data sets. They are easy to format for publication. They are free to use with Microsoft Excel. You can download templates for all of these graphs at Github.



Download at www.Github.com/jdunham76/Mountjoy3/

https://www.github.com/jdunham76/Mountjoy3



Steps to download files from Github.