

**Section 2
Exploratory Data Analysis
(EDA)**

Smart Sampling

Exploratory Data Analysis (EDA)

**EDA is everything you do to understand your data
It includes both objective and subjective analyses.**

Three essential functions of EDA:

- error checking
- understanding physical processes for use in modeling
- statistical validation of results

Topics

- mapping the data
- histogram techniques
- probability-plotting techniques
- correlations among multivariate data
- data transformations

Mound Accelerated Site Technology Deployment

Slide 2-2

Error checking enables you to find the mistakes in the data set - typos, transpositions, etc.

Looking at whether or not the distribution process implies a strong spatial correlation, eg. aerial plume vs guy with wheelbarrow, enables you to develop the appropriate model.

Probability-plotting techniques allow you to identify the different populations in a contaminated area.

Smart Sampling

EDA: Mapping the Data

Plot it! Humans are very good at processing visual data.

- *Look for spatial patterns, correlations with other things*
- *Spot data "busts:" keypunch errors, transposed coordinates*
- *Beware of pitfalls: data clustering, preferential sampling*
- *Use:*
 - Simple pencil and paper posting of values
 - Plots with colors or symbols proportional to value
 - Indicator plots
 - Quick contouring with a "dumb" program

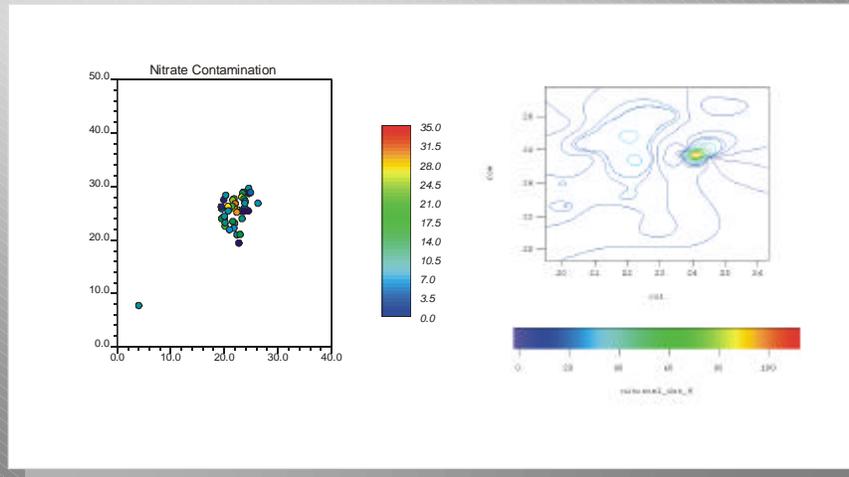
Mound Accelerated Site Technology Deployment

Slide 2-3

Data clustering will distort statistics particularly if the clusters are located around high or low values.

Smart Sampling

Mapping the Data

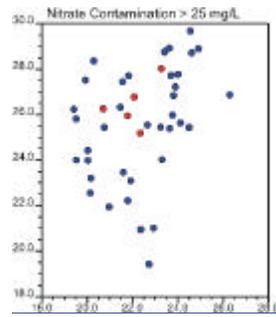
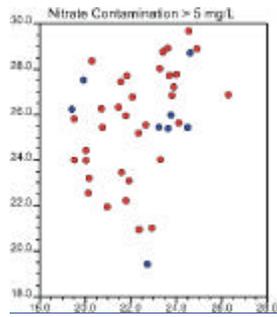


Mound Accelerated Site Technology Deployment

Slide 2-4

The outlying point on the left plot shows you a possible key punch problem. The point anomaly in plot on right shows a "bullseye."

$I(x)=1$ if $Z(x) > Z^*$; $I(x) = 0$ otherwise



Mound Accelerated Site Technology Deployment

Slide 2-5

Data transformations draw the eye to patterns.

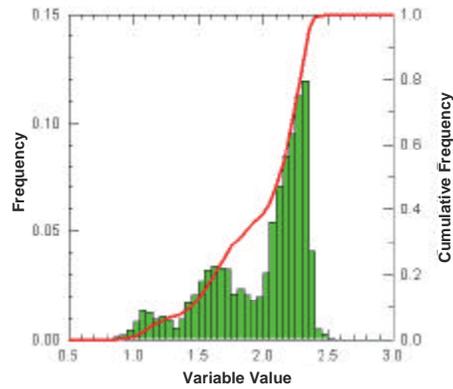
Chris: one of the colors in these plots need to be changed to emphasize difference between the red and purple samples on the viewgraph.

- **Simple Histograms**
 - pdf vs. cdf format
 - Check for outliers (and their cause)
 - Multimodality: evidence for multiple processes
 - Clustering of data or preferential sampling
- *CDF format provides an important conceptual link to downstream modeling*
- *Relationship of the histogram and other “ensemble statistics” to model output*
- *Need to decluster data to remove effects of preferential sampling*

CDF format: By using a uniform distribution to generate random numbers, we can map those numbers, via the cdf, to values that honor the original data and at the same time provide a random component.

Clustered data is a major concern. If it is not taken into consideration, then you have the potential for “double counting” and a faulty histogram.

We are going to reproduce the histogram in the models.



Mound Accelerated Site Technology Deployment

Slide 2-7

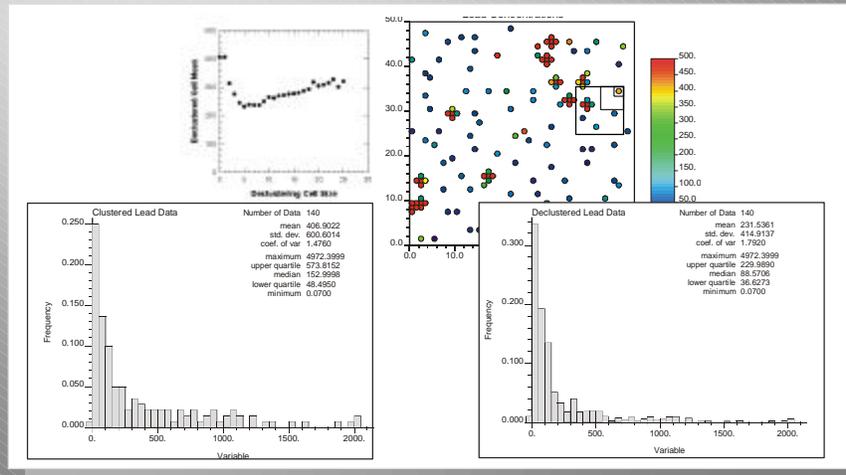
This histogram shows a multimodal distribution.

Cumulative distribution functions (cdfs) have a conceptual link to modeling and simulations. The cdf is the mechanism by which we reproduce the target histogram.

You need to understand the relationship between the input data, the model histogram and other statistical measures of the data.

Smart Sampling

EDA: Impact of Sample Clustering on Histogram



Mound Accelerated Site Technology Deployment

Slide 2-8

This data set shows that every time the samplers got a hot value they took four more samples right next to it. The histogram looks like it has an extreme tail on it. This tail is created by the replicate sampling. This problem can be solved with some relatively simple de-clustering techniques.

The histogram on lower left (without weighting) show replicate sampling of high values.

There are a number of different techniques that you can use to weight the points when de-clustering. You could use a cell-based technique or Kriging, a method of calculating a weighted average

The plot on upper right show weighting done with a cell based technique in which the average of all data within each cell is computed for a number of cell sizes.

The graph on the upper left shows the output of the cell-based technique. Taking the local min of this graph gives us the size of the cell that is the de-clustered values.

On the lower right, you can see that the mean value of the de-clustered data is about half of the value of the original data.

Smart Sampling

EDA: Probability Plotting / Other Exploration

Simple Probability Plots

- plot concentration against probability value rather than against simple frequency
- implicit comparison to theoretical Gaussian distribution
- different underlying populations will plot as straight line segments
- can compare to populations other than Gaussian

Quantile-Quantile Plots

- plot corresponding quantiles of any two populations
- use in understanding cross-correlated variables
- use in "validating" output models

Probability-Probability Plots

- plot corresponding cdf values of two populations

Mound Accelerated Site Technology Deployment

Slide 2-9

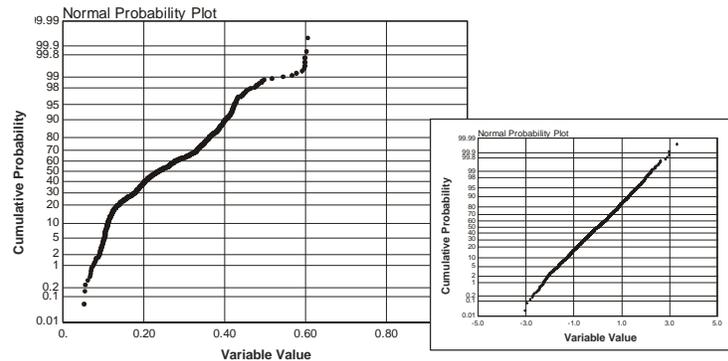
Probability Plots: This provides a simple comparison of a single sample to the normal distribution. *Chris - the following words are taken from your narration. Please rewrite to make sense to you.* You could compare the curve that you get to another non-Gaussian population. You could see if samples from two sites compare statistically, and you could see if you could bring in information from one site to a site where you don't know as much.

Quantile-Quantile Plots: The 25th quantile is the value at which 25% of the sample values are lower. 25% of the values are less than (x) value in one population, 25% of the values are less than (y) value in the other population, we can plot what those values are. This technique is based on empirical estimates, so if you only have 10 samples, each sample accounts for 10% of the data. If you've got 10,000 samples you are much more confident of the relationships you see.

We try to reproduce equivalent quantiles in the realizations/simulations generated for SmartSampling. If working with a small data set, don't try to reproduce much more than the median and the upper and lower quartiles (25%, 75%). If you have thousands of data points, you could conceivably reproduce the 10 deciles in your realizations.

Probability-Probability Plots: *Chris, please add notes*

Plots concentration vs. Gaussian probability



Mound Accelerated Site Technology Deployment

Slide 2-10

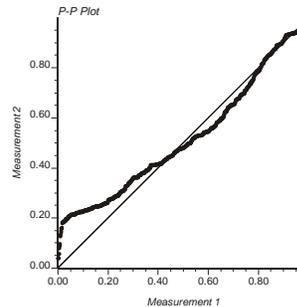
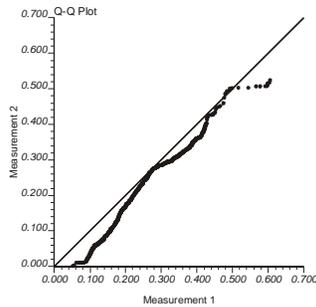
In the plot on the left, the existence of two major line segments implies that you probably have two separate physical distribution processes. The first 'slope' might be the background material, and the second slope might be the contaminant population that was put into the site.

The shape of plot on right shows that you might be looking at background material

Contaminated material should not fall into a perfectly normal curve (like the one plotted on the right.) Such a curve would not be consistent with a typical geological data set.

The plot on the right is a gaussian distribution.

Plot corresponding quantiles or CDF probabilities



Mound Accelerated Site Technology Deployment

Slide 2-11

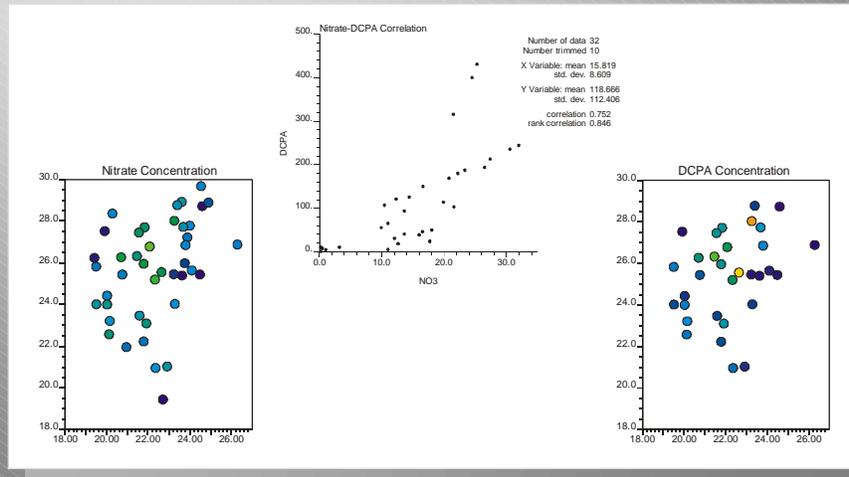
In the plot on the left, the Y axis represents the simulated output model, X axis is data. If this comparison does not fall along the 45°, something is wrong, the statistics are not being reproduced correctly.

We would like to compare equivalent quantiles. We can be more accurate when we have more samples.

Multiple contaminants of concern (COCs) Multiple measurement methods

- For our purposes: Scatterplot Analysis
 - *direct and inverse correlation*
 - *strength of correlation*
 - *correlation coefficient (r)*
 - *coefficient of determination (r²)*
 - *rank-order correlation coefficient*
- **Concept of conditional expectation**

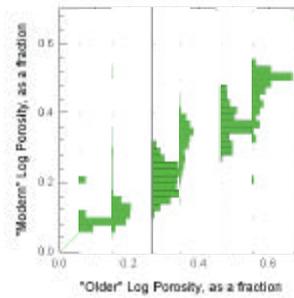
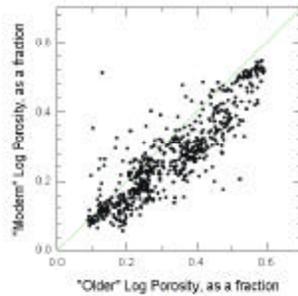
Chris - the following text was taken from your narration. Please rewrite it for your intended meaning and add any notes explaining the “Concept of conditional expectation”. Multiple measurement methods are used to verify correlation of actual and theoretical data.



Mound Accelerated Site Technology Deployment

Slide 2-13

Top plot shows what seems to be a positive correlation between the two variables. This follows since these two contaminants were introduced to the ground water by the same processes; agriculture.



Mound Accelerated Site Technology Deployment

Slide 2-14

In these two plots, we are taking data from the empirical to the theoretical realm. We will start interpreting a scatter plot as a conditional expectation.

If, in this case, I have a measured porosity value of twenty percent, what is the distribution of the other variable? What we've done here is plotted these as a number of histograms, where each different histogram is a function of concentration. If the value of the one variable is twenty percent, then the value I get for the other variable is expressed by this distribution in this exercise. And, if the value of the first variable is sixty percent, I get a very different distribution for the second value.

In theory, each of these distributions should be normal.

Any regression line connects the mean of the conditional distributions. The R-squared value (the correlation coefficient) reflects how spread out things are around that mean value. A perfect 1-to-1 correlation (R-squared = 100%) would have a perfectly straight line with no spike at the mean. Most real data sets have some slop in them that may be related to the physical distribution process.

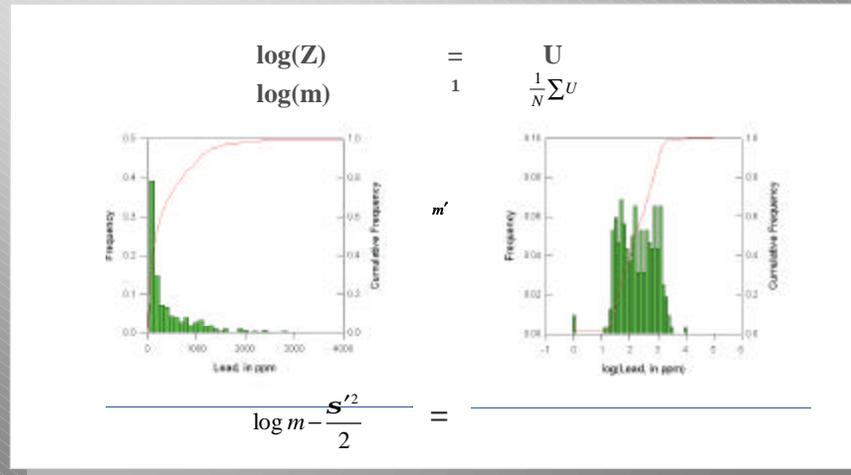
Smart Sampling

EDA: Data Transformations

- A powerful tool for understanding data.
 - reduce numerical artifacts that can obscure relationships
 - simplify portions of numerical modeling
- A “two-edged sword”
 - back-transformation may have negative implications

Examples:

- | | |
|------------------|--|
| - logarithmic | $U = \log(Z)$ |
| - indicator | $I = 1 \text{ if } Z < Z^*; \quad I = 0 \text{ Otherwise}$ |
| - rank-order | $Z \text{ in order } 1, 2, 3, N$ |
| - normal-scores | $m = 0 \quad s^2 = 1$ |
| - uniform-scores | $[0, 1]$ |



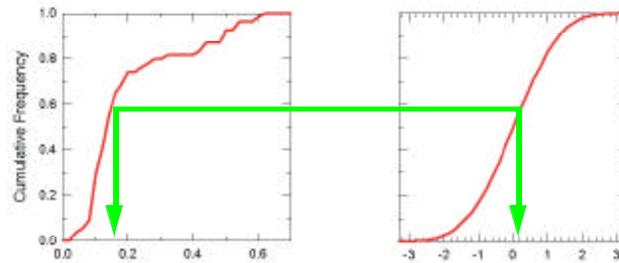
Mound Accelerated Site Technology Deployment

Slide 2-16

When taking log, you need to ask yourself what distortions you're buying into. The mean of the logs is not the log of the means. It is difficult to back out of a logarithmic transformation.

You can't exponentiate the values and expect to get the correct value.

Graphical conceptualization of quantile-preserving process



Mound Accelerated Site Technology Deployment

Slide 2-17

The graph on the left is a population which is not normal, and yet I can transform it to a population that looks exactly normal. Take each value on the left graph, track up to the cumulative frequency, across to the same cumulative frequency on the other graph and project the value onto a pure normal distribution, and track down to emerge into a transformed variable space.

In this way, all of the order relationships are preserved. The highest value is the highest value at the end, the lowest value is the lowest value at the end, the median value is the median value at the end. Even though you have changed the absolute magnitude of the values at the end, you have done nothing to change their spatial relationships. You've still got highs with highs and lows with lows, but computationally it may be much more desirable to work with these transformed values.

This transformation is very useful in the making of the variogram. It ensures that you don't get values that differ by great amounts, while keeping the spatial correlations.

What is a trend?

- *Trends and second-order stationarity?*
- *“Deterministic Geologic Processes”*
 - *Trend analysis and modeling must make geologic sense*
- *Removing a trend - Analysis of residuals*

Stationarity means that the value of whatever you are looking at depends only on the separation (i.e. for things that are separated by two feet over here, the variability is about the same as the value of two things separated by two feet over here. Without regard to the actual location.)

When you have a documented source, you can be comfortable taking out a parabolic or linear trend.

You can take out any trend, but will it make sense?