Data Preparation and Statistical Displays

- Data Cleaning / Quality Control
- Statistics as Parameters for Random Function Models
- Univariate Statistics
- Histograms and Probability Plots
- Q-Q and P-P Plots
- Bivariate Statistics and Distributions
- Exploratory Data Analysis

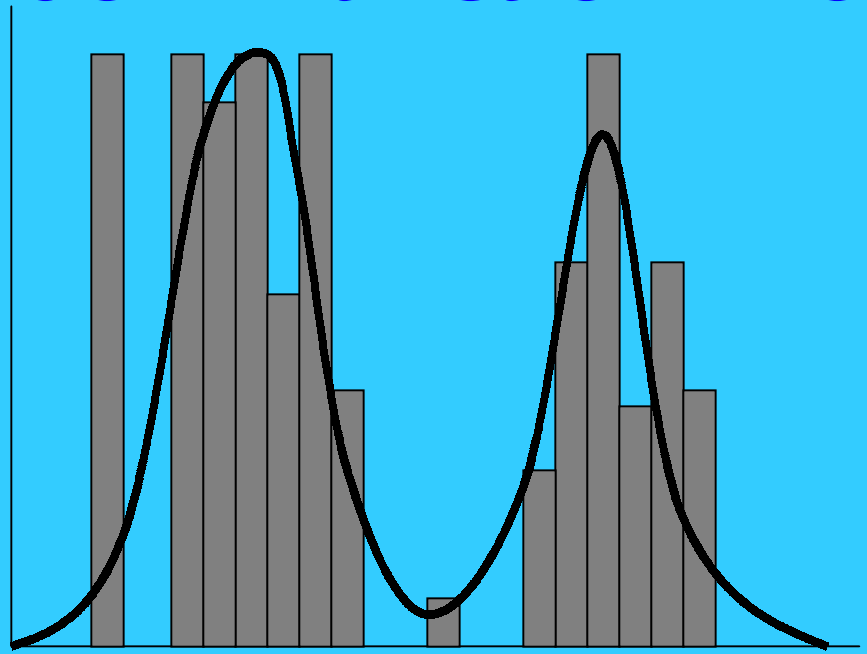
Statistical Analysis of Data

Well	Depth	X	Y	Poro	Perm	Layer	Seq	LogK
10100	7158.9	672.7	2886.0	27.00	554.000	1	51	2.744
10100	7160.4	672.8	2886.0	28.60	1560.000	1	51	3.193
10100	7160.7	672.8	2886.0	30.70	991.000	1	51	2.996
10100	7161.0	672.8	2886.0	28.40	1560.000	1	51	3.193
10100	7161.9	672.9	2886.0	31.00	3900.000	1	51	3.591
10100	7162.3	672.9	2886.0	32.40	4100.000	1	51	3.613

...

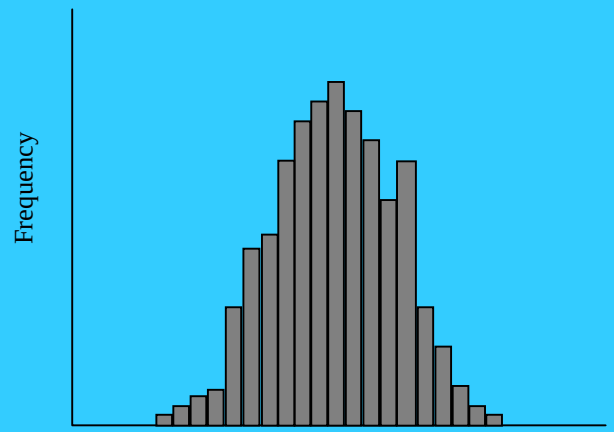
- Goals of Exploratory Data Analysis
  - understand the data: statistical versus geological populations
  - ensure data quality
  - condense information
- Only limited functionality in GSLIB (some special geostat tools) $\mapsto$ supplement with commercial software

Statistics as Parameters of Random Function Models



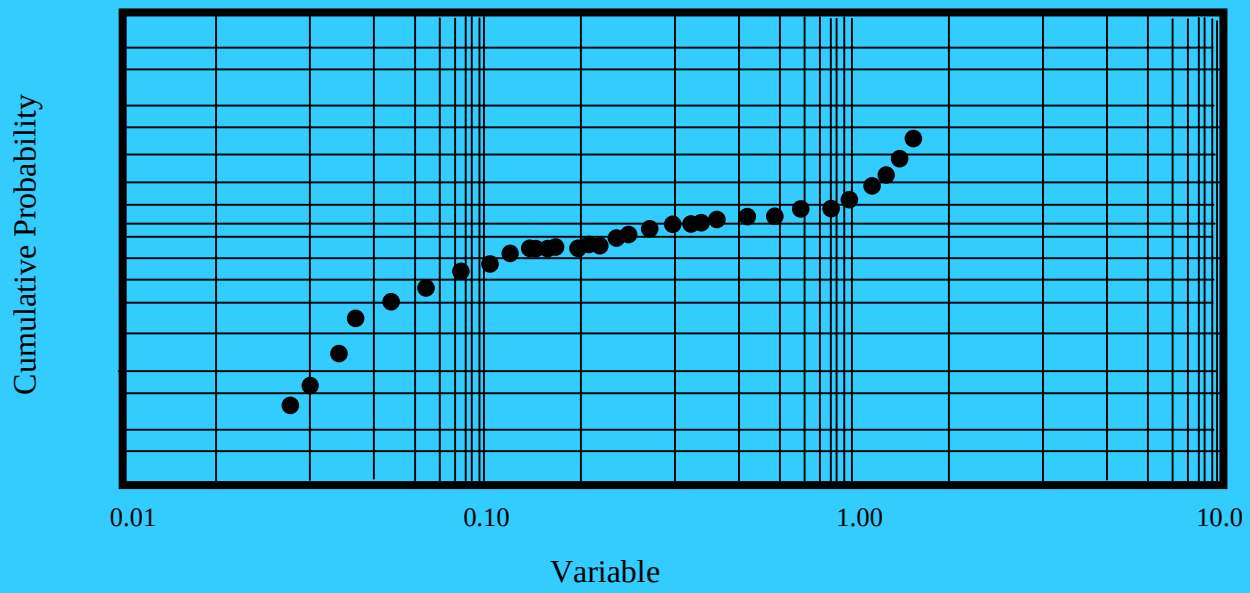
- Not interested in sample statistics ➢ need to access underlying population parameters
- A model is required to go beyond the known data
- Because earth science phenomena involve complex processes they appear as random. Important to keep in mind that actual data are not the result of a random process.

Univariate Description



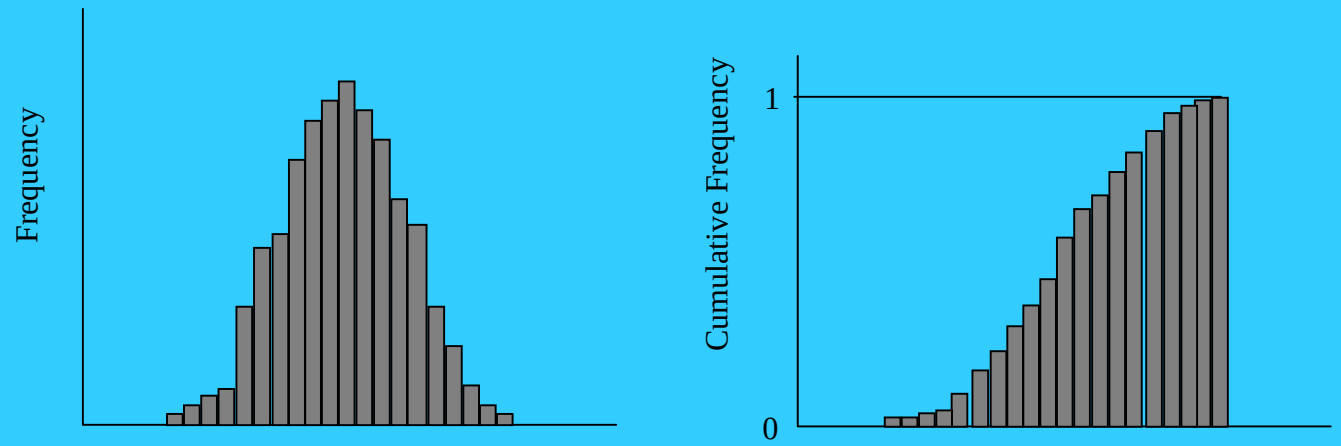
- Histogram: A counting of samples in classes
- Sometimes two scales are needed to show the details (use trimming limits)
- logarithmic scale can be useful
- Summary statistics
  - mean is sensitive to outliners
  - median is sensitive to gaps in the middle of a distribution
  - locate distribution by selected quantiles (e.g., quartiles)
  - spread measured by standard deviation (very sensitive to extreme values)
- GSLIB program histplt

Cumulative Probability Plot



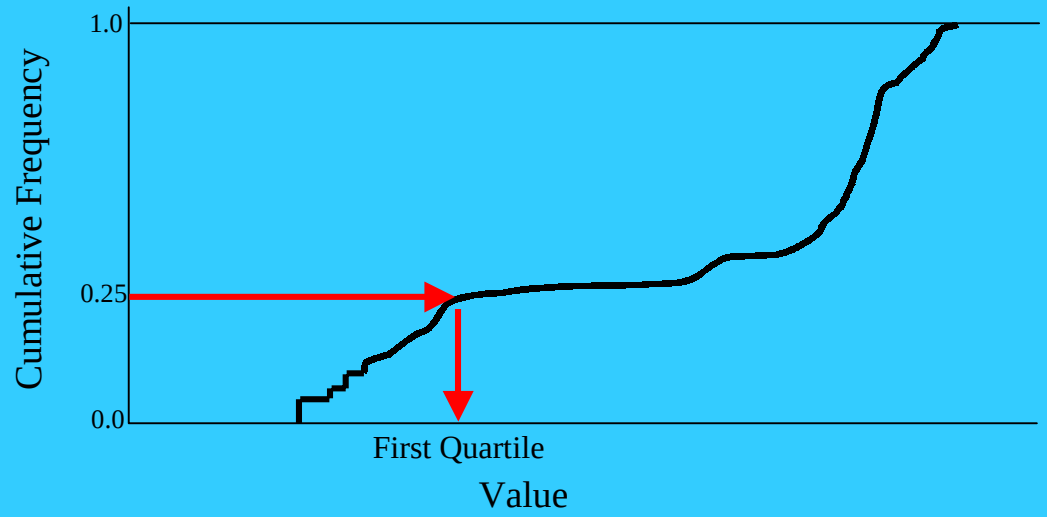
- Useful to see all of the data values on one plot
- Useful for isolating statistical populations
- May be used to check distribution models:
  - straight line on arithmetic scale $\mapsto$ normal distribution
  - straight line on logarithmic scale $\mapsto$ lognormal distribution
  - small departures can be important
  - possible to transform data to perfectly reproduce any univariate distribution
- GSLIB program probplt

Cumulative Histograms



The *cumulative frequency* is the total or the cumulative fraction of samples less than a given threshold

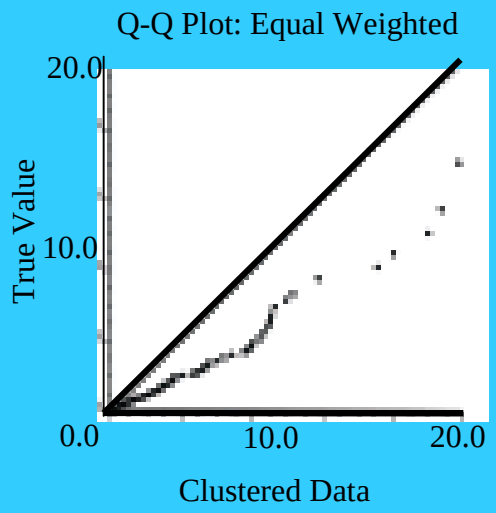
Cumulative Histograms

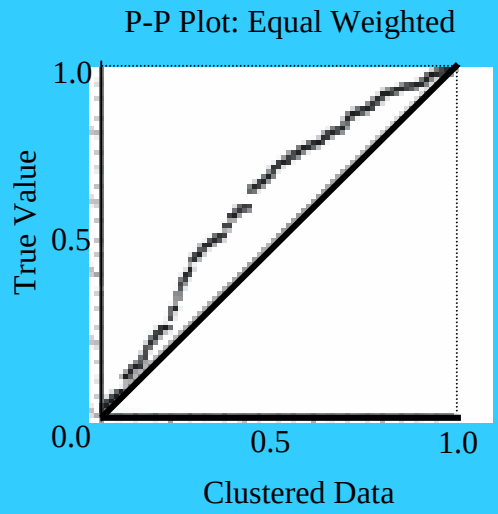


- Cumulative frequency charts do not depend on the binwidth; they can be created at the resolution of the data
- A valuable descriptive tool *and* used for inference
- A *quantile* is the variable-value that corresponds to a fixed cumulative frequency
  - first quartile = 0.25 quantile
  - second quartile = median = 0.5 quantile
  - third quartile = 0.75 quantile

  can read any quantile from the cumulative frequency plot
- Can also read probability intervals from the cumulative frequency plot (say, the 90% probability interval)
- Direct link to the frequency

Q-Q / P-P Plots





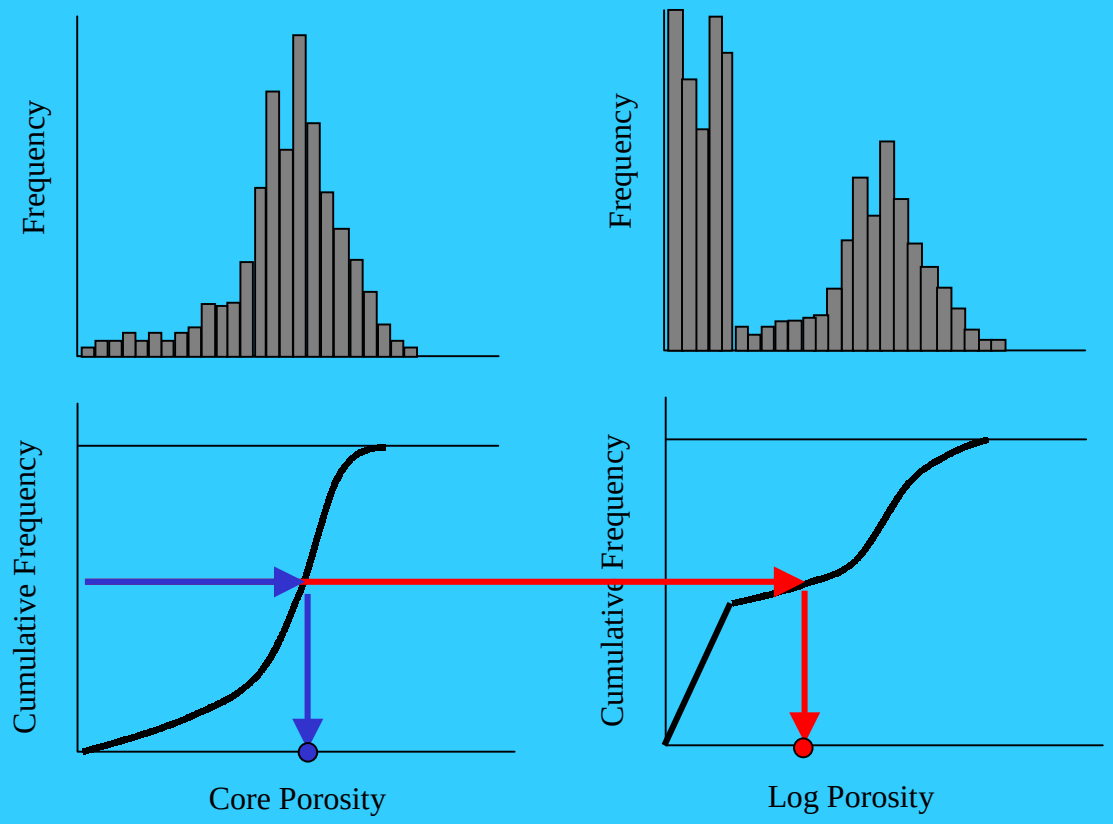
- Compares two univariate distributions
- Q-Q plot is a plot of matching quantiles $\mapsto$ a straight line implies that the two distributions have the same shape.
- P-P plot is a plot of matching cumulative probabilities $\mapsto$ a straight line implies that the two distributions have the same shape.
- Q-Q plot has units of the data, P-P plots are always scaled between 0 and 1
- GSLIB program `qpplt`

Q-Q Plots

Data Set One		Data Set Two	
value	cdf	value	cdf
0.010	0.0002	0.060	0.0036
0.020	0.0014	0.090	0.0250
0.020	0.0018	0.090	0.0321
0.020	0.0022		
0.030	0.0034		
0.030	0.0038		
0.960	0.4998	2.170	0.4964
0.960	0.5002	2.220	0.5036
38.610	0.9962		
40.570	0.9966		
42.960	0.9978		
43.500	0.9982	19.440	0.9679
46.530	0.9986	20.350	0.9750
102.700	0.9998	58.320	0.9964

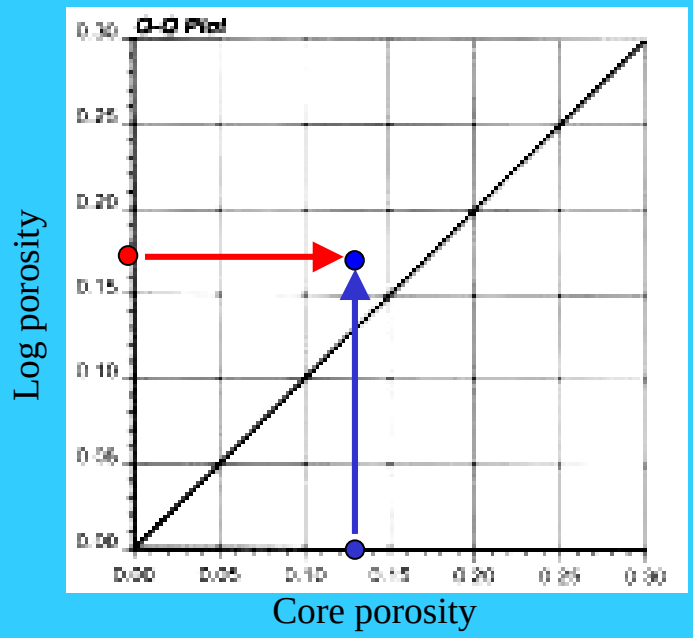
- Sort the values in each data set
- Calculate cumulative distribution function (CDF) for each
- Match according to (CDF) values

Let's Build a Q-Q Plot



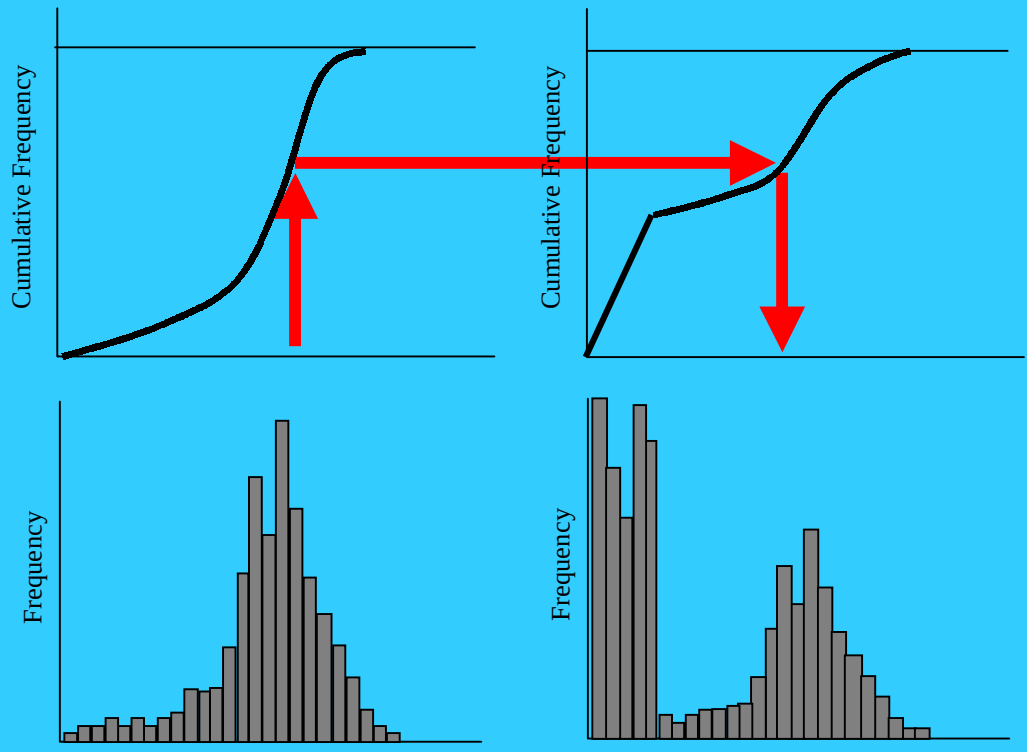
- Histograms of core porosity and log-derived porosity
- Preferential sampling explains difference; these are not “paired” samples so we can not detect bias in samples

Let's Build a Q-Q Plot



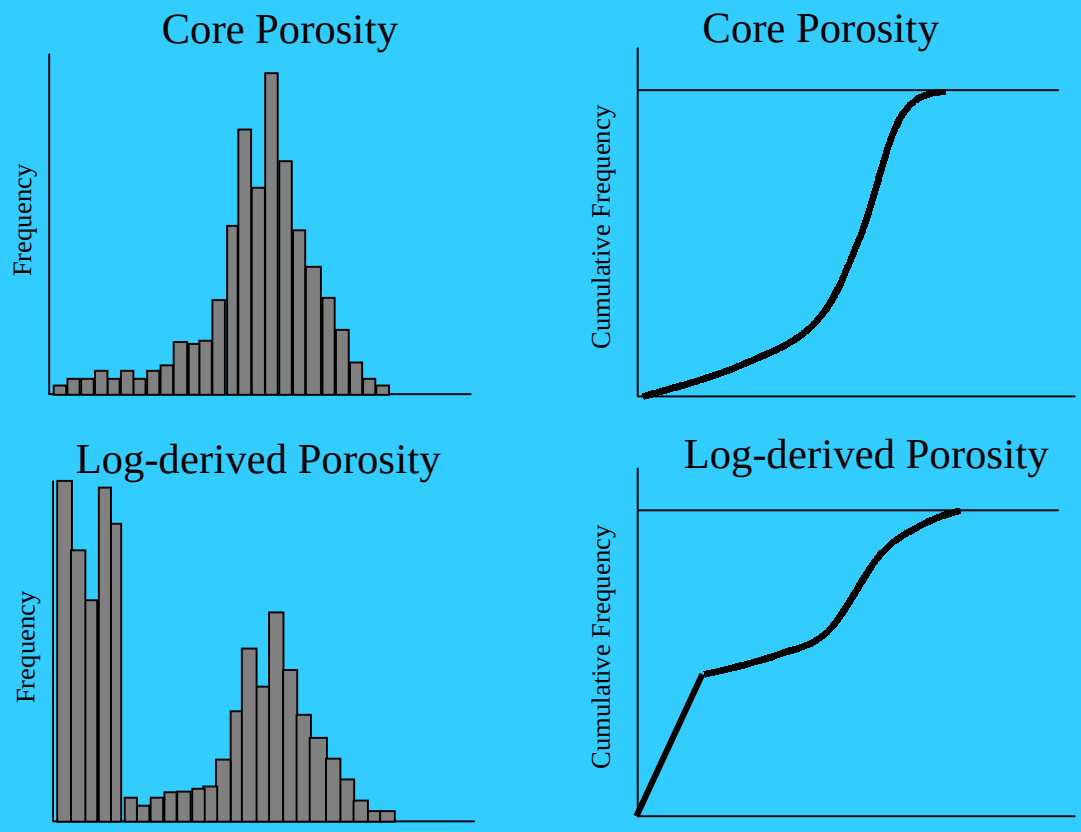
- Read corresponding quantiles from the cumulative frequency plots on the previous page
- Plot those quantiles on the plot

Univariate Transformation



- Transforming values so that they honor a different histogram may be done by matching quantiles
- Many geostatistical techniques require the data to be transformed to a Gaussian or normal distribution

Data Transformation



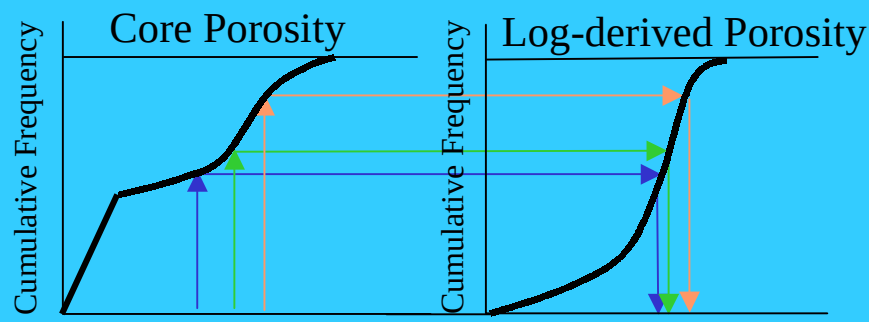
- Histograms of core porosity and log-derived porosity
- These are “paired” samples so we may want to transform the log-derived porosity values to core porosity

Data Transformation

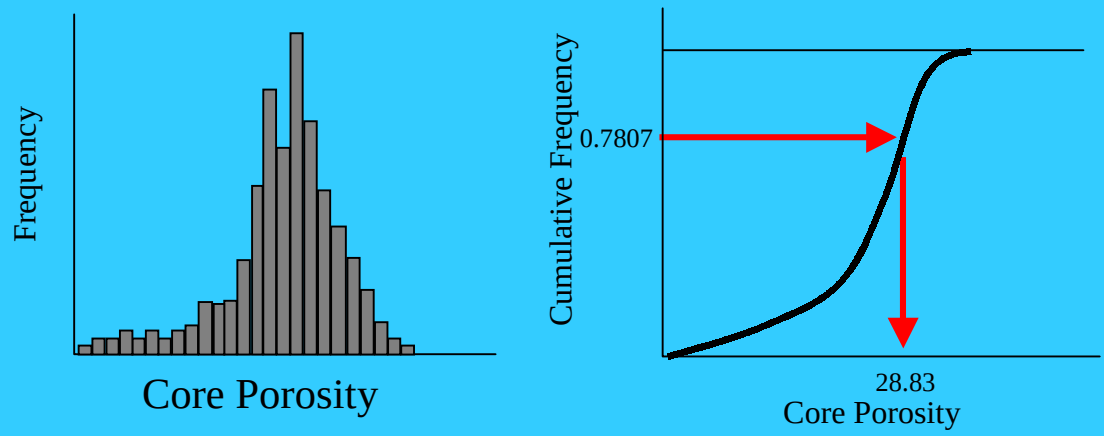
- Use the cumulative frequency plots on the previous page to fill in the following table

Log Porosity		Core Porosity	
10.0		7	
15.0			
20.0		18	
25.0		26	
30.0		29	

- Could we transform 20000 log-derived porosity values according to the 853 paired samples we have?
- Under what circumstances would we consider doing this?
- What problems might be encountered?



Monte Carlo Simulation



- Monte Carlo Simulation / Stochastic Simulation / Random Drawing proceed by reading quantiles from a cumulative distribution

The procedure:

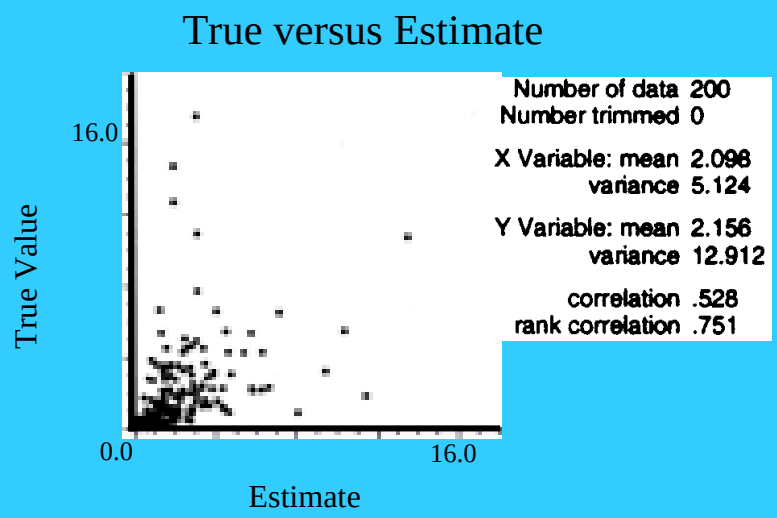
- generate a random number between 0 and 1 (calculator, table, program, **...**
- read the quantile associated to that random number

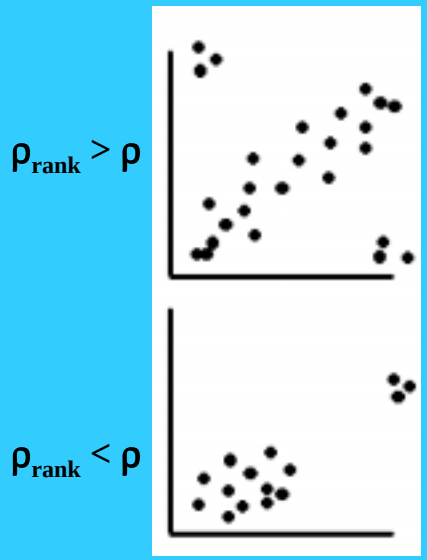
For Example:

Random Number	Simulated Number
0.7807	28.83
0.1562	
0.6587	
0.8934	

...

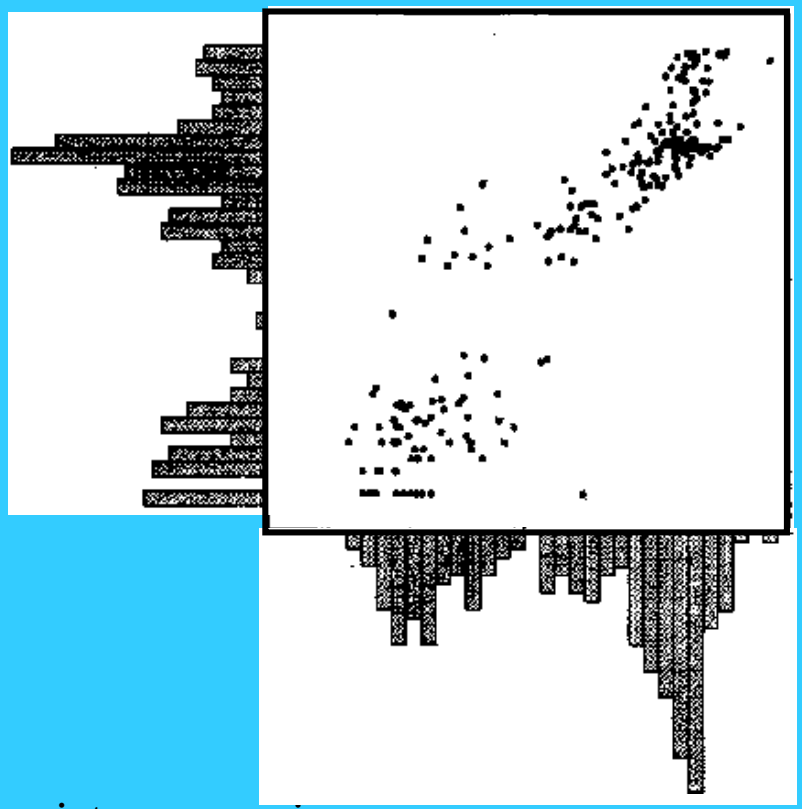
Scatterplots





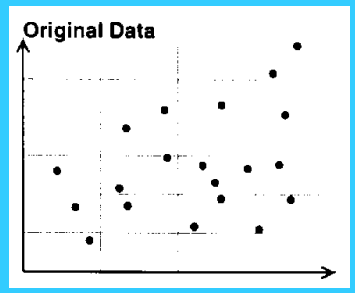
- Bivariate display, estimate-true, two covariates, or the same variable separated by some distance vector
- Linear correlation coefficient ranges between -1 and +1 and is sensitive to extreme values (points away from the main cloud)
- Rank correlation coefficient is a useful supplement:
  - if $\rho_{rank} > \rho$ then a few outliers are spoiling an otherwise good correlation
  - if $\rho_{rank} < \rho$ then a few outliers are enhancing an otherwise poor correlation
  - if $\rho_{rank} = 1$ then a non-linear transform of one covariate can make $\rho = 1$
- GSLIB program `scatplt`

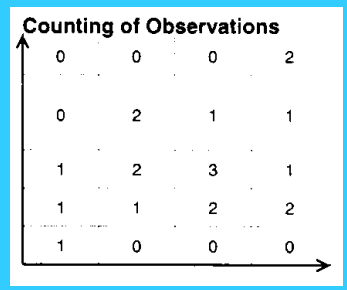
Scatterplots



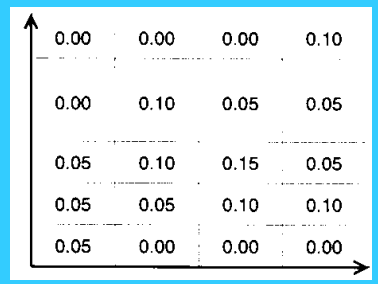
- Look at bivariate summaries
- Marginal histograms

Bivariate Distributions



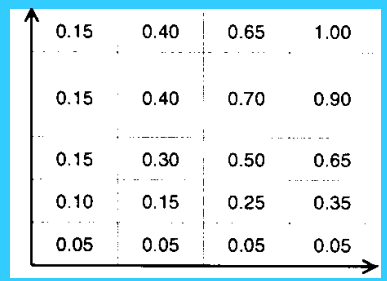


a)



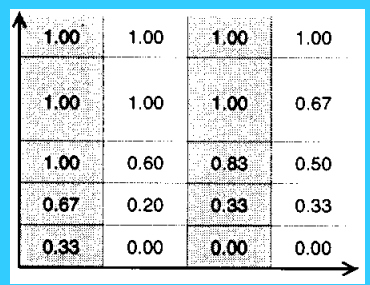
Bivariate histogram

b)



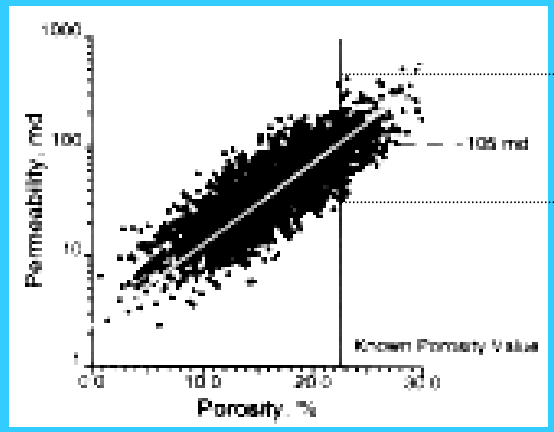
Bivariate cumulative distribution function

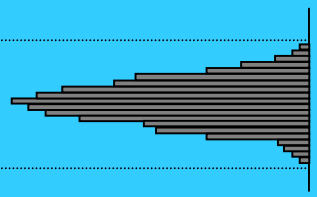
c)



Conditional cumulative distribution function

Conditional Distributions

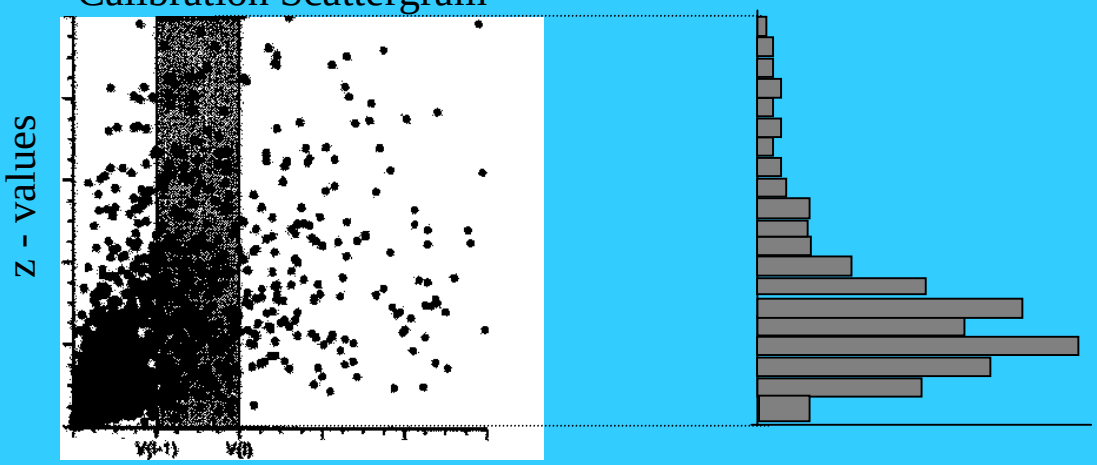




Distribution of possible permeability values at a known porosity value

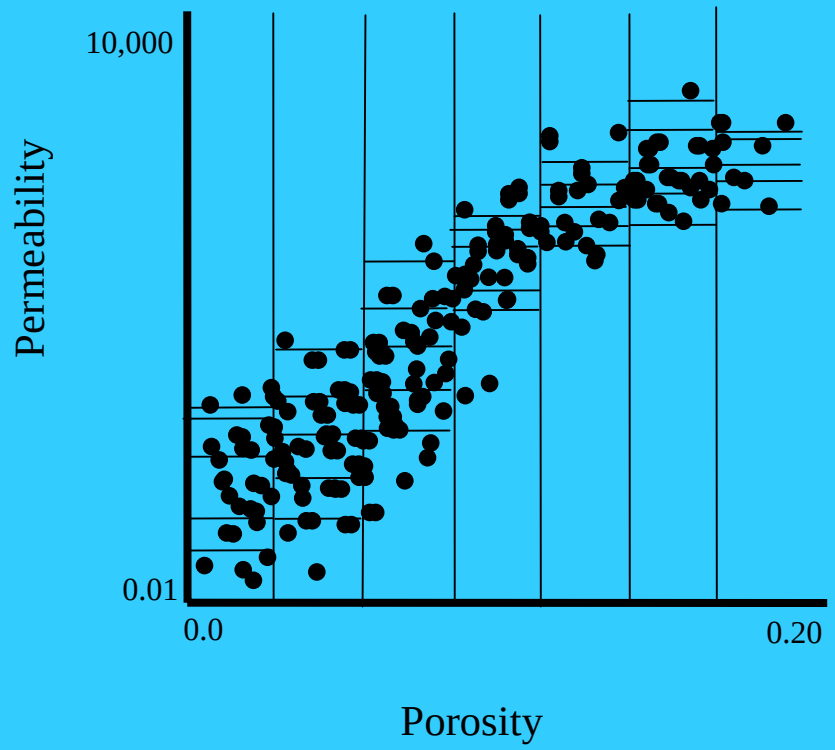
Known primary value

Calibration Scattergram



- Prediction of conditional distributions is at the heart of geostatistical algorithms

Class Definition



- Choose equal probability intervals rather than equal porosity or permeability intervals
- Check for bias in histogram - often there are many more low porosity log data than low porosity core data
- Due to a biased core data set, we may need to set porosity cutoffs based on an even ϕ spacing rather than equal probability

Exploratory Data Analysis

- Plot the data in different ways; our eyes are good at pattern detection
- Choose geological/statistical populations for detailed analyses:
  - populations must be identifiable in wells without core
  - must be able to map these populations (categories)
  - can not deal with too many, otherwise there are too few data for reliable statistics
  - often a decision must be made to pool certain types of data
  - *stationarity* is a property of statistical models and not reality
  - important and very field/data/goals-specific
- Perform statistical analyses within each population:
  - ensure data quality
  - look for trends
  - understand “physics” as much as possible
- Decluster data for geostatistical modeling
- Statistical tools are used throughout a reservoir character