

G606 Geostatistics and Spatial Modeling

Syllabus and Course Lecture Notes

Spring Semester, 2001

© 2001 John Welhan

Contents

Syllabus.....	1
General Information.....	1
Outline of Course Contents.....	2
Detailed Outline of Course Contents.....	3
Basic Statistics Review.....	5
Self-Evaluation Quiz.....	6
Univariate vs. Multivariate Data and Geostatistics.....	7
Week 1 Lecture Notes: Introduction, Definitions, Software, Basic Statistics Review.....	9
Software use during the course.....	18
Generalized geostatistical analysis and modeling sequence.....	19
StatMost software reference.....	21
Week 2 Lecture Notes: Parametric Tests, Nonparametric Tests.....	22
Problem Set I. Statistical summarization.....	26
Summary of hypothesis testing.....	27
Readings.....	28
Week 3 Lecture Material: Bivariate Data, Regionalized Variables, EDA.....	29
Problem Set II. Exploratory Data Analysis 1.....	32
Hypothesis testing of regression parameters.....	33
Week 4-5 Lecture Material: Autocorrelation, Variogram Measures of Spatial Continuity..	34
Problem Set III. Exploratory Data Analysis 2.....	40
Week 6a. Lecture Material: Experimental Variography.....	41
Problem Set IV / V. Spatial Correlation 1 - Experimental Variography.....	41
Words of Wisdom on Experimental Variogram Analysis.....	42
Variogram Analysis Software.....	44
Week 6.b Lecture Material: Modeling of Correlation Structure.....	45
Problem Set VI. Spatial Correlation 2 - Variogram Modeling.....	47
Words of Wisdom on Variogram Modeling.....	48
Week 7-8 Lecture Material: Introduction to Kriging System of Equations.....	49
Problem Set VII. and VIII. Kriging with GeoEas.....	55

Week 9-10 Lecture Material: Practical Implementation of Kriging.....	57
Problem Set IX. Kriging Cross-Validation, Introduction to GSLIB's Kriging.....	60
Week 11-12 Lecture Material: Introduction to Stochastic Simulation.....	61
Problem Set X. Introduction to SGSIM - Gaussian Simulation.....	63
Week 13 Lecture Material: Indicator Variables, Probability Kriging, Indicator Simulation.	64
Problem Set XI. Introduction to Sequential Indicator Simulation.....	67
Other Simulation Techniques: Object-Based and Simulated Annealing.....	70
Week 14 Lecture Material: Scaling, Sample Support.....	71
Week 15 Lecture Material: Evaluation of Uncertainty, Probability of Failure/Success.....	72

Syllabus - General Information

This course is an introduction to the analysis, description and modeling of geospatial data. The practical applications of software tools, underlying theory, and the correct application of these tools is emphasized through the use of various software (including GIS software) to analyze and model data. Although geologic applications and examples are emphasized, the concepts, theory, and software are intended to be equally applicable across disciplines, and students are expected to bring a data set (preferably thesis-related) to class, to be analyzed as a term project.

Prerequisites for the course are an introductory statistics course or permission of the instructor, and familiarity with computer manipulation of data. Knowledge of ArcView or other GIS software is encouraged but not necessary. All course software will be on a PC platform, so that students must be familiar with PC computing. Both Windows and DOS (command line driven) software will be utilized.

Selected material will be on library reserve, and some available for checkout from the instructor; assigned readings will be used to stimulate in-class discussion, and all students will research, present, and lead a class discussion on a published article of their choosing, focusing on concepts and applications of geostatistical analysis, kriging, or stochastic simulation.

Two 80-minute lectures plus one-hour computer lab per week. Office hours: one hour after each class and by appointment. Textbook: Isaaks and Srivastava (1989).

Grading will be based on:

40% - weekly computer lab problem sets

30% - term project, oral presentation, and final report

15% - analysis, oral presentation, and classroom discussion of published literature

15% - class participation in discussion of assigned readings

<u>Week</u>	<u>Material</u>
1 - 3	Univariate statistics review, regionalized variables, exploratory data analysis
4 - 6	Quantification of spatial continuity: variogram analysis
7 - 9	Spatial estimation: kriging, indicator kriging
10 - 13	Spatial simulation (indicator-based, conditional) and applications
14 - 16	Probability mapping; scaling issues; term project presentations

Assigned Readings from: (* on Obeler Library reserve; + check out from instructor; # web access):

General statistics references:

* Till, Roger (1974) *Statistical Methods for the Earth Scientist*; Wiley, NY

* Cheeney, R.F. (1983) *Statistical Methods in Geology*; George Allen & Unwin, Boston

* Koch, G.S. and Link, R.F. (1971) *Statistical Analysis of Geological Data*, Vol. 1, 2; Wiley, NY

Geostatistics (G), kriging (K), stochastic simulation (S), and software (W) references:

* Isaaks and Srivastava (1989), *Introduction to Applied Geostatistics*, Oxford Univ. Press, (G, K; superb)

+ Deutsch and Journel (1997) *Geostatistical Software Library and User's Guide*, Oxford (W)

+ Goovaerts, P. (1997) *Geostatistics for Natural Resources Evaluation*, Oxford (G, K, S; theory focus)

Houlding, S.W. (1999) *Practical Geostatistics*, Springer (G, K, S)

* Clark, I. (1979) *Practical Geostatistics*, Applied Science Publishers (G, K; mining focus)

* Yarus, J.M. and Chambers, R.L. (1994) *Stochastic Modeling and Geostatistics*, AAPG (G,K,S,W)

Pannatier, Y. (1996) *VarioWin: Software for Spatial Data Analysis in 2D*, Springer (W)

Armstrong, M. (1998) *Basic Linear Geostatistics*, Springer (G, K)

Kitanidis, P.K. (1997) *Intro to Geostatistics: Applications in Hydrogeology*, Cambridge U. Press (G, K)

Syllabus - Outline of Course Contents

Overview, Course Topics and Case Study:

Overview of applications and techniques to be covered in the course: univariate and multivariate statistics; spatial continuity analysis; estimation; simulation. Overview of spatial statistics, estimation, and modeling via an example.

Exploratory Data Analysis:

Statistical summarization, analysis; mapping of the data set, histogram and probability distribution, correlation in multivariate data, data transformations (logarithmic, indicator, normal-score, rank-order); software use and applications

Quantification of Spatial Continuity:

Calculation of experimental variograms, fitting models to experimental variogram, concepts of anisotropy and nested structures in variograms, other techniques for defining spatial variability (indicator, covariance), spatial co-variability of multiple variables; application of basic variogram analysis and modeling software.

Spatial Estimation (Kriging):

Review of techniques available for spatial estimation, explanation of the concepts of a 'best' linear unbiased estimate, introduction to the kriging system of equations, use and misuse of kriging variance, application of basic kriging software.

Stochastic Simulation:

Simulation vs. kriging, adaptation of the kriging system of equations to simulation, theory and application of basic gaussian and indicator simulation algorithms.

Scaling and Sample Support:

Impacts of discrepancy between measurement and estimation scales; examples of the effects of scale, accounting for scale discrepancies with analytical techniques, numerical techniques for addressing scale issues (block kriging, averaging techniques).

Application of Analysis of Uncertainty:

Concepts of probability of exceeding a threshold value and probability mapping, incorporation of spatial uncertainty into predicted outcomes of physical processes and human activities; creating probability maps through estimation versus simulation

Detailed Outline of Course Contents:

Overview, Introduction, Course Topics: Presentation and review of applications of the techniques to be covered in the course (univariate and bivariate statistics; spatial continuity analysis; estimation; simulation). Overview of the course via an example.

Key concepts: Basic Univariate & Bivariate Statistics Review in Till (1974) - measurements and scales of (p. 1-6); cf sample size vs population variance; processes, probability (p.7-11); randomness vs Markov processes; transition probability matrices (p.11-15) cf. spatial correlation; describing distributions, population vs sample distribution, pdf vs freq. density function (p.17-25), cdf vs cum. density function; normal and lognormal distribution; standard normal deviate z ; std.normal tables (p.30-45); t-distribution, F-distribution, χ^2 distribution (p.56-75); hypothesis tests (p.56-66); propagation of errors (p.78) param vs non-parametric statistics (p.117 ff.); correlation, closure problem (Davis, p.79-83)

Exploratory Data Analysis:

The need for basic statistical summarization, analysis; mapping of the data set, histogram and probability distribution, correlation between multivariate data, data transformations (logarithmic, indicator, normal-score, rank-order); software applications

Quantification of Spatial Continuity (Variogram Analysis):

Calculation of experimental variograms, fitting models to experimental variogram, concepts of anisotropy and nested structures in variograms, other techniques for defining spatial variability (indicator, covariance), spatial co-variability of multiple variables; application of basic variogram analysis and modeling software.

See Pannatier, Ch. 2 ff.

Spatial Estimation (Kriging):

Review of techniques available for spatial estimation, explanation of the concepts of a "best" linear unbiased estimate, introduction to kriging system of equations, explanation of kriging variance, application of basic kriging software.

See Isaaks and Srivastava: estimation (Ch. 8); random functions (p.218-231) and variogram modeling (Ch. 16); point vs block kriging (pp.278 & 323); intuitive look at ordinary kriging (p.299-313) and cross-validation (Ch. 15); indicator variables, limitations of estimation by averaging (p.417-428)

Key Concepts: Introduction to Estimation, Kriging in Isaaks & Srivastava (1989) tools, methods of spatial correlation analysis (ch. 1); bivariate statistical description (ch. 3); spatial data description by map summaries (p.40-50 & ch. 6 application); descriptors of spatial correlation (p.51-65 & ch. 7 application)

Stochastic Simulation: the similarities and differences between simulation and kriging, application of basic gaussian and indicator simulation software, concept of a transfer function and its use in conjunction with simulation

See chapter 5 introduction in Deutsch and Journel

Scaling and Sample Support:

Impacts of discrepancy between measurement and estimation scales; examples of the effects of scale, accounting for scale discrepancies with analytical techniques, numerical techniques for addressing scale issues (block kriging, averaging techniques).

See Isaaks and Srivastava: sample support p.458-468

Analysis of Uncertainty:

Concepts of probability of exceeding a threshold value and probability mapping, incorporation of spatial uncertainty into predicted outcomes of physical processes and human activities; creating probability maps through estimation versus simulation

Basic Statistics Review

Know and understand all underlined terms and how to calculate or otherwise represent them !

Classical Univariate Statistics: the analysis, modeling, and prediction of a variable on the assumption that all samples are independent of each other (i.e. no regard for a sample's position in relation to other samples - either spatially or temporally), and making predictions on the likelihood of a future outcome based on that variable.

e.g. fitting a normal probability distribution model to a histogram, then determining the probability that a specified threshold will be exceeded in future sampling (examples: ore grade, elk population, areal density of lodge pole pines)

Basic statistical terminology:

(global) population, sample (population) - frequency distribution (= histogram), "pdf", cdf, 'cdf'
expected value (central tendency) - mode / median / mean(s)

dispersion - range / variance / standard deviation / rms deviation / quartiles / skewness / kurtosis

parametric statistics vs. non-parametric statistics - normal distribution (Gaussian); other types

stationarity (single population) - unimodal / bimodal / multimodal

Gaussian distribution statistics:

expected value - arithmetic mean

dispersion - variance/standard deviation

shape - skewness, kurtosis = 0.0

Other general knowledge:

transformations - standard normal deviate / use of Gaussian pdf tables

statistical hypotheses, tests - t-tests of normality, similarity, and mean / K-S test / χ^2 -test

bivariate statistics - regression, variances of variables, covariance, statistical tests of regression

Self-Evaluation Quiz:

Note: if you cannot answer questions (a) through (h), this course may present difficulties!

- a) what is the name for the expected value of a Gaussian probability distribution?
- b) what is the measure of dispersion used to characterize the shape of a bell-shaped probability distribution?
- c) what values of skewness and kurtosis would a normal probability distribution have?
- d) what is a log-normal probability distribution? which of the following most closely approximates the mode of a log-normal distribution: arithmetic mean, geometric mean, harmonic mean?
- e) what percentage of samples fall between the 25th and 75th quartiles in a normal population?
- f) what test could determine if two sets of measurements were drawn from Gaussian populations with approximately the same variance but statistically different means?
- g) a histogram of a sample of water hardness found in SE Idaho wells shows two peaks at about 160 mg/l and about 900 mg/l; is this evidence that hardness can be modeled as a statistically stationary population in this domain?
- h) Given the following histograms of temperature measured in two small adjacent lakes, apply a suitable (quantitative) statistical test to evaluate the hypothesis that both sets of temperatures are drawn from the same population.

Lake A:		Lake B:	
T, °C	Freq.	T, °C	Freq.
12	1	12	1
13	2	13	5
14	2	14	5
15	4	15	4
16	5	16	2
17	4	17	2
18	1	18	1
19	1	19	0
n = 20		n = 20	
mode = ?		mode = ?	
mean = ?		mean = ?	
variance = ?		variance = ?	

- i) In fact, both lakes overlie the same aquifer and are fed by the same ground water source (12 °C) and are the surface expression of the water table. How, then, could the temperature data portray such a different picture?

Univariate vs. Multivariate Data and Geostatistics

Univariate observations - a single, dependent variable measured in an item drawn from a population (e.g. gold assays in rock samples) and analyzed without regard to position

Multivariate observations - a single variable measured with respect to the independent variables of position; or two or more variables measured in individual items without accompanying position information (e.g. gold assays in rock samples referenced to x, y, z position; water levels in a well referenced to temporal “position”; gold, silver, sulfur and copper assays in one or multiple samples, measured with or without regard to position)

Geostatistics - the analysis and modeling of 'spatial (or geospatial)' data (data distributed in space or time); aka 'Spatial Statistics', 'Spatial Data Analysis'; e.g. statistical interpolation via regression; spatial estimation and simulation of a regionalized variable

e.g. chemical rock composition data, physical attributes, and all GIS (spatial) data are multivariate. A univariate variable measured at the same x,y,z coordinates at different times also can be treated as a multivariate variable (e.g. the water level, Z_i , in the same well, i, measured at three different times, t, becomes three variables, $Z_i[t_1]$, $Z_i[t_2]$, $Z_i[t_3]$; time-series measurements of Z at a single location is bivariate, with “position” represented by time as independent variable)

Geospatial data therefore is a type of multivariate data wherein there may be only one variable of interest (the dependent variable) but whose values are related to position (independent variables of location and/or time), or in which multiple dependent variables are related to position

Note that multivariate data can be analyzed as univariate observations with univariate statistics by considering one variable at a time. But where data are multivariate, the interrelationships between the variables can be exploited to learn more of the relevant physical processes, and analyzed using a variety of multivariate statistical methods (e.g. multiple linear regression, generalized analysis of variance, discriminant analysis, factor analysis, canonical correlation, stereographic analysis, cross-variogram analysis, cokriging, etc.)

Geostatistics is a class of statistical methods that considers the interrelationship between a dependent univariate variable and the independent variables of position (x,y,z or time), OR the interrelationships among two or more dependent multivariate variables and position. It can therefore handle both types of data

Geostatistics is useful for identifying the spatial structure of univariate variables, OR the spatial dependencies among multivariate variables

Spatial Modeling predicts behavior on the basis of a variable’s variation at measured locations, OR on the basis of the variation of other variables measured at the same or other locations

When used in a multivariate sense, Spatial Statistics and Spatial Modeling rely on the spatial dependencies among variables, unlike other multivariate methods which rely directly on inter-variable dependencies without regard to spatial dependencies

This course does not deal with other multivariate methods such as multiple linear regression, discriminant analysis, or factor analysis. See Koch and Link (1971), Volume 2, Chapter 10, for an excellent summary of the concepts in (non-spatial) multivariate data analysis and good introductions to these multivariate methods of analysis

Week 1 Lecture Materials - Introduction, Definitions, Software, Basic Statistics Review

1. What is Geostatistics?

- Geostatistics is the name associated with a class of specialized statistical techniques used to analyze and estimate values of a variable which are distributed - and physically correlated - in space or time, as are most geoscience data and GIS data.
- The analysis of such a correlation is usually called a "structural analysis" or "variogram modeling." From a structural analysis, predictions of the value of the variable can be made at unsampled locations using "kriging" or "stochastic simulation." This approach is most useful where the processes responsible for generating the measured variable are unknown or too poorly constrained to permit construction of a quantitative process model to make spatial or temporal interpolations or predictions.
- The overall sequence of steps in a typical geostatistical study include:
 - (a) exploratory data analysis (understanding the spatial character of the variable)
 - (b) structural analysis (determining the spatial correlation or continuity of the data)
 - (c) making estimates (kriging or simulations to predict values at unsampled locations)

2. Some Basic Definitions:

- *probability*: the expectation of an outcome of a random event or measurement
- *stochastic*: synonymous with probabilistic
- *dependent variable*: a qualitative or quantitative measure of a physical attribute
- (*global*) *population*: all possible measurements of a variable or outcomes of a process
- *sample (population)*: a set of measurements or outcomes drawn from a (global) population
- *item (or event)*: a single outcome or measurement (univariate) or related outcomes (multivariate) in a sample; (see Cheeney, p.7): a specimen is an event; a sample is an experiment in which multiple specimens are collected; the population from which the sample is drawn is the ensemble of all possible specimens, about which we attempt to estimate the mean, variance, etc. from the sample's statistical characteristics
- *regionalized variable*: a dependent variable described by spatial or temporal position
- *geostatistical analysis*: evaluation and summarization of a regionalized variable's interrelation with itself or other dependent variables in space or time

3. Additional Definitions:

- *variance*: the variation of a single variable about its mean
- *covariance [as in bivariate regression]*: the joint variation of two correlated variables about their common mean
- (*auto*)*Covariance [as in geostatistics]*: the variation of a single regionalized variable
- *cross-Covariance [as in geostatistics]*: joint variation of two correlated regionalized variables
- *correlation structure*: a statistical description of the Covariance of a regionalized variable
- *kriging*: a statistical weighting procedure producing a best linear unbiased estimate of a regionalized variable at unsampled locations, and the variance or statistical uncertainty of the estimates
- *stochastic simulation*: a probabilistic model depicting the local variability of a regionalized variable superimposed on the regional spatial variability described by kriging

4. Kriging vs Simulation:

- kriging is linear interpolation; best, unbiased estimator of global variability
 - honors global statistics and local data
 - produces smoothed representation
 - quantifies statistical uncertainty at estimated locations
 - used for best estimation of expected values
- simulation is probabilistic representation of local variability
 - honors global and local statistics and local data
 - reproduces local variability
 - used to represent local variability/uncertainty; incorporating other information; quantifying effects of estimation uncertainty via a transfer function
- both methods can incorporate and honor different types of hard data, but only simulation can incorporate soft information and other constraints (eg: physical geometry)

5. Library Reference Sources

- Basic statistics for geologists: Till (1974) and Cheeney (1983) are highly recommended for their clear presentation of concepts and methods, using geologic data as examples.
- Geostatistics: Clark (1979) is an easy-to-read conceptual-level introduction (doesn't touch on simulation). Isaaks and Srivastava (1989) and Hohn (1988) are excellent quantitative books for the beginner or experienced user, covering only variograms and kriging. More advanced mathematical treatments are found in Cressie (1993) and in Journel and Huijbregts (1978).

Although not on library reserve, Deutsch and Journel (1992, 1998) and Goovaerts (1997) are written as companion texts and are highly recommended for the most up-to-date software applications and theory, respectively, for many facets of geostatistics analysis, estimation and simulation, at an intermediate level.

6. Software Overview

- **StatMost** (Windows): general-purpose, univariate / multivariate statistical analysis (non-spatial data); pros: user-friendly interface, flexible, automatic creation of test distributions, good on-line Help; cons: some limitations, errors in Help file figure/table numbering
- **3Plot32** (Windows): summarization, plotting and analysis program for exploratory data analysis in 2-D; pros: easy to learn, good on-line Help, moving window statistical analysis, classified post plots; cons: limited capabilities and uses, some bugs
- **Surfer** (Windows): data plotting, contouring and simple kriging program for 2-D spatial data; pros: easy to learn interface, good graphics and output capabilities, classified post plots, kriging estimation is user-specified (to a degree), capable of kriging with drift, good on-line Help documentation; cons: demo version can't save files, kriging option produces only best unbiased estimator (not kriging variance), user has only limited control over kriging process
- **VarioWin** (Windows): pros: excellent 1- and 2-D variogram calculation and modeling program for spatial data; pros: highly interactive data interface, easy to learn, good graphics, excellent interactive handling of data outliers, good documentation; cons: will not handle 3-D data, no on-line documentation in v.2.2 (although v.2.1 Help can be accessed separately)
- **GeoEas** (DOS): 1- and 2-D variogram calculation / modeling and kriging program for spatial data; pros: powerful yet flexible capabilities, easy to learn menu-driven program structure, univariate statistical analysis, flexible and interactive variogram calculation / modeling, excellent

for interactive 2-D kriging and cross-validation; cons: poor print/graphics output capabilities (DOS interface), will not handle 3-D data, limited handling of data outliers, paper documentation fair to good but hard to find information

- **GSLIB** (DOS): powerful and flexible suite of FORTRAN programs for spatial analysis, estimation and simulation of 1-, 2- or 3-D data; pros: complete functionality for all aspects of variogram analysis / modeling, kriging and simulation; source code allows user modification for custom applications; excellent documentation of all programs in a textbook form that is also a valuable teaching tool for intermediate-to-advanced students; latest program versions include postscript file output, for faster viewing of modeling results; cons: programs must be compiled (slight knowledge of FORTRAN is required to modify array sizes and for compiling/linking); no interactive capability unless programs are modified to make use of FORTRAN graphical routines or are called from custom applications written in Visual Basic or other GUI; graphic output is via 2-D postscript files viewable with third-party software

7. Data and Measurement Scales (Till, p.3-5):

- qualitative to highly quantitative orders of measurement
 - nominal scale = categorization into different classes (classes and scales are arbitrary)
 - eg: different color classes, mineral types
 - ordinal scale = ranking into a sequence of classes (class sizes are arbitrary or constant)
 - (eg: Moh's hardness scale; 1-9 or A-I could be used as long as rank order is preserved; categorization of permeability into high, medium and low groups)
 - interval scale = equal-length classes, arbitrary zero (eg: °C, °F temperature scales;)
 - ratio scale = equal-length classes, true zero (eg: °K temperature scale; permeability; Si content)
-
- nominal and ordinal scales are measured by “attributes” with discrete values; interval and ratio scales are measured by “variables” having continuous range of values
 - nominal scale is historical basis of natural sciences (classification); extremely useful; higher-order scales necessary only when required by the problem at hand
 - ratio scale is the most versatile, permitting the most rigorous statistical tests, but it is not always required; ie. make quantitative measurements only when necessary
 - data can always be collected at a higher-order scale and analyzed on lower-order scales
 - eg: graphical presentation of data: (Till, p.17-22; Cheeney, p.11-17) classification of ordinal and higher data into bins or classes
-
- support 'volume' (size of the representative 1-, 2-, or 3-D space over which a measurement or an estimate is made); e.g. pixel size for spectral images; ore assay on a chip sample vs. a 50-ton truckload sample; contour estimation built on 10-meter vs. 1000-meter grid estimates

8. Purpose of Measurements:

- to obtain descriptive statistics or make statistical inferences (eg: to summarize variability and the nature of spatial relationships, or to make predictions or estimates)
- sample = set of specimens/events/measurements; universal vs. sample population (Till, p.47)
- systematic vs. random samples: practicality (availability) vs. unbiasedness in geodata

- sampling steps (Till, p.51 - “sampling is like religion: all are for it, no one practices it”):
 1. develop a conceptual framework: purpose of the sampling campaign, expected population to be encountered, types of variables (continuous, categorical), expected sources of variability
 2. form a working statistical model based on the conceptual framework (eg: normal pop'n)
 3. choose a suitable sampling plan based on the model that will achieve the stated purpose
 4. decide on the number of samples to collect and the accepted levels of precision / accuracy (repeatability vs truth) of measurements to be made
- types of sampling - regular (gridded or geometric), biased (historical or available), random
- sampling goals differ depending on project goals: target scale of variability, representative of study area or sub-areas, defined sample spacing constraints, measurement vs. analysis support volumes
- Note: the goals of sampling are as varied as the earth is complex; different sampling campaigns within the same project can have very different objectives, and their impacts on spatial data analysis can be enormous! eg: obtaining regional estimates of trace element distribution or environmental contamination may invoke spatially random (unbiased) sampling; locating zones of high-grade ore may bias the collected data towards high values; sampling campaigns may evolve from random or gridded (unbiased) sampling to localized "hot-spot" (biased) sampling
- sampling campaigns that target the high or low end of the data distribution can produce spatially clustered data: the mean, variance and global frequency distribution of the data therefore are biased by the inclusion of a disproportionate number of samples of high or low values. Declustering is not necessary for kriging, which automatically accounts for data clustering, but is very important for simulation because the global frequency distribution is used to estimate local conditions wherever local neighborhood data are lacking. Cell-declustering superimposes regular grids of various cell sizes over the data domain, and assigns a declustering weight to clusters of samples that fall within each cell that are proportional to the inverse of the number of samples within a cell.

The global declustered mean for a given cell size is defined as $m_{declus} = \frac{1}{N} \sum \delta_i x_i$, where N is the number of samples, δ_i is the cell-declustering weight ($\sum \delta_i = N$; $\delta_i > 1$ for widely-spaced data, $\delta_i < 1$ for clustered data). The optimum declustering weights are chosen for the cell size which produces the minimum (or maximum) global declustered mean

- other types of measurements: (see Cheeney, p.9-10)
 - continuous vs discontinuous variables (cf. measurement scales)
 - directional (dip) vs axially-oriented (strike); see Cheeney, p.22-26 for qualitative points; and Till, p.38-43, Cheeney, p.98-106 for quantitative description, examples and references for specific techniques of analyzing and reporting directional data
 - three-dimensional orientation data (Cheeney, chapt. 9)
- Note: (Davis, p.81-83) a closure problem exists for variables whose values are forced to a constant sum (such as ratios, percentages; e.g. chemical analyses of rock); in such situations, some specialized statistical tools are required for analysis (see Koch and Link, ch. 11)

9. Probability: Rdg's Till ch.2-3 (pp.30-37; 56-63; 121-123; 126-131)

- a sample of a population is an outcome of a statistical experiment
- if two or more possible outcomes are possible for each experiment, there is an uncertainty associated with each outcome, i (eg: if the population consists of 50 red and black marbles in a bag and the sample size is 50, then only one possible outcome exists = no uncertainty)
- probability, p , is the proportion of outcome i to all possible outcomes: $p_i = n_i/N$; $N = \sum n_i$
- there is an unknowable "true" probability for the population (= pop'n statistic), which cannot be known with certainty, only estimated (= sample statistic)

- a 'random process' is occurring when the n th occurrence of outcome i is independent of the $(n-1)$ th occurrence (ie. each occurrence has no "memory" of past sampling)
- a Markov process is random but the probability of outcome i , p_i , depends on the outcome of a preceding outcome j , p_j (an outcome in this context can be thought of as a physical state of the system eg: fluvial vs. deltaic vs. lacustrine depositional environment)
 - a Markov chain is a series of possible states, with the probability of transition from state i to state j defined for all possible transitions
 - 1st-order Markov chain: $j = i+1$; n th-order: $j = i+n$
 - a stationary process exists where the transition probabilities are constant in time (space)
 - example: see cyclothem example in Till, p.11-14 for transition probabilities
- the concept of Markov processes is related to the concept of spatial correlation which is central to geostatistics; ie. values of a spatial variable are not randomly distributed but depend on their spatial context (= structural correlation)
- a key concept in geostatistics is that of the random function model, which describes all possible spatial arrangements of the values of a variable (eg: porosity) that satisfy the statistical properties of the variable; this is a purely theoretical concept in that only one of these possible arrangements is ever available to us for sampling (i.e. the earth). Although this is similar to the idea of a "population" in conventional statistics, it is given the name of a "realization" because the random function model allows for an infinite number of possible arrangements or realizations (populations) of porosity. For example, the bag of 50 marbles is the only one we have to sample, so it constitutes the population in classical statistics, but if it is viewed as just one realization of a random function, then there are many possible arrangements or realizations of the 50 red and black marbles in the bag; in spatial data analysis we are never actually concerned with these other possible arrangements, and we use the random function model only as a framework within which to create the tools of geostatistical analysis and modeling)

10. Describing a Sample Distribution: the histogram, pdf, "cdf", cdf

- frequency is the number of specimens/events (see Cheeney, p.13; Till, p.90-91)
- in a frequency histogram, it is the area (not height) of the bars that is proportional to frequency; i.e the bin sizes (class intervals) of the bars can vary; the visual appearance of a histogram can be greatly altered by the choice of class (bin) sizes! (see Cheeney, p.13)
- probability density function (pdf): (Cheeney, p.13-15) if large amounts of data are available, the number of histogram classes can be made arbitrarily large, so that the visual shape of the histogram smooths out and approaches a continuous curve in the limit: this curve is the probability density function (pdf) and is the basis for determining the probability that a variable lies within a given range, or the probability that a variable lies above or below a

specified threshold; most common pdf's are normal (gaussian) and lognormal, and have a closed-form analytical description (see below)

- the cumulative frequency distribution ("cfd") is the cumulative analog of the histogram; the cumulative distribution function (cdf) is the cumulative analog of the pdf: (Cheeney, p.15-17)
- the height of the cdf at a threshold, z , is equivalent to the area beneath its pdf to the left of z ; tabulated values, $\Phi(z)$, for the standard normal distribution are available to define the probability (area under the pdf) that a variable's value is less than z (see Section 1.2 below)
- p-quantile: the value, z , that a stated proportion, p , of samples does not exceed

- measures of central tendency: mode (highest frequency class); median; mean (arithmetic, geometric, harmonic): $m_{arithmetic} = \sum x_i$; $m_{geometric} = [\prod x_i]^{\frac{1}{n}}$; $m_{harmonic} = \frac{1}{\sum [1/x_i]}$
- measures of "dispersion" (shape) about central tendency: range (max, min, interquartile); variance, std.dev.; skewness (ca. $3[\text{mean-median}]/\text{std.dev.} = -1$ to $+1$; see Cheeney, p.21)
- "moments" - 1st (mean) $= \frac{1}{n} \sum x_i$, 2nd (variance) $= \frac{1}{n} \sum (x_i - \mu)^2$, 3rd (skewness) $= \frac{1}{n} \sum (\frac{x_i - \mu}{\sigma})^3$, 4th (kurtosis) $= f[\sum (x_i - \mu)^4]$

- Notes on the cdf: the pdf and cdf exist only for ratio-scale measurements, but a discontinuous "cfd" can be plotted like any histogram (ie. using constant or varying class intervals) for nominal and ordinal data, and used for analysis of categorical (non-parametric) data
- the concept of a cdf is often used interchangeably with that of a "cfd", even where the cdf is unknown, so be aware of its specific usage in a particular context
- the q-quantile (or quantile) on a cdf is the height, q , of the cdf at a given value of the variable; a Q-Q plot therefore compares the shapes of two cdf's (or "cfd"s)
- the p-quantile (or probability) on a cdf is the value of the variable that a proportion, p , of the data does not exceed; thus, a P-P plot compares cumulative probabilities of two cdf's ("cfd"s)

11. Degrees of Freedom: (Till, p.56,57)

- in general, D.F. = number of observations less the number of estimates made from them
- eg: to define the mean, only need n observations, so D.F. = n ; but variance requires an estimate of mean, so its D.F. = $n-1$
- eg: bag of 50 red and black marbles: the population $N = 50$; if sample size $n < 50$, then variance requires estimate of mean, so D.F. = $n-1$; but if $n = 50$ (ie. all of population is represented in the sample), then the mean is no longer estimated (it is known with certainty), so D.F. = $n = N$
- the practical difference between the unbiased estimator of variance ($n-1$) and biased estimator (n) vanishes for practical purposes as n increases

12. Normal Probability Distribution Function (pdf): $y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$

$$(1.1) \quad \text{or} \quad y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) \quad (\text{see Till, p.33})$$

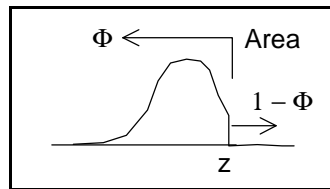
where $z = (x-\mu)/\sigma$ is the "std. normal deviate" (for sample statistics, substitute m and s for μ , σ)

- the standard normal pdf has $\mu = 0$, $\sigma = 1$, so that $y = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$ (see Till, p.37 for $\mu \pm 3\sigma$)
- the total area under *either* curve is 1.0 (ie. the probability that $-\infty \leq x, z \leq \infty$ is 1.0)
- the area under the curve $\in [a \leq x \leq b]$ is equal to the cumulative distribution function (c.d.f.) and is the probability that x lies between a and b :

$$(1.2) \quad \Phi(a, b) = p[a \leq x \leq b] = \int_a^b \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \right] dx$$

- Tabulated values of the standard normal curve list the inverse cdf, $1-\Phi(z)$, of the standardized random variable $z = (x-\mu)/\sigma$ [where $\Phi(z)$ is the standard-normal cdf]
- since $\Phi(z)$ is the probability (ie. area of the pdf from $-\infty$ to z) that the standardized random variable is less than z , then $1-\Phi(z)$ is the probability that it is greater than z :

$$(1.3) \quad 1 - \Phi(z) = p\left[z \leq \frac{(x-\mu)}{\sigma} \leq \infty\right] = \int_z^{\infty} \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \right] dx$$



- tabulated values of $1-\Phi(z)$ (Till, p. 34) range from 0.50 at $z = 0$ to very small values as z increases
- note that only positive z values are tabulated since the normal distribution is symmetrical about a mean of zero; for negative z , the tabulated values of $1-\Phi(|z|) \equiv \Phi(-z)$

Note: values correspond to σ ; e.g. at $z = 1.96$, $1-\Phi(z) = 0.025$ ie. the sum of the two tails is 0.05 so that the probability that values fall within -1.96σ and $+1.96\sigma$ is $1-2\Phi(z) = 95\%$

-Example: if porosity (in percent) is normally distributed with mean, $\mu = 20$ and $\sigma = 2$, the probability that porosity lies between 17.5 and 23 can be found as follows:

-let the variable x represent porosity, so its standardized form is z :

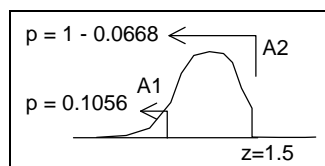
$$x_1 = 17.5, \text{ so } z_1 = (x_1 - \mu) / \sigma = (17.5 - 20) / 2 = -1.25$$

$$x_2 = 23, \text{ so } z_2 = (x_2 - \mu) / \sigma = (23 - 20) / 2 = +1.5$$

- the tabulated value for $1 - \Phi(1.5) = 0.0668$

therefore, area $A_2 = 1 - [1 - \Phi(1.5)] = \Phi(1.5)$

$$= 0.9332 = p\left[\frac{(x-\mu)}{\sigma} \leq z\right] = \text{probability that } x < 23$$



- for $z = -1.25$, look up the value for $1 - \Phi(1.25) = 0.1056$; and since the curve is symmetric about zero, $A_1 = 0.1056 = p\left[\frac{(x-\mu)}{\sigma} \leq z\right]$
- therefore the desired probability = $A_2 - A_1 = 0.8266$
ie. the probability that the porosity lies between 17.5 and 23% is 82.66%

- because of the definition of the standard normal deviate, the value of z is equivalent to the number of standard deviations, σ , away from the mean (see Till, Fig. 3.13)

- Note: for a lognormal distribution, values can be log-transformed and treated like a normal distribution; following the analysis, values of the (non-log-transformed) mean, std. dev'n, confidence limits, etc. can be obtained from the their log-transformed counterparts as:

$$(1.4) \quad x = \exp(\ln[x]) \quad \text{for the non-log value of } x$$

$$(1.5) \quad \mu_x = \exp\left(\mu_{\ln x} + \frac{1}{2}\sigma_{\ln x}^2\right) \quad \text{for estimate of the (non-log) mean of } x$$

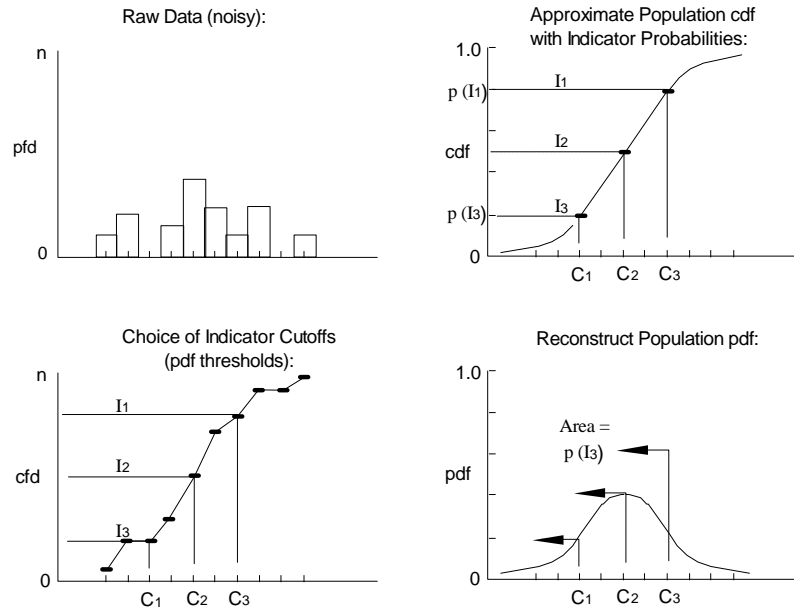
$$(1.6) \quad \sigma_x^2 = \mu^2[\exp(\sigma_{\ln x}^2) - 1] \quad \text{for estimate of the (non-log) std. dev'n of } x$$

- Normal-score transformation: any non-normal distribution (including lognormal ones) can be transformed into a standard normal distribution by a numerical or graphical method known as a normal-score transform, which can be treated as a perfect normal distribution, then back-transformed by an inverse procedure; see Isaaks and Srivastava, p.469-470, Hohn, p.171-172, p.175-185; note that this procedure is always safer than log-transformation, because back-transformation of a log-estimated values introduces a systematic bias in the estimates (Deutsch and Journel, p. 93)

- normalization of the data distribution is not necessary for kriging *per se* but is essential for stochastic simulation of continuous variables based on gaussian-type simulation algorithms. It also circumvents the dual problems of choosing an appropriate (and usually arbitrary) transformation algorithm for an irregular frequency distribution and of appropriately interpreting the back-transform of the linear estimate (as is the case for log-transforms). It makes correlation structure in variogram analysis easier to identify and conceptualize; and, finally, it minimizes chance of numerical instability in solving the kriging matrices

12. Estimating a Distribution with Indicator (non-parametric) Statistics:

- where the form of a cdf is unknown or poorly defined by the available data, one can estimate it by indicator transformations which convert the ratio- or interval-scale data into ordinal-scale classes or single-valued thresholds (known as "indicator cutoffs"); from the relative proportions of data falling below several of these thresholds, the form of the cdf and pdf can be estimated as in the following example:



- this approach is the basis for indicator kriging and indicator simulation. Kriging and Gaussian simulation estimate a continuous variable from a knowledge (or the assumption) of a Gaussian form of its global cdf. Where this is inapplicable, indicator geostatistics are used to 1) represent the cdf of a categorical variable (nominal or ordinal data), 2) represent the non-parametric cdf of a continuous variable, or 3) provide estimates of the distribution of values above or below specified threshold(s)

- see the spreadsheet "Indicator CDF.wk4" for an example of indicator transformation of a continuous distribution

- another type of transformation involves reducing the information content of interval or ratio data to create nominal or ordinal data which can be counted: e.g. classify contaminant concentration into high and low classes; categorize earthquake intensity as microearthquakes ($M < 3$) and "other"; classify permeability measurements into high, medium and low classes; etc. The effect is similar to that of an indicator transformation in that the statistical treatment becomes non-parametric (e.g. see Week 2, Section 5).

13. Introduction to Problem Set Data: refer to data files and details given in class

Software Used During the Course

One of the goals of the course is to introduce and reinforce key concepts use the simplest software for a particular task, and yet to expose students to all available geostatistical software.

Software to be Used in the Course: (GUI software in lower-case; DOS codes in upper case)

Statmost - descriptive statistics, distribution tests, t-tests, normal-score distributions

Spreadsheet - common data import platform; general data manipulation and conversion

3Plot32, Surfer - data post plots, moving window plots, descriptive stat's, intro to .DAT files

HISTPLT, PROBPLT, NSCORE, DECLUS - data analysis, preparation

VarioWin - variogram analysis, modeling

GeoEas, KT3D - kriging, cross-validation

SGSIM, SISIMPDF - simulation

Generalized Geostatistical Analysis and Modeling Sequence

(Note: Windows and DOS programs in lower case, GSLIB FORTRAN programs in upper case; graphical output from all GSLIB programs is viewed with GSViewer Windows software)

a. Exploratory Data Analysis

1. *Create data post plot* 3Plot32, Vario2dp
 - location errors, visual examination (clustering, hi/lo values, etc.)
2. *Evaluate data distribution* Spreadsheet, StatMost, HISTPLT, PROBPLT, SCATPLT
 - global statistical character (normal/skewed, outliers, univariate / bivariate summary)
3. *Compute declustering weights* DECLUS
 - for unbiased statistics, simulations
4. *Transform raw data to normal distribution* Spreadsheet, NSCORE
 - for improving variogram analysis, kriging results, necessary for Gaussian simulation
5. *Trend analysis* Surfer
 - identify possible trends in raw data
6. *Data manipulation (optional)* Spreadsheet, ROTCOORD
 - remove trends in raw data; transform data to one or more indicator variables; rotate or transform raw data coordinates to remove non-orthogonal spatial arrangements for kriging and simulation; repeat 1.1 - 1.5

b. Variogram Analysis and Modeling

1. *Construct experimental variograms* VarioWin, GeoEas
 - identify overall correlation structure, optimal lag classes, anisotropy, data outliers
 - steps in variogram analysis (VarioWin)
 - construct best isotropic variogram (if any), identify best bin parameters
 - construct anisotropic variograms, identify orientations of max/min range (if any)
 - look for internal consistency between alternative measures of correlation
 - if normal-score data or indicators used, look for internal consistency (if any)
2. *Model the variograms' correlation structures* VarioWin, GeoEas
 - identify and fit best random function model(s) for kriging (raw data, indicators, sill) or for simulation (variance / sill constrained to unity for normal-score data)
 - steps in variogram modeling (VarioWin)
 - choose both max and min variograms (if appropriate)
 - assign approximate nugget level to variogram in max correlation direction
 - assign preliminary random function model type
 - assign approximate sill level
 - assign approximate range
 - adjust range, sill, nugget interactively until model fit is appropriate
 - add additional structure(s), if necessary, to improve model fit
 - for n-score data, adjust anisotropy for best model in min correlation direction
 - if semivariogram is noisy, apply above procedure to determine best range(s) from inverted covariance or madogram, sill and nugget from standardized semi-variogram or correlogram (not necessary if n-score data used)

c. Estimation - Kriging

1. *Perform kriging and/or indicator kriging* GeoEas, KT3D, IK3D
 - provide local estimates and/or smooth picture of global trends
 - steps in kriging:
 - specify goal of kriging application
 - decide on type of kriging required for the project goals
 - define the kriging grid array size, based on goals
 - set up the kriging parameters (search radius, number of neighboring data, etc.)
 - perform kriging
 - perform cross-validation
2. *Evaluate the modeling / estimation process* GeoEas, KT3D, IK3D
 - perform kriging cross-validation, for comparison of estimated vs. actual data
3. *If necessary, post-process indicator simulations* POSTIK
 - calculate expected values, variances, exceedance probabilities
4. *Analyze the estimation process* StatMost, 3Plot32, GeoEas, SCATPLT, HISTPLT
 - look for systematic bias and spatial trend between estimated and true data values
5. *If applicable, back-transform transformed variable(s)* BACKTR
 - reproduce original (detrended) variable's range and values

d. Estimation - Simulation

1. *Perform sequential simulation of continuous or indicator variable* SGSIM, SISIM
 - provide estimate of local and global uncertainty and character of spatial variability
 - steps in simulation:
 - define goal of simulation
 - determine type of simulation required to meet goals
 - define the simulation grid array size, based on goals
 - set up the simulation parameters (conditioning data, number of simulations, etc.)
 - perform simulations
2. *Post-process multiple simulations* POSTSIM
 - calculate expected values, variances, exceedance probabilities from n simulations
3. *If necessary, back-transform transformed variable(s)* BACKTR
 - reproduce original (detrended) variable's range and values

e. Post-Estimation Validation and Analysis

1. *If applicable, restore trend surface* Spreadsheet, Surfer, ArcView, etc.
 - reproduce original variable's range and values
2. *Check reproduction of global distribution, variograms, bivariate correlations, etc.*
 - employ any methods listed under 1. and 2. above
3. *Evaluate results, suitability of estimation process* Spreadsheet, Surfer, ArcView
 - overlays and comparisons of original data with estimates; comparison, analysis and evaluation of various estimation procedures and results with original data and/or objectives

StatMost Software Reference

- index to StatMost key features for basic statistical analysis:

(create a new spreadsheet for data analysis)	File New, File Open Spreadsheet
(create a test distribution)	Data Create Distribution
(create a new column from other data)	Data Transform Simple Math
(univariate statistical summary)	Statistics Descriptive General
(create a cumulative freq. distribution)	Statistics Summaries Frequencies
(tests of normality, comparisons of distributions)	Statistics t-test, F-test, Normality Tests
	Statistics Nonparametric, Advanced
(chi-squared tests of nominal data)	Statistics Cross Tabulates
(correlation analysis of paired data)	Statistics Correlations
(estimating C.I. of mean, std.dev'n)	Statistics Confidence Intervals
(parametric tests of sample means)	Statistics Hypothesis Tests
(computation of C.I. of linear regression)	Analyses Regressions Linear
(autocorrelation calculations)	Analyses Time Series Auto Correlation
(histogram plot)	Plot 2D Special Histogram
(normal probability plot)	Plot 2D Statistical, Normal Probability
(plot computed frequencies as cfd)	Plot 2D Curve Step Curve
(graphical summary of highlighted column)	Quickplot Box Plot

Week 2 Lecture Material Parametric Tests, Nonparametric Tests

Parametric Tests are applied to statistical data derived from distributions of a known or assumed form (eg: a normal or Gaussian distribution). They are useful for comparing the means or variances of two populations, for determining whether two samples were drawn from the same or different populations, and for quantifying the confidence or probability that the mean of a sample falls within or outside of specified thresholds. Parametric tests are applicable only to interval or ratio-scale measurements.

1. Student's t-test: (Till, p.56-61)

- this is a statistical test of a statistical hypothesis. A null hypothesis, H_0 , is formulated, such as 'the mean of population 1 = mean of population 2' or the regression slope = zero. The alternative hypothesis, H_a , is the antithesis of H_0 .
- the t-test is a method to compare sample means or to determine whether a population was drawn from a normal population of the same mean, for both known and unknown variances - *but always assuming a normally-distributed population!*
- given a population with mean μ and variance σ^2 , draw a sample of size n , whose sample mean is m and sample variance is s^2
- draw all possible samples of size n and calculate $t = \frac{(\bar{m}-\mu)}{s/\sqrt{n}}$ for each sample, where s/\sqrt{n} can be thought of as the standard deviation standardized by the sample size
- plot the pdf of the t-statistic, which defines Student's - t distribution (Till, p.57)
- the level of significance, α , is defined as the probability of obtaining a value further from the mean than the specified value $|t|$
- a one-tail test is used if the test is formulated as to whether a statistic is greater than or less than a given threshold (ie. when the probability refers to only one side (tail) of the pdf); in that case look up the tabulated t-value for α level of significance
- a two-tailed test is used for a confidence interval (C.I.) or a test of difference or, in general, if the probability being tested refers to the entire pdf regardless of whether the difference is more or less than a specified value; in that case, look up the tabulated t-value for $\alpha/2$ level of significance (eg: if your test is two-tailed at the 95% level, look up the t-value for $\alpha = 0.025$)
- Note: the t-statistic has $(n-1)$ D.F. since we can compute m and s from the data, but we need to *estimate* μ

Example - C.I. estimate: 16 measurements (D.F. = 15) of $\ln K$ define $m = 9.26$, $s = 2.66$
from table, for $n-1=15$, $\alpha/2 = 0.025$, $t_{\alpha/2,15} = 2.131$, therefore:

$$(2.1) \quad -2.131 \leq \frac{m-\mu}{s/\sqrt{16}} \leq +2.131 \quad \text{or, rearranging:}$$

$$(2.2) \quad \bar{m} - 2.131s/\sqrt{16} \leq \mu \leq \bar{m} + 2.131s/\sqrt{16}$$

substituting m and s : C.I.₉₅ = $7.8 \leq \mu \leq 10.7$ ie. 95% likely that μ is within this range

Examples - To compare two sample means: (Till, p.62; Cheeney, p.68) use pooled t-statistic and pooled variance (ie. taking into account the combined sizes of both sample populations) then if the computed t-statistic is less than the tabulated $t_{\alpha/2,15}$ value, the means are $1-\alpha\%$ likely to be from the same population

- Estimate a C.I. for μ : Cheeney, p.66
- see Problem Set II B. "Hypothesis Testing with StatMost" for working with t-tests and hypothesis tests
- Types of t-tests:
 - general t-test: samples are drawn from populations of equal variance; are the means different?
 - unpaired t-test: samples are drawn from populations with different variances; " " "
 - paired t-test: are two sets of outcomes drawn from the same population? (e.g. are the means of duplicate analyses of all samples the same?)
- the power of an hypothesis test: (Till, p.63-65)
 - α is the risk of accepting H_0 when it should be rejected (Type I error), whereas the risk of rejecting H_0 when it is in fact true, is β (Type II error)
 - $1-\beta$ is the "power" of a test of significance; the higher the power, the better the test; but increasing the level of significance ($1-\alpha$) reduces the power of the test ($1-\beta$); *usually, the best compromise between level of significance and power of the test is at $\alpha = 0.05$*
 - see Till, p.63-65 for a discussion of the "power" of an hypothesis test and an example
 - in general, non-parametric tests require fewer assumptions about a population but have a lower power and hence greater risk of Type II error
- 2. Kolmogorov-Smirnov Parametric Test for a Gaussian distribution: (Cheeney, p.62-64)
 - valid only for normal distributions
 - define the D statistic in a "cdf" as the maximum class interval departure from a theoretical cdf
 - for $n > 15$, define the critical D as A/\sqrt{n} for a one-sample test, where $A = 1.22$ for $\alpha = 0.05$ and 1.51 for $\alpha = 0.01$ (Cheeney, p.46)
 - if the calculated D statistic $>$ critical D value, the test sample is not normally distributed
 - Note: StatMost requires that the distribution be transformed to a standard normal deviate prior to testing; Statmost's computed D-value can then be compared with the critical D-value (A/\sqrt{n}) to accept or reject the null hypothesis that the sample distribution was drawn from a normal population
- 3. F-test - Used to Compare Variances: (Till, p.66)
 - sample two normal populations for all possible sample sizes n_1, n_2 ; define $F = s_1^2/s_2^2$ for all possible combinations of n_1, n_2 (ie. an infinite family of F distributions)
 - D.F. = n_1-1, n_2-1 ; one-tailed test, only
- 4. χ^2 -test (Chi-square test):
 - used to test how well a sample distribution fits a theoretical distribution (Till, p.69); this is a parametric test. However, it is more useful for nominal data, in a non-parametric test situation (preferred for situations with $n > 40$: Till, p.121, 124)
 - from a repetitive sampling of a normal distribution, calculate $z = (x-\mu)/\sigma$ for each member of the sample pop'n and define $\chi^2 = \sum_{i=1}^n (z^2)$ for all samples of size n
 - group the sample z-values into r classes and compute the test statistic:

$$(2.3) \quad X^2 = \sum_{i=1}^r \left\{ \frac{(\text{ObservedValue}[ith]Class - \text{ExpectedValue}[ith]Class)^2}{\text{ExpectedValue}[ith]Class} \right\} = \sum \frac{(O-E)^2}{E}$$

- D.F. is defined as $j-k-1$, where k is the number of parameters to be compared against the theoretical distribution (eg: if m, s are to be compared with μ, σ of the normal distribution, then $k = 2$)
- see Till, p. 69-70 example of χ^2 parametric test of goodness-of-fit to a theoretical distribution

Non-Parametric Tests: for distributions of unknown form; for populations of unequal or unknown variances; for nominal or ordinal-scale measurements

5. χ^2 -test: (Till, p.121-124)

- comparing different sample populations where distribution not normal or unknown
- classify n data on a nominal scale, by grouping in a contingency table, regardless of the type of distribution
- eg: a number of high-, med- and low-K measurements (lognormal distributions) made in two different facies: are the two facies the same in terms of their K distributions? (use numerical values in Table 7.4, Till, p.121)

<u>Categories</u>	Number of Measurements in:		<u>Total Number</u>
	<u>Gravel Facies</u>	<u>Coarse Sand</u>	
K>10 ft/d	a	d	a+d
1<K<10 ft/d	b	e	b+e
K<1 ft/d	c	f	c+f
<u>Totals</u>	a+b+c	d+e+f	n

- ie: data have been transformed into nominal measurements, so the only statistical analysis possible is on counting within/between groups
- set up the contingency table, with $i = 3$ rows and $j = 2$ columns and marginal row and column totals, T_i and T_j
- the expected frequency of occurrence of measurements of a given K-class in both facies is the expected value, E :
$$E_i = \frac{T_i}{n}$$
- ie. by chance alone, we would expect that the probability of measuring a low-K value in the two facies combined would be $\frac{(c+f)}{n}$ (frequency of occurrence = $p \cdot n$)
- the expected frequency of occurrence of values in a given K-class in a given facies is the joint probability times the total number of data:
$$E_{ij} = \frac{T_i}{n} \cdot \frac{T_j}{n} \cdot n = \frac{T_i \cdot T_j}{n}$$
- eg: if the two facies were hydraulically identical, a purely random distribution of K-values should exist between, as well as within, facies; so the number of measurements in the high-K category in gravel alone that would be expected by chance is:
$$E_{1,1} = \frac{(a+d) \cdot (a+b+c)}{n}$$
- create null hypothesis H_0 : no significant difference in K distribution of gravel and sand facies
- calculate the test statistic:

$$(2.4) \quad X^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where r, k = no. of rows, columns in contingency table

- D.F. defined as (r-1)*(k-1) = 2 for the above contingency table
- from χ^2 tables, for significance level, α , and D.F.= 2, look up critical value of χ^2
eg: for $\alpha = 0.05$, D.F. = 2, $\chi^2_{(0.05, 2)} = 5.99$
- compare the test statistic, X^2 , with the critical χ^2 value; if $X^2 <$ critical value, the null hypothesis is accepted (ie. 95 times out 100, the variations of a, b, c vs. d, e, f in the contingency table will occur due to chance); conversely, if $X^2 >$ critical value, the null hypothesis is rejected (ie. only 5 times in 100 will the variations be due to random sampling)

Note: see StatMost p.274-276 for detailed use of χ^2 test of RxC contingency tables

- use Statistics | Contingency Table | RxC Table to use built-in chi-square analysis for contingency tables; in StatMost, the contingency table is entered without marginal totals
- again, rather than specify a confidence level, StatMost computes the effective probability corresponding to the computed X^2 statistic
- eg: for Till's p.121 example, the null hypothesis is rejected for confidence levels greater than about 0.02 (See "chi-sq RxC example.dmd" for worked example)

6. Kolmogorov-Smirnov Non-Parametric Test for comparing two populations:

- complex but useful (see Cheeney, p. 45-46; Till, p.125-130 for details and examples)
- a more general, non-parametric test that doesn't assume normally-distributed data, but is less powerful (in a statistical sense) because the probability of erroneous conclusions increases
 - compares the *forms* of two distributions regardless of whether they have different means, variances or skewness (eg: see Cheeney, p.45-46; Till, p.126)
 - use the K-S test to compare two non-normal sample populations, or an unknown sample distribution to a normal distribution
 - because the test does not identify the source of the populations' difference(s), it loses its power for small values of n
- when testing for a simple difference between populations, use a two-tailed test; if testing whether the mean of one population is greater than that of the other, use a one-tailed test
- define the null hypothesis as $m_1 = m_2$ and define the alternate hypothesis as $m_1 \neq m_2$ (two-tail) or $m_1 > m_2$ (one-tail), where m_i = sample mean
- note the sense of direction for population difference in a one-tail test matters (Till, p.128)
- using the same class intervals (bins) for both populations, the data are represented as cumulative relative frequencies
- D = maximum cfd difference in bin i between the two sample populations = $\max[f(x_1)^i - f(x_2)^i]$ for two-tail test, D is maximum bin difference
- critical D values to compare with observed D are determined from the sample sizes of both populations and whether one- or two-tail test: eg: for different sizes of samples, the critical D value for two-tailed tests can be estimated as $A \sqrt{(n_1 + n_2)/(n_1 n_2)}$, where $A=1.36$ for $\alpha=0.05$; $A=1.63$ for $\alpha=0.01$
- Note: StatMost only provides the option of a two-tailed test but calculates the probability of the samples being from the same population; rather than specifying a confidence level α , the value of α is, in effect, back-calculated from the computed D-statistic, providing an estimate of the

minimum confidence level or probability at which the null hypothesis (equal sample means) could be accepted; thus, the computed probability gives more information about the test's level of significance (i.e. how close the test is to a borderline rejection) than a manual calculation; see the "Till p.127 K-S example.dmd" data file for examples of calculated K-S test comparisons

7. Other non-parametric tests (for ordinal data): can be very useful for geologic data
- see Cheeney (Ch.6-7), Till (Ch.7) for examples
 - Mann-Whitney U-test; Kruskal-Wallis; Spearman's rank corr.coeff.; Kendall's- τ
 - useful for ranked (ordinal) data
 - available in StatMost, SPSS, other standard statistical packages
 - before using, be familiar with test procedure, nomenclature, concept; perform a test on known data distributions (eg: use StatMost to generate test data of known distributions, apply tests to learn mechanics and interpretation of test results)

8. Problem Set I. **Statistical Summarization**

- introduction to use of software and class problem set, using demo data set (Walker Lake): univariate summary statistics, box plots, freq. and cum. freq. distributions (histograms, "cfd's"), K-S tests of normality on raw data and log-transformed data

Summary of Hypothesis Testing:

Rationale Behind Statistical Tests - from the size, the shape and dispersion of the sample data, compare the sample distribution to a known distribution or another distribution to determine similarities or differences at a specified level of confidence (probability of being wrong).

Parametric vs. Non-Parametric Tests - if a normal distribution is known or inferred, a parametric t-test is the most powerful; if the parent population distribution is not normal or is not known, a parametric test cannot be applied and non-parametric comparisons have to be applied (eg: K-S).

Type of Tests:

One Sample Tests -

One-Tailed Tests -

1. is the mean $<$ or $>$ a specified value? (t-test)

Two-Tailed Tests -

2. is the sample mean equal to a specified value? (t-test)

Two Sample Tests -

3. are two sample sets drawn from equivalent populations? (t-test or K-S)
t-tests: general, paired, unpaired

Goodness-of-Fit Tests -

4. is the sample population Gaussian? (a two-tailed test) (K-S)

Set up the Null and Alternate Hypotheses:

Null Hypothesis -

case 1. $m <$ specified value or $m >$ specified value

case 2. $m =$ specified value

case 3. $m_{\text{sample population 1}} = m_{\text{sample population 2}}$

case 4. sample's z-score distribution = standard normal distribution

In each case, the Alternate Hypothesis would be the opposite; hence, if the Null Hypothesis is rejected on statistical grounds, the Alternate Hypothesis would be accepted.

Specify the Confidence Level: For one-tailed tests, the confidence level of the test-statistic is the same as the specified confidence level of the test; eg: a 0.05 confidence level is desired, so the confidence level applied to the test-statistic is also 0.05

For two-tailed tests, the test-statistic refers equally to both tails of the outcome, so at a specified confidence level (eg: 0.05), the probability of rejecting the null hypothesis is equally shared by both tails, so that the confidence level applied to the test-statistic is half that specified (eg: 0.025)

Specify the Degrees of Freedom:

case 1., 2. (t-test) D.F. = $n - 1$

case 3. (t-test) D.F. = $n_1 + n_2 - 2$

case 3. (K-S) D.F. = $1.36\sqrt{(n_1 + n_2)/n_1n_2}$ (for $\alpha=.05$); = $1.63\sqrt{(n_1 + n_2)/n_1n_2}$ ($\alpha=.01$)

case 4. (K-S) D.F. = $1.22/\sqrt{n}$ (for $\alpha = .05$); = $1.51/\sqrt{n}$ ($\alpha = .01$)

Readings:

Isaaks and Srivastava (1989): p. 28-38 regression
p. 49-50 proportional effect, skewed data
p. 50-64 spatial continuity introduction
Ch.7 p. 140-163 spatial continuity of Walker lake data

Week 3 Lecture Material Bivariate Data, Regionalized Variables, Exploratory Data Analysis

1. Definitions:

Geostatistics (spatial statistics):

A branch of applied statistics focusing on the characterization of the geospatial dependence of one or more attributes whose values vary over space (in 1-D, 2-D, or 3-D); and the use of that spatial dependence to predict (model) values at unsampled locations.

(time-series analysis - hydrographs, the stock market - is a 1-D relative of spatial statistics)

Prediction (estimation, interpolation, modeling) methods:

Any of a number of methods to produce estimates of a variable at unsampled locations based on values at discrete points. Examples include: tessellation (Theissen polygons, triangular irregular network, Delauney triangulation, etc.), moving average, inverse distance weighting, spline functions, trend surfaces. The geostatistical equivalent is kriging, a statistically- unbiased linear estimator.

Spatial dependence (autocorrelation):

Most physical processes generate spatial variability such that two data values sampled close together tend to be more similar than two values sampled far apart. Where a strong spatial dependence exists, spatial statistical tools can be used to predict (model) values at unsampled locations better than other interpolation procedures.

Bivariate and multivariate dependence (crosscorrelation):

A physical process produces correlated variability in the values of two or more attributes, whose correlation can be used to understand the process and/or to make predictions.

2. Bivariate Correlation:

- an analysis of variation of two variables z^a , z^b drawn from different but related populations
- analysis of bivariate regression is potentially important in spatial data analysis: if an extensively sampled secondary variable (eg: topographic elevation) is correlated to the primary variable we wish to estimate (eg: water table elevation), the spatial cross-correlation between the two variables can greatly help in estimating the primary variable by exploiting the spatial correlation information in the correlated secondary variable if its values are known at other locations where the primary variable is unsampled
- bivariate regression analysis is predicated on the assumption that the distributions of both z^a and z^b are Gaussian; in some cases, this condition can be relaxed e.g. the independent variable can take on discrete values
- however, regression analysis is strictly valid only under the following conditions:
 - both variables are measured with no (or negligible) error
 - both variables are normally distributed
 - the variables are linearly correlated
 - the X values are independent
 - prediction error is homoscedastic (constant variance) and Gaussian
 - prediction errors are independent (not autocorrelated)

- in a bivariate relationship, the overall variance of the two variables can be thought of as composed of three variance components: 1) variance of variable A, 2) variance of variable B and 3) the variance arising from the correlation of A vs. B
- this latter variance is known as the bivariate covariance and is defined by:

$$(3.1) \quad Cov(z_a, z_b) = \frac{1}{n-1} \sum_{i=1}^n \{(z_{a,i} - m_a)(z_{b,i} - m_b)\}$$

- and the correlation coefficient is:

$$(3.2) \quad r = Cov(z_a, z_b) / s_{z_a} s_{z_b}$$

- the value of the square of the correlation coefficient (r^2) represents the fraction of the total variance of z_a and z_b that is due to their linear correlation

- where the population is not bivariate normal, use a non-parametric correlation method on the ranked (ordinal) form of the data, such as Kendall's- τ or Spearman's rank correlation coefficient (see Till, pp. 131-134; Cheeney, Ch. 6), which are based on the difference in rank position of all observations between two samples or between x and y values

3. Hypothesis Test of the Significance of Correlation

- from the definition of the correlation coefficient in equations (3.1) and (3.2), it is apparent that if values of the dependent variable, z_b , are close to their mean (i.e. there is little variation) and the standard deviations of one or both variables, s_{z_a} or s_{z_b} , is relatively large, then the value of r will be small regardless of how close the correlation; in other words, by itself the value of r is a poor estimator of the degree of correlation
- in other words, a more robust test is required to determine statistical significance of a bivariate correlation
- if the two variables, z_a , z_b are normally distributed, then various types of t-tests can be applied to determine if the bivariate relationship between them is statistically significant (see Till, p.86-87)
- for example, to test whether the value of the correlation coefficient represents a statistically significant bivariate correlation, a null hypothesis is set up to represent the situation where the population correlation coefficient, ρ , is zero (ie. the variables are not correlated); ie. $H_0(\rho = 0)$ and the alternative hypothesis is $H_a(r \neq 0)$; this is an example of a two-tailed test, and the t-statistic is defined as:

$$(3.3) \quad t = |r| \sqrt{\frac{(n-2)}{(1-r^2)}} \quad \text{with D.F.} = (n-2)$$

where n is the number of sample data plotted in the regression

- therefore, if the value of |t| calculated from eq'n (3.3) is larger than $t_{(n-2) \alpha/2}$, we can reject H_0 and conclude that a significant bivariate correlation ($\rho \neq 0$) exists at a confidence level of $100(1-\alpha)\%$

4. Regionalized Variables

- unlike a random variable which results from a purely random process (eg: the results of throwing dice), a regionalized variable (or r.v.) is distributed in space (and/or time) with location information attached to each measurement; in other words, the measurement variable (z) no longer represents a statistically independent univariate population, but is part of a multivariate population (x, y, z_{xy}) in which the values of z_{xy} may no longer be strictly independent in a statistical sense; that is, z_{xy} may be correlated to x and y because of the physical process which generated it
- in general, any measurement which is associated with spatial or temporal coordinates is a r.v.
- denote this variable as $z(\mathbf{x})$, where \mathbf{x} designates spatial coordinates (in 1-D, \mathbf{x} is x ; in 2-D, \mathbf{x} is x, y ; etc.); e.g. if the regionalized variable is rainfall amount, each data point is denoted with coordinates $\mathbf{x} = (x, y)$, and $z(\mathbf{x})$ is rainfall amount
- **key concept:** the variability of any regionalized measurement over/in the earth can be viewed as but one possible realization or outcome of a hypothetical random process (a God who throws dice?) which has distributed values of $z(\mathbf{x})$ in just one of an infinite number of possible ways
- **key concept:** in geostatistics, the r.v. is assumed to be the outcome of a physical process (or multiple processes) whose spatial form represents a combination of a structured aspect (eg: lead concentration in contaminated soil due to the contamination process/history) and a random, unstructured aspect (eg: the natural lead content in, and proportions of, feldspar and limestone detritus in the soil); local 'trends' such as contaminant hot spots can be handled within the geostatistical modeling process, but significant regional trends are removed prior to analysis and modeling to ensure that $z(\mathbf{x})$ represents a stationary r.v.; ie. the analysis, estimation and simulation of variation is done with the trend subtracted from the raw data, then the trend is added back in the final estimation (use Surfer or other software to model the trend, then remove it from the raw data); in this course, we will not focus on trend removal (see Koch and Link, 1971, chapt. 9); what constitutes a 'significant' trend is usually a matter of judgement and can be identified in exploratory data analysis or during variogram analysis

5. Walker Lake Data set

- see Isaaks and Srivastava, 1989
- be able to outline the phases / steps in a geostatistical analysis
- the purpose of exploratory analysis: get to know your data, identify sampling patterns, sampling history and possible sampling biases, patterns of variability, bivariate correlations among multiple r.v.'s, etc.
- examine the exhaustive Walker Lake data set, its main features of spatial variability, evidence of heteroscedasticity, correlations of V, U, T
- compare the features of the sample data set; plot the sample data, get to know it, identify patterns, sampling bias, etc. (note that in a real analysis, you will not have access to the exhaustive data set; only God knows what the real situation looks like and you are trying to reconstruct that situation from a few paltry measurements)
- **key concept:** different populations of the r.v.'s may be present (e.g. the T variable may represent rock type); the ability to segregate V, U values into two possible classes ($T = 0$ or 1) raises the question of when population splitting should occur. There are no hard rules, but some guidelines:

- 1) Is the distinction physically meaningful? One should have good reasons for splitting, even if the segregation is done subjectively;
- 2) After splitting into subpopulations, do sufficient data remain within all the subpopulations to justify statistical inference based on the numbers of data points? If some subpopulations have too few data to justify meaningful statistical measures, their segregation may not be useful; and
- 3) What is the goal of the study? Does population splitting contribute to the goal? e.g. estimating the spatial distribution of species proportions from fifty calibration locales in an area with two very different types of land cover: if the distribution of species counts in the calibration sites is statistically indistinguishable between the two land cover types, splitting into subpopulations is unnecessary; if statistically distinct, then splitting must be performed prior to geostatistical analysis.

6. Problem Set II. Exploratory Data Analysis 1

- preliminary spatial analysis (sample locations, contour maps, obvious visual trends, coordinate outliers, sample value outliers, interval [hi/lo] maps, indicator maps) using 3Plot, Surfer, GeoEas, StatMost
- identification of trends, clusters of data, spatial sampling bias (eg: multiple sampling campaigns with targeted/biased purposes)

Example of Hypothesis Testing of Regression Parameters:

A hypothesis test is often used to evaluate the significance of a linear regression fitted to a scatter plot of x, y data. For example, if the range of y values is small compared to the range of x values, even a very tight x-y correlation will produce a low value for the correlation coefficient, r^2 . This is because r^2 is a measure of the proportion of the overall variance of x and y that is due to their correlated component (eg: in a perfect correlation, all of the x-y variance is accounted for by their correlated values, and $r^2 = 1.0$); if the variance of the x values is much greater than the variance of y values, it swamps the contribution of the variance due to correlated x and y values, and r^2 will be low regardless of how good the x-y correlation really is.

An alternative t-test to the one discussed in Equation (3.3) involves a test of the significance of the calculated regression slope. The Null Hypothesis is defined as $b = 0$, where b is the calculated slope of the regression line; in other words, H_0 posits that y varies independently of x and that the x and y values are not correlated. This is a two-tailed test because the Alternate Hypothesis is that b is not zero - ie. b could be either greater than or less than zero - and we are interested in whether the statistic of the x-y correlation is out on either tail of the error curve. If H_0 were rejected, the Alternate Hypothesis, H_a , would be accepted and would conclude that x and y are correlated at the level of significance of the hypothesis test.

In this test, the t-statistic is defined by the formula:

$$(3.4) \quad t = (b - b_o) \frac{s_x}{s_e} \sqrt{n}$$

where b_o is the specified slope (zero in this case), s_x is the standard deviation of the independent variable, and

$$s_e = \sqrt{\frac{n^2 s_y^2 (1-r^2)}{n(n-2)}} \text{ is the standard error of the correlated data}$$

The confidence level defining the critical t-statistic is $\alpha/2$ (two-tailed test) and the test has n-2 degrees of freedom. The Null Hypothesis is rejected if the absolute value of the t-statistic calculated in Equation (3.4) exceeds the tabulated critical t-statistic, $t_{(n-2), \alpha/2}$, and the regression slope is said to be significant at the $100(1-\alpha)$ % level.

Once the regression has been deemed significant, a confidence interval about the least-squares value of the slope can also be determined with the calculated t-statistic. In this case, we wish to determine the values of b_o in Equation 3.4 that produce a calculated value of $|t|$ that is equal to $t_{(n-2), \alpha/2}$. In other words

$$(3.5) \quad C.I. \text{ for slope} = b \pm t_{(n-2), \alpha/2} \cdot \frac{s_e}{s_x} \sqrt{\frac{1}{n}} \text{ at the } 100(1-\alpha) \% \text{ level.}$$

Note that because Equation (3.4) specifies the slope against which the regression slope is tested, this test can be used to compare whether two regression slopes are statistically the same or different.

Week 4-5 Lecture Material Autocorrelation and Variogram Measures of Spatial Continuity

1. Autocorrelation:

- the correlation function defined in equation (3.2) for bivariate data measures the degree of correlation between two related variables (not necessarily regionalized variables)
- within a single 1-D series of spatial measurements, an autocorrelation function can be defined that is a measure of the internal correlation between successive measurements
- for a 1-D series of measurements, the concept of autocorrelation is analogous to the covariance of bivariate data, where the variable z_i^b becomes z_{i+L}^a where L is the offset from position i in the data series; thus, the autocorrelation function is defined as:

$$(4.1) \quad r_L = cov(z, z + L)/s_z^2 = \left[\frac{1}{n-1} \sum_{i=1}^n \{(z_i - m_i)(z_{i+L} - m_{i+L})\} \right] / s_z^2$$

where m represents the mean of the values defined for zero offset and for an offset of L (note that the definition of covariance contained in the numerator of equation 4.1 equals the population variance of z when L = 0)

- this function is calculated for various offsets or separations, L, called "lags" and the value of r_L is plotted at each value of L to form the correlogram (equivalent to the standardized spatial covariance function defined below)
- conceptually, equation 4.1 compares the degree of correlation or similarity between the time-series and its copy, where the copy is shifted by L units and a standardized covariance for the region of overlap is computed; note that at L = 0, the covariance term equals the sample variance and r_L equals 1.0; as L increases, the amount of overlap decreases until the length of record compared is too small to produce reliable estimates of r_L ; n is therefore the number of common data values in the overlapped portion of the data series and its shifted copy
- since the degree of correlation is symmetric for positive and negative lag shifts, only the absolute value of L is plotted
- see Davis, p.235 for examples of different autocorrelative behavior
- a cross-correlation function can be defined for the comparison of two different 1-D time-series using the cross-covariance; the equation is identical, with the appropriate superscripts (z^a and z^b denoting the two different variables) added to the z_i and z_{i+L} terms in equation (4.1) (see Davis, p.240-243)
- Note: for nominal data (e.g. sediment types in stratigraphic sequences), use cross-association (eg: correlation between two sequences of rock types) and a non-parametric test (such as a χ^2 -test) for significance of match (see Davis, p. 247-250)

2. Regionalized Variable vs. Random Function

- aside from random sampling and analytical errors, a geospatial measurement (the regionalized variable, e.g. copper content) is considered to be essentially deterministic (non-random), ie: there exists a single value of porosity, or one possible copper concentration at a point, or a unique water level at any given time in a given water well

- **key concept:** in order to develop spatial correlation statistics from such a variable from which to make geostatistical estimates at unsampled locations, the r.v. is assumed to represent one

statistical sample drawn from an infinite number of possible samples all having identical statistical characteristics; for example, a hydrograph of water level vs. time represents only one possible statistical sample of the distribution of water levels vs. time drawn from an infinite number of possible distributions with the same statistical characteristics

- the fictitious domain of all such possible distributions is known as a Random Function (R.F.) and a single sample of the regionalized distribution of possible porosities or copper contents drawn from it is called a realization of the R.F.
 - a R.F. is a function from which values can be drawn which have a variance about a mean, together with skewness, kurtosis, etc.; each of these statistical measures also depends on spatial position; therefore, to fully specify a R.F.'s statistical properties is a theoretical nightmare and so in practice many simplifications are used to represent the R.F.
 - our task in geostatistics is to infer the nature of the R.F. controlling the spatial distribution of the regionalized variable (eg: the available water level or porosity measurements) so that this function can be used to estimate values of the variable of interest at unsampled locations or times
 - this is analogous to the task of estimating a univariate population distribution (analogous to the R.F.) from a number of statistical samples (analogous to the realizations of the R.F.) drawn from the population; the key difference in geostatistical analysis is that we do not have multiple samples of the R.F. from which to infer its characteristics, only a single sample (the geospatial data representing the regionalized variable) consisting of a limited number of points
 - for Star Trek fans: to use a crude analogy, if we could sample water level vs. time at a particular point in a river in a number of parallel universes, we'd be better able to estimate the underlying R.F., in a manner analogous to the statistician who can draw red and black marbles from a bag many times to estimate the true proportion of marbles in the bag
- **key concept:** the mean state of all possible samples of a r.v. would be equal to the average value of the r.v. within a single sample; this would be an example of ergodicity (from the Greek for “wandering”), in other words, *a single sample would reflect the statistical character of the R.F.*
- the assumption of ergodicity is therefore a crucial one; it is required to infer the properties of the R.F. from a single realization (the measured copper values or porosities); however, non-ergodic behavior is common in the real world, but the criteria for recognizing it are subjective (see discussion under kriging section, Week 7)
 - note that the requirement of ergodicity is not unique to geostatistics: it is an implicit assumption in all inferential statistics: from estimating the proportion of red and black marbles in a bag, to inferring a population distribution from a histogram, to estimating a population mean from the sample mean
- **key concept:** because we only have one realization to work with in geostatistics, one more key concept must be introduced if we hope to use statistics to make estimates at unsampled locations: if homogeneous physical process(es) produced the variability in an r.v. over some area of interest, then the r.v. will demonstrate the same kind of variability over the entire area as it does within smaller subareas; in other words, the R.F. from which the r.v. is drawn is stationary, and statistical homogeneity (stationarity) can be assumed over the entire area; this is equivalent to saying that *if we divide the study area into smaller parts, each part could be considered a different realization of the same R.F.*; if so, then we can generate statistical estimates from each

of the smaller parts as a kind of surrogate for drawing multiple samples from the R.F. (see Pannatier, p.77 and Fig.A.2)

- **note that ergodicity requires stationarity, but that stationarity does not imply ergodicity**

- there are various degrees of stationarity of the underlying process responsible for generating the regionalized variable; the type of stationarity assumed determines the kind of statistical inference that is permitted:

- strict stationarity, in which *all* the random function's parameters are invariant from point to point, is rarely assumed because of the formidable challenge in describing all its parameters
- second-order stationarity exists if the R.F.'s mean and variance are independent of location and the covariance depends only on separation or lag between measured values of the regionalized variable
- the intrinsic hypothesis is the weakest assumption; certain physical processes (eg: Brownian motion) do not have a definable variance or covariance, but the variance of their increments does exist (see definition of semivariogram below), in which case the semivariogram can be defined but other measures of spatial correlation (e.g. covariance) cannot be used

3. Moments

- the expected value of a Random Function $Z(\mathbf{x})$ at any location \mathbf{x} is equal to its mean:

$$m(\mathbf{x}) = E\{Z(\mathbf{x})\}, \text{ assumed constant for a stationary R.F.}$$

(in other words, the R.F. is assumed to have Gaussian pdf characteristics)

- in linear (two-point) geostatistics, there exist three second-order moments:

variance:
$$\begin{aligned} \text{Var}\{Z(\mathbf{x})\} &= E\{[Z(\mathbf{x}) - m(\mathbf{x})]^2\} \\ &= \frac{1}{n} \sum [Z(\mathbf{x}) - m(\mathbf{x})]^2 \end{aligned}$$

covariance:
$$\begin{aligned} C(\mathbf{x}, \mathbf{x}+\mathbf{L}) &= E\{[Z(\mathbf{x}) - m(\mathbf{x})][Z(\mathbf{x}+\mathbf{h}) - m(\mathbf{x}+\mathbf{h})]\} \\ &= E\{Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h})\} - m(\mathbf{x})m(\mathbf{x}+\mathbf{h}) \\ &= \frac{1}{n(L)} \sum [Z(\mathbf{x}) \cdot Z(\mathbf{x}+\mathbf{h})] - m(\mathbf{-h}) \cdot m(\mathbf{+h}) \end{aligned}$$

semivariogram:
$$\begin{aligned} \gamma(\mathbf{x}, \mathbf{x}+\mathbf{L}) &= 1/2 E\{[Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})]^2\} \text{ (valid only if no trend exists)} \\ &= \frac{1}{n(L)} \sum [Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})]^2 \end{aligned}$$

- Note: these second-order moments are not a function of location but are only dependant on lag separation, h

4. Practical Definition of Spatial Correlation Structure for an r.v.

- based on second-order moments
- the term “variogram” is used here as a generic term for a spatial correlation estimator statistic
- specific second-order statistics are defined differently, and are used to better summarize

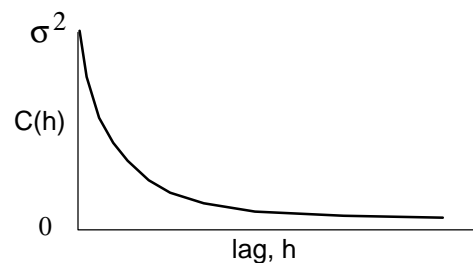
- non-normally distributed data or data with extreme-valued outliers
 - the experimental measure of spatial covariance is defined by:

$$\begin{aligned}
 (4.2) \quad \text{cov}(z_i, z_{i+h}) = C(h) &= \frac{1}{n_h} \sum_{i=1}^{n_h} (z_i - \bar{z}_i)(z_{i+h} - \bar{z}_{i+h}) \\
 &= \frac{1}{n_h} \sum_{i=1}^{n_h} (z_i \cdot z_{i+h}) - \bar{z}_i \cdot \bar{z}_{i+h} \\
 &= \frac{1}{n_h} \sum_{i=1}^{n_h} (z_i \cdot z_{i+h}) - \bar{z}_- \cdot \bar{z}_+
 \end{aligned}$$

where $z_i = i^{\text{th}}$ data value at location (x_i, y_i) , $h = \text{lag}$, n_h is the number of data pairs separated by lag h , and the overbars represent the means of the two endpoints of the lag pairs (also often expressed as - and +)

- the covariance is equal to the sample variance when $h = 0$ ie. at zero lag offset, the values of z_i and z_{i+h} are equal and equation (4.2) equals the definition of variance, thus $C(0) = \sigma^2$; at large values of h , the values of z_i and z_{i+h} are poorly correlated and $C(h) \rightarrow 0$

- graphically, the theoretical covariance function looks like this:



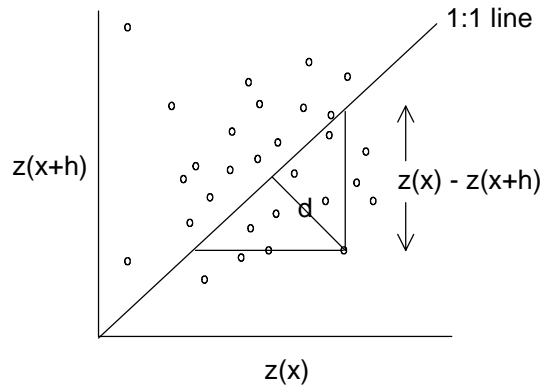
- typically, if the experimental values of $\text{cov}(h)$ level off at large lags, the underlying random function is assumed to be second-order stationary (ie. its mean and variance are independent of location, and covariance depends only on lag separation, h); in this case, the spatial covariance can be expressed as the inverted covariance (sometimes referred to as the non-ergodic covariance, as in the GeoEas software, because it does not assume that $\bar{z}_i = \bar{z}_{i+h}$ in equation 4.2):

$$(4.3) \quad C'(h) = C(0) - C(h), \text{ or } C'(h) = \sigma^2 - C(h)$$

- note that in the presence of a trend, the covariance does not level off and can take on negative values; in that case, second-order stationarity does not exist

- the semivariogram definition is graphically derived from an h -scatterplot (Hohn, p.91-92)
 - plot all values of z_{i+h} vs. z_h that are separated by a given value of h (in practical terms, a range of h values)
 - the moment of inertia, I_m , of the cloud of points about the 45° line is defined as:

$$I_m = \frac{1}{n_h} \sum_{i=1}^{n_h} d_i^2$$



- since the moment of inertia is defined about a 1:1 relationship, a right triangle defines the relationship between d and $[z(x) - z(x+h)]$

- therefore, $2d_i^2 = (z_i - z_{i+h})^2$, and the semivariogram is defined as

$$(4.4) \quad \gamma(h) = I_m = \frac{1}{2n_h} \sum_{i=1}^{n_h} (z_i - z_{i+h})^2$$

- the standardized semivariogram is defined as

$$\gamma_s(h) = \gamma(h)/C(0) \quad \text{where } C(0) \text{ is the sample variance}$$

- similarly, the correlogram is the spatial covariance standardized (divided) by the sample variance, and is exactly analogous to the autocorrelation function for second-order stationarity:

$$\rho(h) = C(h)/C(0)$$

- it takes on a similar form to the other variogram measures when it is expressed as the inverted correlogram:

$$\rho'(h) = 1 - C(h)/C(0)$$

- the madogram is defined as the mean (sometimes median) absolute difference:

$$mad(h) = \frac{1}{n_h} \sum_{i=1}^{n_h} |z_i - z_{i+h}|$$

(for computational definitions of all these statistics, see Deutsch and Journel, p. 45 and Isaaks and Srivastava, p.59)

- these various estimators of spatial correlation are all defined differently but all can be used to inform the parameters of spatial correlation structure during modeling

- except for the madogram, these various estimators are exactly equivalent representations of the underlying R.F. if second-order stationarity exists:

$$(4.5) \quad \rho'(h) = C'(h)/C(0) = 1 - C(h)/C(0) = \gamma(h)/C(0) = \gamma_s(h)$$

- i.e. if the various variogram estimators level off beyond a certain lag then second-order stationarity can be assumed and the inverted covariance, the semivariogram, the standardized semivariogram, and correlogram are all equivalent estimators of spatial continuity

- if the semivariogram does not level off but does not rise faster than the square of h , the random function is not second-order stationary and is said to obey the “intrinsic hypothesis”; in that case, only the semivariogram is a valid estimator of spatial continuity and the other estimators cannot be fitted with a model random function

- if experimental semivariogram values increase as fast as or faster than h^2 , then the intrinsic hypothesis is invalid and the presence of a regional trend is indicated; in order to proceed with variogram analysis, the trend would have to be removed and variogram analysis and modeling performed on the residuals

- become familiar with the following nomenclature: nugget; sill; transition region; range of influence; types of variogram shapes: linear, parabolic, spherical, exponential, gaussian

- Note: in order to analyze variogram structure and to utilize the resulting correlation structure in subsequent kriging or simulation of a geometrically-deformed body (such as a stratified and folded reservoir or ore body, or a stratified aquifer of variable thickness, or a dipping geologic formation), or to avoid numerical instability problems associated with matrices built from coordinate data of different number size (such as x,y in millions of meters vs z in tens of meters), coordinate transformations are applied

- one common type of transformation, utilized for folded or variable thickness geologic bodies, represents the original x,y,z coordinates in stratigraphic coordinates of an equivalent, simpler tabular body:

$$(4.6) \quad z'(x, y) = \frac{\text{top}(x,y) - z(x,y)}{\text{thickness}(x,y)}$$

where z and z' are the original and stratigraphic coordinates, respectively, at location (x,y) ; $\text{top}(x,y)$ is the elevation of the top of the original stratified body; and $\text{thickness}(x,y)$ is the thickness of the geologic body at location (x,y) . This transformation "straightens out" a contorted or variable thickness geologic body and represents it for purposes of correlation structure analysis and modeling as an equivalent tabular body.

- another common transformation is to change the number size of x,y coordinates to match the number size of z coordinates; for example, if the range of z is from 2500 to 3000 ft above sea level, but x,y are in state plane feet with values of the order of 500,000, transform the x,y values as:

$$(4.7) \quad x' = x - x_{\min} \quad y' = y - y_{\min}$$

where x,y and x',y' are the original and transformed coordinates, respectively, and x_{\min}, y_{\min} are the minimum values of the x,y ranges (Note: for 2D data, transformation 4.7 may be necessary for programs such as VarioWin which cannot handle x,y values larger than 5 digits)

5. Computation of Experimental Variograms

- see reading handout (VarioWin's Chapter 2 short tutorial)
- concepts: lag bins, mean lags, overlapping bins, variable bins, directional search parameters
- rules of thumb: minimum data pairs per lag bin ca. 20-30; max. lag ca. 1/2 of max. separation
- see hand-out of variography process

6. Software Comparison

- various variogram programs differ slightly in the manner in which lags are represented for each lag interval, and the flexibility with which lag intervals can be specified; programs which plot lags as their weighted means are preferable because the effects of clustered data on variogram shape can be visually identified
- GeoEas plots weighted mean lags, and allows specification of unequal lag intervals
- VarioWin plots weighted mean lags, and only allows equally-spaced lag intervals
- GamV3 also plots weighted mean lags, and only allows equal lag intervals (but can be recoded to incorporate unequal lags)
- software such as GeoPack plots centered lags, in which the contribution of clustered data cannot be discerned from the correlation function plots

7. Problem Set **III. Exploratory Data Analysis 2**

- clean up the raw Walker Lake data file, evaluate the effect of clustering and create a declustered data set for V using DECLUS, and calculate the normal-score transform of the V data using NSCORE