Introduction to Geostatistics — Course Notes

Ye Zhang

Dept. of Geology & Geophysics

University of Wyoming

Draft date January 12, 2011

Contents

Contents **i**

0.1 Overview . . . iv

1 Overview **1**

1.1 Why Geostatistics? . . . 1

1.2 Geostatistical Prediction . . . 2

1.3 Geostatistics versus Simple Interpolation . . . 4

1.4 Limitations . . . 6

1.5 This Class . . . 9

1.5.1 References . . . 9

1.5.2 Outline . . . 12

1.5.3 Homework, Grades, Tools . . . 12

1.6 Suggested Reading . . . 12

2 Probability Theory Review **15**

2.1 Nomenclature and Notation . . . 15

2.2 Univariate Analysis . . . 16

2.2.1 Introduction . . . 16

2.2.2 Formal Definitions . . . 18

2.2.3 Random Variable Arithmetic . . . 20

2.3 Bivariate Analysis . . . 20

2.3.1 Introduction . . . 20

2.3.2 Bivariate Random Variables . . . 22

2.3.3 Bivariate Arithmetics . . . 23

2.4 Multivariate Analysis . . . 24

2.4.1 Linear Combination of Many r.v. . . . 24

2.4.2 Multivariate Arithmetics . . . 25

2.5 Gaussian Distribution . . . 25

2.6 Central Limit Theorem . . . 26

2.7 Chapter Project . . . 28

2.8 Suggested Reading . . . 31

3 Spatial Analysis **33**
3.1 Conventional Analysis (Nongeostatistical) 33
3.1.1 Data Posting 33
3.1.2 Contour Map 34
3.1.3 Symbol Map 34
3.1.4 Indicator Map 34
3.1.5 Moving Window 35
3.2 Spatial Continuity Analysis (Geostatistical) 37
3.2.1 Experimental Variogram 39
3.2.2 h-Scatterplot 42
3.2.3 Variogram versus Univariate Statistics 44
3.2.4 Higher Dimensions & Statistical Anisotropy 46
3.2.5 Pure Nugget Variogram 51
3.2.6 Standard Deviation of $\hat{\gamma}(\mathbf{h})$ 56
3.2.7 Irregular Data: Variogram Search Envelope 57
3.2.8 Exploring Anisotropy 60
3.2.9 Spatial Continuity Analysis: Outline 65
3.2.10 Issues... 67
3.3 Chapter Project 72
3.4 Suggested Reading 74

4 Variogram Modeling **75**
4.1 Basic Permissible Models 75
4.1.1 Positive Definiteness 75
4.1.2 Basic Variogram Functions 77
4.1.3 1D Variogram Model—Statistical Isotropy 81
4.1.4 Higher Dimensional Model—Statistical Anisotropy 81
4.1.5 Nested Model 82
4.2 Model-Fitting “Rules of Thumb” 83
4.3 Chapter Project 85
4.4 Suggested Reading 89

5 Estimation **91**
5.1 The Problem of Estimation 91
5.2 Nongeostatistical (Deterministic) Estimation 95
5.2.1 Global Estimation 95
5.2.2 Local Estimation 96
5.3 Estimation Criteria 98
5.4 Geostatistical (Probabilistic) Estimation 98
5.4.1 Random Function Models 99
5.4.2 Ordinary Kriging 101
5.4.3 Cokriging & Collocated Cokriging 112
5.4.4 Kriging with moving neighborhood 116
5.5 Chapter Project 117
5.6 Suggested Reading 119

6 Advanced Topics 121
6.1 Cross Validation . . . 121
6.2 Other Estimators . . . 122
6.2.1 Simple Kriging . . . 122
6.2.2 Indicator Kriging . . . 124
6.2.3 Block Kriging . . . 126
6.3 Geostatistical Simulation . . . 127
6.3.1 Cholesky Decomposition . . . 130
6.3.2 Conditional Simulation . . . 133
6.3.3 Sequential Gaussian Simulation (SGS) . . . 136
6.4 Chapter Project . . . 139
6.5 Suggested Reading . . . 141

7 Final Project 143
7.1 Overview . . . 143
7.2 What to hand in . . . 146
7.3 Additional notes . . . 148

8 Geostatistics Final Exam 149

This is the lecture note written & assembled by Ye Zhang for an introductory course in Geostatistics.

Fall 2010
GEOL 5446
3 CREDITS
A-F GRADING
Pre-requisite: Calculus I & II; Linear Algebra; Probability & Statistics; Matlab programming language
Location: ESB1006
Times: TTh (9:35 am ∼ 10:50 pm)
Office hour: M(4:00∼5:30 pm), F(3:00∼4:30 pm), GE 220
Email: yzhang9@uwyo.edu
Phone: 307-766-2981
The syllabus: see handout.

NOTE: The lecture note do not include: (1) solutions to the exercises and projects; (2) proofs to theories and equation derivations. These will be presented only during lectures. So, please do not rely on the notes for everything — class attendance and participation are key to doing well.

0.1 Overview

Geoscientists often face interpolation and estimation problems when analyzing sparse data from field observations. Geostatistics is an invaluable tool that can be used to characterize spatial or temporal phenomena[1]. Geostatistics originated from the mining and petroleum industries, starting with the work by Danie Krige in the 1950's and was further developed by Georges Matheron in the 1960's. In both industries, geostatistics is successfully applied to solve cases where decisions concerning expensive operations are based on interpretations from sparse data located in space. Geostatistics has since been extended to many other fields in or related to the earth sciences, e.g., hydrogeology, hydrology, meteorology, oceanography, geochemistry, geography, soil sciences, forestry, landscape ecology. In this class, both fundamental development of geostatistics and simple, practical applications in the earth sciences will be presented. Exercises and projects are designed to help elucidate the fundamental concepts. Reading assignments will be given illustrating the applications of geostatistics in the particular field of reservoir characterization and modeling.

[1] In this class, we're concerned only with spatial analysis; temporal phenomena might be better understood in a separate class on time series analysis.

Chapter 1

Overview

What is geostatistics? Data analysis and spatial continuity modeling (Journel, 1989). Establish quantitative measure of spatial correlation to be used for subsequent estimation and simulation (Deutsch, 2002). The following introduction and overview materials are based on compilation of several source materials (see full references in Sec. 1.5.1).

1.1 Why Geostatistics?

Classic statistics is generally devoted to the analysis and interpretation of uncertainties caused by limited sampling of a property under study. Geostatistics however deviates from classic statistics in that Geostatistics is not tied to a population distribution model that assumes, for example, all samples of a population are normally distributed and independent from one another. Most of the earth science data (e.g., rock properties, contaminant concentrations) often do not satisfy these assumptions as they can be highly skewed and/or possess spatial correlation (i.e., data values from locations that are closer together tend to be more similar than data values from locations that are further apart). To most geologists, the fact that closely spaced samples tend to be similar is not surprising since such samples have been influenced by similar physical and chemical depositional/transport processes.

Compared to the classic statistics which examine the statistical distribution of a set of sampled data, geostatistics incorporates both the statistical distribution of the sample data *and* the spatial correlation among the sample data. Because of this difference, many earth science problems are more effectively addressed using geostatistical methods. As stated by Marc Cromer (*in* Geostatistics for environmental and geotechnical applications, 1996, ASTM International, edited by Rouhani et al.):

Geostatistical methods provide the tools to capture, through rigorous examination, the descriptive information on a phenomenon from sparse, often biased, and often expensive sample data. The continued examination and quantitative *rigor of the procedure provide a vehicle for integrating qualitative and quantitative understanding by allowing the data to "speak for themselves". In effect, the process produces the most plausible interpretation by continued examination of the data in response to conflicting interpretations. ... The application of geostatistics to environmental problems (e.g., groundwater contaminant cleanup) has also proven a powerful integration tool, allowing coordination of activities from field data acquisition to design analysis. For example, data collection is often incomplete, resulting in uncertainties in understanding the problem and increasing the risk of regulatory failure. While this uncertainties can often be reduced with additional sampling, the benefits must be balanced with increasing cost. ... Thus,* ***geostatistics offers a means to quantify uncertainty****, while leveraging existing data to support sampling optimization.*

1.2 Geostatistical Prediction

The goal of geostatistics is to predict the possible spatial distribution of a property. Such prediction often takes the form of a map or a series of maps. Two basic forms of prediction exist: **estimation** (Figure 1.1) and **simulation** (Figure 1.2). In estimation, a single, statistically "best" estimate (map) of the spatial occurrence is produced. The estimation is based on both the sample data and on a model (variogram) determined as most accurately representing the spatial correlation of the sample data. This single estimate or map is usually produced by the kriging technique. On the other hand, in simulation, many equal-likely maps (sometimes called "images") of the property distribution are produced, using the same model of spatial correlation as required for kriging. Differences between the alternative maps provide a measure of quantifying the uncertainty, an option not available with kriging estimation.

Geostatistics has played an increasing role in both groundwater hydrology and petroleum reservoir characterization and modeling, driven mainly by the recognition that heterogeneity in petrophysical properties (i.e., permeability and porosity) dominates groundwater flow, solute transport, and multiphase migration in the subsurface. Geostatistics, by transforming a sparse data set from the field into a spatial map (kriging estimation), offers a means to recreate heterogeneity to be incorporated into numerical flow and transport modeling. On the other hand, by transforming a sparse data set into multiple spatial maps (unconditional/conditional simulations), it offers a means of evaluating the uncertainties on modeling due to the uncertain nature of each map (Figure 1.3). In both reservoir simulation and groundwater modeling, for example, Monte Carlo simulation is a popular technique. Note that this uncertainty reflects our lack of knowledge about the subsurface, though the geological "groundtruth", albeit unknown, is deterministic and certain.

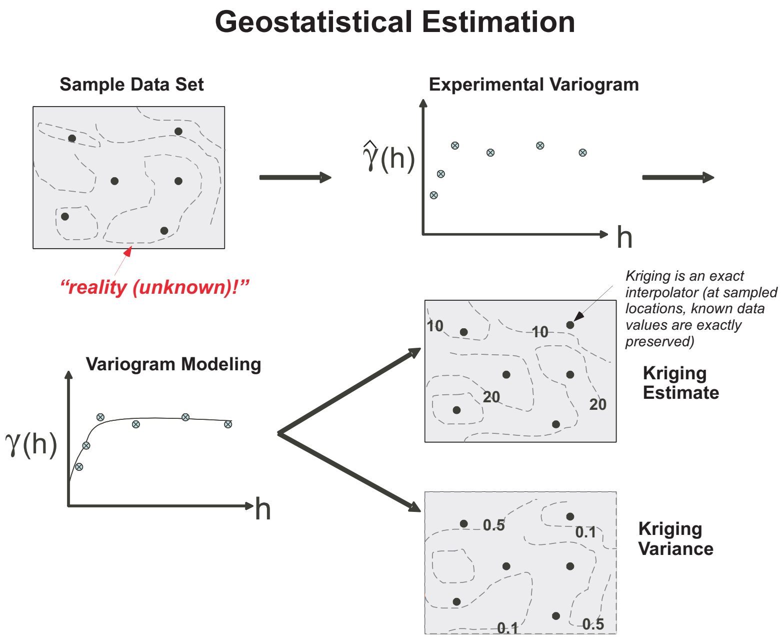


Figure 1.1: Geostatistical estimation workflow.

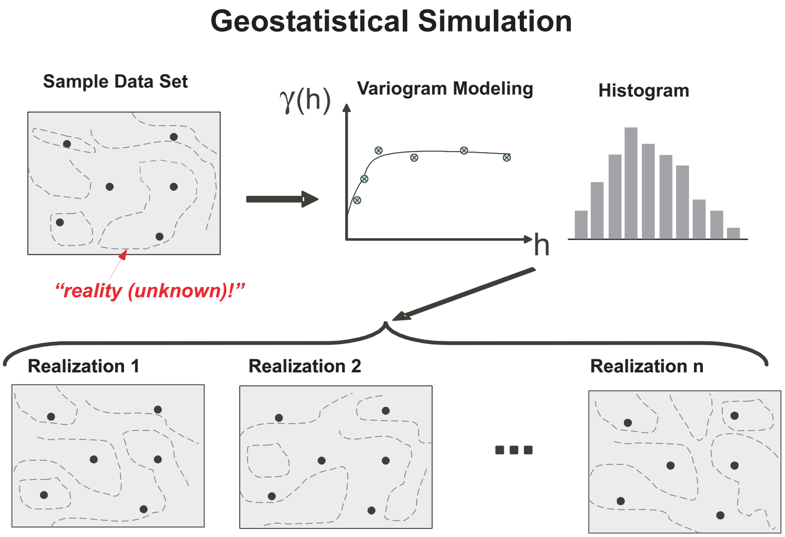


Figure 1.2: Geostatistical simulation workflow.

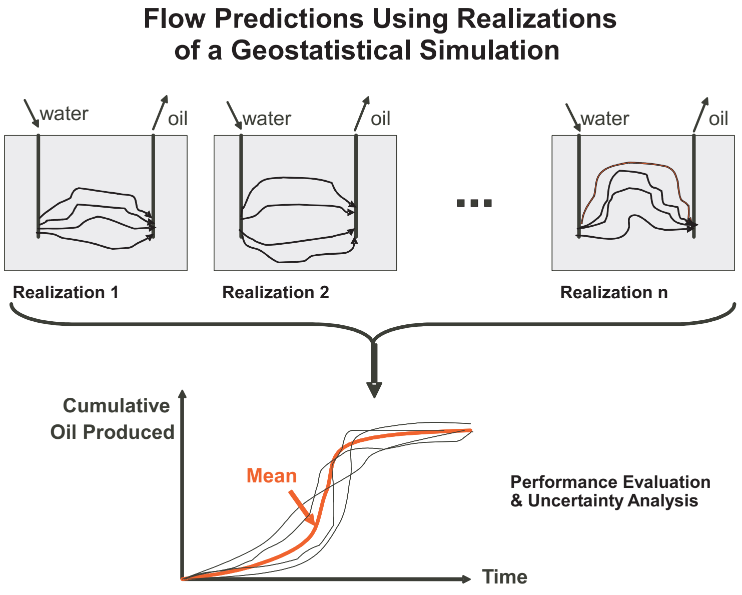


Figure 1.3: Flow predictions based on the realizations generated in a geostatistical simulation, e.g., Figure 1.2.

1.3 Geostatistics versus Simple Interpolation

In geostatistical estimation, we wish to estimate a property at an *unsampled* location, based on the spatial correlation characteristics of this property and its values at existing sampled locations. But, why not just use simple interpolation? How is spatial correlation incorporated in the geostatistical approach? A simple example may illustrate this point more clearly (Figure 1.4): we know permeability at n sampled locations, we wish to estimate the permeability at an unsampled location, z_0. Using inverse distance, the unknown value can be evaluated as:

$$z_0 = \sum_{i=1}^{n} w_i z_i \qquad (\textit{estimate})$$

$$w_i = \frac{1/d_i}{\sum_{i=1}^{n}(1/d_i)} \qquad (\textit{weight})$$

We can see that the above relation is a linear estimator, i.e., z_0 is a weighted sum of the n known values. Each weight (w_i) (assigned to a known z_i) is determined by the distance of the known data point to the unknown data point. For $n = 7$, for example, the weights can be calculated easily as shown in Figure 1.5.

Using this scheme, the weights assigned to points 1, 2, 4, 6 are all equal to 0.2. However, from the understanding of geology, we realize that permeability

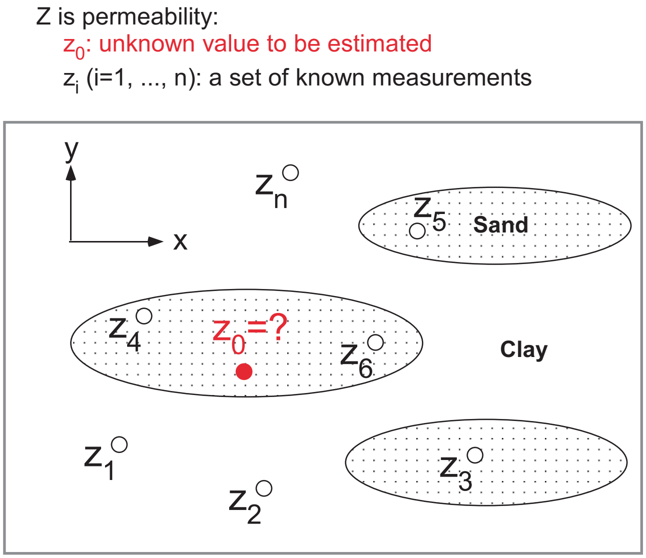

Z is permeability:
z_0: unknown value to be estimated
z_i (i=1, ..., n): a set of known measurements
y
x
z_n
z_5
Sand
z_4
z_0=?
z_6
Clay
z_1
z_2
z_3

Figure 1.4: Estimation of the unknown permeability z_0 based on a set of known values of permeability at n locations.

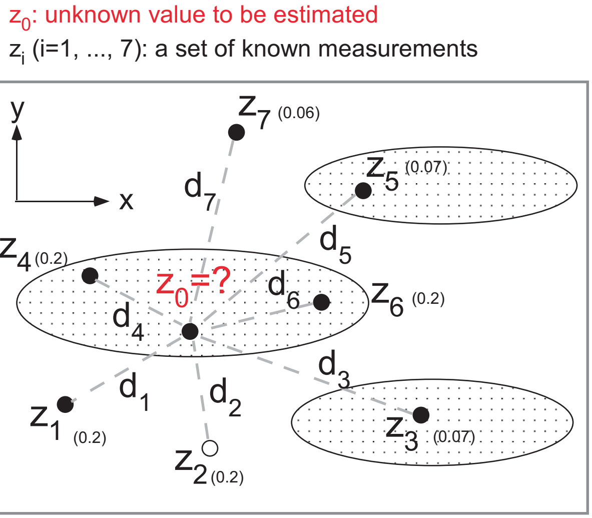


Figure 1.5: Estimation of the unknown z_0 given 7 known values. Numbers in parenthesis are weights assigned to the known values based on inverse distance.

within the elongated sand body should be more similar in the lateral direction. Thus, points 4 and 6 should be given higher weights than points 1 and 2. This is obviously not the case when using inverse distance. Thus, in conventional interpolation methods (e.g., inverse distance, inverse distance squared), information on spatial correlation is not incorporated. On the other hand, geostatistical estimation considers *both* distance and spatial correlation. In general, geostatistical estimation consists of 3 steps: (1) examining the similarity between a set of sample (known) data points via an experimental variogram analysis; (2) fitting a permissible mathematical function to the experimental variogram; (3) conducting kriging interpolation based on this function. In the above example, the spatial correlation will be revealed by the more similar values of z_4 and z_6 (step (1)). It will be modeled via step (2) (variogram modeling). Then, using kriging, we'll find that the weights assigned to points 4 and 6 will increase (those of 1 and 2 will decrease accordingly since the total weight must sum to 1.0) (step (3)). In kriging, based on the new weights, a best linear unbiased estimate of z_0 is obtained. Further (though sometimes optional depending on the goal of the study), uncertainty in the estimated field is additionally evaluated. In this class, we'll use many exercises to illustrate how to conduct a geostatistical study.

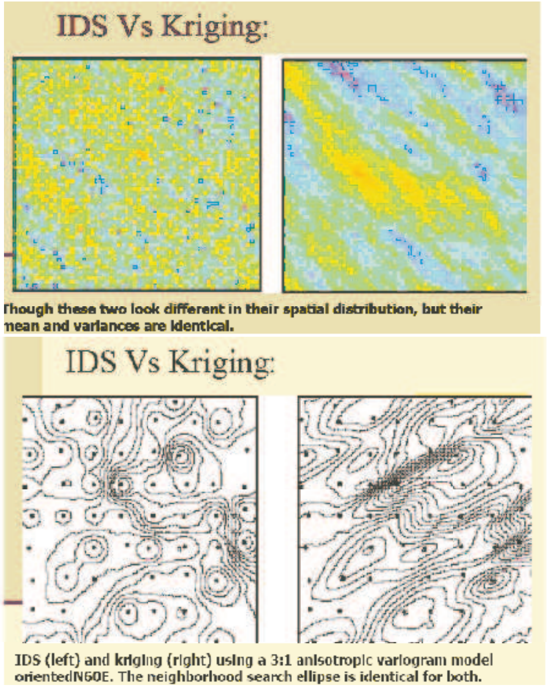
Given the same set of sampled data, interpolation results using IDS (d_i is replaced by d_i^2) and kriging can look drastically different (Figure1.6). However, does this mean that kriging is the preferred interpolation method regardless of the data? It turns out, there are situations when the sampled data are simply not *good* for kriging (we'll explore this aspect when we look at the "pitfalls" of conducting a variogram analysis). Given such data—either too unreliable or too sparse and widely spaced to capture the spatial correlation of the true property field, the conventional IDS may give just as good result. The decision of which method to use is in a way data-driven. Usually, an increase in sample quality or density will affect which method may be the most appropriate for the study.

1.4 Limitations

What is not geostatistics?

Interestingly, geostatistics models mathematical objects, not geological objects. For example, given a set of spatial measurements of isopach values, a geologist can create various contour maps based on his/her understanding of the underlying geology (Figure 1.7). This process is best described as *pattern recognition*—the geologist has an existing idea of the underlying geology when doing the interpretation. Geostatistics, however, does not recognize pattern, rather, it is based on a set of mathematical principles.

As stated by *AndréJournel* (1989), "geostatistics is an art, and as such, is neither completely automatable nor purely objective". In an experiment conducted by the US EPA, 12 independent geostatisticians were given the same dataset and asked to perform the same kriging. The 12 results were very different due to widely different data analysis conclusions, variogram models, choices of



IDS Vs Kriging:

Though these two look different in their spatial distribution, but their nean and variances are identical.

IDS Vs Kriging:

IDS (left) and kriging (right) using a 3:1 anisotropic variogram model orientedN60E. The neighborhood search ellipse is identical for both.

Figure 1.6: Estimation results from IDS and Kriging, based on the same set of sample data.

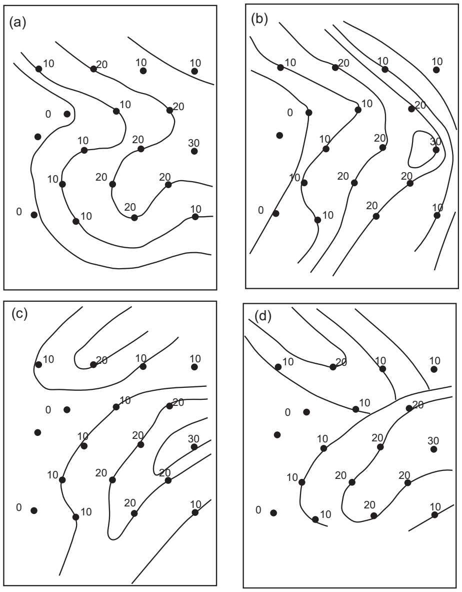


Figure 1.7: A geologist's interpretations of isopach data (after Chiles & Delfiner, 1999): (a) meandering channel; (b) in-fill channel; (c) paleovalleys; (d) barrier bar and tidal channel.

kriging type, and search strategy. As stressed by Journel, “that there are no accepted universal algorithm for determining a variogram/covariance model, that cross-validation is no guarantee that an estimation procedure will produce good estimates at unsampled locations, that kriging needs not be the most appropriate estimation method, and that the most consequential decisions of any geostatistical study are made early in the exploratory data analysis”.

In this class, I will repeatedly emphasize the importance of understanding our data, via exploratory analysis, trend analysis, error identification, and dealing with sampling issues and non-stationarity. From both the environmental engineering and petroleum reservoir modeling literature, I present “rules of thumb” or “best practice” guide that is recommended by experts in the field. Further, it is my recommendation that before you embark on a geostatistical study, you should research the literature for analysis conducted on similar data in the past. You can often learn a lot from past studies and hopefully, you can try to avoid pitfalls that others had stumbled upon before you. In the end of this class, I will present a lecture on literature search and point to further resources that you can use to solve your own problems.

Thus, geostatistics is not a black box. Without understanding its fundamental assumptions and limitations, an untrained person is more likely use it incorrectly. As summarized by Journel (1989): Geostatistics is a tool: it cannot produce good results from bad data. It cannot replace common sense, good judgment, or professional insight. Throughout the course, I’ll pay equal attention to its limitations as well as its useful applications. In practice, as more data become available, the geostatistical procedure often need to be repeated, the data re-analyzed or reinterpreted.

Another point to make is that estimation or simulation based on variograms cannot very well capture curvilinear features, e.g., curved channels (Figure 1.8). To overcome such limitations, recent development includes multiple point statistics (where correlation is characterized among multiple data points and then incorporated into simulations), pluralGaussian simulation (several correlated populations can be superimposed), and hybrid or hierarchical approaches (e.g., kriging is used to create property distribution within a higher-order geobody created via either deterministic or stochastic means, often object-based). These are currently areas of active research.

1.5 This Class

1.5.1 References

In this class, a fairly rigorous mathematical treatment is presented. This course is thus designed at the upper undergrad and graduate level, appropriate for the level of rigor contained herein. Course lecture is the key, though most materials are assembled based on several textbooks, tutorials, and lecture notes, each with its own emphasis:

- Engineering Geostatistics, Course Notes, Randal Barnes, 2000, Depart-

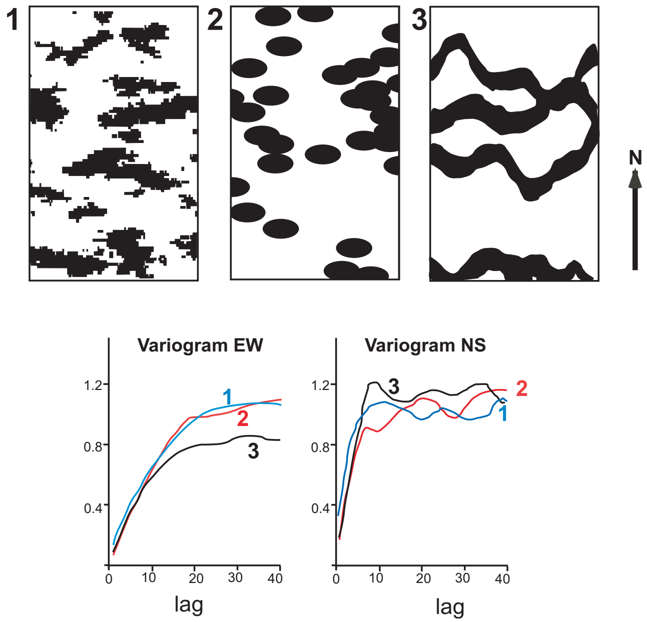


Figure 1.8: Experimental variograms computed for 3 different types of geological heterogeneity (from Caers and Zhang, 2002), illustrating the limitations of using variograms to describe heterogeneity with non-linear features.

ment of Civil Engineering, U of Minnesota.

- An Introduction to Applied Geostatistics, Isaaks & Srivastava (I&S), 1989, Oxford University Press.

- Geostatistical Reservoir Modeling, Clayton Deutsch, 2002, Oxford University Press.

- Fundamentals of Geostatistics in Five Lessons, Andre Journel, 1989, Short Course in Geology, vol 8, Presented at the 28^{th} International Geological Congress, Washington, D. C.

- Introduction to Geostatistics, Application in Hydrogeology, Peter Kitanidis, 1997, Cambridge University Press.

- GSLIB: Geostatistical Software Library and User's Guide, Clayton Deutsch & Andre Journel, 2nd Edition, Oxford University Press, 1997.

Bear in mind that the current course can only serve an introductory purpose: it cannot hope to cover every aspect of the subject as presented in the references, nor will we have time to explore many advanced topics, as they are being continuously developed and refined in the literature. In particular, the topics of this course are limited to *stationary* random space function (RSF) (stationarity here, roughly speaking, means that the mean, variance, and variogram do not change with position in the data field; if we have time, we'll cover Simple Kriging which is not based on assuming stationary RSF). Although there are geostatistical estimation methods developed for non-stationary RSF, the most widely used ones are based on stationary RSF. Wikipedia has a listing of the major kriging techniques used in practice:

http://en.wikipedia.org/wiki/Kriging

To further be exposed to the power of the geostatistical analysis, we might have a guest lecturer to give a talk about the reservoir simulation workflow.

Finally, all lectures are rooted in a fairly rigorous mathematical framework. Hopefully, such an approach will better prepare you for the more advanced topics or doing independent research. The exercises are designed to help you understand both the strength of the geostatistical methods and the various pitfalls you may encounter when working with raw data and the suggested solutions. The suggested reading list at the end of each chapter presents either example applications of geostatistics in different geoscience specialties or select topics specific to geostatistical reservoir simulation. They are not specific to the topics of each chapter, however, i.e., most papers are assuming you're already familiar with the fundamentals. Some of these papers come from the papers compiled in: Geostatistics for environmental and geotechnical applications, Shahrokh Rouhani et al. (editors), 1996, ASTM publication. Some come from excerpts of the textbook Geostatistical Reservoir Modeling by Deutsch (2002). I will also post additional papers on a ftp site that you can access.

1.5.2 Outline

The outline of this course is:

1. Probability Theory Review
2. Spatial Analysis
3. Experimental Variogram
4. Variogram Modeling
5. Geostatistical Estimation (Kriging & Co-Kriging)
6. Geostatistical Simulation (Unconditional & Conditional)
7. Advanced Topics

1.5.3 Homework, Grades, Tools

In addition to exercises and projects, reading assignments are given, which may be expected to be discussed during the class meeting if we have time. For some paper assignments, students are expected to produce a short (15 minute) powerpoint presentation on what has been learned from these papers. Those who do not show up in class or fail to participate in the exercises may expect F. Tools for simple exercises include ruler, calculator, Excel, Matlab. For more complex projects, we'll use software packages such as Surfer (kriging estimation) and Gslib (stochastic simulation).

1.6 Suggested Reading

Besides the above textbooks, other reading materials may come from:

1. Geostatistics for environmental and geotechnical applications: a technology transferred, M. V. Cromer, *in* Rouhani et al. (1996).
2. Describing spatial variability using geostatistical analysis, R. M. Srivastava, *in* Rouhani et al. (1996).

However, it is my belief that depending on the type of research you do, do literature search, and focus on papers that have similar aspects to your problems. This might help you not get lost in the sea of the ever expanding geostatistical literature!

Some final thoughts: This is a-graduate level class on a challenging subject. Instead of learning how to use some software, the course emphasizes fundamental and quantitative understanding. So, be prepared to think hard. Work out the exercises and projects yourself. Theoretical rigor is emphasized because I believe that if you do research related to spatial analysis, fundamental aspects are important. You simply cannot hope to understand many literature papers or produce quality results if you're not exposed to a systematic study of the fundamental principles lying behind software applications. However, if you're only interested in applied problems, you may feel that theories and accompanying derivations and programming (you're required to write a few Matlab codes) are too tedious and not of interest to you. For those with such a view, please consider taking an alternative class with a more applied emphases. In this course, a series of chapter projects are designed using Surfer which shows the typical steps involved in applications of kriging, though solving applied problems is not the focus. Make sure you sign up the class for the right reason.

Chapter 2

Probability Theory Review

We review some basic concepts and results of probability theory that are of relevance. It is meant to be a refresher of concepts covered in a separate statistics course. However, in case you have not taken such courses, the concepts and results will be explained using simple examples.

2.1 Nomenclature and Notation

1. Important nomenclature:
   - $Pr[A]$—the probability of event A occurring;
   - $Pr[\overline{A}]$—the probability of event A *not* occurring;
   - $Pr[A \cap B]$—the probability of event A *and* event B both occurring;
   - $Pr[A \cup B]$—the probability of event A *or* event B occurring;
   - $Pr[A|B]$—the probability of event A occurring given that event B has occurred.
2. Axioms:
   - $0 \leq Pr[A] \leq 1$;
   - $Pr[\Omega] = 1$, Ω is the union of all possible outcomes.
3. Conditional Probability:
   $Pr[A|B] = \frac{Pr[A \cap B]}{Pr[B]}$
4. Independence:
   Events A and B are statistically independent if and only if:

   $$Pr[A \cap B] = Pr[A] \cdot Pr[B]$$

   In words, if events are independent, the probability of their join occurrence is simply the product of their individual probabilities of occurrence.

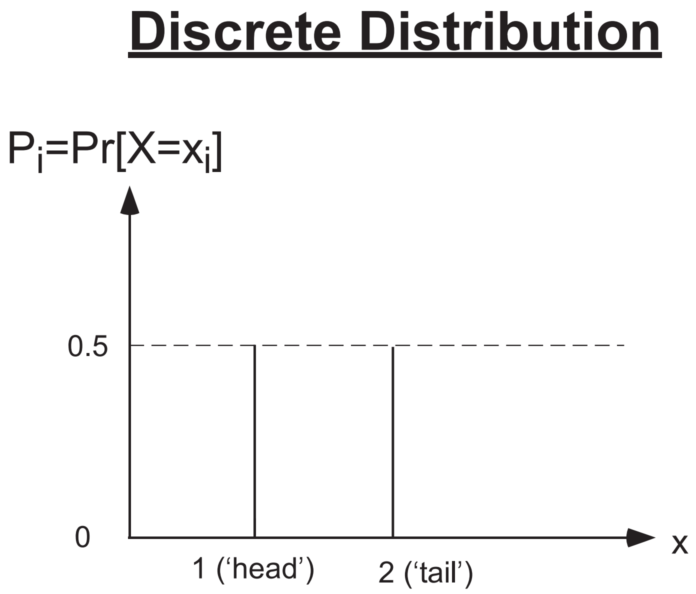


Figure 2.1: Outcomes of experiments of tossing a coin: discrete r.v. and its probability distribution.

2.2 Univariate Analysis

2.2.1 Introduction

One may define the *probability* of an event as between 0 and 1, representing the chance or relative frequency of occurrence of the event. The probabilities of all possible (mutually exclusive) events of an experiment must sum to 1. In practice, the outcomes of experiments are assigned numerical values, e.g., when tossing a coin, 1 and 2 can be assigned to the outcome of "head" and "tail", respectively (Figure 2.1). Such numerical values can be represented by a *Random variable* (r.v.). Two types of r.v. exist: discrete and continuous. Discrete examples include the outcome of tossing a coin (head or tail), the grades of this course (A, B, C, D, F); continuous examples include the height of all men in the U.S. (ranging from, say, 4 ft to 7 ft), the grades of a class (e.g., 0.0∼100.0 points). In this class, a r.v. is expressed with a upper-case letter, e.g., X. The numerical value of a particular outcome is designated as the lower-case equivalent "x".

The probability of a r.v. occurring at any possible value (discrete r.v.) or within a range of values (continuous r.v.) is described by its probability distribution. For discrete r.v., its distribution is also discrete (Figure 2.1):

$$P_i = Pr[X = x_i], \qquad i = 1, \ldots, n. \tag{2.1}$$

Table 2.1: Grade distribution of a class. An example of a discrete r.v.

Grades	Number of People	$X = x_i$
A	2	6
B	12	5
C	20	4
D	7	3
E	3	2
F	1	1

In this case, $P_1 = Pr[X = 1] = 0.5$, $P_2 = Pr[X = 2] = 0.5$, and $P_1 + P_2 = 1$. In the discrete r.v., its distribution is just the frequency (or proportion) of occurrence.

Exercise 1: *Calculate the probability distribution of the grade in a class, see Table 2.1. Steps: (1) assign each grade a numerical value (a discrete r.v.); (2) calculate the proportion; (3) plot the probability. You can either do it by hand with a calculator, or using Excel. Does the total probability sum up to 1? We often call such a diagram "histogram".*

Now, can you calculate the probability of *any* person with a grade less than or equal to A ($X <= 6$)? What is the probability for B ($X <= 5$), and so on? We can also plot this probability out. We see that it rises from near 0.0 ($X <= 1$) to 1 ($X <= 6$). Such a plot represents the (discrete) cumulative probability for a discrete r.v.

Similarly, histogram and the cumulative probability can be constructed for a continuous r.v. In this case, for a sample data set: $\{x_1, x_2, \ldots, x_n\}$ (outcomes of experiments or realizations of X), we use "bins" to do the trick. Each bin represents a range of values that X may fall in. Similar to what was done for the discrete r.v., we can also calculate the proportion within each bin and plot it out. For the sample set, a few other key statistics are also of interest: mean (μ), variance (σ^2) (standard deviation: σ), and coefficient of variation ($CV = \sigma/\mu$):

$$\mu = (1/n)\Sigma_{i=1}^{n} x_i \tag{2.2}$$

$$\sigma^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \mu)^2 \tag{2.3}$$

Why "n-1"? (Hint: it constitutes an unbiased estimator. However, due to time constraint, we will not go into the subjects on estimator, e.g., maximum likelihood versus unbiasedness.)

Exercise 2: *A sample data set of a continuous r.v.: the thickness (X; m) of an aquifer is measured along the horizontal distance (di; m) (Table 2.2). For the thickness, calculate the mean, variance and CV, calculate and plot the histogram and cumulative distribution.*

Table 2.2: Aquifer thickness along a 1D distance.

d_i	1	2	3	4	5	6	7	8	9	10	11
x_i	56	57	55	54	49	43	37	36	39	37	41
d_i	12	13	14	15	16	17	18	19	20	21	22
x_i	41	36	33	40	44	53	53	54	51	48	54
d_i	23	24	25	26	27	28	29	30	31	32	33
x_i	63	65	63	63	53	50	50	54	49	43	43
d_i	34	35	36	37	38	39					
x_i	47	47	50	53	61	61					

Can you imagine fitting a curve (a mathematical function) to the histogram and cumulative distribution? Loosely speaking [1], such functions will be, respectively, the probability density function (pdf: $f_X(x)$) and the cumulative distribution function (cdf: $F_X(x)$) (Figure 2.2).

2.2.2 Formal Definitions

For a continuous r.v., the cdf is defined as:

$$F_X(a) = Pr\left[X \leq a\right] \tag{2.4}$$

where a is a constant (non-random). For example, if $F_X(x) = 1 - e^{-x}, x \geq 0$, then $F_X(1) = 0.62$ means that the probability that X takes a value smaller than 1 is 0.62. Since $F_X(x)$ is a probability, it must be non-negative. For $F_X(x)$, the corresponding pdf is defined as:

$$\int_{-\infty}^{a} f_X(x)dx = Pr\left[X \leq a\right] \tag{2.5}$$

Clearly, $F_X(x)$ and $f_X(x)$ are related:

$$\int_{-\infty}^{a} f_X(x)dx = F_X(a) = Pr\left[X \leq a\right] \tag{2.6}$$

and[2]

$$\frac{dF_X(a)}{dx} = f_X(a) \tag{2.7}$$

The pdf of a r.v. must satisfy:

$$f_X(x) \geq 0 \qquad \int_{-\infty}^{\infty} f_X(x)dx = 1 \qquad Pr[a \leq x \leq b] = \int_{a}^{b} f(x)dx \tag{2.8}$$

[1] Strictly, pdf and cdf apply to the underlying population of X, not a function fitted to the sample histogram and cumulative distribution which can be biased for a finite sample set.

[2] $\frac{dF_X(a)}{dx} = \frac{d}{dx}\int_{-\infty}^{a} f_X(x)dx = f_X(a)$; Note this relation only holds when $F_X(x)$ is differentiable with respect to x (i.e., it's slope exist). When $F_X(x)$ is non-differentiable, the pdf function $f_X(x)$ does not exist.

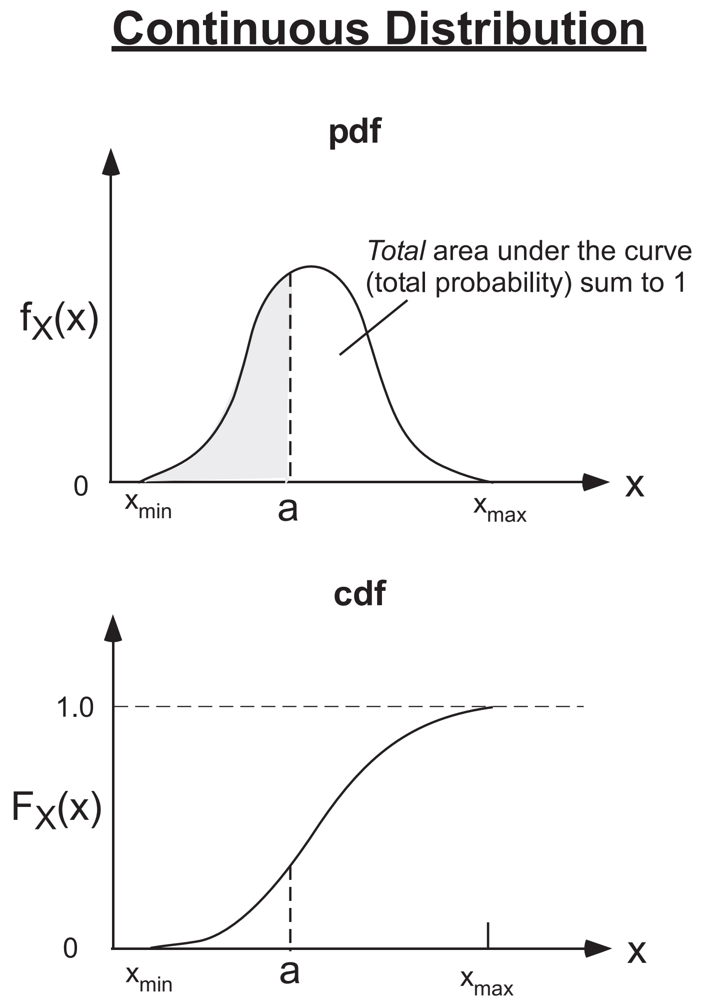


Figure 2.2: pdf and cdf functions (curves) of a continuous r.v.

The mean is defined as:

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx \tag{2.9}$$

Here, we introduce the expectation operator: $E[\cdot]$. Equation 2.9 can be generalized to define the expected value of a function: g(X) (some known, non-random function of the random variable X):

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx \tag{2.10}$$

The variance (a special case of $g(X)$) is defined as:

$$\sigma^2 = Var[X] = E[(X - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) dx \tag{2.11}$$

Notice that we write the formula for variance directly following equation (2.10), the variance is written out within the expectation as a function of X. Thus, though we introduce a variance operator, it can alternatively be expressed by the expectation operator. Clearly, the variance is just the mean squared deviation of X from its expected value (μ).

An extremely useful relationship between mean and variance is (proof given in class):

$$\sigma^2 = E[X^2] - \mu^2 \tag{2.12}$$

Note that proofs are not written in the course notes. This is to encourage class attendance, besides, the proofs tend to be long and tedious and just require too much typing!

2.2.3 Random Variable Arithmetic

From equation 2.10, we can obtain some useful properties of the expectation operator (proofs given in class): a, b are constants.

$$E[aX + b] = aE[X] + b = a\mu + b \tag{2.13}$$

$$Var[aX + b] = a^2 Var[X] = a^2\sigma^2 \tag{2.14}$$

The above relations always hold regardless of the type of distribution function for X. They are very useful when trying to understand the properties of a scaled random variable, e.g., what are the $Std[aX + b]$ and $CV[aX]$ $(a > 0)$?

2.3 Bivariate Analysis

2.3.1 Introduction

In the previous section, we look at the statistical measures of a single r.v. However, correlation can often exist between two r.v. For example, the height and

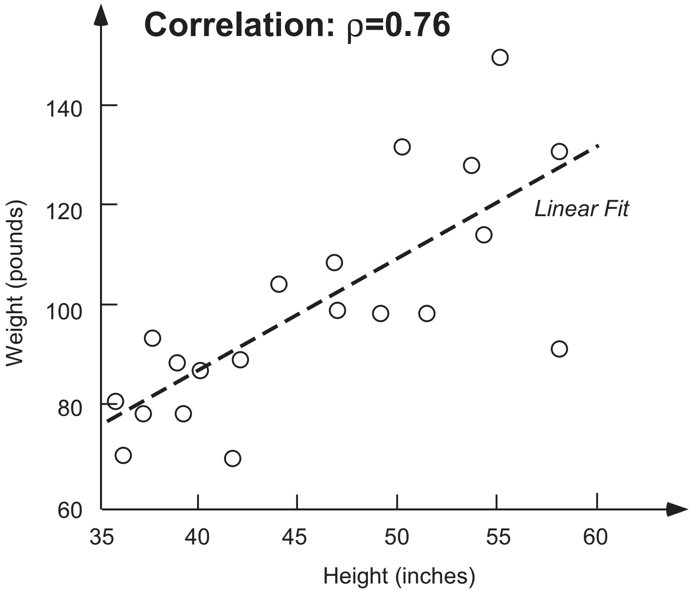


Figure 2.3: An example of positive correlation: height and weight scatter plot for a junior high school class.

weight of people are often correlated—a set of sample data is shown for students in a junior high school (Figure 2.3). In this case, the weight increases with increasing height for which we say a positive correlation exists between the two variables. To investigate correlation, a scatter plot is often used, e.g., for each person, the height and weight is cross-plotted. Often, some sort of fit is attempted. Here, we see a linear function fitted to the scatter plot. However, to quantitatively evaluate correlation, a correlation coefficient (r_{XY}) is often used:

$$\rho_{XY} = \frac{1}{n-1}\Sigma_{i=1}^{n}\left(\frac{x_i - \mu_X}{\sigma_X}\right)\left(\frac{y_i - \mu_Y}{\sigma_Y}\right) \tag{2.15}$$

As defined previously, μ_X (or μ_Y) is the mean of X (or Y) in its univariate distribution. ρ_{XY} varies between -1 (perfect negative correlation: Y=-X) to 1 (perfect positive correlation: Y=X). When $r_{XY} = 0$, we say the two variables are not correlated. In this example, $r_{XY} = 0.76$, thus there is a certain amount of positive correlation between weight and height.

The correlation between two r.v. is the cornerstone of geostatistics: one r.v. is a geological/hydrological/petrophscial property at one spatial location, the second r.v. can be the (1) same property at a different location (auto-correlation studies; kriging); or, (2) a different property at a different location (cross-correlation studies, co-kriging). To develop the fundamental geostatistical equations, a formal definition is needed and introduced next.

2.3.2 Bivariate Random Variables

Given two r.v. X and Y, the bivariate distribution function is defined as $F_{XY}(x,y) = Pr[(X \leq x) \cap (Y \leq y)]$ $(0 \leq F_{XY}(x,y) \leq 1)$. The marginal distribution of X and Y is related to the bivariate function: $F_X(x) = F_{XY}(x, +\infty)$, $F_Y(y) = F_{XY}(+\infty, y)$.[3] X and Y are statistically independent when: $F_{XY}(x,y) = F_X(x)F_Y(y)$.

The bivariate density function is defined as: $f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$ when it exists. Similar to the univariate density function, we have:

$$Pr[(a \leq X \leq b) \cap (c \leq Y \leq d)] = \int_{x=a}^{b} \int_{y=c}^{d} f_{XY}(x,y)dxdy \tag{2.16}$$

And, we have total probability sum to 1:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1 \tag{2.17}$$

As with marginal distribution function, by definition, each r.v. also has associated marginal density function:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy$$
$$f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx$$

Similar to the distribution function, X and Y are statistically independent when: $f_{XY}(x,y) = f_X(x)f_Y(y)$.

Similar to the univariate analysis, the expectation of a function of two random variables, $g(X,Y)$, is given:

$$E[g(X,Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{XY}(x,y)dxdy$$

The covariance between X and Y (σ_{XY}) measures how well the two variables track each other: when one goes up, how does the other go on average? By definition, the covariance between X and Y is given as:

$$\sigma_{XY} = Cov[X,Y] = E[(X - \mu_X)(Y - \mu_Y)] =$$
$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y) f_{XY}(x,y)dxdy = E[XY] - \mu_X \mu_Y$$

Clearly, compared to the expectation of $g(X,Y)$, the covariance is a special case: $g(X,Y) = (X - \mu_X)(Y - \mu_Y)$. The unit of covariance is the product of

[3] The marginal distribution functions are nothing more than the distribution function defined in the previous section for each univariate r.v.; the term "marginal" is added only when working with 2 or more r.v., so as not to confuse the univariate (or marginal) distribution function with the bivariate or multi-variate distribution function.

the unit of r.v. X and unit of r.v. Y. The covariance between a r.v. X with itself is equal to its variance:

$$Cov[X,X] = E[(X-\mu_X)(X-\mu_X)] = E[(X-\mu_X)^2] = \sigma_X^2 = Var[X]$$

The correlation (or correlation coefficient) between X and Y is a dimensionless, normalized version of σ_{XY}:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad (-1 \leq \rho_{XY} \leq 1)$$

Note that this corresponds to the correlation defined in the introduction by equation(2.15). Actually, equation(2.15) is a (discrete) *estimator* of the above relation which is defined in terms of expectation and continuous function, i.e., the joint pdf of X and Y: $f_{XY}(x,y)$. By comparing these two equations, we can see that an estimator of the covariance can be defined as:

$$\sigma_{XY} = \frac{1}{n-1}\Sigma_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y)$$

If X and Y are independent, then they are uncorrelated and their covariance σ_{XY} (thus ρ_{XY}) is zero (proof given in class). However, zero covariance does *not* imply statistical independence (think of $Y = |X|$, it can be calculated that $\sigma_{XY} = 0$, but X and Y are obviously not independent with each other). The covariance is best thought of as a measure of *linear* dependence.

2.3.3 Bivariate Arithmetics

Let a, b, c be known constants (they are not random). For the univariate random variable, X, we have:

$$\begin{aligned} \mu_X &= E[X] \\ \sigma_X^2 = Var[X] &= E[(X-\mu_X)^2] \end{aligned}$$

For a second univariate random variable, Y, we have:

$$\begin{aligned} \mu_Y &= E[Y] \\ \sigma_Y^2 = Var[Y] &= E[(Y-\mu_Y)^2] \end{aligned}$$

When there exists correlation between X and Y, we have (by definition):

$$\sigma_{XY} = Cov[X,Y] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_X)(y-\mu_Y)f_{XY}(x,y)dxdy$$

Note $f_{XY}(x,y)$ is the joint pdf function.

The following rules always hold regardless of the underlying distribution functions for X and Y (proofs given in class):

$$\begin{aligned} E[aX + bY + c] &= a\mu_X + b\mu_Y + c \\ Var[aX + bY + c] &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY} \\ E[aXbY] &= abE[XY] \\ Var[aXbY] &= (ab)^2 Var[XY] \\ Cov[aX, bY] &= abCov[X, Y] \end{aligned}$$

Special cases include:

$$\begin{aligned} E[X + Y] &= \mu_X + \mu_Y \\ E[X - Y] &= \mu_X - \mu_Y \\ Var[X + Y] &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \\ Var[X - Y] &= \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY} \end{aligned}$$

Clearly, if X, Y are independent, $\sigma_{XY} = 0$, thus $Var[X+Y] = Var[X-Y] = \sigma_X^2 + \sigma_Y^2$, and $E[XY] = \mu_X\mu_Y$ (try proving this yourself).

2.4 Multivariate Analysis

2.4.1 Linear Combination of Many r.v.

Extending the bivariate arithmetics into multivariate analysis, we can get another host of relationships. Let X_i (i=1,N) be N random variables with unspecified distribution. Let a_i (i=1,N) be N known constants. The N-multivariate joint pdf of X_i is $f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N)$. If X_i are mutually independent, $f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \ldots f_{X_N}(x_N)$. Particularly relevant definitions and properties are listed below (note the similarity when comparing to the previous univariate and bivariate definitions):

$$\begin{aligned} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N) dx_1 dx_2 \ldots dx_N = 1 \\ E[g(X_1, X_2, \ldots, X_N)] = \\ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(x_1, x_2, \ldots, x_N) f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N) dx_1 dx_2 \ldots dx_N \\ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N) dx_2 dx_3 \ldots dx_N = f_{X_1}(x_1) \end{aligned}$$

where $f_{X_1}(x_1)$ is the marginal pdf of X_1. Note that the first two formulas integrate N times; the last formula integrate N-1 times.

In this class, we'll only use what's applicable to the study of geostatistics for which we study linear combination of many r.v. [4]

2.4.2 Multivariate Arithmetics

The following rules always hold (proofs given in class):

$$E[\sum_{i=1}^{N} a_i X_i] = \sum_{i=1}^{N} a_i E[X_i]$$
$$Var[\sum_{i=1}^{N} a_i X_i] = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j Cov[X_i, X_j]$$

Remember, $Cov[X_i, X_i] = Var[X_i]$. Thus, if X_i are mutually uncorrelated, we have (proof given in class):

$$Var[\sum_{i=1}^{N} a_i X_i] = \sum_{i=1}^{N} a_i^2 Var[X_i]$$

Exercise 3: *Sums, Products, and Normalization. Consider a set of N independent, identically distributed (i.i.d.) random variables* $\{X_1, X_2, \ldots, X_N\}$, *with mean and variance:* $E[X_i] = \mu_X$, $Var[X_i] = \sigma_X^2$. *Defined 3 derived r.v.: (1)* $P_i = N \cdot X_i$, $\forall i = 1, 2, \ldots, N$, *what is the mean and variance of* P_i*? (2)* $S = \sum_{i=1}^{N} X_i$, *what is the mean and variance of* S*? (3)* $Z_i = \frac{X_i - \mu_X}{\sigma_X}$, $\forall i = 1, 2, \ldots, N$, *what is the mean and variance of* Z_i*?*

2.5 Gaussian Distribution

The importance of the Gaussian distribution (or normal distribution) as a model of quantitative phenomena in the natural sciences is due to the central limit theorem (next). Many physical phenomena (like photon counts and noise) can be approximated well by the normal distribution. While the mechanisms underlying these phenomena are often unknown, the use of the normal model can be theoretically justified by assuming that many small, independent effects are additively contributing to each observation.

If the density function (pdf) of a random variable X is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < +\infty$$

[4] Not coincidentally, Kriging system is developed based on the linear combination of many random variables. We'll get back to these important relationships once we reach the point of deriving the kriging equations.

X is said to have a Gaussian distribution: X$\sim N(\mu, \sigma)$.

The Gaussian distribution has a well-known "bell" shape, e.g., Figure 2.4. It is symmetric around the mean. Can you approximately draw the shape of the pdf for Y, if $Y = (X - \mu_X)/\sigma_X$ (in the previous exercise, we determined that $Y \sim N(0, 1)$)? What is the pdf for Y? ($f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}; -\infty < y < +\infty$) Y is called a *standard* Gaussian random variable.

The distribution function (cdf), recall equation(2.6), is by definition: $F_X(a) = \int_{-\infty}^{a} f_X(x)dx = Pr\left[X \leq a\right]$. For Gaussian distribution, however, $F_X(x)$ cannot be expressed by elementary functions, as the above integral cannot be analytically integrated for the Gaussian pdf $f_X(x)$. Often, a "Normal Table" is given for the standard Gaussian r.v. $Y = (x - \mu_X)/\sigma_X$.

Though geostatistical analysis is *not* based on the assumption of data normality, some estimation and simulation tools work better if the distribution is close to normal. It is thus of interest to determine if one's data is close to normal. A *normal probability plot* is a type of cumulative frequency plot that helps decide this question (normal probability paper can be purchased in most engineering supply stores). On a normal probability plot, the y-axis (cumulative frequency) is scaled in such a way that the cumulative frequency will plot as a straight line if the distribution is Gaussian.

Many variables in earth sciences have distributions that are not close to normal. It is common to have many small values and a few large ones. Though the normal distribution is inappropriate to model such asymmetric distribution, a closely related distribution, the *lognormal distribution function*, can sometimes be a good alternative. A random variable (Y) satisfy a lognormal distribution if its log-transform (e.g., $Z = \ln Y$, or $Z = log_{10}Y$) satisfies a normal distribution. A normality test is thus performed on the log-transformed variable Z.

2.6 Central Limit Theorem

The Central Limit Theorem is often stated as:

> Let $\{X_1, X_2, \ldots, X_N\}$ be N random variables, each with a finite variance. Let Y be a r.v. defined as $Y = \sum_{i=1}^{N} X_i$, then, under a set of general conditions, $\lim_{N\to\infty} Y \Rightarrow$ Normal Distribution.

The practical importance of the theorem is that may physical phenomena in nature arise from a number of additive variations. The distributions of the individual variations are often unknown, however, the histogram of the summed variable is often observed to be approximately Normal.

A more restrictive, but more useful version of the theorem is stated as:

> Let $\{X_1, X_2, \ldots, X_N\}$ be N independent, identically distributed random variables, each with a finite variance. Let Y be a r.v. defined as $Y = \frac{1}{N}\sum_{i=1}^{N} X_i$ (the arithmetic mean of X_i), then, under a set of general conditions, $\lim_{N\to\infty} Y \Rightarrow$ Normal Distribution.
>
> Y$\sim N(\mu_Y, \sigma_Y)$, $\mu_Y = \mu_X$, and $\sigma_Y = \sigma_X/\sqrt{N}$.

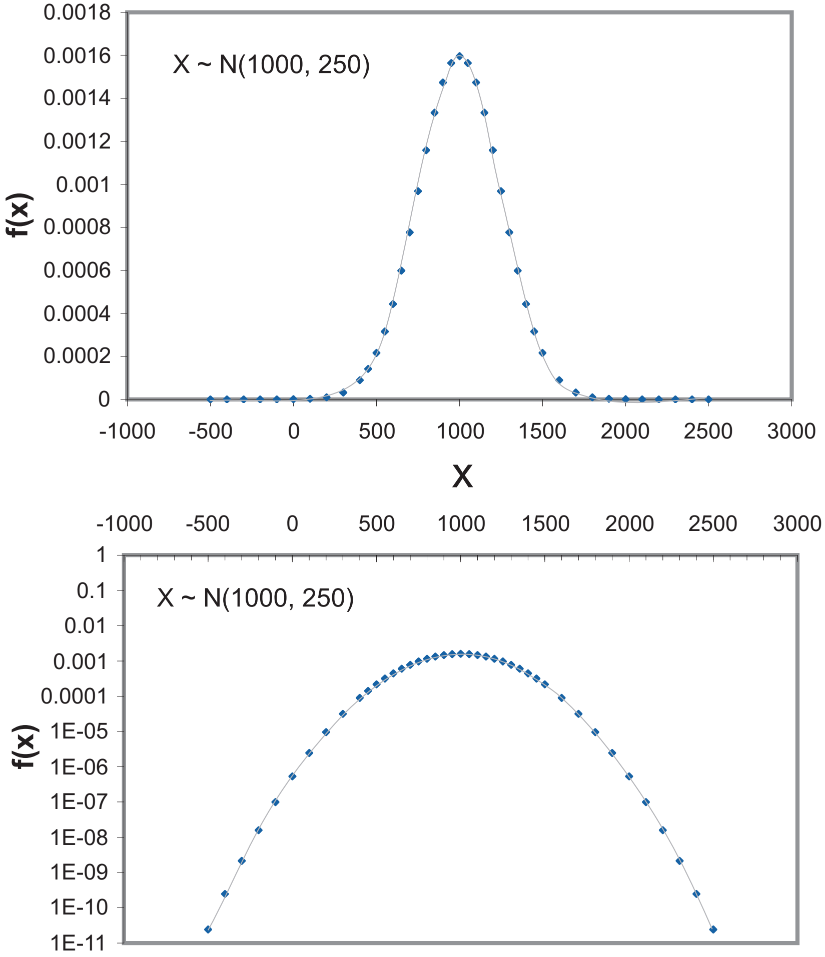


Figure 2.4: An example of Gaussian distribution: X∼ $N(1000, 250)$. In the bottom plot, f(x) is in log scale.

81	77	103	112	123	19	40	111	114	120
82	61	110	121	119	77	52	111	117	124
82	74	97	105	112	91	73	115	118	129
88	70	103	111	122	64	84	105	113	123
89	88	94	110	116	108	73	107	118	127
77	82	86	101	109	113	79	102	120	121
74	80	85	90	97	101	96	72	128	130
75	80	83	87	94	99	95	48	139	145
77	84	74	108	121	143	91	52	136	144
87	100	47	111	124	109	0	98	134	144

Figure 2.5: Location map and values of 100 V measurements.

2.7 Chapter Project

In the Walker Lake area in Nevada, the concentration (V; in ppm) of a arsenious contaminant in soil has been measured on a 10 × 10 m^2 grid (Figure 2.5). At each measurement point, a second contaminant, PCE (U; in ppm) has also been measured (Figure 2.6). All values are rounded off to integers—you can report your results in integer (data source: Isaaks & Srivastava (1989)). In this project, we will investigate both the univariate statistics of V and U, and their bivariate correlation.

In the univariate analysis, tasks include:

(1) Calculate the mean, variance, and standard deviation of V.

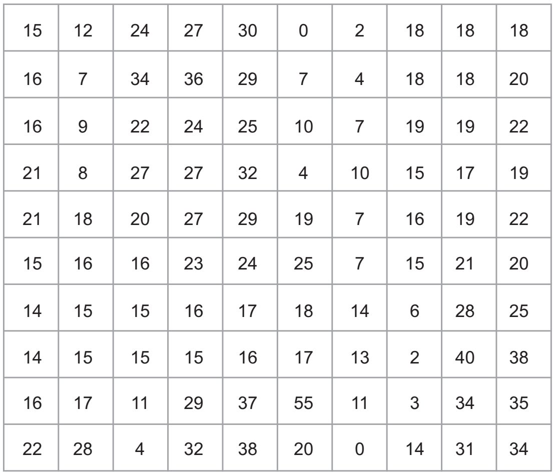
(2) Plot the histograms of V. Hint: By inspecting the V values (Figure 2.5), we find that it ranges from 0 to $\sim$ 150. We can construct a frequency table to count the number of V occurring in these intervals: $0 \le V < 10$, $10 \le V < 20$, ..., $130 \le V < 140$, $140 \le V < \infty$. Each "number" can also be converted to a frequency (%) by dividing it with the total data count (100). You can do it by hand, or use Excel-Tool-Data Analysis-Histogram, and set up a "bin" to do it.

(3) Plot the cumulative frequency of V. Hint: Based on the frequency table of V, calculate the frequency of V (number/data count) that occurs in intervals: $V < 10$, $V < 20$, ..., $V < 140$, $V < \infty$.

(4) Plot the cumulative frequency of V on a normal probability paper. Determine if V falls close to a normal distribution.

(5) Calculate the mean, variance, standard deviation of U. Construct its histogram and cumulative frequency U. Plot the cumulative frequency of U on a normal probability paper.

The above univariate analysis can be used to describe the distribution of



15	12	24	27	30	0	2	18	18	18
16	7	34	36	29	7	4	18	18	20
16	9	22	24	25	10	7	19	19	22
21	8	27	27	32	4	10	15	17	19
21	18	20	27	29	19	7	16	19	22
15	16	16	23	24	25	7	15	21	20
14	15	15	16	17	18	14	6	28	25
14	15	15	15	16	17	13	2	40	38
16	17	11	29	37	55	11	3	34	35
22	28	4	32	38	20	0	14	31	34

Figure 2.6: Location map and values of 100 U measurements.

an individual random variable. However, a very limited view is obtained if we analyze a multivariate data set one variable at a time. Some of the most interesting feature of earth science data are the relationship between variables. To explore the relation between U and V, a bivariate analysis is necessary. A scatter plot is the most useful mean of detecting both data correlation and potential errors.

(6) Construct a scatter plot of U and V and calculate the correlation coefficient. Do you see any correlation? Is it positive (larger U corresponds to larger V) or negative (larger U corresponds to smaller V)? Calculate the correlation coefficient: ρ_{UV}

(7) Change the last value of V to 14 (by accident—commonly happening during data input). Now look at the scatter plot. Calculate the correlation coefficient.

Linear Regression: Using Excel-Chart-Add Trendline, a best-fit line function can be plotted to the scatter plot of (6) (Figure 2.7). Excel also gives the value of R^2—in effect, a square of ρ_{UV} (is it the same as your ρ_{UV} calculated by hand?). However, for small V, we note that this best-fit line extends into negative U values which is not physically reasonable. Thus, the regression line can not be used for prediction blindly. In this case, what would the value of U be given V of around 5 ppm? Commonly, it is appropriate to set U (at V=5 ppm) either 0, or, consider using other forms of non-linear regression. For more discussions on linear and non-linear regression (e.g., conditional expectation), see p. 33-39, Isaaks & Srivastava (1989).

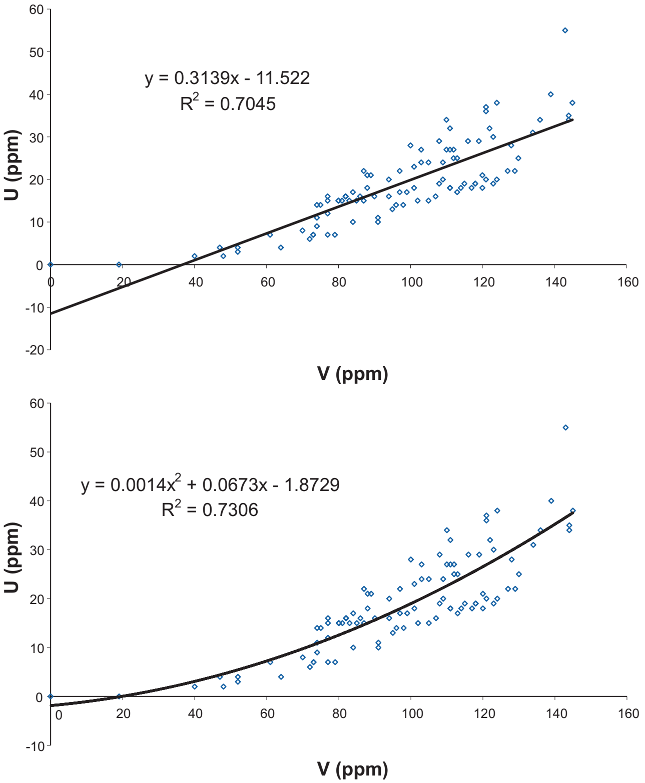


Figure 2.7: Scatter plot of U and V fitted with linear and nonlinear functions. Note that the non-linear function seems a slightly better fit. However, this particular 2$^{\text{nd}}$ order polynomial does not guarantee positive U at small V values.

2.8 Suggested Reading

1. Geostatistical estimation: kriging, S. Rouhani, *in* Rouhani et al. (1996).

2. Modeling spatial variability using geostatistical simulation, A. J. Desbarats, *in* Rouhani et al. (1996).

   I will also post additional papers on the class ftp site. Stay tuned.