Statistical Tools and Concepts

2

Abstract

Mineral resource estimation requires extensive use of statistics. In our context, *statistics* are mathematical methods for collecting, organizing, and interpreting data, as well as drawing conclusions and making reasonable decisions based on such analysis. This chapter presents essential concepts and tools required throughout the book.

2.1 Basic Concepts

A conventional presentation of statistics includes the notion of a *population* that is the virtually infinite collection of values that make up the mineral deposit. A *sample* is a representative subset selected from the population. A good sample must reflect the essential features of the population from which it is drawn. A r*andom sample* is a sample where each member of a population had an equal chance of being included in the sample. The *sample space* is the set of all possible outcomes of a chance experiment, for example a drilling campaign. The *event of a sample space* is a group of outcomes of the sample space whose members have some common characteristic. *Statistically independent events* are such that the occurrence of one event does not depend on the occurrence of other events. Sampling mineral deposits rarely fits nicely in the framework of representative samples from a statistical population; nevertheless, many concepts and tools from conventional statistics are used routinely.

Inductive statistics or *statistical inference* is attempted if the sample is considered representative. In this case, conclusions about the population can often be inferred. Since such inference cannot be absolutely certain, the language of probability is used for stating conclusions. *Descriptive statistics* is a phase of statistics that describes or analyses a given sample without inference about the population. Although our goal in mineral resource estimation is almost always inference, we use many descriptive statistics for viewing, understanding, and evaluating data.

An essential concept in statistics is *stationarity*, that is, our choice of data to pool together for common analysis. Chapter 6 describes stationarity more formally, but the concept is that data must be grouped together before any statistical calculations are attempted. Ideally, a decision of how to group the data can be made on the basis of clear geological controls, as discussed in Chap. 4. Some of the statistical tools presented in this chapter are useful to help make a choice of stationarity, but most assume that the decision has already been made and the data have been assembled into reasonable groups.

In most cases we consider continuous variables that are mass or volume fractions, that can take any value between a minimum (0%) and maximum (100%). We sometimes consider categorical or discrete variables that can take specific values from a closed set. A typical categorical variable would be lithology or mineralization type.

Statistical tools are used for several reasons, including (1) an improved understanding of the data and the mineral deposit, (2) to ensure data quality, (3) to condense information, and (4) to make inferences and predictions. In general, we are not interested in the statistics of the samples. Our goal is to go beyond the limited sample to predict the underlying population. Additionally, creative visualization of data is an important component of mineral resource estimation, partly because of its usefulness as a tool to understand data, but also to help validate spatially distributed models.

There are many good references on basic statistics. One accessible reference is Lapin (1983). This book uses a few notation conventions. Lowercase letters (x, y, z,…) denote actual values such as a measured value or a specified threshold. Uppercase letters (X, Y, Z,…) denote a random variable (RV) that is unknown. We characterize the uncertainty in a random variable with a probability distribution. A random

M. E. Rossi, C. V. Deutsch, *Mineral Resource Estimation,* DOI 10.1007/978-1-4020-5717-5_2,

variable could be the grade at an unsampled location denoted $Z(\mathbf{u})$ where $\mathbf{u}$ represents a location coordinates vector. A set of random variables is called a random function (RF). The set of grades over a stationary geologic population $\mathcal{A}$ is a random function $\{Z(\mathbf{u}), \mathbf{u} \in \mathcal{A}\}$.

2.2 Probability Distributions

Probabilities are closely associated to proportions. A probability of 0.8 or 80% assigned to an event means that the proportion of times it will occur, in similar circumstances, is 0.8 or 8/10 or 80%. The *similar circumstances* relates to our decision of stationarity. In some cases we calculate the probabilities directly through proportions. For example, the probability for a mineral grade within a particular geologic unit to be less than a particular threshold could be calculated by counting the number of samples below the threshold and dividing by the total number of data.

There are many cases, however, when probabilities cannot be calculated from proportions. This is particularly true for conditional probabilities, that is, probability values given certain a set of data events. Consider the probability that a mineral grade be less than a particular threshold given one measurement 50 m away that is twice the threshold and another measurement 75 m away that is just below the threshold. In such cases, we do not have multiple replications to calculate an experimental proportion. We must rely on probabilistic models and well established probability laws.

Probability distributions are characterized as parametric or non-parametric. A parametric distribution model has a closed analytical expression for the probability, and is completely determined by a finite number of parameters, as for example the Gaussian distribution model with parameters mean (m) and standard deviation (s) that control the center and spread of the distribution, respectively.

It is common to consider probability distributions that relate to one continuous or categorical variable at a time. Such distributions are called univariate distributions. Two examples: (1) the probability for a continuous variable to be less than a particular threshold, or (2) the probability for a particular lithology to prevail at a certain location. When we consider probability distributions of more than one variable at a time, then we call them multivariate distributions. The distribution of two variables is a bivariate distribution. For example, the probability of one grade being less than one threshold and a second grade being less than another threshold is a bivariate probability.

There are a large number of references for probability and basic statistics. Some general statistical ones and also some related to spatial data include Borradaile (2003); Davis (1986); Koch and Link (1986); Ripley (1987); and Rohatgi and Ehsanes Saleh (2000).

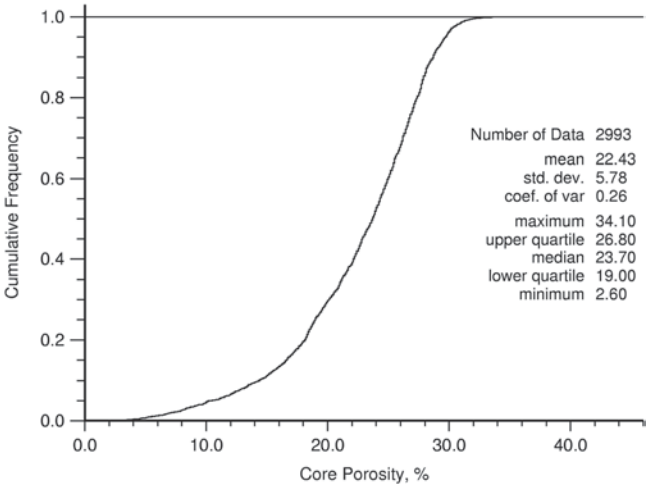


Fig. 2.1 Cumulative distribution of 2,993 data values. The cumulative frequency or probability is the probability to be less than the threshold value

2.2.1 Univariate Distributions

The cumulative distribution function (CDF) is the universal way to express a state of incomplete knowledge for a continuous variable. Consider an RV denoted by Z. The CDF $F(z)$ is defined as:

$$F(z) = Prob\{Z \leq z\} \quad \in [0,1]$$

The lowercase z denotes a threshold. $Prob\{\cdot\}$ denotes a probability or proportion. An example CDF is shown on Fig. 2.1; the z-variable is between 2 and 35 and is most probably between 20 and 30.

A cumulative histogram is an experimental CDF based on the data. It is useful to see all of the data values on one plot and sometimes can be used to isolate statistical populations. Cumulative histograms do not depend on a bin width, and can be created at the resolution of the data.

An important challenge is to determine how representative each sample is of the actual mineralization. This issue is discussed in more detail in Chap. 5. It is also important to determine whether the distribution of all samples adequately represents the actual grade distribution in the deposit, or whether certain weighting should be applied.

The interval probability of Z occurring in an interval from a to b (where $b > a$) is the difference in the CDF values evaluated at values b and a:

$$Prob\{Z \in [a,b]\} = F(b) - F(a)$$

The probability density function (PDF) is the derivative of the CDF, if it is differentiable. Applying the fundamental theorem of calculus, the CDF can be obtained by integrating the PDF:

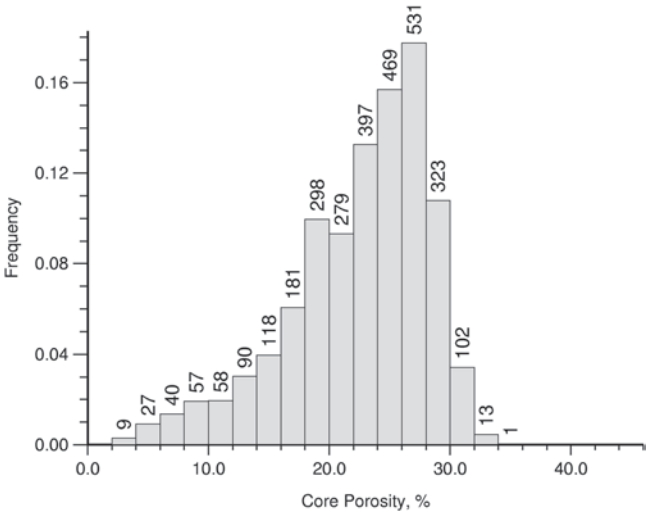


Fig. 2.2 Histogram of 2,993 data values. The common representation of the histogram has constant bin widths; the number of data in each bin is labeled on this histogram

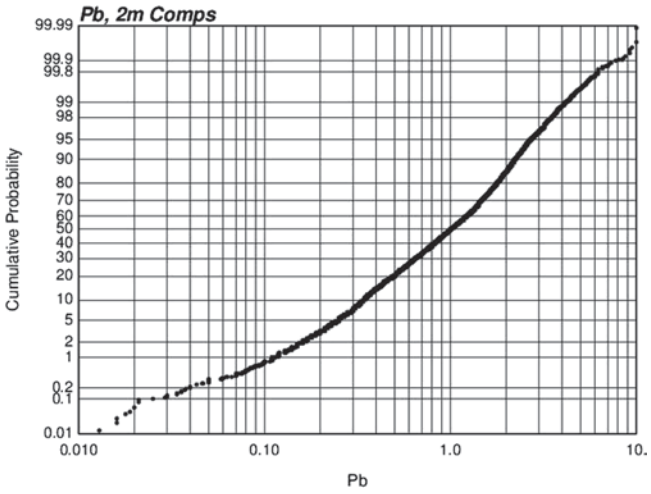


Fig. 2.3 An example of a probability plot. Data is lead concentration, on 2 m composites, on a logarithmic scale

$$f(z) = F'(z) = \lim_{dz \to 0} \frac{F(z+dz) - F(z)}{dz}$$

$$F(z) = \int_{-\infty}^{z} f(z)dz$$

The most basic statistical tool used in the analysis of data is the histogram, see Fig. 2.2. Three decisions must be made: (1) arithmetic or logarithmic scaling—arithmetic is appropriate because grades average arithmetically, but logarithmic scaling more clearly reveals features of highly skewed data distributions; (2) the range of data values to show—the minimum is often zero and the maximum is near the maximum in the data; and (3) the number of bins to show on the histogram, which depends on the number of data. The number of bins must be reduced with sparse data and it can be increased when there are more data. The important tradeoff is reduced noise (less bins) while better showing features (more bins).

The mean or average value is sensitive to extreme values (or outliers), while the median is sensitive to gaps or missing data in the middle of a distribution. The distribution can be located and characterized by selected quantiles. The spread is measured by the variance or standard deviation. The coefficient of variation (CV) is the standard deviation divided by the mean; it is a standardized, unit-less measure of variability, and can be used to compare very different types of distributions. When the CV is high, say greater than 2.5, the distribution must be combining high and low values together and most professionals would investigate whether the pool of data could be subset based on some clear geological criteria.

Sample histograms tend to be erratic with few data. Sawtooth-like fluctuations are usually not representative of the underlying population and they disappear as the sample size increases. There are techniques available for smoothing the distribution, which not only removes such fluctuations, but also allows increasing the class resolution and extending the distributions beyond the sample minimum and maximum values. Smoothing is only a consideration when the original set of data is small, and artifacts in the histogram have been observed or are suspected. In practice, sufficient data are pooled to permit reliable histogram determination from the available data.

The graph of a CDF is also called a probability plot. This is a plot of the cumulative probability (on the Y axis) to be less than the data value (on the X axis). A cumulative probability plot is useful because all of the data values are shown on one plot. A common application of this plot is to look at changes in slope and interpret them as different statistical populations. This interpretation should be supported by the physics or geology of the variable being observed. It is common on a probability plot to distort the probability axis such that the CDF of normally distributed data would fall on a straight line. The extreme probabilities are exaggerated.

Probability plots can also be used to check distribution models: (1) a straight line on arithmetic scale suggests a normal distribution, and (2) a straight line on logarithmic scale suggests a lognormal distribution. The practical importance of this depends on whether the predictive methods to be applied are parametric (Fig. 2.3).

There are two common univariate distributions that are discussed in greater detail: the normal or Gaussian and the lognormal distributions. The normal distribution was first introduced by de Moivre in an article in 1733 (reprinted in the second edition of his *The Doctrine of Chances*, 1738) in the context of approximating certain binomial distributions for large n. His result was extended by Laplace in his book

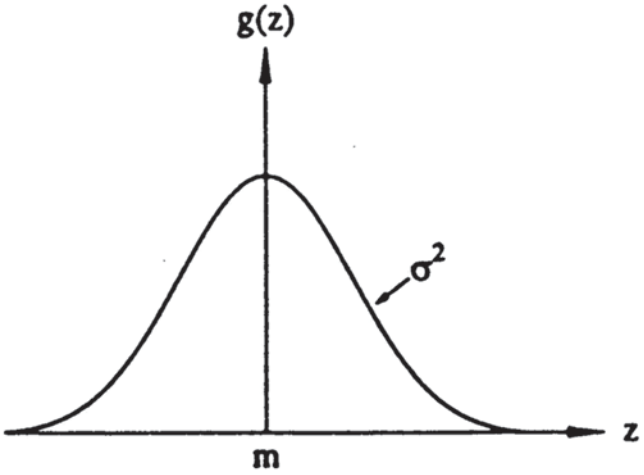


Fig. 2.4 A sketch of a normal or Gaussian distribution

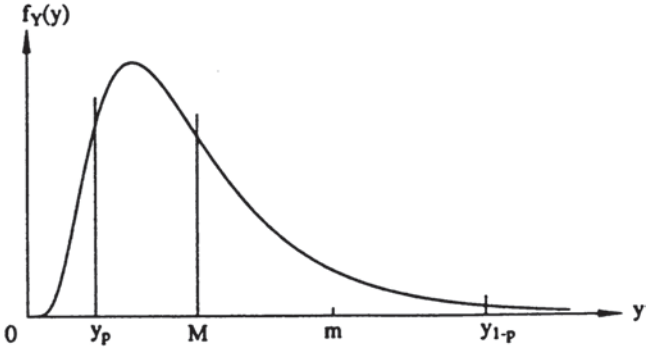


Fig. 2.5 A sketch of a lognormal distribution

Analytical Theory of Probabilities (1812), and is now called the Theorem of de Moivre-Laplace. Laplace used the normal distribution in the analysis of errors of experiments. The important method of least squares optimization was introduced by Legendre in 1806. Gauss, who claimed to have used the method since 1794, justified it rigorously in 1809 by assuming a normal distribution of the errors.

The Gaussian distribution is fully characterized by its two parameters, the mean and the variance. The standard normal PDF has a mean of zero and a standard deviation of one. The CDF of the Gaussian distribution has no closed form analytical expression, but the standard normal CDF is well tabulated in literature. The Gaussian distribution has a characteristic symmetric bell shaped curve about its mean; thus the mean and median are the same, see Fig. 2.4.

The lognormal distribution is important because of its history in spatial statistics and geostatistics. Many earth science variables are non-negative and positively skewed. The lognormal distribution is a simple distribution that can be used to model non-negative variables with positive skewness. A positive random variable is said to be lognormally distributed if $X = ln(Y)$ is normally distributed (Fig. 2.5). There are many grade distributions that are approximately lognormal. These distributions are also characterized by two parameters, a mean and a variance, although three-parameter lognormal distributions have been used in mining, see for example Sichel (1952). Lognormal distributions can be characterized by either their arithmetic or their logarithmic parameters.

The Central Limit theorem (see for example Lapin 1983) states that *the sum of a great number of independent equally distributed (not necessarily Gaussian) standardized random variables (RV) tends to be normally distributed, i.e. if n RV's Z_i have the same CDF and zero means, the RV tends toward a normal CDF, as n increases towards infinity.* The corollary of this is that the product of a great number of independent, identically distributed RV's tends to be normally distributed. The theoretical justification of the normal distribution is of little practical importance; however, we commonly observe that the distribution of grades becomes more symmetric and normal-like as the volume of investigation becomes large—the randomness of the grades is averaged and the results tend to a normal distribution.

2.2.2 Parametric and Non-parametric Distributions

A parametric distribution model has an analytical expression for either the PDF or the CDF, as for the Gaussian density function and the Lognormal distribution. Parametric distributions sometimes relate to an underlying theory, as does the normal distribution to the Central Limit Theorem. There are many parametric distributions that are used in different settings including the lognormal, uniform, triangular, and exponential distributions. Modern geostatistics makes extensive use of the Gaussian distribution because of its mathematical tractability. The lognormal distribution is important as well, but mostly from an historical perspective. In general, however, modern geostatistics is not overly concerned with other parametric distributions because data from any distribution can be transformed to any other distribution including the Gaussian one if needed. Adopting a parametric distribution for the data values may be the only option in presence of very sparse data; a non-parametric distribution is used when there are sufficient data.

There is no general theory for earth science related variables that would predict the parametric form for probability distributions. Nevertheless, certain distribution shapes are commonly observed. There are statistical tests to judge whether a set of data values follow a particular parametric distribution. But these tests are of little value in resource estimation because they require that the data values all be independent one from another, which is not the case in practice.

Parametric distributions have three significant advantages: (1) they are amenable to mathematical calculations, (2) the PDF and CDF are analytically known for all z values, and (3) they are defined with a few parameters. The primary disadvantage of parametric distributions is that, in general, real data do not conveniently fit a parametric model. However, data transformation permits data following any distribution to be transformed to any other distribution, thus capitalizing on most of the benefits of parametric distributions.

Most data distributions are often not well represented by a parametric distribution model. Sometimes distributions are characterized as non-parametric, that is, all of the data are used to define the distribution with experimental proportions; a parametric model for the CDF or PDF is not required. In this case, the CDF probability distribution may be inferred directly from the data, and therefore non-parametric distributions are more flexible. The CDF is inferred directly as the proportion of data less than or equal to the threshold value z. Thus, a proportion is associated to a probability.

A non-parametric cumulative distribution function is a series of step functions. Some form of interpolation may be used to provide a more continuous distribution $F(z)$ that extends to arbitrary minimum $z_{\min}$ and maximum $z_{\max}$ values. Linear interpolation is often used. More complex interpolation models could be considered for highly skewed data distributions with limited data.

2.2.3 Quantiles

Quantiles are specific Z values that have a probabilistic meaning. The p-quantile of the distribution $F(z)$ is the value z_p for which: $F(z_p) = \text{Prob}\{Z \leq z_p\} = p$. The 99 quantiles with probability values from 0.01 to 0.99 in increments of 0.01 are known as percentiles. The nine quantiles at 0.1, 0.2, …, 0.9 are called deciles. The 3 quantiles with probability values of 0.25, 0.5 and 0.75 are known as quartiles. The 0.5 quantile is also known as the median. The cumulative distribution function provides the tool for extracting any quantile of interest. The mathematical inverse of the CDF function is known as the quantile function:

$$z_p = F^{-1}(p) = q(p)$$

The *interquartile Range* (IR or IQR) is the difference between the upper and the lower quartiles: IR = q(0.75) − q(0.25) and is used as a robust measure of the spread of a distribution. The s*kewness* sign is the sign of the difference between the mean and the median (m-M) that indicates positive skewness or negative skewness.

Quantiles are used for comparing distributions in various ways. They can be used to compare the original data distribution to simulated values, compare two types of samples, or compare assay results from two different laboratories. A good way to do this is with a plot of matching quantiles, that is, a quantile-quantile (Q-Q) plot (Fig. 2.6). To generate a Q-Q plot, we must first choose a series of probability values pk, k = 1, 2, …, K; then, we plot $q_1(p_k)$ versus $q_2(p_k)$, k = 1, 2, …, K.

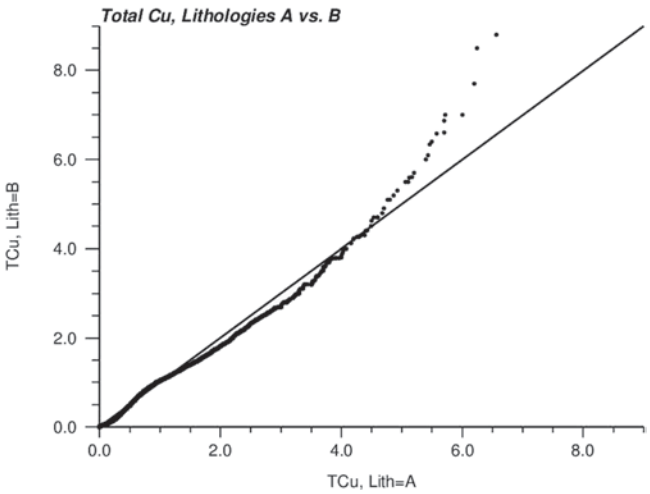


Fig. 2.6 An example of a Q-Q plot. The data is total copper, corresponding to two different lithologies

If all the points fall along the 45° line, the two distribution are exactly the same; if the line is shifted from the 45°, but parallel to it, the two distribution have the same shape but different means; if the slope of the line is not 45°, the two distributions have different variances, but similar shapes; and if there is a nonlinear character to the relationship between the two distributions, they have different histogram shapes and parameters.

The P-P plot considers matching probabilities for a series of fixed Z values. The P-P plot will vary between 0 and 1 (or 0 and 100%), from minimum to maximum values in both distributions. In practice, Q-Q plots are more useful because they plot the values of interest (grades, thicknesses, permeabilities, etc.), and it is therefore easier to conclude how the two distributions compare based on sample values.

2.2.4 Expected Values

The expected value of a random variable is the probability weighted average of that random variable:

$$E\{Z\} = m = \int_{-\infty}^{+\infty} z dF(z) = \int_{-\infty}^{+\infty} z f(z) dz$$

The expected value of a random variable is also known as the mean or the first moment. The expected value can also be considered as a statistical operator. It is a linear operator.

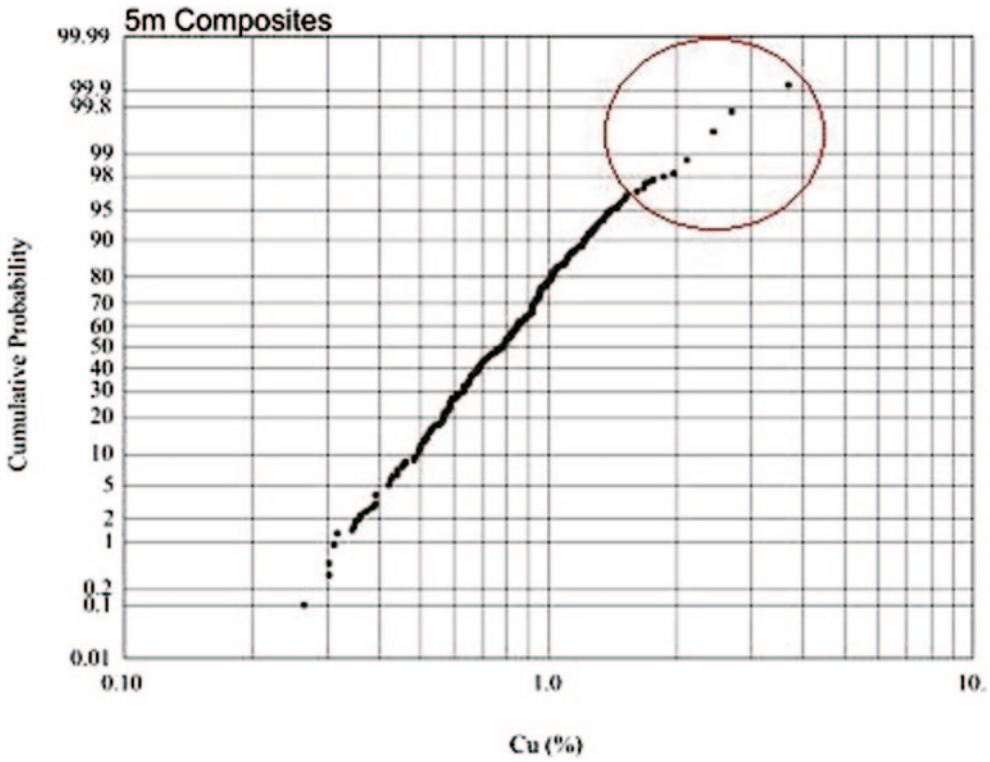


Fig. 2.7 Probability plot with outliers identified

The expected value of the squared difference from the mean is known as the variance (σ^2). It is written:

$$\begin{aligned} Var\{Z\} &= E\{[Z - m_z]^2\} = \sigma^2 \\ &= E\{Z^2 - 2Zm_z + m_z^2\} \\ &= E\{Z^2\} - 2m_z E\{Z\} + m_z^2 \\ &= E\{Z^2\} - m^2 \end{aligned}$$

The square root of the variance is the standard deviation (σ or s). The standard deviation is in the units of the variable. It is common to calculate a dimensionless coefficient of variation (CV), that is, the ratio of the standard deviation divided by the mean.

$$CV = \sigma / m$$

As an approximate guide, a CV less than 0.5 indicates a fairly well behaved set of data. A CV greater than 2.0 or 2.5 indicates a distribution of data with significant variability, such that some predictive models may not be appropriate.

There are additional measures of central tendency aside from the mean. They include the median (50% of the data smaller and 50% larger), the mode (the most common observation), and the geometric mean.There are also measures of spread aside from the variance. They include the range (difference between the largest and smallest observation), the interquartile range (described above), and the mean absolute deviation (MAD). These measures are not used extensively.

2.2.5 Extreme Values—Outliers

A small number of very low or very high values may strongly affect summary statistics like the mean or variance of the data, the correlation coefficient, and measures of spatial continuity. If they are proven to be erroneous values, then they should be removed from the data. For extreme values that are valid samples, there are different ways to handle them: (1) classify the extreme values into a separate statistical population for special processing, or (2) use robust statistics, which are less sensitive to extreme values. These options can be used at different times in mineral resource estimation. As a general principle, the data should not be modified unless they are known to be erroneous, although their influence in spatial predictive models may be restricted.

Many geostatistical methods require a transformation of the data that reduces the influence of extreme values. Probability plots can sometimes be used to help identify and correct extreme values, see Fig. 2.7. The values in the upper tail of the distribution could be moved back in line with the trend determined from the other data. An alternative consists of *capping* whereby values higher than a defined outlier threshold are reset to the outlier threshold itself. The high values could be in-

terpreted as a separate population altogether (see for example Parker 1991). There are a number of methods to deal with outliers at the time of variography and resource estimation.

In general, outliers or extreme values are considered on a case-by-case basis with sensitivity studies and considering their impact on local and global resource estimates.

2.2.6 Multiple Variable Distributions

Mineral resource estimation commonly considers multiple variables. The multiple variables could be geometric attributes of the deposit or grades such as thickness, gold, silver, or copper grades. They could be the same grade sampled at different locations. Bivariate and multivariate statistics are used in these cases. There are many references to multivariate statistics, such as Dillon and Goldstein (1984).

The cumulative distribution function and probability density function can be extended to the bivariate case. Let X and Y be two different RVs. The bivariate cdf of X and Y, $F_{XY}(x, y)$ and the pdf of X and Y $f_{XY}(x, y)$ are defined as

$$F_{XY}(x,y) = Prob\{X \le x, \text{ and } Y \le y\}$$

and

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$$

We could also define a bivariate histogram, that is, divide the range of the X and Y variables into bins and plot bivariate frequencies. It is more common to simply plot a scatterplot of paired samples on arithmetic or logarithmic scale. Figure 2.8 shows an example from the oil sands in Northern Alberta, Canada, after transformation to a Gaussian variable.

The means and variances of each variable are used as summary statistics. The covariance is used to characterize bivariate distributions:

$$Cov\{X,Y\} = E\{[X - m_X][Y - m_Y]\} = E\{XY\} - m_X m_Y$$
$$= \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} (x - m_X)(y - m_Y) f_{XY}(x,y) dy$$

The unit of the covariance is the product of the units of the two variables, for example, g/t Au multiplied by thickness in meters. Since these units are hard to understand or interpret, it is common for the covariance to be standardized.

The covariance describes whether the bivariate relationship is dominated by a direct or an inverse relationship, see Fig. 2.9. The product of $[X - m_X][Y - m_Y]$ is positive in quadrants II and IV; it is negative in quadrants I and III. The expected value is the average of the product over all pairs. The example of Fig. 2.9 has a positive covariance, while the example in Fig. 2.8 has a negative covariance because the relationship is dominated by an inverse relationship.

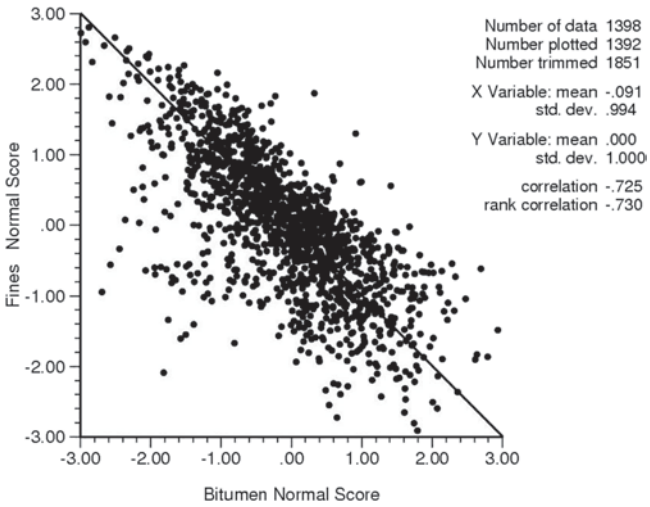


Fig. 2.8 Scatterplot of Bitumen vs. Fines Gaussian variables

The correlation coefficient between random variables X and Y is defined as the covariance between X and Y divided by the standard deviations of the X and Y variables:

$$\rho_{XY} = \frac{Cov\{X,Y\}}{\sigma_X \sigma_Y}$$

The correlation coefficient is a dimensionless measure between −1 (a perfect inverse linear relationship) and +1 (a perfect direct linear relationship). Independence between the two variables means that the correlation coefficient is zero, but the reverse is not necessarily true. A covariance or correlation coefficient of zero means there is no dominant direct or inverse relationship, but the variables may be related in a nonlinear manner.

Second order moments like the variance and covariance are significantly affected by outlier data. Some outlier pairs can destroy an otherwise good correlation or enhance an otherwise poor correlation, see Fig. 2.10. The sketch on the left illustrates a case where some outliers would make an otherwise good correlation appear low; the sketch on the right shows a case where a few outliers make an otherwise poor correlation appear high.

The *rank* correlation is more robust with respect to outliers, and is obtained by calculating the correlation coefficient on the rank order of the data. Each data variable is replaced by its rank position in the dataset, and then the correlation coefficient is calculated using the rank positions.

It is common for both correlation coefficients to be shown on experimental cross plots as in Fig. 2.8 where a direct comparison of the two correlation coefficients can be made. Their difference highlights whether there are data features,

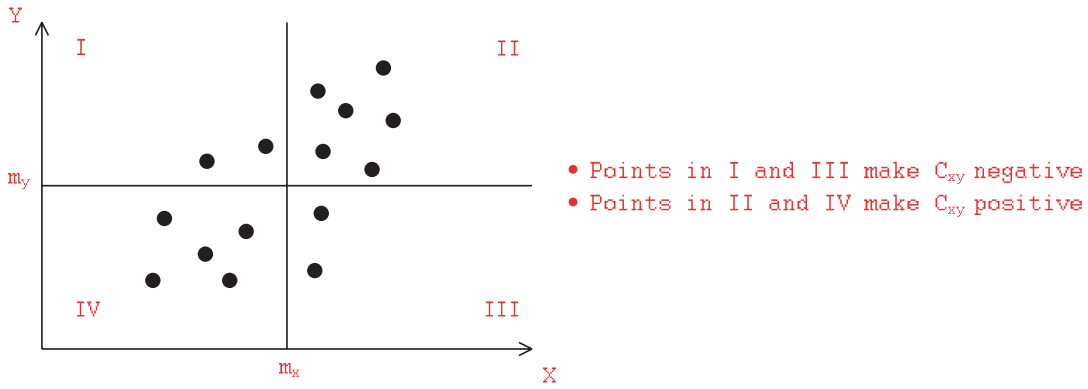


Fig. 2.9 Schematic cross plot with the mean of X drawn as a vertical line, the mean of Y drawn as a horizontal line and the four quadrants numbered

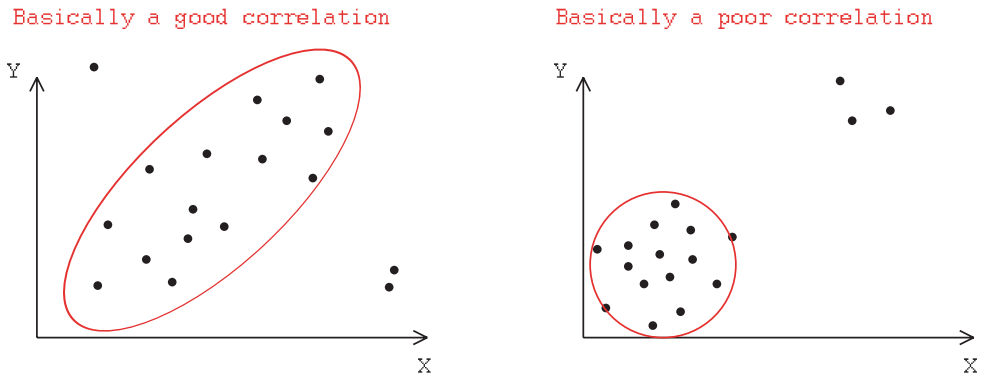


Fig. 2.10 Two schematic scatterplots where outlier data damage a good correlation or enhance a poor correlation

such as outliers, that render the linear correlation measure less useful. Classical least-squares regression requires traditional covariances and not those calculated on a transform of the data. Therefore, rank correlations should only be used for data exploration.

As with the univariate case, scatterplot smoothing is possible and sometimes necessary if the amount of original information is insufficient to characterize the bivariate distribution.

2.3 Spatial Data Analysis

This section describes a series of tools used to better understand spatial distributions. There are several tools that can be used, and are applied in the process called Exploratory Data Analysis, see for example Isaaks and Srivastava (1989).

Posting the data on a variety of cross-sectional or projection views provides clues as to the collection of the data and potential clustering. Posting the values colored differently for values above and below different grade thresholds provides an assessment as to the continuity of high and low grade trends.

Contour maps are used for understanding trends. These can be made by hand or the computer and are used to help in the description of trends. Contouring is typically done on two-dimensional planes defined according to the grid coordinates in plan, cross-sectional, and longitudinal views. It is common to rotate the locations to a local coordinate system prior to any spatial analysis, such that the main coordinate axes are approximately matched with the general orientation of the deposit.

Symbol maps may be more convenient than grade posting maps. A symbol represents some significant aspect of the data, for example drill hole data obtained in different campaigns, by different drilling methods, or at different points in time.

Indicator Maps are a particular form of a symbol map, where a binary variable is used to observe the presence or absence of certain characteristics such as data above and below certain thresholds or presence or absence of specific geologic variables.

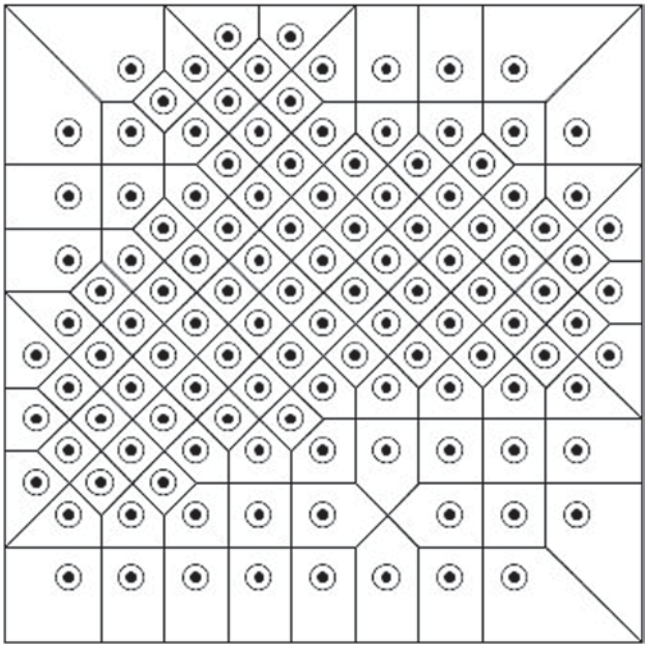


Fig. 2.11 An example of 122 samples with their polygonal areas of influence

2.3.1 Declustering

Data are rarely collected randomly. Drill holes are often drilled in areas of greatest interest, for example high grade areas that will be mined early in the production schedule. This practice of collecting more samples in areas of high grade should not be changed because it leads to the greatest number of data in portions of the study area that are the most important. There is a need, however, to adjust the histograms and summary statistics to be representative of the entire volume of interest.

Declustering techniques assign each datum a weight based on closeness to surrounding data $wi, i=1, \ldots, n$. These weights are greater than 0 and sum to 1. The experimental distribution and all summary statistics are calculated with the weights instead of a constant $1/n$.

The polygonal declustering method (Fig. 2.11; Isaaks and Srivastava 1989) is perhaps the simplest, and assigns each weight proportional to the area or volume of interest of each sample. Studies have shown that this approach works well when the limits to the area of interest are well defined and the ratio of the largest to smallest weight is less than 10 to 1.

The nearest-neighbor declustering technique is commonly used in resource estimation, and is like the polygonal method. The difference is that it is applied to a regular grid of blocks or grid nodes. The closest datum of the set being declustered is assigned to each block. Because it works on the same blocks that are used to estimate resources, it is more practical in resource estimation.

The technique of cell declustering is another commonly used declustering technique (Journel 1983; Deutsch 1989). Cell declustering works as follows:

1. Divide the volume of interest into a grid of cells $l=1, \ldots, L$.
2. Count the occupied cells L_o and the number of data in each occupied cell n_{lo}, $lo=1, \ldots, L_o$.
3. Weight each data according to the number of data falling in the same cell, for example, for datum i falling in cell l, the cell declustering weight is:

$$w_i = \frac{1}{n_l \cdot L_o}$$

The weights are greater than zero and sum to one. Each occupied cell is assigned the same weight. An unoccupied cell simply receives no weight.

Figure 2.12 illustrates the cell declustering procedure. The area of interest is divided into a grid of $L=36$ cells, with $L_o=33$ occupied cells. The number of data in each occupied cell is established by arbitrarily moving data on the grid boundaries to the right and down.

The weights depend on the cell size and the origin of the grid network. It is important to note that the cell size for declustering is *not* the cell size for geologic modeling; it simply defines an intermediate grid that allows assigning a declustering weight.

When the cell size is very small, each datum is in its own cell and receives an equal weight. When the cell size is very large, all data fall into one cell and are equally weighted. Choosing the optimal grid origin, cell shape, and size requires some sensitivity studies. It is common to choose the cell size so that there is approximately one datum per cell in the sparsely sampled areas or, if available, to choose it according to an underlying, quasi-homogeneous sampling grid.

The sensitivity of the results to small changes in the cell size should be checked. If the results change by a large amount, then most likely the declustering weights are changing for one or two anomalously high or low grades.

Since it is generally known whether over-sampling occurs in high- or low-valued areas, the weights can be selected such that they give the minimum or maximum declustered mean of the data. The declustered mean versus a range of cell sizes should be plotted, and the size with the lowest (Fig. 2.13, data clustered in high-valued areas) or highest (data clustered in low-valued areas) chosen. Care should be taken not to over-fit the minimum. The correct cell size should be approximately the spacing of the data in sparsely sampled areas. This qualitative check can be used to ensure that a too-large or too-small cell size is not chosen.

The shape of the cells depends on the geometric configuration of the data, as it is adjusted to conform to the major directions of preferential sampling. For example, if the samples are more closely spaced in the X direction than in the Y direction, the cell size in the X direction should be reduced.

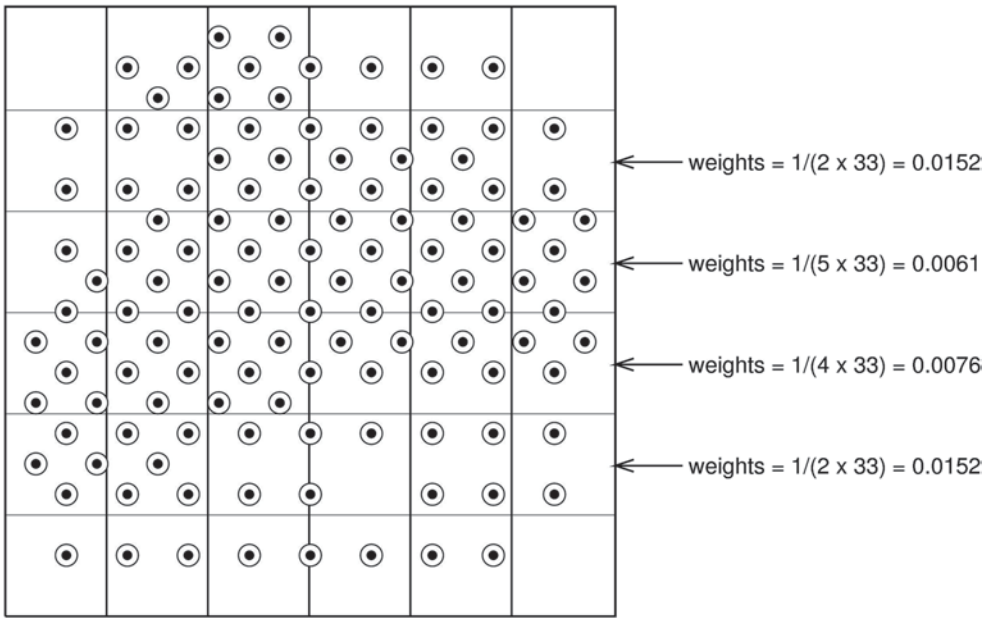


Fig. 2.12 Illustration of the cell declustering method

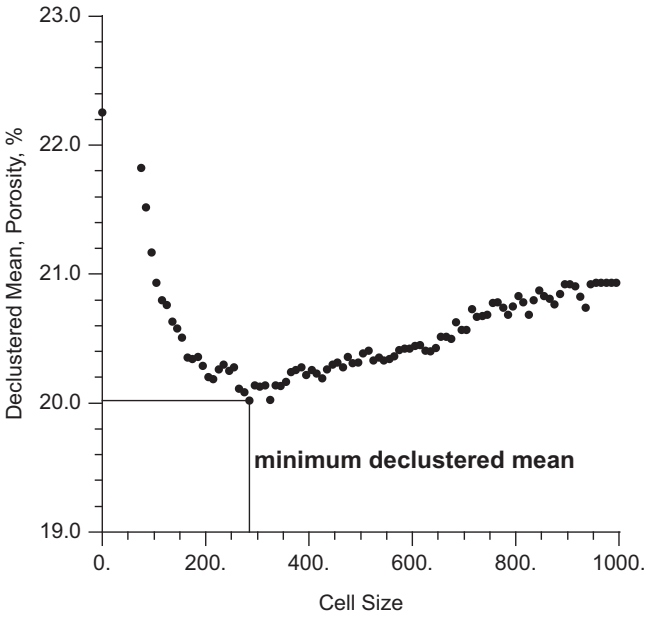


Fig. 2.13 Plot of declustered mean vs. cell size

The origin of the cell declustering grid and the number of cells L must be chosen such that all data are included within the grid network. Fixing the cell size and changing the origin often leads to different declustering weights. To avoid this artifact, a number of different origin locations should be considered for the same cell size. The declustering weights are then averaged for each origin offset.

Declustering assumes that the entire range of the true distribution has been sampled. If this is not the case, then the data is biased and debiasing techniques may be required. These techniques include trend modeling for debiasing and debiasing using qualitative data, subjects that are not covered in this book.

2.3.2 Declustering with Multiple Variables

Declustering weights are determined on the basis of the geometric configuration of the data; therefore, only one set of declustering weights is calculated in presence of multiple variables that have been equally sampled. However, different declustering weights will need to be calculated when there is unequal sampling. For example, there are sometimes different sets of Copper and Molybdenum samples in a Cu-Mo porphyry deposit, which would require two sets of declustering weights.

Declustering weights are primarily used to determine a representative histogram for each variable; however, we also require the correlation between multiple variables. The same set of declustering weights can weight each pair contributing to the correlation coefficient (Deutsch 2002).

2.3.3 Moving Windows and Proportional Effect

Moving windows are used to understand the local spatial behavior of the data, and how it may differ from global statistics. The process is to lay over the volume of interest a grid of cells, which may or may not be partially overlapping, moving them over the entire domain or deposit, and obtaining statistics within them. Overlapping windows are typically used when there are few data within the window to provide reliable statistics (Goovaerts 1997; Isaaks and Srivastava 1989).

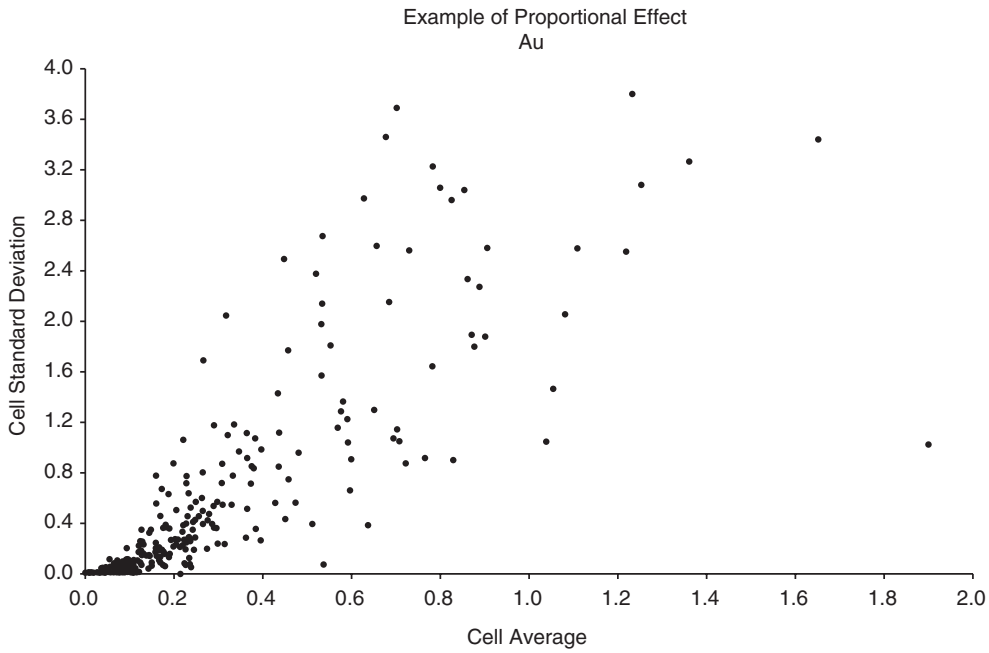


Fig. 2.14 An example of proportional effect from a West-African gold deposit. Cell averages and standard deviations are both in g/t

The most common statistics analyzed are the mean and standard deviations of the data within the windows.

A plot of the mean versus standard deviation calculated from moving windows of data can be used to assess changes in local variability, see Fig. 2.14 for an example. Generally, positively skewed distributions will show that windows with higher local mean usually exhibits higher local standard deviation. This is the proportional effect described by various authors, for example David (1977) and also Journel and Huijbregts (1978). The proportional effect is due to a skewed histogram, but it may also indicate spatial trends or a lack of spatial homogeneity. Proportional effect graphs are sometimes used to help determine homogeneous statistical populations within the deposit (see Chap. 4).

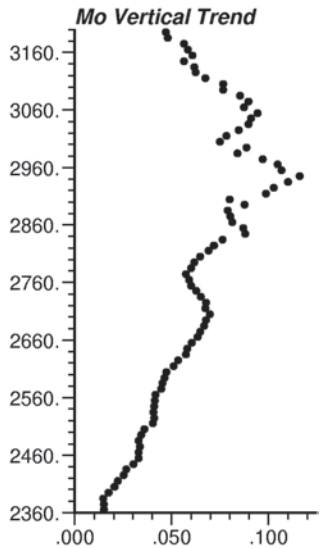


Fig. 2.15 Example of a molybdenum vertical trend

2.3.4 Trend Modeling

Trend modeling is applied when a trend has been detected and is assumed to be well understood. While some geostatistical estimation methods are quite robust with respect to the presence of trends, such as Ordinary Kriging (Chap. 8; Journel and Rossi 1989), there are many others, most notably simulation (Chap. 10) that are quite sensitive to trends.

The trend is modeled as a deterministic component plus a residual component. The deterministic component is removed and then the residual component is modeled either through estimation or simulation techniques. Finally, the deterministic trend is added back. In such a model, the mean of the residual and the correlation between the trend and the residual should be close to 0.

The drill hole data is typically the source for trend detection. In some cases where the geological environment is well understood, trends can be expected and modeled without the drill hole data, but this should only be attempted when there is no other option. Large scale spatial features can be detected during several stages of data analysis and modeling. Sometimes a simple cross-plot of the data against elevation may show a trend, as in the example of Fig. 2.15. In other cases, simple contour maps on cross-sections, longitudinal sections, or plan views are enough to identify and model trends. Moving window averages can also provide an indication of whether or not the local means and variances are stationary. If there are notable changes in the local

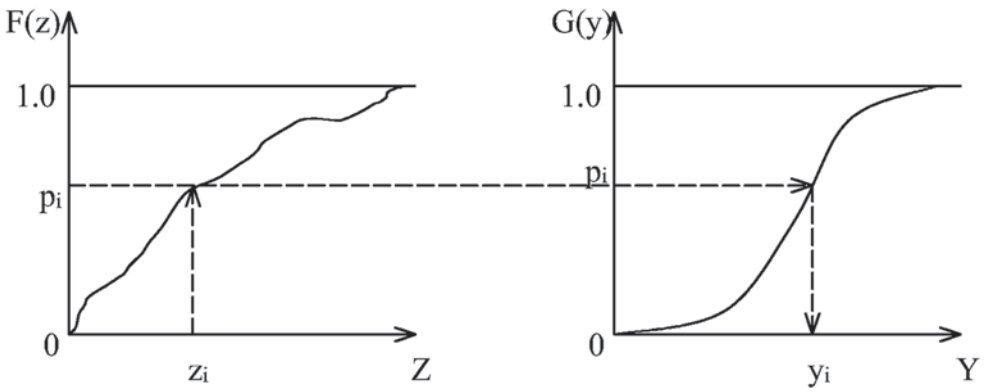


Fig. 2.16 Data transformation using Cumulative Distribution Functions

mean and variance of reasonably large subdivisions within the domain, as in Fig. 2.14, then a spatial trend model may be required.

Although the identification of a trend is subjective, it is generally accepted that the trend is deterministic and should not have short scale variability. It should be identified from features that are significantly larger than the data spacing, i. e., domain-wide. This sometimes can be evident from the experimental variogram that may show a trend in any one or more directions. The experimental variogram continues to increase above the variance of the data as the lag distance increases (Chap. 6; Journel and Huijbregts 1978). This usually indicates that the decision of stationarity should be revisited, and consider whether the domain should be subdivided or a trend considered.

2.4 Gaussian Distribution and Data Transformations

Gaussian distributions are commonly used due to their convenient statistical properties. The Gaussian distribution is derived from the Central Limit Theorem, which is one of the most consequential theorems in statistics.

A univariate Gaussian distribution is fully characterized by its mean (m) and standard deviation (σ). The probability density function is given by:

$$g(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-m}{\sigma}\right)^2\right]$$

It is common to transform data to a Gaussian distribution. There are many instances where the prediction of uncertainty at un-sampled locations becomes much easier with a Gaussian distribution.

The simplest method to transform any distribution into a Gaussian distribution is a direct quantile-to-quantile transformation, whereby the CDF of each distribution is used to perform the transform. This is known as the Normal Scores (NS) transform, see Fig. 2.16. The NS transform is achieved by quantile transformation:

$$y = G^{-1}(F(z))$$

which is back-transformed by

$$z = F^{-1}(G(y))$$

The expected values should not be back transformed unless the distribution is symmetric.

A variable Z is non-standard Gaussian when the standardized variable Y is standard Gaussian. A non-standard Gaussian value is easily converted to/from a standard Gaussian value.

$$y = \frac{z - m_Z}{\sigma_Z} \qquad z = y \cdot \sigma_Z + m_Z$$

The normal score transform is rank preserving and reversible. The disadvantages of performing such a transform are that the significance of the numbers themselves is less clear, more difficult to interpret, and also that the distribution parameters cannot be back transformed directly due to the nonlinearity of the process.

Spikes of constant values in the original distribution can cause problems. Gaussian values are continuous and ties (equal values) in the original distribution must be resolved prior to transforming the data. There are two different methods commonly used to break the ties or despike. The simpler method is to add a small random component to each tie, which is the most common approach used in popular software packages, such as the GSLIB programs (Deutsch and Journel 1997). A better alternative is to add a random component based on local averages of the data (Verly 1984), which ranks the ties based on the local grades of nearby data. Although more onerous in terms of time and computer effort, it is justified when the proportion of original data with the same values is significant. Typical drill hole data from Au epithermal deposits can show a significant number of values at or below the laboratory's detection limit, sometimes as much as 50 or 60 %, in which case despiking is better accomplished using the local averaging method. Of course, an alternative is to separate the barren or un-mineralized material into its own stationary population.

This is reasonable when the spatial arrangement of the barren material is predictable.

2.5 Data Integration and Inference

The prediction of spatial variables requires consideration of multivariate distributions of values at different locations. Inference requires the combination of sample data to estimate at an unknown location. The calculation of conditional distributions is accomplished by application of Bayes' Law, one of the most important laws in statistical theory.

Bayes' Law provides the probability that a certain event will occur given that (or conditional to) a different event has already occurred. The mathematical expression for Bayes' Law can be written as:

$$P(E_1 \mid E_2) = \frac{P(E_1 \text{ and } E_2)}{P(E_2)}$$

with E_1 and E_2 being the events, and P representing probabilities.

If E_1 and E_2 are independent events, then knowing that E_1 occurred does not give additional information about whether E_2 will occur:

$$P(E_1 \mid E_2) = P(E_1)$$
$$P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2)$$

Direct inference of multivariate variables is often difficult, which leads us to use the multivariate Gaussian model, mostly because it is straightforward to extend to higher dimension. The bivariate Gaussian distribution is defined as:

$$(X,Y) \rightarrow N(0,1,\rho_{X,Y}),$$
$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\left[-\frac{1}{2(1-\rho^2)}\left(x^2-2\rho xy+y^2\right)\right]}$$

The relationship between the two variables is defined by a single parameter, the correlation coefficient, and in the XY cross-plot the probability contours are elliptical. The conditional expectation of Y given an event for X is a linear function of the conditioning event:

$$E\left\{Y \middle| X = x\right\} = m_Y + \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(x - m_x)$$

The conditional expectation follows the equation of a line, $y = mx + b$, where m is the slope (correlation coefficient) and b is the intercept (mean).

The conditional variance is independent of the conditioning event(s). This is an important consideration that will influence some of the geostatistical methods to be described later, and is written as:

$$Var\left\{Y \middle| X = x\right\} = \sigma_Y^2\left(1 - \rho_{X,Y}^2\right)$$

For a standard bivariate Gaussian distribution (that is, both variables, X and Y have a mean = 0 and variance = 1.0) the parameters are:

$$E\left\{Y \middle| X = x\right\} = \rho_{X,Y} \cdot x$$
$$Var\left\{Y \middle| X = x\right\} = 1 - \rho_{X,Y}^2$$

The extension to multivariate distributions is straightforward, and can be written as:

$$\mathcal{N}(\mathrm{x};\mu,\textstyle\sum) = \frac{1}{(\sqrt{2\pi})^d \left|\sum\right|^{1/2}} \cdot \exp\left[-\frac{1}{2}(\mathrm{x}-\mu)^T \textstyle\sum^{-1}(\mathrm{x}-\mu)\right]$$

where d is the dimensionality of x. Note that μ is a (d × 1) vector and Σ is a (d × d) positive definite, symmetric variance-covariance matrix. The expression $|\Sigma|$ is the determinant of Σ. μ is the mean of the distribution and Σ is the covariance matrix. The i-th element of μ expresses the expected value of the i-th component in the random vector x; similarly, the (i, j) component of Σ expresses the expected value of $\mathrm{x}_i\,\mathrm{x}_j$ minus $\mu_i \mu_j$. The diagonal elements of Σ are the variances of the corresponding component of x.

The multivariate (N-variate) Gaussian distribution possesses some extraordinary properties (Anderson 1958; Abramovitz and Stegun 1964):

1. All lower order N-k marginal and conditional distributions are Gaussian.
2. All conditional expectations are linear functions of the conditioning data:

$$E\{X_i \mid X_j = x_j, \forall j \neq i\} = \sum_{j \neq i} \lambda_j x_j$$
$$= \varphi(x_j, j \neq i) = [x_i]^*_{SK}$$

3. All conditional variances are homoscedastic (data-values-independent):

$$E\left\{\left[X_i - \varphi(x_j, j \neq i)\right]^2 \mid X_j = x_j, \forall j \neq i\right\}$$
$$= E\left\{\left[X_i - \varphi(x_j, j \neq i)\right]^2\right\}$$

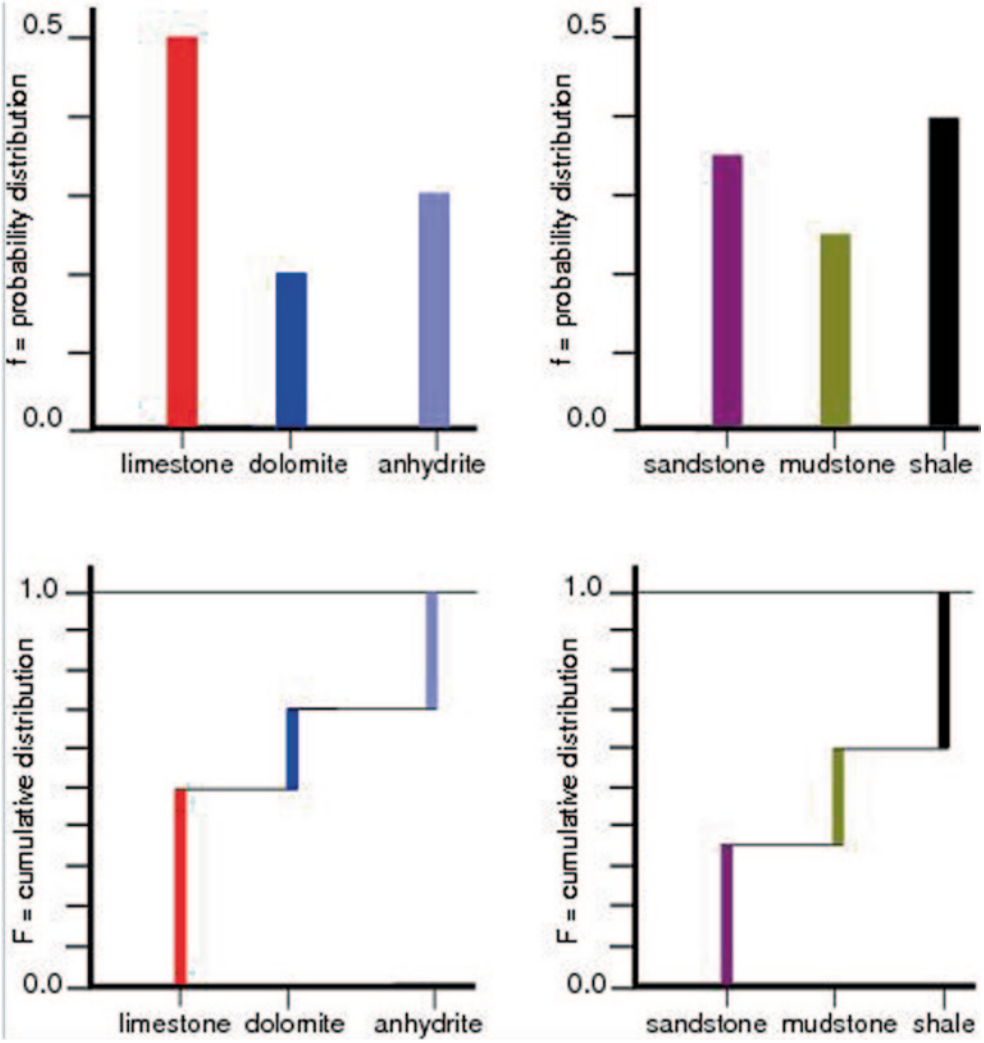


Fig. 2.17 PDFs and CDFs for categorical variables

Conditional expectations are linear functions of the data. All linear combinations of Gaussian variables are also Gaussian, and in particular, averages are Gaussian. Also, conditional variances are data-values-independent, a property called *homoscedasticity*.

In geostatistics, it is common to assume that the normal scores of grade variables are multivariate Gaussian within geologically defined domains. This is done for convenience since the simple (co)kriging method provides exactly the mean and variance of all conditional distributions, as described in Chaps. 8–10.

Performing a univariate normal score transformation guarantees a univariate Gaussian distribution, but there is no guarantee of a multivariate Gaussian distribution. The transformation does not remove nonlinearity or other constraints. The proportional effect and heteroscedasticity is largely removed by the transformation, but then it is reintroduced by the back transformation. Transforming a multivariate distribution is rarely done in mineral resource estimation because of the complexity and requirement for many data.

Categorical Variables The probability distribution of a discrete or categorical variable is defined by the probability or proportion of each category, that is, p_k, $k=1, \ldots, K$, where there are K categories. The probabilities must be non-negative and sum to 1.0. A table of the p_k values completely describes the data distribution. Sometimes, however, it is convenient to consider a histogram and cumulative histogram as shown below (Fig. 2.17):

The cumulative histogram is a series of step functions for an arbitrary ordering of the discrete categories. Such a cumulative histogram is not useful for descriptive purposes but is needed for Monte Carlo simulation and data transformation. In general, but not always, the ordering does not matter. The cases where the ordering affects the results will be discussed later in the book.

Consider K mutually exclusive categories s_k, $k=1, \ldots, K$. This list is also exhaustive; that is, any location $\boldsymbol{u}$ belongs to one and only one of these K categories. Let $i(\mathbf{u}; s_k)$ be the indicator variable corresponding to category s_k, set to 1 if location $\boldsymbol{u}$ in s_k, zero otherwise, that is:

$$i(\mathbf{u}_j; s_k) = \begin{cases} 1, & \text{if location } \mathbf{u}_j \text{ in category } s_k \\ 0, & \text{otherwise} \end{cases}$$

Mutual exclusion and exhaustivity entails the following relations:

$$i(\mathbf{u}; s_k) \cdot i(\mathbf{u}; s_{k'}) = 0, \quad \forall k \neq k'$$

$$\sum_{k=1}^{K} i(\mathbf{u}; s_k) = 1$$

The mean indicator for each category s_k, $k=1, \ldots, K$ is the proportion of data in that category:

$$p_k = \overline{i(\mathbf{u}; s_k)} = \frac{\sum_{j=1}^{N} w_j i(\mathbf{u}_j; s_k)}{\sum_{j=1}^{N} w_j}$$

The variance of the indicator for each category s_k, $k=1, \ldots, K$ is a simple function of the mean indicator (Journel 1983; Deutsch 2002):

$$Var\{i(\mathbf{u}; s_k)\} = \frac{\sum_{j=1}^{N} w_j [i(\mathbf{u}_j; s_k) - p_k]^2}{\sum_{j=1}^{N} w_j} = p_k (1.0 - p_k)$$

The variance would be used to standardize variograms for quicker interpretation and comparison across different categories.

2.6 Exercises

The objective of this exercise is to review some mathematical principles, become familiar with some notation, work with some common probability distribution models and gain some experience with declustering. Some specific (geo)statistical software may be required. The functionality may be available in different public domain or commercial software. Please acquire the required software before beginning the exercise. The data files are available for download from the author's website—a search engine will reveal the location.

2.6.1 Part One: Calculus and Algebra

Question 1: Consider the following function $(aX+bY)(X+Y)$. Calculate the derivative of this function with respect to X and Y.

Question 2: Calculate the integral for the function below:

$$\int_0^5 \frac{1}{2}x^2 + x^3 - \frac{1}{4}x^5 dx$$

Question 3: Consider the three matrices below:

$$\mathbf{A} = \begin{bmatrix} 5 & 2 \\ 2 & 3 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 2 & 3 \end{bmatrix}$$

What is the result of **AB**, $\mathbf{AC}^{\mathbf{T}}$, and **(AB)C**?

2.6.2 Part Two: Gaussian Distribution

Consider the standard Gaussian or normal distribution that is of extraordinary importance in statistics and geostatistics because it is the limit distribution of the central limit theorem and is mathematically tractable.

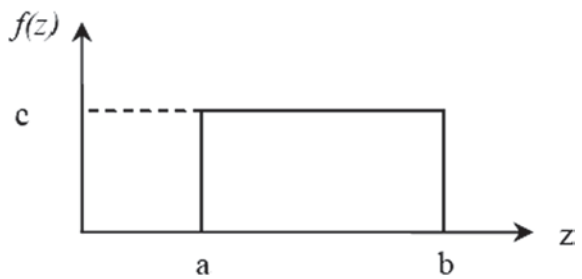
Question 1: Verify that the sum of independent random variables tends toward a normal distribution. Consider (1) setting up a grid of 100 rows by 10 columns in Excel with uniform random numbers between 0 and 1, (2) create an 11th column with the sum of the 10 first columns, (3) plot a histogram of the 11th column, and (4) comment.

Question 2: What is the mean and variance of a probability distribution that is uniform between 0 and 1? The central limit theorem tells us that the mean of 10 values added together should be this mean multiplied by 10—check against Question 1 and comment. The central limit theorem would also tell us that the variance is multiplied by 10—check against Question 1 and comment.

Question 3: Create a 12th column in your spreadsheet with the sum (the 11th column) minus the mean divided by the standard deviation, that is, $y12 = (y11 - m)/\sigma$. Plot a histogram and calculate the statistics of this *standardized deviate*. Comment on the results.

2.6.3 Part Three: Uniform Distribution

Consider the uniform distribution specified below:



Question 1: Write the definition and equation for the cumulative distribution function (cdf) of

the uniform distribution above. Draw the corresponding cdf to the probability density function (pdf) above.

Question 2: What is the value of c that makes $f(z)$ a licit probability distribution? Write your answer in terms of a and b.

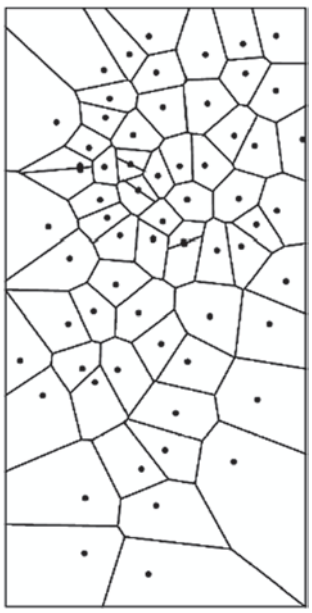
Question 3: What is the expected value (or mean) of the variable Z in terms of a, b, and c? Solve the integral.

Question 4: What is the variance of the variable Z in terms a, b, and c? Solve for the expected value of Z_2 and solve for the variance using $\sigma 2 = E\{Z2\} - [E\{Z\}]^2$.

Question 5: What is the 90% probability interval? Write out the function corresponding to the cdf and solve for the 5th and 95th quantiles.

The objective of this exercise is to become familiar with the different ways to use declustering to infer a representative probability distribution. Declustering software and the specified datasets are required.

2.6.4 Part Four: Small Declustering



Consider the 2-D data in `red.dat` (see right). The 67 drill hole intersections have a hole ID, location, thickness, four grade values, and a rock type. The area is from 20,100 to 20,400 in the northing direction and –600 to 0 in elevation. The rock type is simply a flag that specifies below or above –300 m. There is a difference below that elevation that warrants our attention.

Question 1: Plot a location map of the thickness and the gold grade. Plot a histogram of all the gold grades without any declustering weight.

Question 2: Setup and run polygonal declustering to get a map that looks like the one to the right. Plot a declustered histogram of the gold grades.

Question 3: Cell declustering is widely used because it is robust in 3-D and is less sensitive to edge effects. Run cell declustering for a range of cell sizes—explain your choice of parameters. Plot the declustered mean versus cell size, choose a cell size, and justify your choice. Compare results to those obtained above.

2.6.5 Part Five: Large Declustering

Consider the 3-D Au/Cu data in `largedata.dat`. This data will be used in some subsequent exercises. We need declustered distributions for the two variables in all rock types.

Question 1: Consider cell declustering on a by-rock type basis and considering all of the data together. Compare the results and comment on the preferred approach. Prepare a reasonable set of plots to support your conclusions including the declustered mean versus cell size plot(s) and tables of declustered mean and standard deviation values.

Question 2: Assemble the reference distributions for subsequent modeling (based on your chosen method).

References

Abramovitz M, Stegun I (1964) Handbook of mathematical functions. Dover Publications, New York, p 1046

Anderson T (1958) An introduction to multivariate statistical analysis. Wiley, New York

Borradaile GJ (2003) Statistics of earth science data. Springer, Heidelberg

David M (1977) Geostatistical ore reserve estimation. Elsevier, Amsterdam

Davis JC (1986) Statistics and data analysis in geology, 2nd edn. Wiley, New York, p 646

de Moivre A (1738) The doctrine of chances: or, a method for calculating the probabilities of events in play, 2nd edn. printed by H. Woodfall, London

Deutsch CV (1989) DECLUS: A FORTRAN 77 program for determining optimum spatial declustering weights. Comput Geosci 15(3):325–332

Deutsch CV (2002) Geostatistical reservoir modeling. Oxford University Press, New York, p 376

Deustch CV, Journel AG (1997) GSLIB: geostatistical software library and user's guide, 2nd edn. Oxford University Press, New York, p 369

Dillon W, Goldstein M (1984) Multivariate analysis: methods and applications. Wiley, New York, p 587

Gauss CF (1809) Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium. English translation by C. H. Davis, reprinted 1963, Dover, New York

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York, p 483
Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, New York, p 561
Journel AG (1983) Non-parametric estimation of spatial distributions. Math Geol 15(3):445–468
Journel AG, Huijbregts ChJ (1978) Mining geostatistics. Academic Press, New York
Journel AG, Rossi ME (1989) When do we need a trend model? Math Geol 22(8):715–739
Koch G, Link R (1986) Statistical analysis of geological data, 2nd edn. Wiley, New York
Lapin LL (1983) Probability and statistics for modern engineering. PWS Publishers, Boston, p 624
Laplace PS (1812) Théorie analytique des probabilités. Printed in 1814 by Mme. Ve. Courier, Paris
Legendre AM (1806) Nouvelles Methodes pour la Determination des Orbites des Cometes. F. Didot, Paris
Parker HM (1991) Statistical treatment of outlier data in epithermal gold deposit reserve estimation. Math Geol 23:125–199
Ripley BD (1987) Spatial statistics, 2nd edn. Wiley, New York
Rohatgi VK, Ehsanes Saleh AK Md (2000) An introduction to probability and statistics. Wiley, New York
Sichel HS (1952) New methods in the statistical evaluation of mine sampling data. Trans Inst Min Metall Lond 61:261
Verly G (1984) Estimation of spatial point and block distributions: the multigaussian model. PhD Dissertation, Department of Applied Earth Sciences, Stanford University