<<<Back to Table of Contents

On to Part G: Familiarization with spatial analysis>>>

Geoscience after IT: Part F

Familiarization with quantitative analysis

T. V. Loudon
British Geological Survey, West Mains Road, Edinburgh EH9 3LA, U.K.
e-mail: v.loudon@bgs.ac.uk

Postprint of article in Computers & Geosciences, 26 (3A) April 2000, pp. A41-A50

Abstract - Numbers, measurement and calculation extend our view of the world. Statistical methods describe the properties of sets of quantitative data, and can test models (particularly the model that observed relationships arose by chance) and help us to draw conclusions. Links between spatial and quantitative methods, through coordinate geometry and matrix algebra, lead to graphical representations for visualizing and exploring relationships. Multivariate statistics tie into visualization to look at pattern among many properties.

Key Words - Statistics, matrix algebra, visualization, multivariate analysis.

1. Background

Computing, in the sense of calculation, is a small part of IT applications in geoscience. But, even in traditional aspects of geology, the quantitative representation of spatial entities and models is important. To understand why, we need to look at some basic mathematical concepts, and see how numbers relate to properties and programs to physical processes. Readers with a mathematical background may prefer to skip at least the first part of this chapter, and those without may wish to skip detail irrelevant to them.

2. Measurement and number

Real numbers form a continuous sequence in an exact order. Given any two numbers, say 1.415 and 1.416, you can find as many numbers as you need between them, for example 1.4151 or 1.4150032. You can compare any two numbers to see if they are equal or if one is larger or smaller than the other. This leads to a **scale of measurement**. The continuous sequence of numbers is analogous to situations in the real world. For example, you can compare the thickness of any two beds of sandstone to see if they are the same or if one is thicker or thinner than the other. To avoid carrying sandstone around, you can measure their thicknesses by comparison with a standard tape marked in millimeters. Measurements enable comparisons between any two bed thicknesses. You can measure more than one property. For example, you could also measure the maximum grain size for the beds of known thickness and compare the two sets of numbers (bed thickness and grain diameter), pair by pair.

We can use numbers in several other ways that rely on different aspects of their properties and so must be handled differently (see Krumbein and Graybill, 1965 or Davis, 1973 for a fuller account). We can assign numbers quite arbitrarily as object identifiers. An identification number, such as an **accession number** (numbers from a given range issued in sequence), need bear no relationship to the object's properties. Indeed it is easier to issue unique identifiers by separating identification from description. A program or a user who knows the identifier can find the object by consulting a numerical index. Integer numbers are appropriate for identifiers.

The **subject classification** on a book or library shelf, such as a UDC number, is rather less arbitrary. Similar numbers apply to related subjects, and the number hierarchy (hundreds, tens, units) reflects subject subdivisions (part H section 2). Numbers with decimal fractions are convenient in book classification. One can insert additional subdivisions without limit, simply by adding more digits after the decimal point. By shelving books or arranging object identifiers in numerical order we bring together those on the same subject for convenient searching or browsing. A sequence of numbers can represent a strict order of categories. Typical **ordered categories** are Mohs' scale of mineral hardness and the Richter scale of earthquake intensity. The larger the number, the higher the value, but the steps between successive values are not equal.

Measurement compares a property of some object with a standard scale. Intervals are equal, although the zero value may be quite arbitrary. For example, most scales of temperature, unlike the Kelvin scale, place zero at a convenient but arbitrary point. It would therefore be foolish to say that 20^{o}C is twice as hot as 10^{o}C. Nevertheless, we can reasonably say that the increase of temperature from 0^{o} to 10^{o} is half that from 0^{o} to 20^{o}. Other physical properties, such as length, have an obvious and unique zero value, and there is no difficulty in adding, subtracting, multiplying and dividing those quantitative measurements.

The number field can then lead to a useful model at a deeper level than categorization. Equations can mimic real physical relationships. For instance, physicists can write equations describing the relationships of temperature with the pressure and volume of a closed body of gas. Equations imply the ability to calculate and maybe predict. Aspects of physical systems have direct analogs in well-known arithmetic operations. This astonishing correspondence between the physical world and mathematics is the basis for **mathematical modeling** (F 3).

The quantitative approach introduces a new mode of thinking. Instead of seeing the subject of investigation as a set of discrete objects, such as formations and rock types, we view it as a continuum, with characteristics that we can measure and compare as they vary from place to place. Gravity and aeromagnetic surveys, satellite imagery, or regional geochemical studies of stream sediments are examples. If objects are seen as 'things' represented by nouns, and processes resemble verbs, then quantitative measurements are more akin to adjectives, describing the properties or composition of the objects.

It is tempting to wonder how far this mode of thinking can extend. Could we, for instance, replace our rather arbitrary classifications of geological objects by a more quantitative view where we measure continual change. This is considered further in J

2.3, but classification is basic to science (J 2.1) and descriptions with adjectives and no nouns have little meaning. I argue later (L 6.3) that while, with IT support, the scope of quantitative studies will surely continue to expand, different modes of thought are complementary, each adding to the overall understanding. The more important role of IT may be to ensure that information of all kinds is readily available to the investigators. If this is correct, the scientist (or a multidisciplinary team) needs to understand and use an appropriate combination of methods and modes of thought.

Before collecting measurements, it makes sense to consider their intended applications. This is the next topic. For detail, see Griffiths (1967), Krumbein and Graybill (1965), Davis (1973) or Swan and Sandilands (1995).

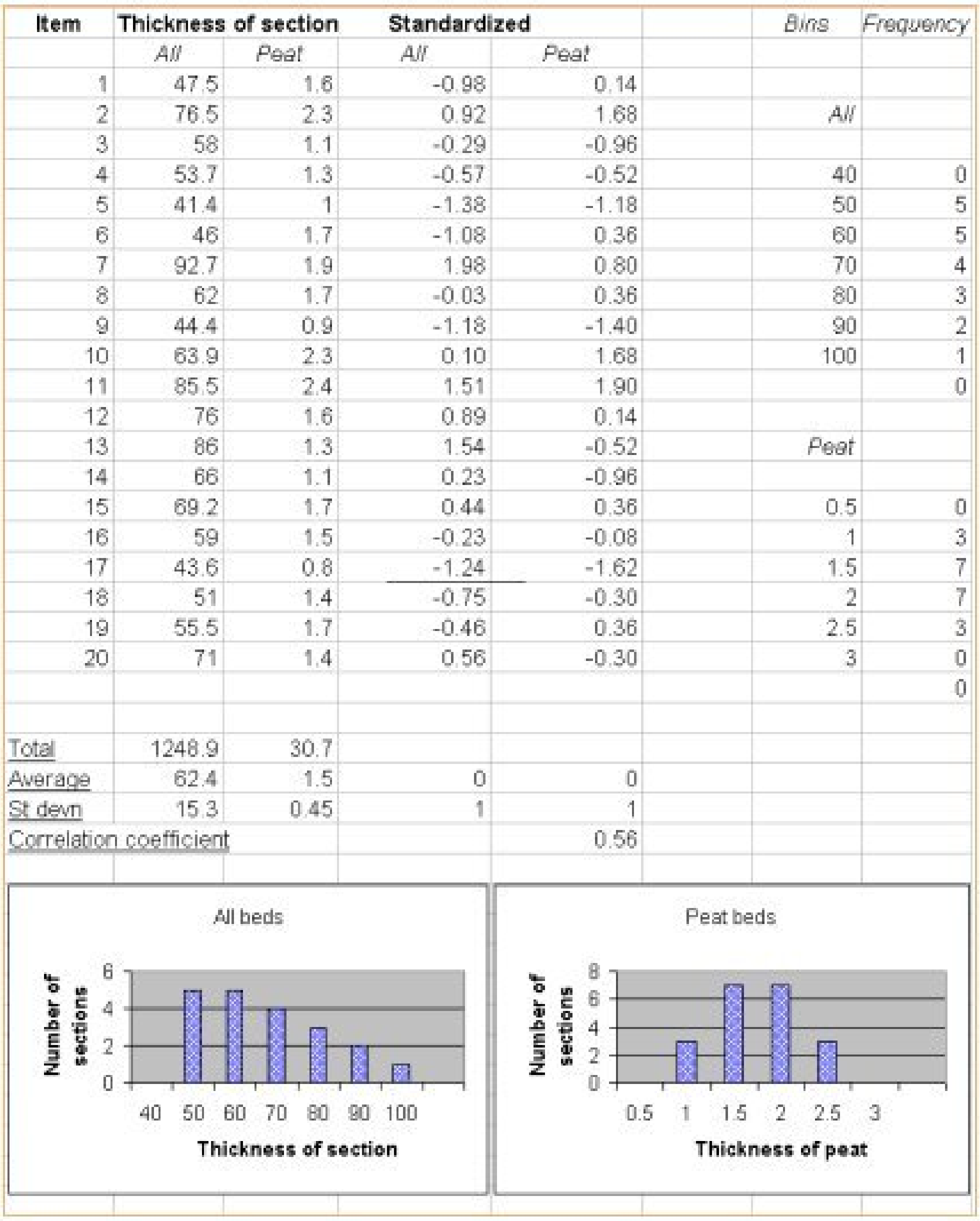
3. Descriptive statistics

We can manipulate numbers with simple operations of addition, subtraction, multiplication and division. They take us beyond individual comparisons to the properties of entire sets of measurements, and to general statements about relationships, say between grain size and bed thickness. **Statistics** (the branch of mathematics that deals with collecting, analyzing, interpreting and presenting numerical data) addresses these topics. One requirement is to characterize a set of measurements, like bed thickness, by fewer numbers that reflect the properties of the set as a whole. Important **statistics** (the measures or values calculated using the science of statistics) include the average value, also known as the **mean**. It is calculated by adding the measurements together and dividing the total by the number of measurements. We can measure the spread of values around the mean by the **variance** (the mean squared deviation from the mean) or by its square root - the **standard deviation**.

Statistics leads on from the description of a single **variable**, that is, a set of measurements of a single property, to explore the relationships between pairs of variables measured at the same point, such as bed thickness and grain size. An obvious approach would be to multiply each pair of measurements together and take their average (the mean cross-product). But the mean and standard deviation of each variable would greatly affect the result, and these have been measured already. Instead, we can **standardize** each variable by subtracting the mean from each value and dividing the result by the standard deviation. The mean cross-product of the transformed variables is known as the **correlation coefficient**, which has a value somewhere between +1 and -1. The extreme values are 1 if the bed thickness increases precisely as the grain size increases, and -1 if one decreases precisely as the other increases. The value is 0 if one variable shows no relationship to the other.

There are two general points here. One is that statistics measure different properties separately. Having calculated the mean, we remove its effects in calculating the next property, the standard deviation. We remove the effects of both in calculating the correlation coefficient. As a consequence, we can compare variation in a sequence of thick beds with that in a sequence of thin beds, and can judge whether the correlation of bed thickness and grain size is more pronounced in sandstone or in siltstone.

The other general point is that we are not dealing with sharply defined relationships. If we had measured the properties a few millimeters away, or made twice as many measurements, the results would have been different. If the processes of deposition had changed, with stronger currents, deeper water or different grain composition, the results would again differ. Statistical methods can measure the uncertainties of sampling and imperfect knowledge of the process. Their success depends on the skill with which the data are sampled and analyzed and on appreciation of the subject matter.



Item	Thickness of section		Standardized			*Bins*	*Frequency*
	All	*Peat*	*All*	*Peat*			
1	47.5	1.6	-0.98	0.14			
2	76.5	2.3	0.92	1.68		*All*	
3	58	1.1	-0.29	-0.96			
4	53.7	1.3	-0.57	-0.52		40	0
5	41.4	1	-1.38	-1.18		50	5
6	46	1.7	-1.08	0.36		60	5
7	92.7	1.9	1.98	0.80		70	4
8	62	1.7	-0.03	0.36		80	3
9	44.4	0.9	-1.18	-1.40		90	2
10	63.9	2.3	0.10	1.68		100	1
11	85.5	2.4	1.51	1.90			0
12	76	1.6	0.89	0.14			
13	86	1.3	1.54	-0.52		*Peat*	
14	66	1.1	0.23	-0.96			
15	69.2	1.7	0.44	0.36		0.5	0
16	59	1.5	-0.23	-0.08		1	3
17	43.6	0.8	-1.24	-1.62		1.5	7
18	51	1.4	-0.75	-0.30		2	7
19	55.5	1.7	-0.46	0.36		2.5	3
20	71	1.4	0.56	-0.30		3	0
							0
Total	1248.9	30.7					
Average	62.4	1.5	0	0			
St devn	15.3	0.45	1	1			
Correlation coefficient				0.56			

Fig. 1. Calculation of simple statistics with a spread-sheet. The total thickness of Pleistocene and Recent sediments were recorded at twenty boreholes, together with the thickness of peat in each. Simple statistics were calculated with a Microsoft Excel spreadsheet to examine the frequency distributions and their statistical correlation. See also Fig. 3.

Statistics are normally calculated by computer, particularly if the datasets are large. Good programs are readily available. The most flexible, although not the easiest to use, are subroutine libraries. The computer program normally calculates the values of statistical parameters using mathematical shortcuts. However, for teaching or exploratory purposes, spreadsheets and bar charts show intermediate steps and their effects on the individual items. For instance, they can show the original measurements converted to standard deviations from the mean (columns D and E of Fig. 1) and the user can examine the measurements in a local framework that may clarify relationships.

Some statistical programs help the user by providing an account of each method, a description of the algorithm, and examples of its use. The examples are unlikely to refer to geoscience, but it can be helpful to take an example as a template, and replace its variables and data with your own. An excellent range of textbooks is available on statistical methods and their applications. I have no plans to add to them, but do wish to point out the place of such techniques in geoscience investigations and to indicate some assumptions that constrain their application. The calculations of mean and standard deviation make no such assumptions. Their interpretation, however, raises many questions. The properties of the sets of beds constitute the **population** (D 4), as opposed to the actual measurements, which constitute a **sample** of the population. Sampling theory helps to clarify the link, so that conclusions about the population can be drawn from the sample, if appropriate sampling procedures (D 4) have been followed.

Circumstances determine the appropriateness of statistics. For example, an average thickness calculated from 49 siltstone beds and one very much thicker conglomerate bed would not be helpful. The result would alter greatly if we arbitrarily included another thick bed. A better procedure would be to study the thickness of conglomerate separately. Statistical measures make sense only for a clearly defined and coherent population. The **frequency distribution**, that is the pattern of relative frequencies of each measured value, can be examined on a bar chart or frequency plot (Fig. 1). Ideally, the frequencies are greatest in the center and fall off on either side to give the symmetrical bell-shaped frequency distribution of the so-called **normal distribution** (Fig. 2). A surprising number of actual distributions approximate to this, perhaps after a simple transformation such as replacing the original values by their logarithms (F 5). It then makes sense to describe the distribution as a whole with a few numbers, such as mean and standard deviation. Otherwise, **robust statistics**, described in most modern statistics texts, offer a less complete means of description but make fewer assumptions.

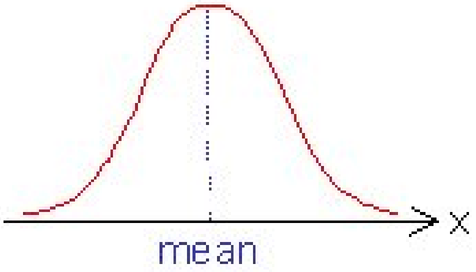


Fig. 2. Bell-shaped frequency curve of a normal distribution. The values of a variable that are deflected from their expected value (the mean) by many random events, might be expected to show this type of distribution.

With some assumptions about the distribution and sampling scheme, it is possible, for example, to calculate the likely population mean and variance from the sample, and the probability of their lying within a particular range. A technique known as **analysis**

of variance can show how mean values relate to sources of variation. For instance, if the ratio of Ca to Mg were determined in a number of samples, it could be of interest to see how it varied between formations, or between lithologies, or between analytical laboratories. With a carefully designed investigation, analysis of variance might be able to separate out the effects of each. Examination of the frequency distribution may however suggest a more complex situation, such as populations of different characteristics being sampled together. Descriptive statistics could then mislead by obscuring the real complexity.

Presumably measurements are made in order to draw some conclusions or to check some hypothesis. The conclusions must refer to something beyond the measurements themselves. Not “here are fifty beds that I have measured”, but rather “these beds are noticeably thicker than their counterparts farther east, and the beds are thicker and the grain size coarser towards the base of the succession”. Hypotheses about directions of sediment movement or deepening of the basin might in turn have prompted an interest in these findings. The hypotheses must be linked to the more general concepts in which they are embedded, and may lead to a mathematical model.

The analogy between the number field and the measurement of properties extends to the **mathematical model** - an analogy between mathematical operations (operating on the numbers) and physical processes (affecting the properties). Thus, adding together the thickness of beds in a vertical section is equivalent to finding their total thickness, a reflection, perhaps, of the total deposition of sediment at that point. Dividing the total by the number of beds to find the average is equivalent to recreating the original number of beds, but all of the same thickness. If nothing else, this may remind us that there is nothing magical about calculation. However, mathematical operations can mimic aspects of quite complex physical operations, often surprisingly well.

If you develop one or more quantitative models before or during data collection, you can statistically compare the predictions of the models with the observed data to see if they conflict. This somewhat negative view is characteristic of scientific argument. If the observed values lie within the range of expected values, they give no indication that the model fails to match reality, and the model may in consequence be accepted. Acceptance is always tentative, for there may be other models that would also fit and would be more realistic in other ways. Quantitative models can help to investigate simple aspects of the process, such as: is it likely that the data reflect purely random events? Griffiths (1967) showed how salutary this approach can be. They can also be designed to throw light on the deep structures of the process (Wendebourg and Harbaugh, 1997). Science always seeks to disprove, and if data conflict with the predictions of the model, this could be taken as disproof of the model’s validity.

The **random model**, in which events proceed on the basis of chance alone, is widely used in statistics. It addresses the question of whether the results of analysis might merely be a consequence of random variation. If this can be ruled out, the argument for an alternative explanation is strengthened. A statistical model may have a **deterministic** element, giving precise values calculated from the mathematical model representing the physical process. It may also include a random element, reflecting the unpredicted, chance events (although as pointed out in J 2.3, some non-linear deterministic systems are inherently unpredictable). The random element makes it possible to specify not only the most likely values of the outcome of the model, but also a range within which the values are likely to occur, taking random effects into account.

If the investigation is more than simply an initial gathering of ideas, and the data will be widely shared, then the investigator should describe the procedures and state how observations and measurements were made (D 4). Another scientist following the same sampling scheme and procedures would not expect to obtain identical data, but would expect the overall statistics to match within the calculated margin of error. An account of the sampling scheme can also help others to decide how far to rely on the results. By invoking statistical arguments, the process of geoscience investigation and reasoning can be made more rigorous and repeatable.

Quantitative methods have another role to play in geoscience. Measurements may be made in an exploratory way, without clear ideas about which of a range of hypotheses is being tested, but with hopes of unearthing a familiar pattern or detecting an interesting relationship (Tukey, 1977). Many geoscience investigations are concerned, not with the operation of physical systems, but with geological events and the environment in which they occurred. Sets of measurements may throw light on spatial patterns and interrelationships that guide the formulation of hypotheses. In geoscience, data are commonly displayed as a map or set of maps. Computer visualization studies (Cleveland, 1993, Gallagher, 1995) address the wider question of how to display data to detect pattern (G 2). They normally start with quantitative data and explore graphical means of conveying their significance to the human eye and brain. Field mapping faces the same problem of detecting pattern among a host of interrelated properties. It too can be seen as a form of spatial visualization. Statistical methods can then help in another way. They may offer concise summaries (like the mean) and reveal relationships that otherwise might not be noticed. They are unlikely to offer rigorous proof of the conclusions, but might point you towards a conclusion that you could test or support by other lines of argument more familiar in geoscience.

4. Matrix algebra and spatial data

The development of **coordinate geometry** by Descartes in the early 17th century gave the basis for spatial visualization of quantitative data, but also meant that spatial data could be brought within a quantitative framework. Position in space is measured by **coordinates** - distances from a zero point known as the **origin** along each of a set of **axes** at right angles. In consequence, spatial data can be manipulated, analyzed and managed on the computer. A range of computer techniques can be applied to information that would normally be recorded on maps and cross-sections. Those of us who think more readily in pictures than in numbers can make use of the correspondence between numbers and position, and between algebraic and geometric operations, as an easy way to gain an understanding of statistical methods. Many geoscientists may find it easier to visualize quantitative techniques as manipulation of points in variable space rather than as manipulations of numbers.

The link between computation and space, between algebra and geometry, is perhaps most obvious in matrix algebra, which enables a sequence of related operations to be written as one operation. Matrix algebra is an extensive and complex study in its own right, and is widely used in quantitative geoscience. Most computer users who are not programmers themselves, can understand the results without understanding the details of the method. A few words of explanation here may make the process less mysterious to those who lack the mathematical background.

A table of quantitative values, such as those set out in a spreadsheet, can be regarded as a **matrix**. A single row of the matrix can be referred to as a row **vector**, and a single column as a column vector. The individual values or **elements** of the matrix are referred to in the spreadsheet by a letter indicating the column and a number indicating the row. In matrix algebra, the notation is slightly different, with the row and column both indicated by numbers. Letters are used, not to designate a specific column, but as placeholders that can be replaced by any number. Algebraic statements using the placeholder (or **index**) are thus quite general, and not tied to any particular numeric values. The matrix as a whole can be referred to by a name, in algebra usually a capital letter in bold type. The element has the same name, in lower case with the row and column numbers as suffixes. Thus the element x_{ij} is in row *i* and column *j* of the matrix **X**. A typical notation in a programming language is X(i,j) where X is a name that could be several characters in length.

Matrices of the same size, that is, the same number of rows and columns, can be added. **Z**=**X**+**Y** means that each $z_{\mathrm{ij}} = x_{\mathrm{ij}} + y_{\mathrm{ij}}$. Subtraction follows the same pattern. Multiplication, **Z** = **X**.**Y**, requires that the number of columns in **X** is equal to the number of rows in **Y**. The element z_{ij} is found by adding the products of the elements in the *i*th row of **X** with the corresponding elements in the *j*th column of **Y**.

$$z_{\mathrm{ij}} = x_{\mathrm{i1}}.y_{\mathrm{1j}} + x_{\mathrm{i2}}.y_{\mathrm{2j}} \;.\;.\;.\; x_{\mathrm{in}}.y_{\mathrm{nj}}$$

The ellipsis (. . .) indicates continuation of the preceding series in the same way, and n is the number of columns in **X**.

A problem frequently encountered in statistics and in some geophysical topics is that of solving a set of simultaneous equations, which could be written as **A** = ***x***, where ***x*** is a column vector. In matrix notation, the general solution is $\mathbf{A}^{-1}$. Thus matrix algebra is useful where each new value is dependent on several existing values in some systematic way. It provides a more compact and so more powerful notation than writing out each individual operation.

Returning to coordinate geometry, let us suppose that *x*, *y* and *z* are variables holding the latitude, longitude and elevation of a point in space. They can first be brought to the same units, say meters from an origin at mean sea level on the lower left corner of a map sheet. A set of points within the map sheet, say where samples had been collected, could be numbered with an index *i*, which takes the values 1, 2, 3 . . . *n*. As similar types of measurement are made for each sample, it is convenient to refer to them all in the same way. Each has an *x*, *y* and *z* value to indicate its location. To identify each sample, it can be given a suffix; thus the *i*th sample is located at x_{i}, y_{i}, z_{i}. The three values together, placed in brackets (x_{i}, y_{i}, z_{i}), form a vector, in the sense of a set of numbers referring to the properties of an object. In this case, because the elements of the vector are geometrical coordinates, (x_{i}, y_{i}, z_{i}) also denotes a vector in the geometric sense - a line from the origin (0,0,0) with length, orientation and direction.

As *x*, *y*, and *z*, once they are measured in the same units, refer to similar things, it is convenient to refer to them with the same letter, say x_1, x_2, x_3 (or x_{j}, where *j*=1,2,3). The values of *x* then have two suffixes, and the values can be arranged as a table with rows *i*, numbered from 1 to *n* and columns *j*, numbered from 1 to 3. In Fortran, the matrix is referred to by the more general term of an **array**. It is one-dimensional if it has one suffix, two-dimensional if it has two, and so on, whether or not this is its geometric dimension. The geometric operations that might be applied to these vectors are described in G 4. Their algebraic equivalents may involve changing each of the values of each row vector (x_{i1}, x_{i2}, x_{i3}) in a way that depends on all three of the current values. If the corresponding values after the operation are called y, then we can write

$$y_{i1} = ax_{i1} + bx_{i2} + cx_{i3}$$

with similar equations for y_{i2} and y_{i3} , making three in all. Rather than referring to the constants, such as *a*, *b* and *c*, with separate letters, they can be seen as a matrix in their own right with three rows and three columns, say **T**. The transformation of the entire data matrix **X** to new values **Y** can then be written as **Y** = **XT**. Some important geometric operations have equivalents in matrix algebra that can be implemented on a computer system. As described in G 4, the transformation matrix **T** can represent familiar operations of moving objects about and changing their shape. It is a basic tool in creating computer maps and multidimensional spatial models.

5. Multivariate statistics

The link between numbers and space, between algebra and geometry, works in both directions. Spatial features can be represented by numbers; quantitative data points can be visualized as a cloud of dots. They can be manipulated in a space where the coordinate axes, at right angles to one another, are marked off in the units of measurement of the individual variables. The units refer to any measured property, such as bed thickness, grain size, gravity or uranium content. There is no limit to the number of axes, but we have trouble visualizing more than three at a time, as we appear to live in a three-dimensional world. Visualization may help us to understand the statistical relationships of a set of variables (see Cook, 1998). Statistics is concerned not just with single variables and comparison of different sets of measurements of the same variable, but also with the relationships between different properties of the same objects. This leads to techniques of **multivariate** analysis (a variate is a variable that exhibits some random variation).

Given a set of quantitative data, say, a collection of measurements of fossil shells, statistical methods are a guide to possible conclusions. Many different properties could be measured on each shell, such as length, breadth, thickness, length of hinge-line, and number of growth lines. We might wish to investigate how the properties are related to one another. We might also need some means of measuring the characteristics of the set of measurements as a whole, to compare them with another set from a different locality. The task has moved from comparing individual measurements to that of comparing aggregates of measurements, that is, a set of distinct items gathered together.

A starting point, however, is to look at the characteristics of each variate in terms of statistics such as the mean and standard deviation (F 3). Each variate can then be **standardized** to zero mean and unit standard deviation. This frame of reference may make it easier to compare their relative variation. The cloud of standardized data points is centered on the origin and has equal spread or dispersion along each axis. Measures, such as skewness and kurtosis, based on the third and fourth powers of the deviations from the mean, can be calculated to assess the symmetry and shape of the frequency distribution of each variable (F 3). However, there is no substitute for their visual inspection with a bar chart, histogram or scatter diagram.

Some frequency distributions are quite unevenly distributed about the mean, such as the grain size of sediments or thickness of beds in a vertical section. A **log transformation**, which compresses the scale at the higher end, can bring the distribution to a more tractable form. Many other transformations are possible, and may be justified simply because the subsequent analysis is more straightforward. It is more satisfying if there is a physical justification for the transformation. For example, if an organism doubles in size each year, the distribution of size at random times will be logarithmic, reflecting a multiplicative rather than an additive process. Replacing the original measurements by their logarithms converts the numbers to a simple straight-line distribution.

Statistical reasoning, as opposed to description, tends to assume that variates approximately follow the so-called **normal distribution** - the familiar bell-shaped curve of Fig. 2. Under a number of assumptions, it is possible to compare the actual sample with that expected from a random set of events, and determine the likelihood of this "**null hypothesis**" being incorrect. A number of excellent textbooks, including some for geoscientists (for example, Davis 1973), give fuller information on these powerful methods.

The relationship between two variates can be measured by the correlation coefficient (F 3), or by a regression equation. The **regression** equation predicts the value of one variable (y) from that of another (x). The regression equation describes a line, using the formula $y=a+bx$, selecting the values of a and b to minimize the sums of squares of deviations between the measured and calculated values of y (see Fig. 3). The average squared deviation is a measure of the closeness of fit of the data to the line. The line that best fits the data could be regarded as a mathematical model of the relationship. As before, transformations that give the individual distributions a shape like the normal distribution are helpful. The null hypothesis of no correlation can be tested, given a number of assumptions, and only if it fails is there a need for a scientific explanation of the relationship.

The situation becomes more interesting where several variates are measured on the same items. You may be able to visualize the data as a cloud of dots in n dimensions, each dot representing one item, and each axis representing the standardized measurements for one variate. This is easiest with 2 or 3 variates, but the calculations are similar in higher dimensions. Each axis is regarded as independent of the others, which in geometry means that they are shown at right angles to one another. View the cloud in any two dimensions, and a correlation may be apparent that can be measured by the correlation coefficient. The correlation coefficients can be calculated for each pair of variates separately and shown as an n x n matrix, where n is the number of variates. There may be underlying processes that affect many of the variates together. Possibly there are several such processes affecting different variates in different ways. The matrix of correlation coefficients may throw light on the structure of the data which in turn might suggest underlying causes.

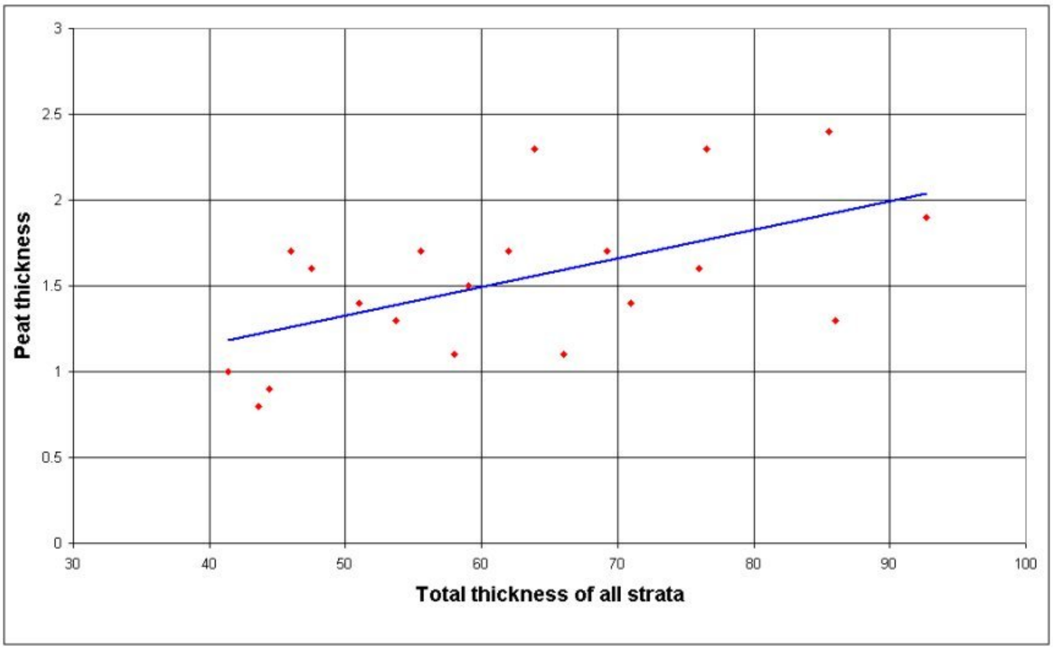


Fig. 3. Line of best fit between two variables. The data of Fig. 1 are plotted here to examine their correlation, and a regression line added. The chart was prepared from the spreadsheet with Microsoft Excel software.

One procedure for analyzing the correlation matrix is known as **principal component analysis.** It simply rotates the cloud of points in *n* dimensions (try visualizing it in three), until the largest variance is along one axis (the first principal axis), the greatest remaining variability along the second, and so on. The least variance is that around the final principal axis. The number of principal axes is the same as the original number of variates. They are still at right angles, but have been rotated together, as described, to a new orientation. When the points are referred to the new frame of reference (the principal axes), the new variates are known as the principal components. It is likely that most of the total variance will be accounted for by the first few components. If the remainder show no interesting pattern and appear merely to reflect random effects, they can be disregarded. The result from the principal component analysis (PCA) program is thus a smaller set of new variates and a statement of how much of the variance each represents. The relative contribution of each of the original variates to each principal component is defined. The challenge for the geoscientist is to decide whether the principal components reflect underlying causal processes, or if not, how they might be accounted for. For example, the measurements of many aspects of the size and shape of fossil shells might be related to a few features of the environment like wave energy, nutrient availability, depth and clarity of water. This approach has been extended and elaborated as **factor analysis** (see Reyment, Jöreskog, 1993).

As well as the correlation coefficient between two variables, we noted the regression equation as an alternative way of looking at the relationship. This again is not limited to two variates. An equation $y = a + bx_1 + cx_2 + . . . + gx_n$ representing a straight line in *n* dimensions, can be fitted to *n* variates x_1 to x_n, so that the value of the selected variable *y* can be predicted from all the *x* variates, minimizing the total sum of squares of differences between its measured values and the values predicted by the equation.

Unlike PCA, which treats all variates alike, **multiple regression** focuses on one variate, with all the others seen as contributing to its variation.

As an aside, if the number of terms in a regression equation is greater than the number of data points, additional information is required to give a unique equation. Methods of **linear programming** achieve this by introducing an objective function that must be maximized to yield the greatest benefit to the system. This has applications in allocating raw materials to different products, as in models that allocate chemical elements to the mineral constituents of a crystallizing igneous rock. The more usual statistical case is overspecified, with many more data points than terms in the equation, and the least-squares criterion just mentioned, or a similar alternative, is used to arrive at a unique surface.

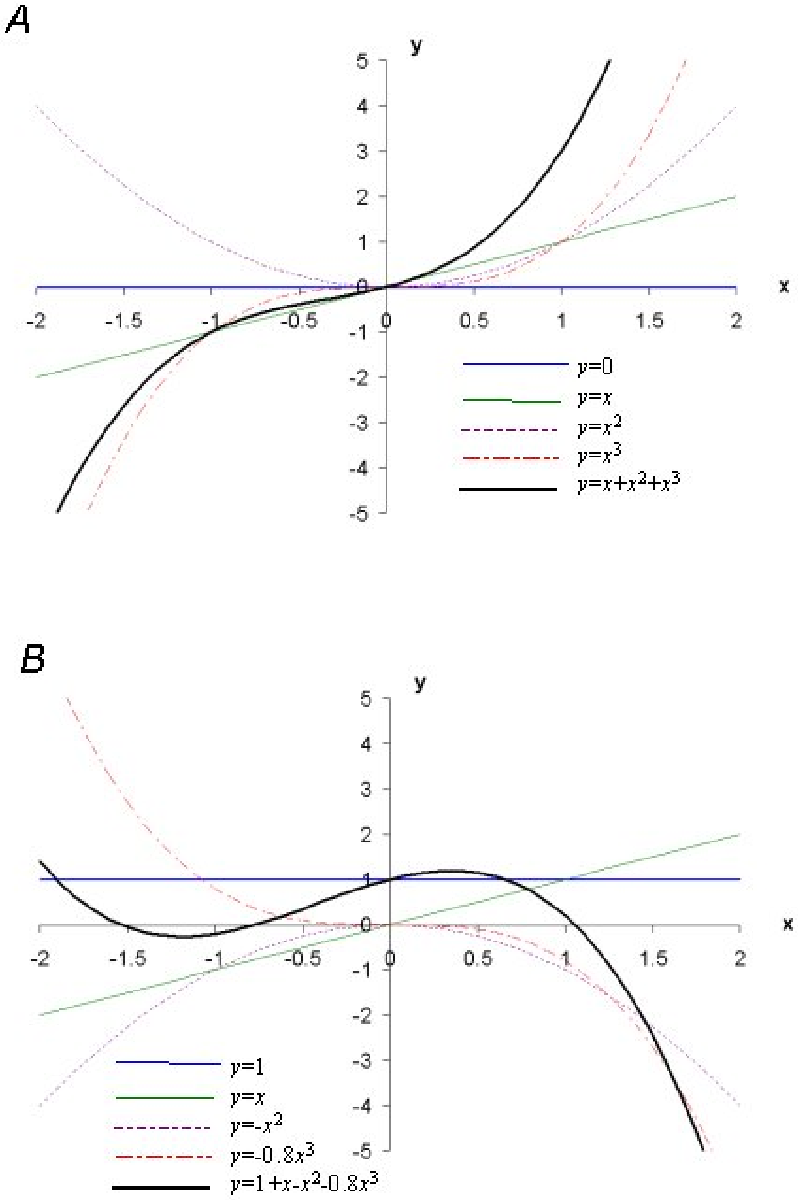
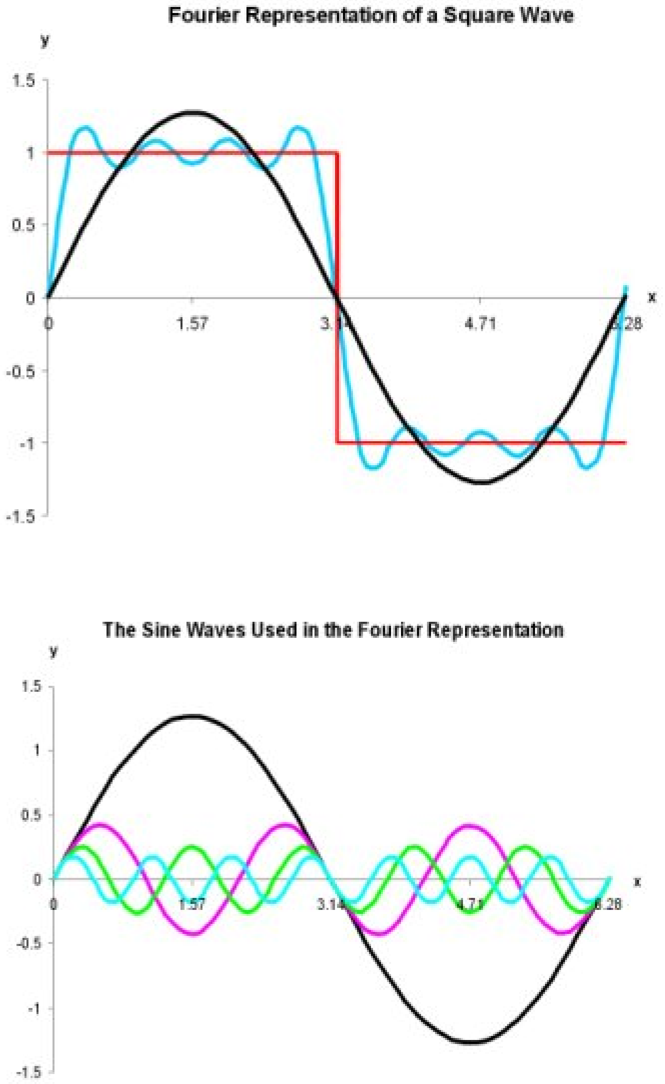


Fig. 4. Generation of polynomial curves. The basis functions for a cubic polynomial, and the combined curve from adding them together, are shown in A. In B, the coefficients are altered to give a different curve, which is still smooth, has the same number of inflection points, and heads for plus or minus infinity at each end.

The form of the regression equation suggests how it is possible to fit curves other than straight lines to a relationship between two variables *x* and *y*. From the values of *x*, it is a straightforward task to calculate x^2, x^3, . . . x^n. We can then write an equation similar to that given above:

$$y = a + bx + cx^2 + . . . + gx^n$$

If we look at a graph of the powers of *x* (Fig. 4), we see that adding them in different proportions can generate quite complicated curves.



Fourier Representation of a Square Wave

The Sine Waves Used in the Fourier Representation

Fig. 5. Complex periodic curve. The upper diagram shows how sine waves can be combined to approximate even awkward shapes, such as a square wave. A single wave offers a first approximation, which can be improved by combining it with appropriately weighted harmonics, shown individually in the lower diagram.

Many natural sequences are periodic, retracing a sequence again and again, like rotations of the Earth around the Sun. This type of sequence can be mimicked mathematically by a series in which, instead of an *x* value increasing along a line, we take an angle θ measured out by a radius rotating around a circle from 0 to 360°

repeatedly. As it rotates, the sine of the angle θ changes from 1 to 0 to -1 and back again. A complex periodic curve (see Fig. 5) can be generated by taking, not powers of x, but sines of multiples of the angle θ:

$$y = a + b\sin\theta + c\sin 2\theta + \ldots + g\sin n\theta$$

The power series and the sine and cosine series have the mathematical property that successive terms have a smaller influence as the series continues. They are therefore suitable for approximating to an arbitrary curve. The slope and curvature at any point are readily calculated from the equations, and the form of the power series is suited to statistical calculation. They are not appropriate, however, for extrapolating the relationship beyond the data points. The periodic curve repeats itself indefinitely, and the power series heads off to infinity.

Statistical tests (to which power series are well suited) can measure the "goodness of fit", that is, how well the curve fits the data, compared with expectations from a random relationship. The test is based on a number of assumptions, notably that a random sample has been obtained from an underlying distribution that is close to the normal, bell-shaped, curve. It follows that the sample is expected to be drawn from a single, homogeneous population. There are situations where we suspect that this is not the case. Our sample could have been drawn, without our knowing it, from populations that were formed in different ways by different processes, and they might have quite different properties. It is convenient, therefore, to have a means of searching for different groups within the dataset.

Cluster analysis does this by looking for the most similar items in a dataset, combining them as one item, looking for the next most similar items, and so on until all the items in the dataset are combined. The similarity of two items, or its opposite, can be measured in various ways, such as the distance between them in standardized variate space. If the dataset is homogeneous, the clustering will proceed uniformly. If there are a number of natural groups or clusters, then the clustering is more likely to proceed with sudden breaks. Closely similar items are brought together, followed by a break before the next most similar items are found. This is a hierarchical process with clusters of clusters amalgamating as bigger clusters. Further examination of the characteristics of each cluster may suggest why they fall into groups (different species, different environments, different weathering, and so on). Cluster analysis can point to the existence of non-homogeneous populations, and lead to better analysis. If it is known before the investigation that several groups are present, **discriminant analysis** (see Davis, 1973) provides equations for assigning new items to appropriate groups. Techniques of this kind are used in **numerical taxonomy**, where measurements of sets of properties are the basis for classification. They may not be appropriate where objects are classified on the basis of an underlying qualitative model, as is often the case in geoscience (J 2.3).

Most multivariate techniques can simply be regarded as arithmetic transformations of a set of numbers, and do not necessarily require any underlying assumptions. The results may suggest ideas to a geoscientist who can then proceed to test them by other means. Calculating the descriptive statistics is then purely an exploratory exercise. Visualization, by displaying patterns through interactive graphics, follows this approach (see Cleveland, 1993). However, statistical tests of significance, and indeed any conclusions that depend on the numbers themselves, almost certainly imply some assumptions. Perhaps the most important and the most difficult requirement is to ensure that the items recorded (the sample) are truly representative of the population about which the conclusions are drawn. This applies of course not just to quantitative measurements but to any observation of the natural world. There is, however, a danger that in the course of carrying out the complex manipulations of the data, original constraints and limitations are forgotten. The subject matter is all important.

6.References

Cleveland, W.S., 1993. Visualizing Data. Hobart Press, Summit, New Jersey, 360pp.

Cook, R. D., 1998. Regression Graphics: Ideas for Studying Regressions through Graphics. Wiley, New York. 349 p.

Davis, John C., 1973. Statistics and Data Analysis in Geology: with Fortran Programs. Wiley, New York, 550pp.

Gallagher, R.S. (Ed), 1995. Computer Visualization, Techniques for Scientific and Engineering Analysis. CRC Press, Boca Raton, 312pp.

Griffiths, J. C., 1967. Scientific Method in Analysis of Sediments. McGraw-Hill, New York, 508pp.

Krumbein, W.C., Graybill, F.A., 1965. An Introduction to Statistical Models in Geology. McGraw-Hill Inc., New York, 475pp.

Reyment, R.A., Jöreskog, K.G. (Eds.), 1993. Applied Factor Analysis in the Natural Sciences. Cambridge University Press, New York, 371pp.

Swan, A.R.H., Sandilands, M., 1995. Introduction to Geological Data Analysis. Blackwell Science, Oxford, 446pp.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, Mass., 499pp.

Wendebourg, J., Harbaugh, J.W., 1997. Simulating Oil Entrapment in Clastic Sequences. Computer Methods in the Geosciences, 16. Pergamon, Oxford, 199pp.

Disclaimer: The views expressed by the author are not necessarily those of the British Geological Survey or any other organization. I thank those providing examples, but should point out that the mention of proprietary products does not imply a recommendation or endorsement of the product.

<<<Back to Table of Contents

On to Part G: Familiarization with spatial analysis>>>