AAPG Computer Applications Workshop Notes

J. H. Doveton
J. C. Davis
R. J. Sampson

1975

Kansas Geological Survey
Open-File Report OFR 75–15

1. INTRODUCTION

SCALES OF MEASUREMENT

Scientific observations are related to four scales of measurement which are named nominal, ordinal, interval and ratio (listed in order of increasing information content).

Nominal Scale

Assignment to discrete categories which have no implicit ordering and no metrically defined boundaries; e.g., rock types, fossils.

Ordinal Scale

Discrete categorization with an inherent ordering; e.g., Moh's scale of hardness, stratigraphic age scale.

Interval Scale

Continuous or discrete numerical measurement in which distances between objects can be measured, but cannot be related to an absolute zero; e.g., phi size scale, structural elevation.

Ratio Scale

Continuous or discrete measurements with a definitive absolute zero; e.g., mass, density, length.

(1) Nominal and ordinal scales apply to non-metric discrete categorical data; interval and ratio scales are for metric discrete and continuous measurements.

(2) The greater information content of the higher grade scales extends the range of permissible statistics used to summarize the data and the precision of statistical inference based on them.

Reference

Griffiths, J.C., 1960, Aspects of measurement in the geosciences: Mineral Industries (Penn State), v. 29, no. 4, p. 1-8.

STATISTICS

Descriptive statistics

A variety of measures aimed at summarizing the characteristics of data sets (means, variances, correlations, etc.) together with pictorial representations of the data distributions (histograms, scatter plots, etc.).

Inferential statistics

The process of making generalizations or predictions concerning the phenomenon under study based on raw measurement variation and relationships between measured variables. Conclusions are drawn from limited information and used for making decisions under uncertainty. The logic is inductive, as inferences concerning the general are derived from a study of the observational particular.

All the values of interest (the universal set) are termed the population, for which summary measures are precise characterizations of the studied variables. These measures (mean, variance, etc.) are the parameters of the population.

It is usually only practical to measure a limited sample of the total population. Statistical measures of a sample are known as sample statistics and are estimates of the parameters of the parent population.

The sample must be representative of the total population in order for sample statistics to provide unbiased estimates of parameters. Random sampling provides a means by which every object in the population has an equal chance of being selected in the measured sample.

Parameters are conventionally denoted by Greek letters; sample estimates by Roman.

Univariate statistics are concerned with summarization and inferential analysis of a single variable measured on a sample of objects. Multivariate statistics marks an extension to several variables of measurement on each object and is the numerical description of variable interrelationships and inferences drawn from them.

The choice of descriptive and inferential methods is dictated largely by the scale of measurement of the observational variables and the geometric form of their distribution.

2. MEASURES OF CENTRAL TENDENCY

A statistic that expresses a typical or average value of a distribution is a measure of centrality and is the most basic descriptor of a distribution. Listed in order of increasing precision of information, the most commonly used statistics are the mode, median, and mean.

Mode (Mo)

The mode is the most frequently occurring value in the distribution. It can be applied to all four measurement scales and is the only available centrality measure for nominal data. The mode represents the most typical value. A distribution may have several modes or the mode may have marginally greater frequency than other categories with the result that the mode may not be a very stable estimate of centrality.

Median (Md)

The median is the point on the variable reference scale at which the distribution is divided into equal halves, with 50% of the data having lower values than the median, 50% higher. The measure is applicable to the ordinal, interval, and ratio scales. The median is preferred to the mode for ordinal scale data since it utilizes the order information in the distribution, which the mode does not.

Mean

The mean is the arithmetic average which equals: $\frac{\text{the sum of the values}}{\text{the number of values}}$.

If N is the total number of objects in the population, then the population mean μ equals

$$\frac{\sum_{i=1}^{N} X_i}{N}$$

Similarly, for a sample of smaller size n, the sample mean $\bar{X}$ equals

$$\frac{\sum_{i=1}^{n} X_i}{n}$$

From a mechanical viewpoint, the mean corresponds to the center of gravity of the distribution. The mean is the most sensitive measure of centrality and is applicable only to interval and ratio scaled data. The real power of the mean, as compared with the mode and median, is that it has strong theoretical relationships with other statistics used in basic

inference procedures. The mean is the expected value of the distribution, which will be observed as a long-term average. Symbolically, $\mu = E(X)$. By contrast, the mode and median are essentially descriptive statistics which are used primarily in the nominal and ordinal scales where calculation of the mean is precluded.

3. MEASURES OF VARIABILITY

Statistics of variability or dispersion express the degree of spread of values in a distribution about the central measure (mode, median or mean). A dispersion statistic indicates how well the central measure represents the distribution. If dispersion is small, the central measure is close to most of the distribution values; if dispersion is large, the central measure is only an intermediate representation of a wide range in values.

The dispersion statistics variation ratio, interquantile range, and variance are matched with their compatible measures of central tendency, the mode, median, and mean, respectively.

Variation ratio

Linked with the mode and the most common dispersion measure for nominal data. The variation ratio, v, indicates the degree to which the mode represents a sample;

$$v = 1 - \frac{f}{n}$$

where f = frequency in modal category and n = total number of values in the sample.

Interquartile range

Used in conjunction with the median primarily for ordinal data and strongly asymmetric continuous distributions. The range is the distance between the smallest and largest values in the distribution. The decile range is the distance between C_{10} (the point demarcating the bottom 10% of the distribution) and C_{90} (the corresponding point cutting the distribution at the top 10%). Similarly, the quartile range is the distance between C_{25} and C_{75}.

Variance

The variance and standard deviation (square root of the variance) are measures of variability used with the mean and are appropriate for continuous data (interval and ratio scale).

If the mean is subtracted from all the values in the distribution, the resulting arithmetic deviations are distances from the centroid of the distribution. The sum of all the deviations is necessarily zero if positive and negative signs are retained. When absolute values of the deviations are summed and divided by the number of values, a mean absolute deviation (MAD) may be calculated. MAD is restricted to a role

as a descriptive statistic and is intractable for statistical inference computations. Instead, the sum of the squared differences between the mean and the observations are divided by the total number of observations to obtain the variance.

In the case of a population, the variance, $\sigma^2 = \frac{\Sigma (X_i - \mu)^2}{N}$ and the standard deviation, $\sigma = \sqrt{\sigma^2}$.

For a sample, the variance, $s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$ and the standard deviation, $s = \sqrt{s^2}$. Just as the mean represents the center of gravity of the distribution, the standard deviation corresponds to the radius of gyration. Notice particularly that the denominator in the variance expression is N for the population and (n-1) for the sample. The reason for this is that in the sample calculation, the sample mean is almost certain to be closer to sample values than the population mean. In most instances we do not know the population mean but are forced to estimate it with $\bar{X}$. To correct for the bias that is introduced by this estimate we divide by (n-1) which corresponds to the number of degree of freedom (q.v.). When n is large, the difference between the biased and unbiased variance estimates becomes arithmetically trivial, but the distinction is important in small samples.

The procedure for sample variance computation is simplified by use of the following formula:

$$s^2 = \frac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}$$

If the biased variance estimate is used (in cases of large n), this formula reduces to:

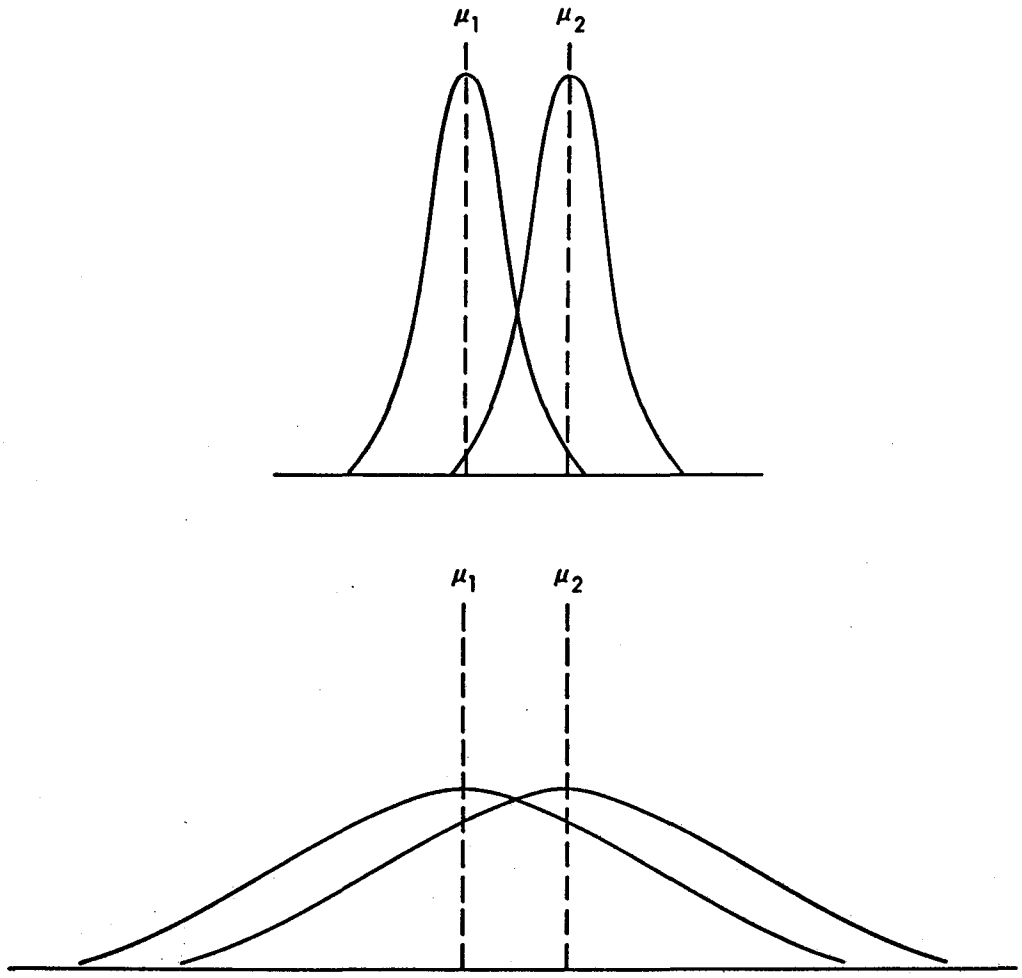
$$s^2 = (\frac{1}{n} \Sigma X^2) - (\bar{X})^2$$

which may be remembered by the phrase "mean square minus square mean."

The mean and variance are of central importance as both descriptive statistics and key elements in inferential statistics.

(1) They are used as the basic descriptors of all discrete and continuous distributions.

(2) Statistical hypotheses directed towards the distinction or similarity of sample distributions take account of both the mean and the variance (e.g., t-test, F-test). Both measures are necessary for these operations as indicated in the figure.



Overlap of two pairs of distributions with equivalent means but different variances.

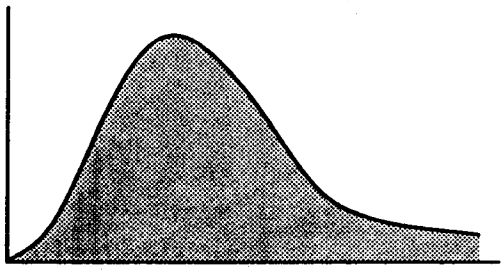
(3) For samples described by several variables, the means, variances and covariances (variation of two variables together) are used as the fundamental geometrical descriptors of the locations and dispersion of multivariate distributions. Matrix algebra operations are used to define principal axes in this high-order dimensional space as summaries of total variation as in principal component analysis or for linear classification as in discriminant function analysis.

4. MOMENTS OF A DISTRIBUTION

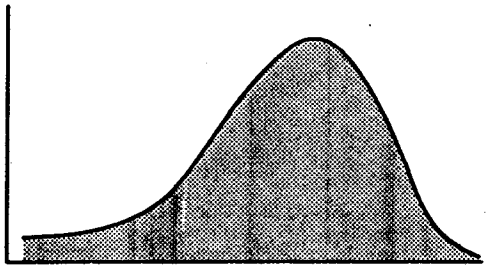
The mean and variance are appropriate measures of location and dispersion of metric-scaled distributions. Higher order measures may be calculated to characterize the symmetry of the distribution (skewness) and its "peakedness" (kurtosis). The mean, variance, skewness and kurtosis are often known as the first four moments of a distribution.

The kth moment, $m_k = \frac{1}{n}\sum(X-\bar{X})^k$. The first moment about the origin is the mean. The second moment about the mean is the variance. ("Second" merely designates the power to which the deviations are raised.)

The third moment about the mean is the skewness, $m_3 = \frac{1}{n}\sum(X-\bar{X})^3$. For a symmetrical distribution, $m_3 = 0$. If the value of m_3 is greater than zero, the distribution has a positive skew or is "skewed to the right;" if less than zero, the distribution has a negative skew or is "skewed to the left."



positive skew



negative skew

The units used to measure X influence the size of m_3 and so a dimensionless measure of skewness, Sk is computed by the ratio:

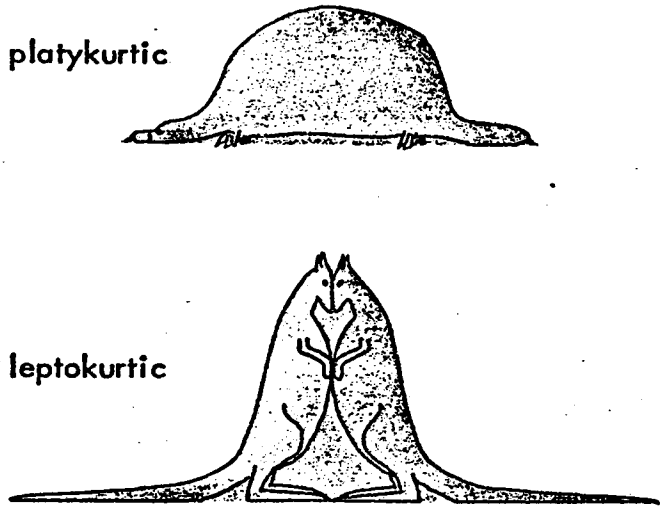
$$Sk = \frac{m_3}{\sqrt{m_2^3}}.$$

The fourth moment about the mean, m_4 is used to measure the kurtosis. Dimensionality is eliminated by computing kurtosis, Kt, by the ratio:

$$Kt = \frac{m_4}{m_2^2}.$$

The standard distribution used for purposes of comparison is the normal distribution which has a kurtosis of three and a skewness of zero.

Distributions with higher kurtosis are more "peaked" and are termed leptokurtic; those with lower kurtosis are "flatter" and are platykurtic. The cartoon by the famous statistician "Student" (pen-name of W.S. Gosset) is an aide-memoire for this terminology; kangaroos are featured since they are noted for "lepping" (the platypus speaks for itself).



The sample measures of skewness (Sk) and kurtosis (Kt) are often denoted by $\sqrt{b_1}$ and b_2, respectively. They constitute statistical estimates of the population parameters $\sqrt{\beta_1}$ and β_2 for which sampling distributions applicable to the normal distribution are tabulated by Pearson and Hartley (1954).

Reference

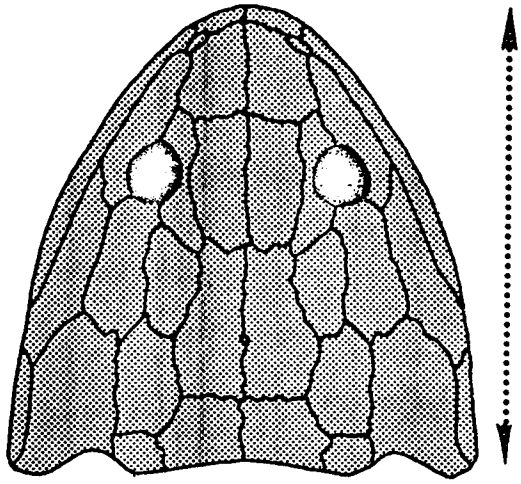
Pearson, E.S., and Hartley, H.O., 1954, Biometrika tables for statisticians, v. 1: Cambridge Univ. Press, New York, 238 p.

Example: Skull lengths of fossil amphibians

Skull dimensions of the amphibian *Trimerorhachis insignis* were measured from fossil remains in the Lower Permian Wichita Group by Olson (1953). The table contains the skull lengths of 33 specimens, representing a statistical sample of the hypothetical population of all *T. insignis* individuals. The lengths collectively describe a distribution whose moments provide concise measures useful in paleontologic interpretation. The most basic conceptual model applicable to this distribution corresponds to the hypothesis of simple random variation about an average skull size and is numerically described by the normal distribution. Deviation from normal distribution parameters of skewness and kurtosis may be selectively interpreted in terms of mortality factors, sexual dimorphism, taphonomic mechanisms, etc.

Skull lengths of Trimerorhachis insignis specimens (mm.)

53	60	61	66	70
71	72	73	76	76
77	78	78	80	81
85	90	91	92	92
93	93	98	100	104
106	108	116	120	120
123	138	139		



(data and figure after Olson, 1953)

n = 33

m_1 = 90.3 ($\bar{X}$)

m_2 = 465.1 (biased s^2)

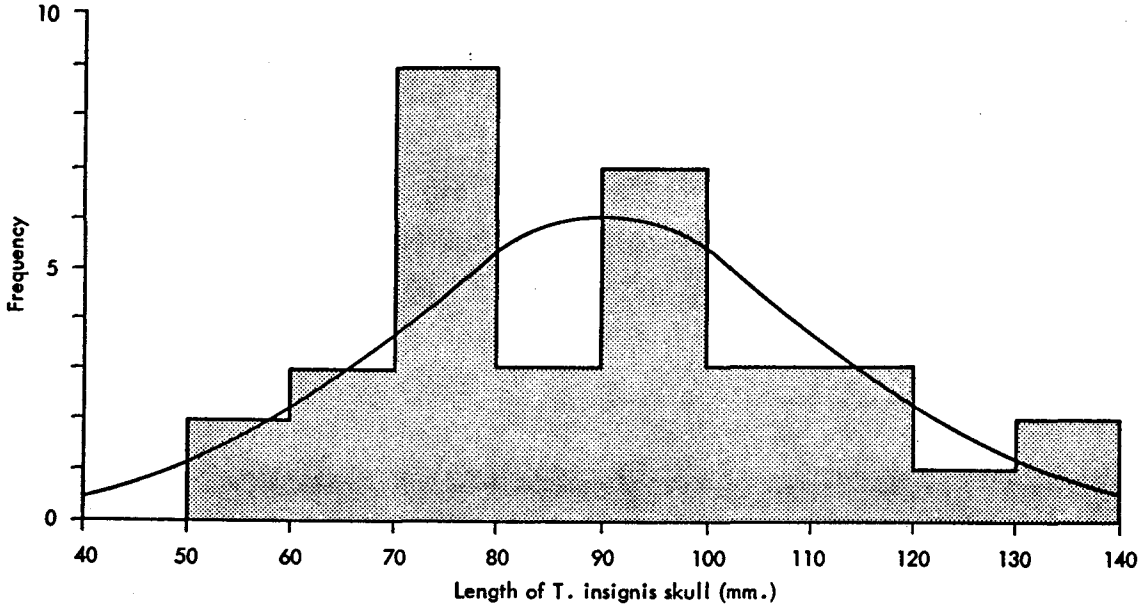
m_3 = 5506.9 Sk = 0.55 ($\sqrt{b_1}$)

m_4 = 566638.1 Kt = 2.62 (b_2)

In summary, the skull length distribution is positively skewed and platykurtic. A histogram of the raw measurements shows a distinctly bimodal character and that the peak of a fitted normal distribution coincides with the intermodal low. It is highly likely that the distribution represents samples drawn from two distinctive skull populations. This property is recognized by Olson (1953) who attributes the differentiation to sexual dimorphism where one group had shorter snouts and nasal tracts (relative to skull width) than another group.

Reference

Olson, E.C., 1953, Integrating factors in amphibian skulls: Jour. Geol., v. 61, p. 557-568.



Measures of skewness and kurtosis are widely applied in the characterization of sediment size analyses (together with the mean and standard deviation) in terms of textural indices. Sedimentologists often used broadly equivalent graphic measures of these descriptors based on cumulative interquantile statistics (median, quartiles, etc.) rather than the moment formulae used in this text. The statistical moments have the advantage of being related to the entire sample distribution rather than isolated percentile cuts and are grounded in statistical sampling theory. However, as noted by Folk (1968), moment computation is infeasible in cases of "open-ended" distributions, such as caused by residual pan fractions of unknown grain size.

The four moments of sand-size distributions have been extensively studied as textural characteristics that aid in the distinction of environments of deposition. Friedman (1961) found that dune and river sands tended to be positively skewed; beach sands to be negatively skewed. Beach sands are further differentiated from river sands by better sorting with consequent lower standards deviations. Dune sands are best distinguished by a combination of moments, and marginal cases are resolved by computation of a mean grain size ratio of quartz and a selected heavy mineral. Kurtosis did not appear to be diagnostically linked with environment. (These remarks relate to a phi-measurement scale of grain size.)

References

Folk, R.L., 1968, Petrology of sedimentary rocks: Texas Hemphill's Book Store, Austin, Texas, 170 p.

Friedman, G.M., 1961, Distinction between dune, beach, and river sands from their textural characteristics: J. Sed. Pet., v. 31, no. 4, p. 514-529.

Exercise 4.1: Grain sizes on Mazatlan beach

Mazatlan is a major Mexican fishing port situated at the mouth of the Gulf of California. Large cuspate beaches are developed along the coast near Mazatlan, built up by longshore drift of northward-moving currents. They can be considered to be reasonably typical examples of a "normal" oceanic beach environment. The beach sands are dominantly

Mazatlan beach quartz grain a-axis measurements

f = number of grains with valves
s = a-axis length (mm.)
$\phi = -\log_2 s$

f	s	ϕ
2	0.6	+0.737
2	0.7	+0.515
3	0.8	+0.322
2	0.9	+0.152
11	1.0	0.000
12	1.1	-0.138
5	1.2	-0.263
9	1.3	-0.379
1	1.4	-0.485
5	1.5	-0.585
1	1.7	-0.766
1	1.8	-0.848
1	2.0	-1.000
2	2.2	-1.138
1	3.0	-1.585
1	3.2	-1.678
1	3.6	-1.848
Total 60		

composed of quartz grains, shell shards and rock fragments. A composite sample broadly representative of one of the Mazatlan beaches was collected by pooling spot samples of five randomly selected quartz grains from each of twelve stations located on two traverses made at right angles to the strand line. The length of the a-axis of each grain was measured and the results are summarized in the table above.

Compute the first four moments of the a-axis length distribution, together with the dimensional ratios $\sqrt{b_1}$ and b_2 using the phi-scale measurements. How would you characterize this distribution in words? If the a-axis is considered a satisfactory index of overall grain size, how do these results compare with Friedman's findings on relationships between moments and depositional environments?

5. PROBABILITY

Founding Fathers

Gerolamo Cardano (1501-1576), an Italian physician, wrote "Liber de Ludo Alae" (Book on Games of Chance), essentially a gamblers' manual, but considered to be the first book on probability. A profuse writer on a variety of subjects, Cardano was constantly caught up in brawls and was jailed for publishing the horoscope of Christ. Cardano predicted the date of his death and, finding himself still alive on the designated day, committed suicide to maintain his reputation. Galileo Galilei (1564-1642) devoted some time to probability problems also connected with the results of throwing dice.

A series of correspondence between Blaise Pascal (1623-1662) and Pierre de Fermat (1608-1665) started the first formal analysis of probability. Their work was sparked by an acquaintance of Pascal, the Chevalier de Méré, who was encountering distinct problems in his gambling career. De Méré had been winning steadily by betting that he could throw a six in four throws of a die. Becoming more ambitious, he had bet that he could throw a double-six in 24 throws of two dice. Mystified by his losses in the long run, de Méré consulted Pascal. Pascal demonstrated that the break-even point occurred at 24.6 throws. The chevalier denounced science as a swindle.

Fundamentals

The probability that an event will occur is registered on a scale ranging from zero (absolute impossibility) to one (absolute certainty). *A priori* probabilities can be set in advance of the occurrence of the event in circumstances where the external physical constraints are exactly known (e.g., games of chance). Empirical probabilities are measured as a frequency ratio from an observed trial series where:

$$\text{probability of event} = \frac{\text{total number of occurrences of event}}{\text{total number of trials}}.$$

For a finite trial series, the probability is a sample estimate, denoted by P, of the population probability, Π.

1. If events A and B are possible outcomes in a trial series and cannot occur simultaneously, they are said to be mutually exclusive. Then the probability that either A or B will occur is the sum of their separate probabilities: P(A or B) = P(A) + P(B). This is the additive rule of probability.

2. If events A and B are not mutually exclusive but are independent of one another then the joint probability that they will both occur simultaneously is the product of their separate occurrence probabilities: P(A and B) = P(A) X P(B). This is the multiplicative rule of probability. In this case, the occurrence of the two events overlaps and the additive rule becomes modified to: P(A or B) = P(A) + P(B)- P(A and B).

3. When the occurrence of events A and B are dependent to some degree, then their joint probability of occurrence is conditional and: P(A and B) $\neq$ P(A) X P(B).

These concepts are reviewed in the following illustrative example:

The Mississippian "B" is a unit of weathered chert and shale that occurs in the subsurface of Stafford County, Kansas and contains several small oil and gas fields. 124 wells were drilled in the unit, of which 33 were producing wells and 91 were dry holes. In this series of wells the probability of success:

$$P(\text{producer}) = 33/124 = 0.27$$

Conversely, the probability of failure:

$$P(\text{dry}) = 91/124 = 0.73$$

Since these outcomes are mutually exclusive then, by the additive rule of probability:

$$P(\text{producer or dry}) = 0.27 + 0.73 = 1.00$$

As the probabilities sum to one, the two outcomes are exhaustive.

The thickness of the Mississippian "B" in each well was measured and the results summarized as:

Thickness (feet)	Number of wells
0 to 20	52
20 to 40	64
Greater than 40	8

The empirical probabilities that are derived from these frequencies by dividing the well total of 124 are:

$$P(0\text{-}20) = 0.42$$

$$P(20\text{-}40) = 0.52$$

$$P(>40) = 0.06$$

The two types of "event," namely the outcome of the well and the thickness of the section it contains, may be summarized as joint frequencies of occurrence in a contingency table:

Thickness (feet)	Outcome		Thickness Totals
	Producer	Dry	
0-20	5	47	52
20-40	22	42	64
>40	6	2	8
Outcome Totals	33	91	

The rows and columns of the table separate events that are mutually exclusive. The marginal probability of any thickness range is found by summing the appropriate row and dividing by the grand total (124). (Similarly, the marginal probability of a well outcome is the column total divided by the well total). Any given thickness range and well outcome are not

mutually exclusive events and may be either dependent phenomena (relationship between oil occurrence and thickness) or independent (no relationship).

(1) If there is no relationship between oil and unit thickness (independence) then the unconditional joint probability is given by the multiplicative rule:

e.g. P(producer and 0-20 feet thickness) = 0.27 X 0.42 = 0.11

(2) The empirical joint probability is the joint frequency divided by the grand total. Then:

P(producer and 0-20 feet thickness) = 5/124 = 0.04

(3) Clearly, there is a lower observed probability of finding oil in a section less than twenty feet thick than would be expected if there was no relationship between oil accumulation and thickness of the unit.

The conditional relationship between oil occurrence and unit thickness may be examined by computing the empirical joint probabilities and contrasting these with the probabilities that would be expected for the situation of independence:

Thickness (feet)	Outcome	
	Producer	Dry
0-20	0.04 / 0.11	0.38 / 0.31
20-40	0.18 / 0.14	0.34 / 0.38
40	0.05 / 0.02	0.01 / 0.04

Joint probabilities:

Empirical /

/ Expected for independence

The table shows that thicker sections are more prospective for oil exploration than are thin developments. The demonstration of this simple relationship is useful in future prospect evaluation in the area, both by directing attention to the mapping of thickness trends in the Mississippian "B" and enabling a conditional probability estimate to be made of the outcome of any future wildcat. If a prospect is located where the unit is over 40 feet thick, then the conditional probability of a successful well is $6/8 = 0.75$.

1. If events A and B are possible outcomes in a trial series and cannot occur simultaneously, they are said to be mutually exclusive. Then the probability that either A or B will occur is the sum of their separate probabilities: P(A or B) = P(A) + P(B). This is the additive rule of probability.

2. If events A and B are not mutually exclusive but are independent of one another then the joint probability that they will both occur simultaneously is the product of their separate occurrence probabilities: P(A and B) = P(A) × P(B). This is the multiplicative rule of probability. In this case, the occurrence of the two events overlaps and the additive rule becomes modified to: P(A or B) = P(A) + P(B)- P(A and B).

3. When the occurrence of events A and B are dependent to some degree, then their joint probability of occurrence is conditional and: P(A and B) $\neq$ P(A) × P(B).

These concepts are reviewed in the following illustrative example:

	K/Ar Age Interval (m.y.)			
Rock Type	10-70	71-130	131-190	Rock Totals
Quartz Diorite	37	25	24	86
Granodiorite	31	54	31	116
Quartz Monzonite	33	39	21	93
Granite	20	9	10	39
Others	10	7	9	26
Age Totals	131	134	95	360

Contingency table of K/Ar ages of representative plutonic rocks from the Canadian Cordillera, adapted from Petö (1974). The numbers are frequencies of occurrence drawn from a total sample of 360 rocks.

The contingency table summarizes the joint frequencies of occurrence of two types of "event," namely rock type and radiometric age.

The range of rock types is comprised of events that are mutually exclusive. (A given rock cannot simultaneously be a granite and a granodiorite.) The same constraint applies to the set of age categories.

The marginal probability of any particular rock type being selected from the sample is found by summing the appropriate row and dividing by the grand total: e.g., P(granodiorite) = 116/360 = 0.32. (Similarly, for age category events, marginal probabilities are found by using the column totals as numerators: e.g., P(71-130) = 134/360 = 0.37.) By the additive rule: P(granodiorite or granite) = P(granodiorite) + P(granite) = 0.43.

Any given rock type and age category are not mutually exclusive events and may either be dependent phenomena (relationship between intrusive type and age) or independent (no relationship).

(1) If there is no relationship between age and rock type (independence) then the unconditional joint probability is given by the multiplicative rule: e.g., P(granodiorite and 71-130 m.y.) = 0.32 x 0.37 = 0.12.

(2) The empirical joint probability is the joint frequency divided by the grand total. Then: P(granodiorite and 71-130 m.y.) = 54/360 = 0.15.

(3) Clearly there is a higher observed probability of granodiorites of Cretaceous age than would be expected if granodiorites had been intruded more or less evenly through time and were equally available at outcrop. The expected joint frequency, assuming no relationship, is found by multiplying the unconditional joint probability by the grand total and is 43.2. This figure contrasts with the observed frequency of 54. Whether the difference in these two totals are statistically significant (rather than a freak variation compatible with the sample size) is a matter that must be formalized as a statistical hypothesis and tested by methods described later.

Reference

Petö, P., 1974, Plutonic evolution of the Canadian Cordillera: Geol. Soc. Amer. Bull., v. 85, p. 1269-1276.

Exercise 5.1 : Age of plutonic rock types in Canadian Cordillera

Compute the marginal probabilities of the plutonic rock type and age categories in the Canadian Cordillera example. Prepare a table of expected unconditional joint frequencies of rock types and ages and compare them with the observed frequencies. What (if any) appears to be the apparent relationship between igneous intrusive and age?

29717

(1)

USE OF THE BINOMIAL PROBABILITY DISTRIBUTION TO PREDICT THE RESULTS OF DRILLING PROGRAMS IN FRONTIER AREAS

The binomial probability distribution describes the outcomes of certain types of simple games of chance, where the odds of winning are known in advance. It [can] [also] be used to forecast the probability of success in drilling programs, if several assumptions ~~can be~~ are made. These assumptions seem most reasonable when applied to rank wildcat exploration ~~of~~ in relatively virgin ~~areas~~ basins; hence, the binomial distribution is often used to predict the outcomes of drilling programs in frontier areas and offshore concessions.

Under the assumptions of the binomial distribution, each wildcat must be classified as either a discovery or a dry hole. Successive wildcats are presumed to be independent; that is, success or failure of one well will not influence the outcome of the next well. (This assumption is difficult to justify in most circumstances, as a discovery [will] [usually] affect [the selection] subsequent drilling ~~locations~~ sites. A ~~long~~ protracted succession of dry holes will also cause a shift in an exploration program.) The probability of a discovery is assumed to remain unchanged. (This assumption is reasonable at the initiation of exploration, but becomes increasingly tenuous during later phases when a large proportion of the fields in a basin have been discovered.) Finally, the binomial is appropriate when a fixed number of holes will be drilled during an exploratory program, or during a single time period (perhaps a budget cycle) for which the forecast is being made.

29712

The following statements illustrate the development of the binomial model as applied to exploratory drilling.

1. The probability that a hole will result in a discovery is p.
2. Therefore, the probability that a hole will be dry is 1-p.
3. The probability that n wildcats will all be dry is

$$P = (1-p)^n \quad \odot$$

4. The probability that the nth wildcat will be a discovery but the preceding (n-1) wildcats will all be dry is

$$P = (1-p)^{n-1} p \quad \odot$$

5. The probability of one discovery in a series of n wildcats is

$$P = n(1-p)^{n-1} p \quad \odot,$$

since the discovery can occur on any of the n wildcats.

6. The probability that (n-r) dry holes will be drilled, followed by r discoveries, is

$$P = (1-p)^{n-r} p^r \quad \odot$$

27733

7. However, the (n-r) dry holes and the r discoveries may be arranged in $\binom{n}{r}$ combinations, or equivalently in $\frac{n!}{(n-r)!r!}$ ways. So, the probability of having r discoveries in a drilling program of n wildcats is

$$P = \frac{n!}{(n-r)!r!} (1-p)^{n-r} p^r \quad \odot$$

This is an expression of the binomial distribution, giving the probability of r successes in n trials, when the probability of success in a single trial is p.

The probability of a discovery, p, is often estimated as the company's (or industry's) success ratio in the exploration area. In the case of virgin provinces, the probability may be estimated as the success

27732

ratio in more mature areas judged to be geologically similar, or as the worldwide success ratio. Since the binomial probability can easily be calculated or read from tables or charts, several different values of p may be used in an evaluation. This will provide a range of possible outcomes, and the results of an initial exploratory campaign may indicate which of the assumed values of p is most appropriate.

The basic parameters of the binomial distribution may be found from the probability and the sample size. The mean or expected number of discoveries is simply

$$\overline{X} = np \ .$$

The variance of the binomial distribution is

$$s^2 = np(1-p) \ ,$$

which is a measure of the variation expected in a large number of exploratory programs having the same number of holes and the same success ratio.

The following example develops probabilities associated with a five-well exploration program in a virgin basin where the success ratio is anticipated to be about 10%.

1. What is the probability that the five-well program will be a total failure, with no discoveries? (Such an outcome is called "gambler's ruin.")

   $n = 5$

   $r = 0$

   $p = 0.10$

$$P = \binom{n}{r} \cdot p^r \cdot (1-p)^{n-r} = \binom{5}{0} \cdot 0.10^0 \cdot 0.90^5$$

Now,

27731

4

$$\begin{bmatrix} n \\ r \end{bmatrix} = \frac{n!}{(n-r)!r!} = \frac{5!}{5!0!} = \frac{120}{120} = 1$$

so,

$$P = 1 \cdot 0.10^0 \cdot 0.90^5$$

$$= 1 \cdot 1 \cdot 0.59$$

$$= 0.59 \odot$$

The probability that a five-well program will be a total failure is 59%.

2. If only one hole is a discovery, it may pay off the costs of the entire drilling program. What is the probability that one well will come in during the five-well exploration campaign?

$$P = \begin{bmatrix} 5 \\ 1 \end{bmatrix} \cdot 0.10^1 \cdot 0.90^4$$

$$\begin{bmatrix} 5 \\ 1 \end{bmatrix} = \frac{5!}{4!1!} = \frac{120}{24} = 5$$

$$P = 5 \cdot 0.10^1 \cdot 0.90^4$$

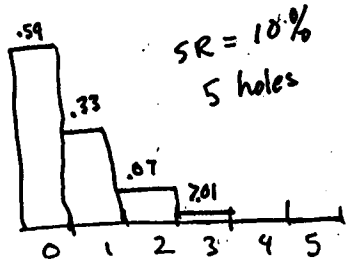
$$= 5 \cdot 0.10 \cdot 0.656$$

$$= 0.328$$

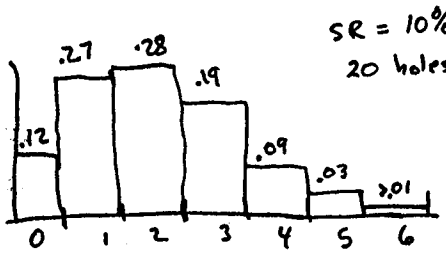
The probability that one well will be successful is 33%.

We can express the possible outcomes of the drilling program in a histrogram by calculating the probabilities of none or one to five discoveries. Note that the odds of <u>some</u> success are 1.00 - 0.59 = 41%.

(ilo.) B-1



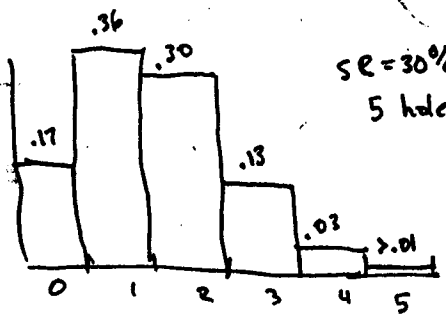
Note that if the number of holes in the drilling program is increased the probability of a small number of successes also increases. For example, the histogram below shows the possible outcomes for a 20-well exploration effort. The odds of some success are now increased to 1.00 - 0.12 = 88%. ZBP, STC ICR



(illo.) B-2

Similarly, if the success ratio is greater, the probability of a small number of discoveries is also greater. The histogram below shows outcomes for a five-hole drilling program in an area where the success ratio is expected to be 30%. The odds for some success are 1.00 - 0.17 = 83%.

(illo.) B-3



6. BINOMIAL DISTRIBUTION

"Binomial" means "consisting of two names or classes." The binomial is a discrete probability distribution that specifies the frequency of different combinations of events in situations where there are only two possible outcomes of an event in any one trial. A basic condition of the distribution is that successive trials must be independent.

(1) If p is the probability that an event occurs ("success") and q, that it does not occur ("failure"), then $q = 1 - p$; there are two alternative results for any trial and the system is binomial.

(2) The probability of an event not occurring in n trials is q^n.

(3) The probability of an event occurring on the nth trial (but on none of the previous) is $q^{n-1}p$. However, the probability of the event occurring only once in n trials (regardless of which trial) is $nq^{n-1}p$, since the trials are independent and there are n different positions for this success.

(4) The probability of there being n-r failures followed by r successes is $q^{n-r}p^r$. However, the n-r failures and r successes may be arranged in $\frac{n!}{(n-r)!r!}$ ways which are mutually exclusive. So the probability of n-r failures and r successes (without regard to order) is $\frac{n!}{(n-r)!r!}q^{n-r}p^r$.

(5) The formula $\frac{n!}{(n-r)!r!}q^{n-r}p^r$ is the description of the binomial distribution. Substitution of n, p, q and successive integer values of r into the formula yields probability values of the occurrence of combinations of r successes and n-r failures.

(6) The area under the histogram of this distribution is equivalent to the summation of the probabilities of all possible combinations and equals one.

(7) Anyone familiar with the binomial theorem will recognize that the distribution is equivalent to the terms of the expansion of $(q+p)^n$.

(8) The population mean of the distribution $\mu = \pi N$.

(9) The population variance, $\sigma^2 = N\pi(1-\pi)$.

(10). For a sample, p is an unbiased estimator of π and $\bar{X} = pn$.

(11). Similarly, the sample estimate of the variance of the distribution, $s^2 = npq$.

(12) The standard deviation of the sample estimates of p about their population parameter π, is known as the standard error σ_e:

$$\sigma_e = \sqrt{\frac{\pi(1-\pi)}{n}}$$

If π is unknown then we must make a maximum likelihood estimate ("best guess") of s_e by inserting p in place of π when:

$$s_e = \sqrt{\frac{pq}{n}}$$

Geological Applications

The binomial distribution is widely used in the computation of constituent volume percentages made in petrographic point counting. The process of point counting is implicitly statistical, since measurements of grain counts on a thin section are used as statistics of a random sample, from which generalizations are made regarding the parameters of the parent population (the mineral proportions in the total rock). The point-count model is essentially multinomial as the combination occurrences of m different constituents are described by an m-nominal distribution. However, by considering each constituent independently, a binomial model may be used to describe the occurrence or non-occurrence of a constituent at each successive count.

The sources of error arising from the point counting procedure are:

(1) operator error - misidentification of constituents by the petrographer;

(2) specimen error - the degree to which the thin-section deviates from being representative of its host rock (petrofabric heterogeneity); and

(3) counting error - the deviation between the sample estimates of the areal proportions of the constituents from the true or population parameters of these quantities.

The binomial distribution model is directed towards the evaluation of the counting error.

Example: Estimating standard error in point counting

A traverse count of 100 points is made across a petrographic thin-section. The total number of times mineral A is encountered is recorded as 15, as opposed to the alternative B (other constituents). Then, $n = 100$, $p = 0.15$, $q = 0.85$. p is a sample estimate of the true proportion, π and the maximum likelihood estimator of its standard error about π,

$$s_e = \sqrt{\frac{0.15 \times 0.85}{100}} = 0.036$$

The quantity s_e is a measure of the accuracy of the point count estimate of the actual proportion of mineral A in the total rock, and is clearly a function of both n and π. Longer traverses will reduce counting error; high volume constituents (up to 50%) will require longer traverses than minor constituents for equivalent standard errors (measured as absolute percent).

Example: Estimation of proportions in point-counting

Mineral A is believed to constitute 20% of the total volume of a homogeneous rock type. If a single traverse is made of 50 counts, what is the probability that the number of mineral A counts will correspond precisely to the true volume figure?

$$n = 50, \quad \pi = 0.20, \quad (1-\pi) - 0.80$$

For an estimate identical with the population proportion, the traverse must encounter precisely 10 grains, when $p = \pi$.

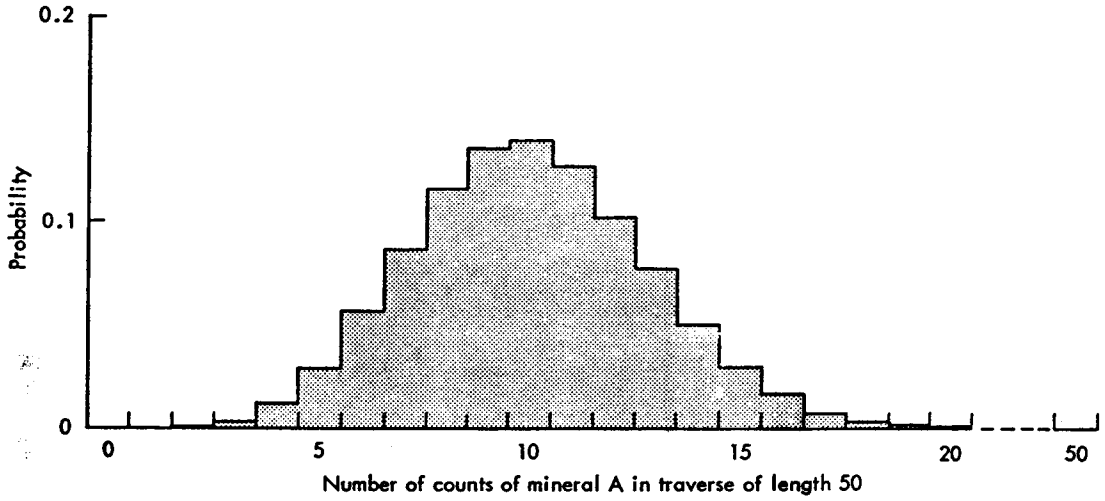
Applying the binomial distribution formula to this case:

$$P_{10} = \frac{50!}{40!10!} \; 0.80^{40} \cdot 0.20^{10} = 0.14$$

In other words, if we recorded the results of a hundred different traverses of 50 point-counts, we would arrive at a true estimate in only 14 cases.

The probabilities that apply to all the other possibilities the traverse might encounter (no mineral A, 1, 2, 3... 50 counts of A) may be computed in a similar fashion and combined in the descriptive histogram shown above. The high dispersion of the distribution illustrates that the short traverse length involves a high standard error (0.057). Longer traverse lengths will constrict the distribution relative to the scale and result in lower standard error. Since measurement error can only be eliminated completely by measuring the entire population, the

onus rests on the investigator to set a counting error that he is prepared to live with and estimate a minimum point-count total that is compatible with this figure.



References

Chayes, F., 1956, Petrographic modal analysis: John Wiley and Sons, Inc., New York, 131 p.

Griffiths, J.C., 1967, Scientific method in analysis of sediments: McGraw-Hill Book Co., New York, 508 p.

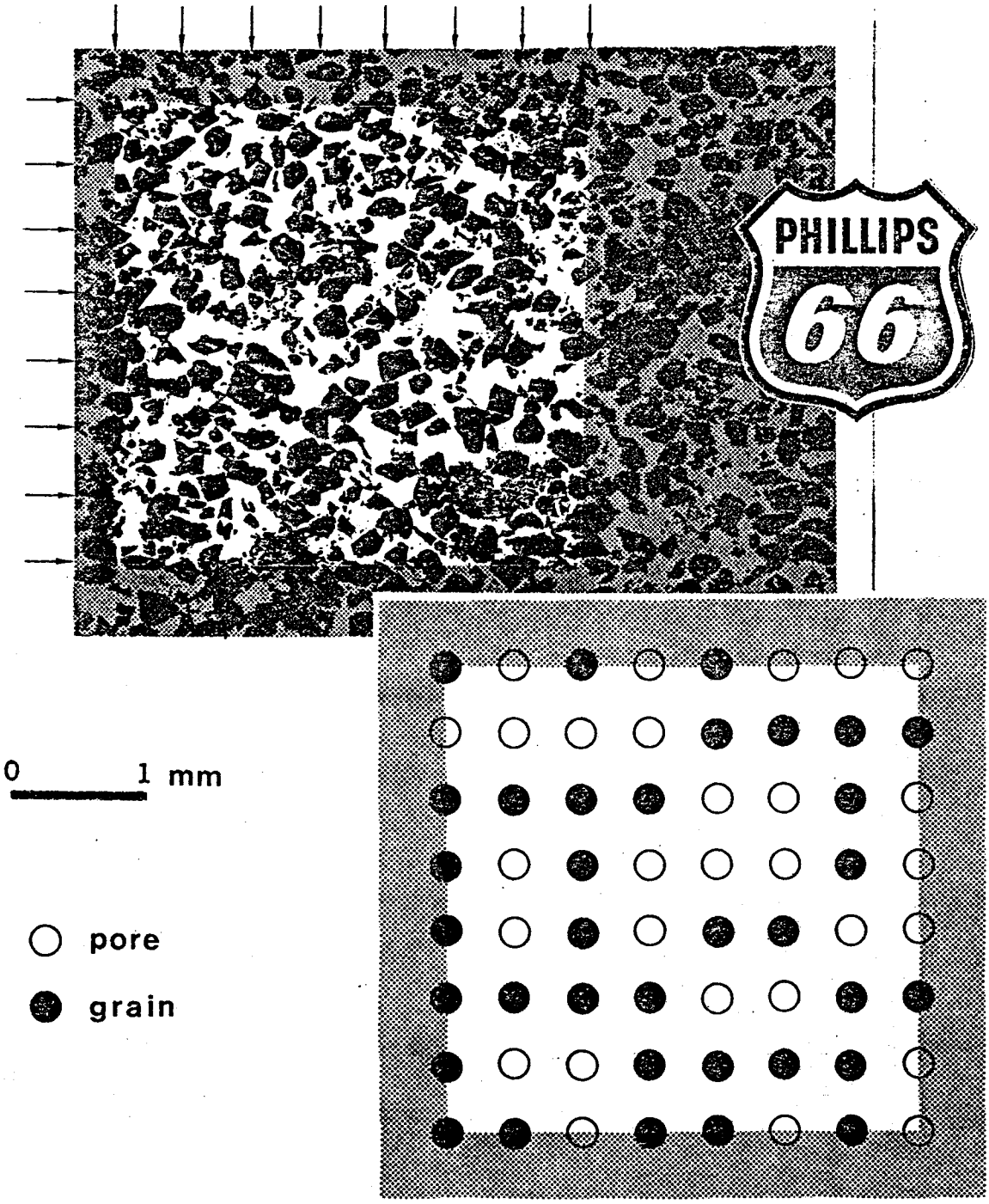
Romig, H., 1947, "50-100 Binomial Tables": John Wiley and Sons, Inc., New York, 172 p.

Exercises

6.1: In order for a binomial distribution to be strictly applicable, the condition that successive trials are independent must be fulfilled. What are the implications of this condition with regard to mineral grain size and the spacing between successive point-counts?

6.2: A rock sample contains 10% mineral A, by volume. (a) What is the probability of making a thin-section traverse of 30 point-counts without encountering mineral A. (b) How many points would have to be counted before one could be 99% certain of recording at least one instance of mineral A?

6.3: The Bartlesville Sandstone is a Pennsylvanian deltaic sand complex containing many alluvial channel units which form good reservoir rocks



Bartlesville Sandstone thin-section with index diagram of pore/grain count locations

and are prolific oil producers in Oklahoma. (The origins of Phillips 66 as a major oil corporation are historically linked with the Bartlesville Sandstone which was the first major producing formation exploited by the company.)

A petrographic thin-section of a Bartlesville reservoir sandstone core is illustrated and shows dark grains of quartz and white zones of pore space. Point-count results from a superimposed 8 X 8 square grid are shown graphically on the index diagram.

(a) What is the sample estimate of the porosity? Compute a maximum likelihood estimate of the standard error of this porosity figure.

(b) Count the frequencies of zero, one and two pore counts for 32 independent pairs of adjacent grid nodes. What is the mean and standard deviation of this empirical distribution? Compute binomial model estimates of the distribution frequencies, the mean and standard deviation, using the sample porosity estimate. How does the binomial model compare with the empirical data and what are the implications?

(c) An effective porosity of 18.3% was measured in the parent core sample, using brine saturation. Discuss the relationship between this figure and the petrographic porosity estimate.

7. NORMAL DISTRIBUTION

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY ◆ IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING ◆
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

Origins

In *Approximatio ad Summam terminorum binomii* $(a+b)^2$ *in Seriem Expansi,* De Moivre (1667-1754) derived what is now known as the normal distribution. He considered the extreme case of the binomial distribution when n (the number of trials) approaches infinity. In the limit, the curve becomes a continuous probability distribution. De Moivre did not see much practical future for the normal curve, but it was applied to scientific problems by Laplace (1749-1827) and Gauss (1777-1855), who realized that the distribution described a curve of errors generated by repeated physical measurements (such as locations of fixed stars). The Victorians extended the use of the curve into many fields to describe distributions of natural data and considered it to be an almost universal law of variation. The best known application is to I.Q. measurements.

The normal distribution equation: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-1/2 \left(\frac{X-\mu}{\sigma}\right)^2}$

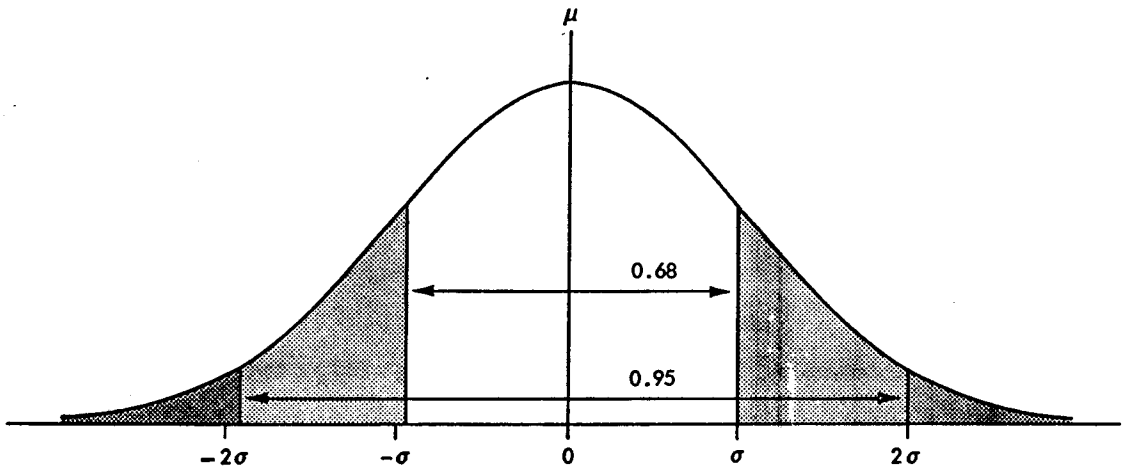
Computations may be simplified by use of the transformation:

$$\log_e f(x) = -\log_e \sqrt{2\pi} - \log_e \sigma - \frac{1}{2} \left(\frac{X-\mu}{\sigma}\right)^2$$

where f(x) is the probability density at the value X of the measured variable and π is the trigonometric constant. In almost all practical applications, this formula is bypassed through use of statistical tables of the standard normal distribution.

Any normal distribution is uniquely specified by μ and σ (or their sample estimates, $\bar{X}$ and s). They are symmetrical about the mean and range from $-\infty$ to $+\infty$.

Since the variable, X is continuous, probabilities must be computed for ranges of X, rather than discrete values, and these correspond to segmented areas under the curve. The total area under the normal curve sums to one.



The normal distribution with mean of zero and standard deviation of one is the standard normal distribution and is used as a reference. Empirical observational data may be matched against the distribution after standardization to a Z-score, by the transformation,

$$Z = \frac{X-\bar{X}}{s}$$

Areas of the standard normal distribution between specified Z values are derived from published statistical tables and represent the probabilities of occurrence within their range.

Example: More on amphibian skulls

In a previous example, the distribution of skull lengths of *Trimerorhachis insignis* was related to a normal distribution by consideration of the sample estimates of skewness and kurtosis. The initial supposition that the sample might be described by a normal model was weakened by both the values of the computed moments and the bimodality

of the skull lengths. A hypothetical normal distribution that has the same mean (90.3) and unbiased sample variance (479.7) as the amphibian skulls may be generated in the following manner for purposes of comparison:

(1) Compute the Z-scores of the class limits of the histogram, using the sample mean and standard deviation.

(2) Consult tables of the standardized normal distribution and list the values of the area between zero (the standardized mean) and the Z-score of each class limit.

(3) Calculate the area in each class as the difference in area between each pair of limits.

(4) Multiply each class area by the total number in the sample to obtain the number of skulls in each class as expected for the fitted normal distribution.

These computational steps are numerically summarized in the following table.

Normal curve fitted to amphibian skull length data

f_o	l_1	l_2	Z_1	Z_2	a_1	a_2	δ_a	f_e
0	(−∞)	50	(−∞)	−1.84	0.500	0.467	0.033	1.09
2	50	60	−1.84	−1.38	0.467	0.416	0.051	1.68
3	60	70	−1.38	−0.93	0.416	0.324	0.092	3.04
9	70	80	−0.93	−0.47	0.324	0.181	0.143	4.72
3	80	90	−0.47	−0.01	0.181	0.004	0.177	5.84
7	90	100	−0.01	+0.44	0.004	0.170	0.174	5.74
3	100	110	+0.44	+0.90	0.170	0.316	0.146	4.82
3	110	120	+0.90	+1.36	0.316	0.413	0.097	3.20
1	120	130	+1.36	+1.81	0.413	0.465	0.052	1.72
2	130	140	+1.81	+2.27	0.465	0.488	0.023	0.76
0	140	∞	+2.27	∞	0.488	0.500	0.012	0.40
33			TOTALS				1.000	33.01

f_o and f_e are observed and normal curve expected frequencies;
l_1 and l_2 are lower and upper interval limits;
Z_1 and Z_2 are Z-scores of l_1 and l_2
a_1 and a_2 are cumulative areas under the standard normal curve from 0 to Z_1 and Z_2;
δ_a is area of normal curve between Z_1 and Z_2.

Exercise 7.1: More on the Mazatlan beach

Refer to the table of a-axis lengths measured on quartz grains from the Mazatlan beach sample. Plot a histogram of the length distribution with phi-scale class limits of -2.2, -1.8, -1.4 ... +1.0. Using the sample mean and unbiased standard deviation, compute frequencies in each class that would be expected for a normal distribution with the same mean and standard deviation. Sketch the fitted normal curve on the observational histogram. Compare the visual relationships between the histogram and normal curve with the sample estimates of skewness and kurtosis.

The Central Limit Theorem

This most important theorem states that if a random sample of size n is drawn from any population with mean μ and variance σ^2, then as n becomes larger, the distribution of sample means about the population mean approaches a normal distribution. The key words are "any distribution," which extend the use of the normal distribution from analysis of raw measurements thought to be normally distributed to parameter estimation from data described by almost any kind of distribution.

The limiting normal distribution of sample means from any data can be set to the standard normal distribution by transforming the deviations of the sample means from the population mean to Z-scores by the formula:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

This is equivalent to the Z-transformation formula for raw data cited previously. The ratio $(\sigma/\sqrt{n})$ has been substituted for 's', and is the standard deviation of the mean of sample size n about the population mean. The ratio is known as the standard error of the mean (q.v. binomial distribution). The value of n is held to be "large" when greater than 30, at which point s (sample standard deviation) becomes a reasonable estimate of σ and may be substituted in the formula as a maximum likelihood estimate of the standard error.

Confidence Intervals

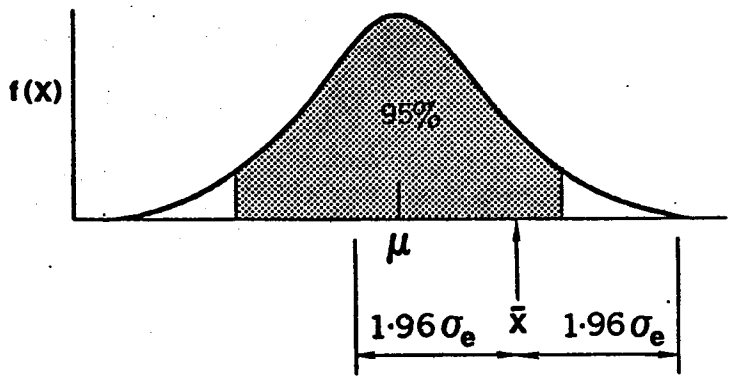
The standard error of the mean may be used to make a probability statement concerning the location of the true population mean (or parameter) based on a sample estimate. Since the standard error is:

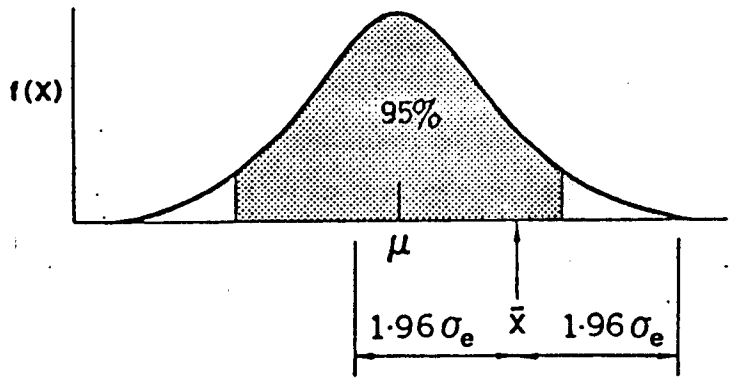
$$s_e = \frac{s}{\sqrt{n}} \simeq \frac{\sigma}{\sqrt{n}} \quad \text{for } (n > 30)$$

This represents a standard deviation of the distribution of estimates of the mean drawn from samples of size n. The interval,

$$\mu \pm 1.96\ s_e$$

contains 95% of the sample means. This statement is useless in itself since, if μ is already known, no further qualification is needed. However, the statement can be transposed so that the parameter is related to an interval about a sample mean. As 95% of the sample means fall within a range of $\mu \pm 1.96\ s_e$ it follows that for any particular sample mean, we can be confident that the range $\bar{X} \pm 1.96\ s_e$ contains the population mean 95% of the time. The range $\bar{X} \pm 1.96\ s_e$ is called the 95% confidence interval. By altering the constant 1.96 to 2.58, the range becomes the 99% confidence interval. Any percentage probability level may be used and its appropriate constant is Z drawn from tables of the standard normal distribution. However, the 95% and 99% levels are the most widely used.





As an example of the use of the confidence interval, the thicknesses of the Mississippian "B" in producing wells and dry holes may be considered. From a sample size of 33 wells, the estimate of the mean thickness of productive interval is 31.1 feet. Assuming that the sample is a random selection, can a statement be made concerning the true mean of all possible productive sections? It is not known whether the distribution of section thicknesses is normal or non-normal. However, this point is immaterial since the distribution of sample means will be normal. The sample size is greater than 30 and so the standard error may be estimated as:

$$s_e = s/\sqrt{n} = 11.4/\sqrt{33} = 1.98$$

and

$$X_p \pm 1.96\, s_e = 31.1 \pm 1.96 \times 1.98$$

$$= 31.1 \pm 3.9$$

It may therefore be said that the population mean of all productive section thicknesses is contained in the range between 27.2 and 35.0 feet. The probability that this conclusion is incorrect is 5%.

By a similar computation the 95% confidence interval may be calculated for the mean thickness of non-productive sections:

$$s_e = 10.2/\sqrt{91} = 1.07$$

$$X_d \pm 1.96\, s_e = 20.5 \pm 1.96 \times 1.07$$

$$= 20.5 \pm 2.1$$

Note that the confidence interval is more constricted in this case, even though the thickness standard deviations are fairly similar. The reason for this is that there is a much larger sample of dry wells with the result that the sample mean is a better estimate of its true population value.

In cases where the sample size is less than about 30, the sample size is said to be "small." The reason for this seemingly arbitrary boundary will be shown in the t-distribution section (q.v.). When $n < 30$ the substitution of s for σ in the standard error formula involves sampling errors which become too large to be ignored. At this level, a t-distribution is used as the appropriate sampling distribution in the place of the normal distribution. Computation of the confidence interval is made in the usual manner with replacement of the Z value by the corresponding value of t appropriate for the sample size.

Confidence intervals are widely used in many statistical procedures such as linear regression and correlation as a way to estimate a range within which the true population parameter is most likely to occur, as related to the location of its sample estimate.

As an example of the use of the confidence interval, the skull lengths of *Trimerorhachis insignis* may be considered. From a sample size of 33 skulls the estimate of the mean skull length is 90.30 mm. Assuming that the sample is a random selection, can a statement be made concerning the true mean of all *T. insignis* skull lengths? Regardless of whether the distribution of skull lengths is normal or non-normal, the distribution of sample means will be normal. Since the sample size is greater than 30, the standard error may be estimated as:

$$s_e = \frac{s}{\sqrt{n}} = \frac{21.9}{\sqrt{33}} = 3.81$$

and

$$\bar{X} \pm 1.96\ s_e = 90.30 \pm 1.96 \times 3.82$$

$$= 90.30 \pm 7.47$$

It may therefore be said that the population mean of all *T. insignis* skull lengths is contained in the range between 82.83 and 97.77 mm. The probability that this conclusion is incorrect is 5%.

In cases where the sample size is less than about 30, the sample size is said to be "small." The reason for this seemingly arbitrary boundary will be shown in the t-distribution section (q.v.). When $n < 30$ the substitution of s for σ in the standard error formula involves sampling errors which become too large to be ignored. At this level, a t-distribution is used as the appropriate sampling distribution in the place of the normal distribution. Computation of the confidence interval is made in the usual manner with replacement of the Z value by the corresponding value of t appropriate for the sample size.

Confidence intervals are widely used in many statistical procedures such as linear regression and correlation as a way to estimate a range within which the true population parameter is most likely to occur, as related to the location of its sample estimate.

8. HYPOTHESIS TESTS

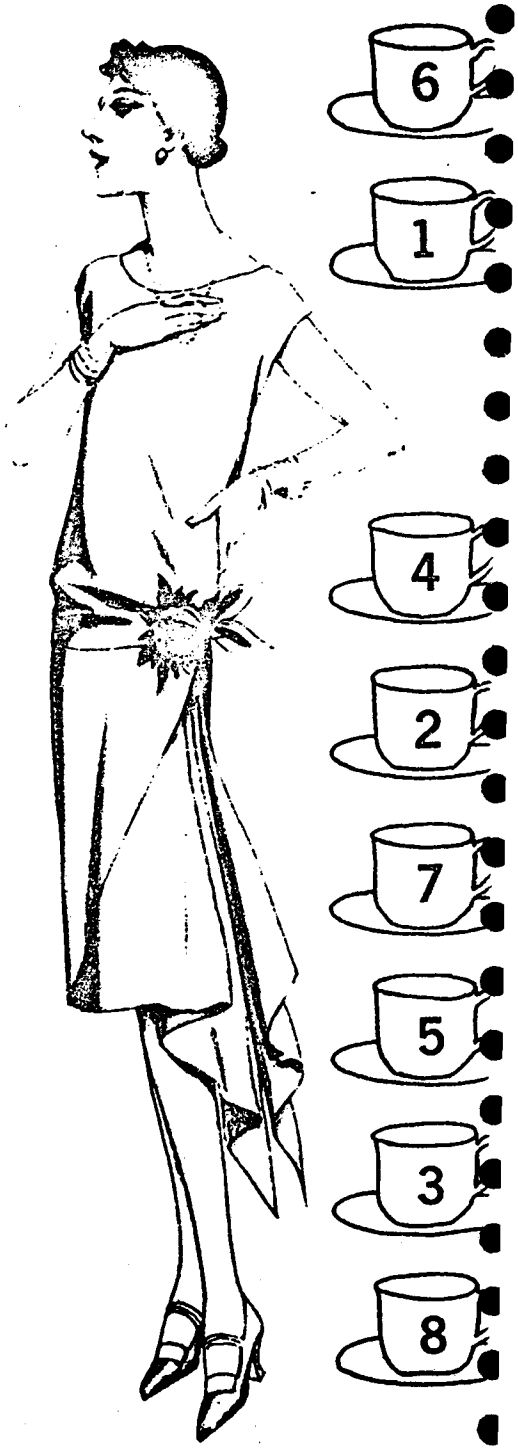
Introduction

In the preceding text, sample statistics have been computed as estimates of population parameters and compared with those of theoretical distributions in an informal fashion. Conclusions concerning the applicability of the theoretical models have been made a matter of subjective judgement that ideally takes into account the difference between the observed and model prediction values in conjunction with the size of the observed sample. The procedure of drawing generalized conclusions based on limited observational data is formalized in the application of inferential statistics. A null hypothesis is postulated and accepted or rejected according to whether or not a computed test statistic exceeds or falls short of its tabulated critical value at a set level of significance.

Sir Ronald A. Fisher (1890-1962) set statistical inference on a firm theoretical basis starting from the premise that almost any research project is an exercise in inductive logic. The results of a few controlled experiments are extended to generalizations concerning the phenomenon under study. The translation from the particular to the general inevitably involves some degree of uncertainty, since the true value of any population parameter is not known until its entire content is measured. However, this degree of uncertainty may be calculated by using statistical procedures rooted in probability theory. A prerequisite of this approach is that experiments must be designed in such a way as to be strictly appropriate for statistical analysis. This condition can be met in modern environment analyses by the intelligent application of sampling procedures, but is more difficult to satisfy in paleoenvironment studies where nature is often the arbiter on which "experiments" are preserved in the stratigraphic record.

Fisher devised the method of Analysis of Variance and published his ideas in the classic work, "The Design of Experiments" (1935) which initiated a revolution in research methods and analysis of experimental results. Fisher's text opens with probably the most famous illustrative example in inferential statistics, the "tea-tasting experiment." It is supposed that there is a lady who claims she can tell the difference

between a cup of tea in which milk has been added after the tea, as opposed to a cup in which milk has been introduced before the tea. Eight cups of tea are prepared, four one way, four the other, and presented to her in random order. She is further told there are four cups of each type. If the lady identifies all eight cups correctly, it is by no means certain that she can tell the difference between the two types of tea. There is a 1/70 probability that this outcome could arise by chance alone. Therefore, if the null hypothesis that she can not discriminate is rejected, there is a 0.014 chance of this conclusion being mistaken. This homely parable provides a philosophical starting point for further development of inferential statistical theory.



There is historical irony in the timing of Fisher's pioneer work in experimental design and inference. In 1936, Fisher demonstrated that, barring an "absolute miracle of chance," the experimental figures reported in Gregor Men del's classic paper on genetics were deliberately faked. This conclusion is now generally accepted, although suspicion is directed at Mendel's well-meaning gardeners rather than the monk himself. The irony is compounded by the fact that the statisticians who founded the journal Biometrika at the turn of the century were passionate anti-Mendelians who supported a "blending" mechism of inheritance. Had they but known it, these statisticians had it in their power to set the science of genetics back by fifty years. By 1936, enough research had been undertaken to thoroughly confirm the basic tenets of genetics. (This saga and its ramifications are entertainingly described by Koestler, 1971.)

References

Fisher, R.A., 1935, The Design of Experiments: Oliver and Boyd, Edinburgh, 245 p.

_____, 1936, Has Mendel's work been rediscovered?: Annals of Science, v. 1, p. 115-137.

Koestler, A., 1971, The Case of the Midwife Toad: Vintage Books Edn., Random House, New York, p. 53-57.

Example: Porosity of the Bartlesville Sandstone

In the Bartlesville Sandstone porosity study, comparison was made between the estimate based on optical point-count (45.3%) and the brine-saturation measurement (18.3%) made from the entire core slice. The two values appear to be distinctly different, but this observation must be tempered by an awareness of the small sample size (64 counts) used for the point-count estimate. A finite probability exists that it would be wrong to reject the optically measured porosity as a valid estimate of the core measurement (equating this with the population parameter).

The null hypothesis is postulated that the point-count measure, p, is a sample estimate of a population parameter π of 18.3%. This is expressed symbolically as: $H_0 : p = \pi$ (the null hypothesis); while $H_1 : p \neq \pi$ is the alternative or motivated hypothesis.

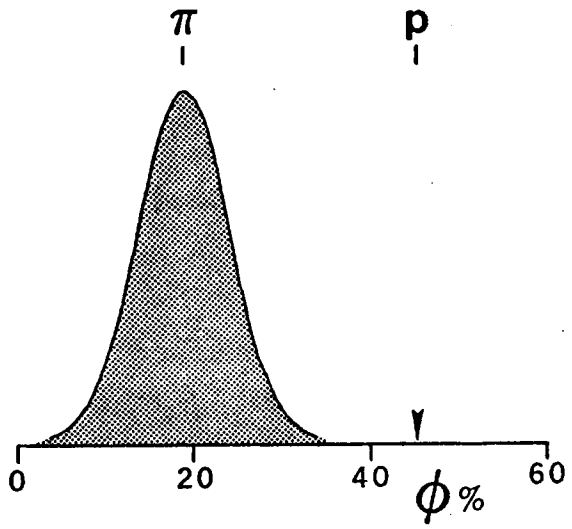
Now, by the Central Limit Theorem, the sampling distribution of porosity means in samples of size 64 are approximately normally distributed about the parameter mean (0.183) with a standard deviation that is the standard error of the mean.

$$\sigma_e = \sqrt{\frac{\pi(1-\pi)}{n}} = 0.048$$

The point-count estimate, 0.453, has a deviation of 0.270 from the parameter mean. Translated into a Z-score (standard deviation units) this distance is:

$$Z_p = \frac{\text{deviation}}{\text{standard error}} = \frac{0.270}{0.048} = 5.63$$

From the table of the standard normal distribution it can be seen that the probability of a sample estimate (size 64) with a Z-value greater than 5.63 is vanishingly small.



Rather than compute a probability of the sample statistic being drawn from the theoretical sampling distribution, a probability value is chosen as the critical cut-off value and is termed a level of significance and denoted by α. This figure is selected by the investigator as discussed later, but is conventionally specified as 0.05 or 5%. (Computed statistics that correspond to probabilities less than this figure are considered to be significant and the null hypothesis is rejected.)

The two complementary hypotheses of this example are non-directional, since the null hypothesis equates the sample statistic with the population parameter, while the alternative specifies the sample statistic to be significantly different (either greater or less). If one wished to use the motivated hypothesis that the sample statistic was significantly greater than the population parameter, then the null hypothesis would be that the sample statistic either equalled or was less than the parameter. In this case the hypothesis would be directional. A two-tailed test is used for non-directional hypotheses; a one-tailed test for directional hypotheses. If a 0.05 significance level is used, then the sample statistic must lie in the extreme 5% area of the sampling distribution in either case. However, in the one-tailed situation, this corresponds to a 5% area in one of the tails of the distribution; in the two-tailed case, this is the sum of the 2 1/2% tails at either end of the distribution.

two-tailed 5% one-tailed 5%

In this example, the test is two-tailed and the critical test value of Z that corresponds to each 2 1/2% extreme is ±1.96 (drawn from the table of the standard normal distribution). The computed Z-value of the sample mean must be either greater than 1.96 or less than -1.96 for the null hypothesis to be rejected, which is, in fact, the case. If the test had been one-tailed and the directional null hypothesis:

$$H_0 : (p-\pi) \leqslant 0$$

then the tabulated critical value at significance level 0.05 would be -1.65 and a Z-score less than this figure would be needed for rejection of the null hypothesis.

The preceding example is an illustration of the Z-test (or normal test) and is applicable for comparison of a sample mean with a population mean when the sample is large (>30) and the population standard deviation is known. (The population standard error was assumed to conform to a binomial distribution for pedagogic purposes.)

Levels of significance

The significance level is commonly denoted α (significance a level of 0.05, or 5%, is $\alpha = 0.05$). If a result is significant at the 5 percent level, then the probability of the result arising as a consequence of random sampling fluctuations is less than 5 percent. The lower the percentage figure quoted, the more reliable are conclusions drawn from the result. Prior to any experimental analysis, a significance level must be set by the investigator to be used as an arbiter. The selection of this value varies in different fields of investigation so that, for example, the statistical analysis of trials with newly developed drugs is geared to minutely low levels, since the chance of being wrong about the harmful effects of a test drug may mean death for some unfortunate individual. Significance levels of 0.05 and 0.01 are

most widely used in geological studies and are inherited from biological statistics where these figures have the force of tradition. Selection of a significance level after the experimental results are calculated is shameless gerrymandering since it will obviously reflect the investigator's bias for acceptance or rejection of a hypothesis.

Type I and Type II errors

In any decision concerning the validity or non-validity of the null hypothesis, there is always a residual uncertainty regarding the "rightness" of the decision, which may be subdivided into two types of possible error as shown in the table:

DECISION

REALITY	Accept H_0	Reject H_0
H_0 is true	No error	Type I error
H_0 is false	Type II error	No error

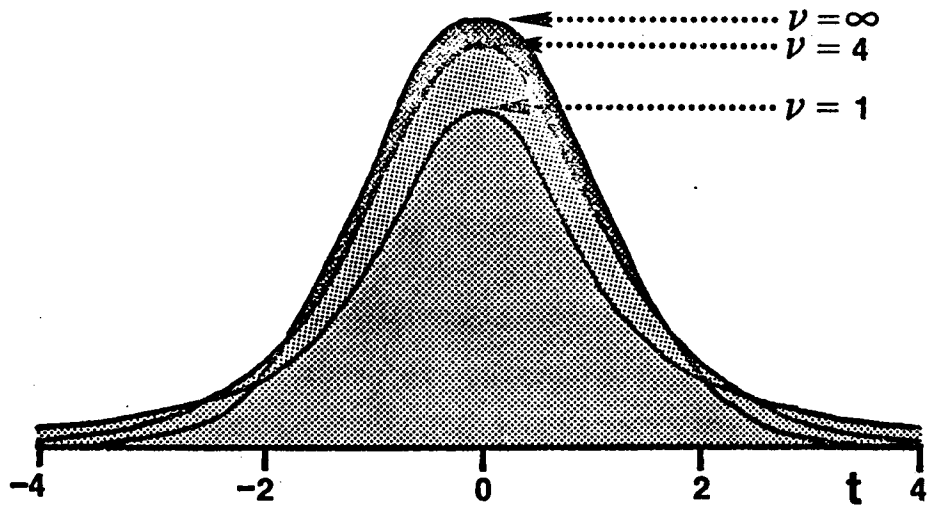
Type I error is denoted by α, is the level of significance, and is the probability of rejecting the null hypothesis when it is true. The investigator has control over this value because he can specify the amount of error he is prepared to live with, which is then the level of significance he applies in hypothesis testing. Type II error is denoted by β and is the probability of accepting the null hypothesis when it is false. The complementary probability $(1-\beta)$ is called power. As α decreases, β increases.

The specification of α is an easier task than the regulation of β. Consequently, statistical hypotheses are normally cast in a form of negativism. The hypothesis that the researcher hopes to reject is set as the null hypothesis and the Type I error is specified at a low figure. It follows that there is a low probability of being wrong if the null hypothesis is rejected, and a higher probability of being wrong if the null hypothesis is accepted. This strategy is strictly conservative in the sense that the odds are preferentially stacked against the researcher's basic wishes. If the computed statistic results in rejection of the null hypothesis under these conditions, a shade more confidence is associated with the reliability of the conclusion.

9. STUDENT'S t-DISTRIBUTION

W.S. Gosset (1876-1937) reported pioneer statistical work under the penname "Student," as his employers, the Guinness Brewery of Dublin, did not at that time allow its staff to publish research openly. (The name of the brewery is now strongly associated with the statistics of the world's extremes through sponsorship of the "Guinness Book of Records.")

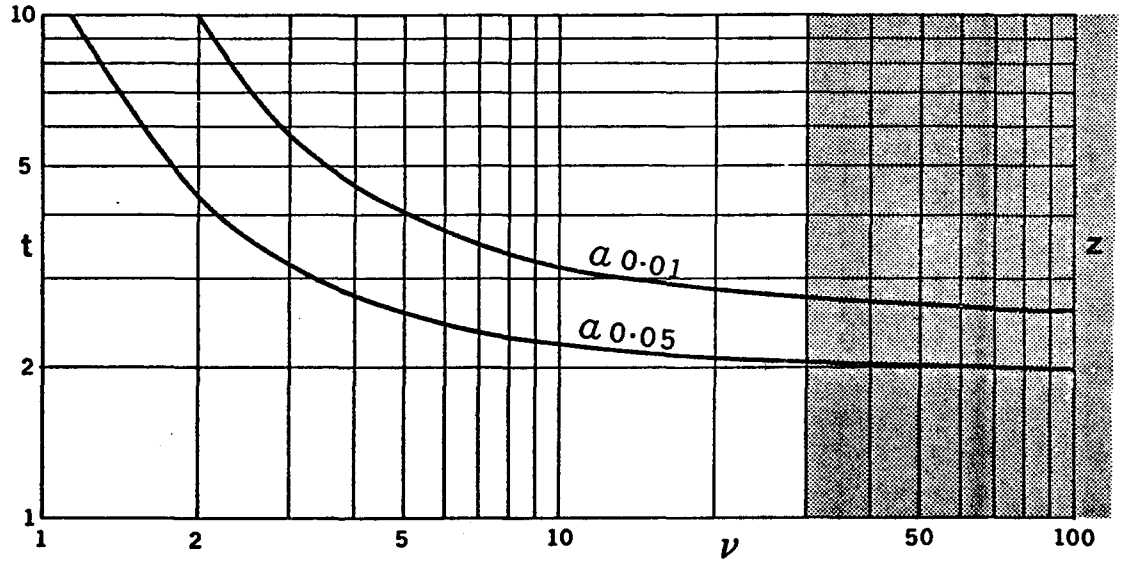
In 1908, Student investigated the distribution of the quantity $(\bar{X}-\mu)/(s/\sqrt{n})$ for small samples. By the central limit theorem, the distribution of this variable approaches the standard normal distribution for large samples and the variable is designated as a Z-score. For small samples ($n > 30$), the quantity is denoted by the variable, t. As n grows larger, the t-distribution approaches the normal distribution, being effectively equivalent when n is about 30 and becoming exactly coincident at an infinite size sample.



The t-distribution is symmetrical and standardized with a mean of zero but, unlike the standard normal distribution, its standard deviation is greater than one. The reason for this is that s is used in the formula in place of σ. As s tends to underestimate σ the result is that the distribution of t is more dispersed in small samples.

The formula for the density function of t is complex but is numerically summarized in standard statistical tables according to the number of degrees of freedom ν and selected probability values corresponding to areas of the distribution. The degrees of freedom are equal to the sample size minus one, while significance levels can be specified to delimit critical areas under the distribution. (Remember that a distribution area is related to a selected significance level according to whether the hypothesis test is one-tailed or two-tailed.)

The graph shows why a sample size of 30 is commonly chosen as a pragmatic limit between "small" and " large" samples. When $n > 30$, the values of t rapidly stabilize as constant values which are the Z-scores of a normal distribution. The values shown are for a two-tailed test.



The t-distribution is used for the t-test of the null hypothesis that the means of two samples are equal and is applicable under the conditions:

(1) The measured variables are normally distributed.

(2) The population variance of both samples is the same.

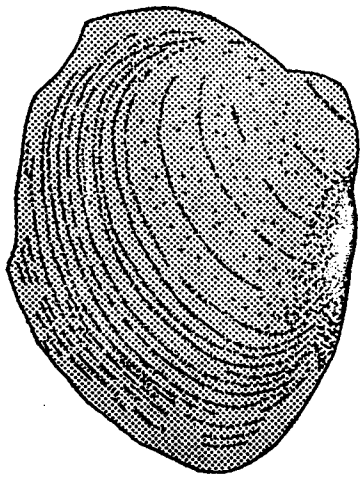
The first condition is an ideal which can be relaxed to a fair degree since moderate deviations from normality do not significantly change the sampling distribution of t. Because of this property, the t-distribution is said to be robust. Fulfillment of the second condition by two samples may be checked by computation of an F-ratio as a test of the null hy-

pothesis that the two sample variances are estimates of a common population variance (see F-test). If the two variances are not equivalent, a modified version of the t-test may be applied as described in the second example.

Example: Samples with unknown but common population variance

Growth rates of modern marine bivalves were measured in specimens collected from Canadian Arctic and Sub-Arctic waters by Andrews (1972). The aim of the study was to establish a relationship (if any) between sea-water temperature and salinity level with bivalve growth rates. Systematic dependencies could then be applied to the interpretation of temperature fluctuations in the Late Quaternary based on fossil shell measurements. The relevant sample statistics relating to growth rates in mm per year for *Serripes groenlandicum* specimens were:

Arctic	Sub-Arctic
$n_1 = 3$	$n_2 = 4$
$\bar{X}_1 = 5.0$	$\bar{X}_2 = 6.6$
$s_1 = 0.56$	$s_2 = 0.68$



Serripes groenlandicum
(after a photograph by Andrews, 1972)

The null hypothesis is postulated that there is no difference in mean growth rate of the bivalve in the two provinces, $H_0 : \mu_1 = \mu_2$.

The alternative hypotheses are non-directional so that the appropriate test statistic is two-tailed.

It may be possible to argue on zoological grounds that the variances of each sample are likely to be equivalent in that, while the growth rate may vary with marine conditions, variations about the mean will be approximately similar. Even if this is not strictly true, the actual difference between the measured sample variances is small enough to ensure that any necessary correction is fairly minor.

As a first step, the variation in both samples is pooled in a common estimate of the population variance they represent by the formula:

$$s_p^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = 0.56$$

$$s_p = 0.75$$

The best estimate of the standard error of the difference in the means of the two samples is:

$$s_e = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.57$$

The difference between the two means is:

$$\bar{X}_2 - \bar{X}_1 = 1.6$$

Then,

$$t = \frac{\text{Difference of means}}{\text{Standard error of differences}} = \frac{\bar{X}_2 - \bar{X}_1}{s_e} = 2.81$$

The number of degrees of freedom, $\nu = n_1 + n_2 - 2 = 5$. The critical value of t with $\nu = 5$ and $\sigma = 0.05$ in a two-tailed test is 2.57.

The null hypothesis that there is no difference between the mean growth rates is therefore rejected. The alternative hypothesis is accepted that the annual growth rate of *Serripes groenlandicum* differs between Arctic and Sub-Arctic waters.

Reference

Andrews, J.T., 1972, Recent and fossil growth rates of marine bivalves, Canadian Arctic, and Late Quaternary Arctic marine environments: Palaeogeogr., Palaeoclimatol., Palaeoecol., v. 11, p. 157-176.

Example: Samples with unknown and probably unequal population variances

Trace element concentrations in non-detrital carbonate fractions were determined in core samples from wells penetrating the Upper Devonian Sturgeon Lake Reef and its off-reef lateral equivalent by Chester (1965). Sturgeon Lake is an oil-productive Leduc Formation reef in Central Alberta and is dominantly dolomite with subsidiary limestone. The lateral off-reef facies is designated the Ireton Formation and is a shale succession with minor limestone beds. The purpose of the study was to determine the worth of trace elements as facies indicators. In this example, consideration is directed to the variation in vanadium content between reef and non-reef facies. Vanadium analyses for individual core samples were summarized as grand means for each well. While the individual analyses would be expected to be lognormally distributed (q.v.), the distribution of well means will probably conform to a normal distribution in the limit (the central limit theorem). The sample statistics of reef and non-reef well means are:

Reef wells	Non-reef wells
$n_1 = 6$	$n_2 = 13$
$\bar{X}_1 = 2.67$	$\bar{X}_2 = 8.08$
$s_1 = 1.03$	$s_2 = 3.06$

$$H_0 : \mu_1 = \mu_2$$

In this case, $s_1 \neq s_2$ and so they are not pooled in a common estimate of a population σ. The best estimate of the standard error of the difference of means is then:

$$s_e = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.95$$

The approximate number of degrees of freedom is given by the formula:

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

$$\nu \approx 14$$

Then,

$$t = \frac{\bar{X}_2 - \bar{X}_1}{s_e} = 5.69$$

The test is two-tailed and the critical value of t at $\nu = 14$ and $\alpha = 0.05$ is 2.15. The null hypothesis is rejected and the motivated hypothesis accepted that there is a significant difference between the mean vanadium content in reef and non-reef well non-detrital fractions.

Reference

Chester, R., 1965, Geochemical criteria for differentiating reef from non-reef facies in carbonate rocks: Am. Assoc. Petroleum Geologist Bull., v. 49, no. 3, p. 258-276.

Exercise 9.1: Albertan dinosaurs

Upper Cretaceous beds exposed in the Red Deer River valley in the vicinity of Drumheller, Alberta are among the most prolific dinosaur bone localities in the world. Complete and fragmentary dinosaur skeletons are found both within the Oldman Formation and the overlying Edmonton Formation, the two units being separated by the marine Bearpaw Shale. The tabulated data records the stratigraphic depths below the base of the Bearpaw, at which specimens of the ceratopsian (horned dinosaur) genera *Centrosaurus* and *Chasmosaurus* have been discovered in the Oldman Formation between Steveville and Deadlodge Canyon (from Dodson, 1971). The ID number for each skeleton corresponds to an official quarry designation whose locations are marked on a Canadian Geological Survey map (Sternberg, 1950).

Centrosaurus		*Chasmosaurus*	
ID	D	ID	D
5	173	9	118
33	144	24	129
42	152	31	164
65	245	37	95
66	197	41	150
74	167	49	109
94	222	83	203
95	221		
100	292		
104	256		
107	232		

ID: Quarry number

D: Stratigraphic depth below base of Bearpaw Shale (feet)

If the frequency of fossil occurrence within the range of a dinosaur genus can be adequately modelled by a normal distribution, then the depth statistics of different genera may be compared analytically.

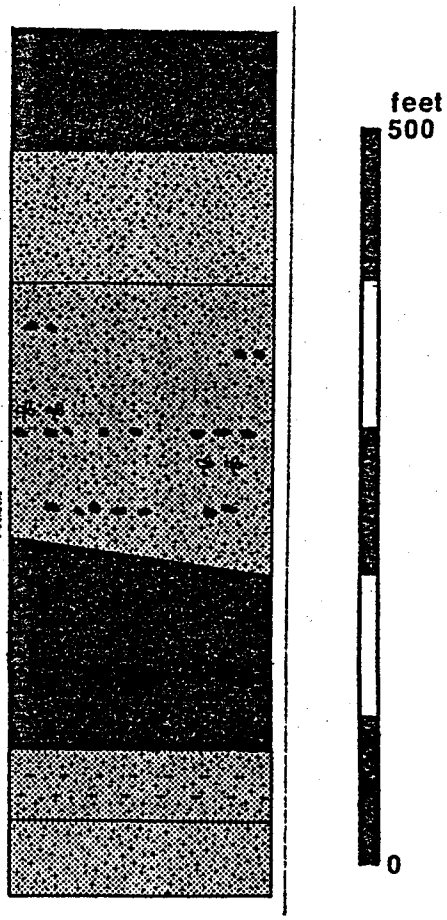
(1) Test the null hypothesis that the mean depths of *Centrosaurus* and *Chasmosaurus* skeletons are common estimates of the same parameter. Do not assume equal population variances.

(2) Compute and plot a normal distribution fitted to the stratigraphic depth variation for each genus, using the sample totals, means and variances. What is the critical stratigraphic level at which a fossil skeleton has an equal chance of being either of the two genera?

References

Dodson, P., 1971, Sedimentology and taphonomy of the Oldman Formation (Campanian), Dinosaur Provincial Park, Alberta (Canada): Palaeogeogr., Palaeoclimatol., Palaeoecol., v. 10, p. 21-74.

Sternberg, C.M., 1950, Steveville west of the 4th meridian, with notes on fossil localities: Geol. Surv. Can., Map 969A.



Centrosaurus skull excavated at Quarry 42 with location referenced to local stratigraphic succession. (Section adapted from Dodson, 1971; photograph courtesy of the National Museums of Canada, Ottawa).

10. F-DISTRIBUTION

Suppose a large number of samples of size n_1 are drawn randomly from a normal population having a variance σ^2. Another large set of samples of size n_2 is taken from the same population. Sample variances are computed for each sample in both sets. Every variance s_1^2 of the first set is then divided by every variance s_2^2 of the second set to form the ratios s_1^2/s_2^2. Since the values of both s_1^2 and s_2^2 change from sample to sample, the ratio also changes from one pair of samples to another pair. The continuous frequency distribution of these variance ratios is the F-distribution.

The F-distribution was originally devised by R.A. Fisher in the 1920's, and subsequently modified into a more convenient form by George Snedecor, who named the distribution in Fisher's honor. The F-distribution is widely used as a test criterion in the analysis of variance and in regression.

The ratio of variances from two samples will have an F-distribution provided:

(1) both samples are selected randomly from their respective populations;

(2) both populations are normal;

(3) the population variances are the same; i.e.,

$$\sigma_1^2 = \sigma_2^2 .$$

The form of the F-distribution is determined by the size of the samples used to calculate the variances. These are the degrees of freedom, designated ν_1 and ν_2, (q.v.). The first degree of freedom is associated with the variance in the numerator of the ratio and is n_1-1. The second degree of freedom is that associated with the denominator of the ratio and is n_2-1. Although two variance ratios can be calculated for any pair of variances (i.e., s_1^2/s_2^2 or s_2^2/s_1^2), by convention the larger variance is placed in the numerator.

The null hypothesis of an F-test is that the variances calculated from two samples represent estimates of the same population variance. That is,

$$H_0 : \sigma_1^2 = \sigma_2^2 .$$

If this is true, the ratio of the sample variances should be nearly equal to one, but because of sampling fluctuations, the ratio will depart from this ideal. As the larger variance is placed in the numerator, an F-test provides a probabilistic answer to the question: "Does the ratio of these two sample variances exceed that expected if both samples had been randomly selected from the same population?".

Example: Shapes of drainage basins in Kentucky

Geomorphic measurements were made on a collection of randomly selected drainage basins in Kentucky by Krumbein and Shreve (1970). Among other variables, the basin shape was characterized by the ratio of the largest circle that could be inscribed within the outline of a basin, to the smallest circle that could be circumscribed around the basin. Ninety of the basins were selected as having 10th magnitude streams, which is essentially a count of the number of sources in the basin. A second sample of size 90 was chosen from third order basins, order being defined by the number of successive levels of stream junctions from the sources to the point where a stream joins another of equal or higher order. We may hypothesize that these are random samples from the same normal population.

	Basin shape (ratio D_I/D_c)	
	Basins of magnitude 10	Basins of order 3
Minimum ratio	0.24	0.24
Maximum ratio	0.80	0.88
$\bar{X}$	0.55	0.52
s^2	0.0108	0.0139
s	0.1040	0.1178

The hypothesis to be tested is

$$H_0 : \sigma_1^2 = \sigma_2^2 .$$

The test statistic is

$$F = \frac{s_1^2}{s_2^2} = \frac{0.0139}{0.0108} = 1.29 .$$

From tables of the F-distribution, we can determine that the critical region for rejection with a 5% probability of making a Type I error (rejecting a hypothesis when it is true), for $\nu_1 = 89$ and $\nu_2 = 89$ degrees of freedom, is 1.49. We cannot conclude from these measurements that there is a difference in the shape of drainage basins in this area, whether the basins are defined by magnitude or order.

Reference

Krumbein, W.C., and Shreve, R.L., 1970, Some statistical properties of dendritic channel networks: Tech. Rept. 13, Office of Naval Research, ONR Task No. 389-150, 117 p. (Available from NTIS, Arlington, VA, as document AD 705 625.)

11. ANALYSIS OF VARIANCE

Analysis of variance is the name applied to a host of statistical techniques whose purpose is to subdivide the total variability in a set of data into components which can be attributed to different sources. This is usually done by designing the sampling pattern in such a manner that individual observations can be combined in different ways and regarded as composite observations. The total variation among the individual observations is thereby split into the variation between the various composites. The null hypothesis is that the different variances all arise from the same parent population. If this hypothesis is rejected, this indicates that there are systematic components to the variation within the data.

One-way analysis of variance

The simplest form of analysis of variance is a one-way design, used to assess a single source of possible variation. This is basically an extension of the t-test to the situation where more than two samples are being compared. In an experiment to assess the analytical consistency of five different laboratory teams, a single block of calcareous shale was ground, homogenized, and split into 20 samples. Four samples were randomly selected and given to each group for determination of $CaCO_3$ content.

Team (j)	Measured $CaCO_3$ (i)				Team means ($\bar{X}_j$)	Team sums (ΣX_j)	Team sum of squares $(\Sigma X_j)^2$	$(\Sigma_j X_j^2)$
1	11	18	24	15	17	68	4624	1246
2	26	25	22	11	21	84	7056	1904
3	13	19	22	10	16	64	4096	1114
4	26	21	19	22	22	88	7744	1962
5	19	27	28	22	24	96	9216	2358

Grand mean ($\bar{\bar{X}}$)....20

Grand total $\left(\sum_{j=1}^{5}\sum_{i=1}^{4} X_{ij}\right)$....400

Grand sum of squares $\left(\sum_{j=1}^{5}\sum_{i=1}^{4} X_{ij}^2\right)$....8586

The total variation in the data set is given by the sum of the squared deviations of the individual analyses around the grand mean:

$$SS_T = \sum_{j=1}^{5} \sum_{i=1}^{4} (X_{ij} - \bar{\bar{X}})^2 = 586$$

However, this can be equated to

$$SS_T = \sum_{j=1}^{5} \sum_{i=1}^{4} [(X_{ij} - \bar{X}_{.j}) + (\bar{X}_{.j} - \bar{\bar{X}})]^2$$

because all that has been done is to add and subtract the team means to each term. If this expression is expanded, some terms will cancel out leaving

$$SS_T = \sum_{j=1}^{5} \sum_{i=1}^{4} (X_{ij} - \bar{X}_{.j})^2 + \sum_{j=1}^{5} \sum_{i=1}^{4} (\bar{X}_{.j} - \bar{\bar{X}})^2$$

$$586 = 402 + 184$$

The first of these terms can be called SS_W, because it is the sum of squares within the analyses made by each team. The second term can be called SS_B, because it is the sum of squares between the analytical teams.

SS_B essentially is the variance of the team means, if it is divided by the number of teams minus one (5-1=4). This variance is an estimate of the random variation in the original material, plus any bias by one or more of the analytical teams that may exist. Because the team means have been subtracted out of SS_W, it estimates only the random variation of the parent population. This variance estimate is based on the n original observations, less the m team means which were also estimated from the same data. Therefore, the two variances can be compared by an F-test; if they are found to be the same, they both estimate the population variance and any bias by the analytical teams is negligible. On the other hand, if the variances are different, one or more of the teams is producing analyses which are systematically different from those produced by the other teams.

Statistics calculated for an analysis of variance are traditionally presented in an ANOVA (ANalysis Of VAriance) table. The ANOVA for a one-way analysis has the form

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F-test
Between treatments	$SS_B = \Sigma\Sigma(\bar{X}_{.j} - \bar{\bar{X}})^2$	m-1	$MS_B = \frac{SS_B}{m-1}$	$\frac{MS_B}{MS_W}$
Within treatments	$SS_W = \Sigma\Sigma(X_{ij} - \bar{X}_{.j})^2$	n-m	$MS_W = \frac{SS_W}{n-m}$	
Total	$SS_T = \Sigma\Sigma(X_{ij} - \bar{\bar{X}})^2$	n-1		

Here, m is the number of "treatments" (in this case, different analytical teams), and n is the total number of observations. For the carbonate analyses,

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F-test
Between teams	184	4	46.0	1.72
Within teams	402	15	26.8	
Total	586	19	30.8	

The test statistic follows an F-distribution with 4 and 15 degrees of freedom. The 5% critical value for such a distribution is F = 3.06. Therefore, the ANOVA does not suggest that there is any difference in the analytical results of the five laboratory teams.

More advanced designs

By arranging the data table so rows represent differences in one type of treatment and columns represent differences in another type of treatment, a two-way ANOVA can be constructed. Each row-column combination forms a cell; if more than one observation is used to represent each cell, these are referred to as replicates. It is then possible to test for differences between the row treatments, differences between the column treatments, and for row-column treatment interactions.

Still more complex designs can be created, by nesting levels of treatment inside one another, or examining the effects of three treatments simultaneously by a Latin-square ANOVA. Here, different conditions

of three different treatments are assigned randomly to the cells formed by rows and columns, subject to the constraint that each level of each treatment appears only once in any row or column. Obviously, such an experiment must be carefully designed prior to the collection of the data.

A key point is the role of randomization, which means that the experimental material must be assigned to the different treatments without bias. This insures that the effects of any other possible systematic sources of variation, which are not being considered, are confounded or spread over all the treatments rather than being concentrated in a few of them.

Exercise 11.1: Recovery of strip-mined land.

A number of different coal seams have been strip-mined in southeastern Kansas since the late 1800's. The overburden above these coals, now in spoils piles, may have different physical and chemical characteristics which effect the ease with which the mined land can be reclaimed. In particular, the pH is important because the acidity of the material in the spoils piles determines the amount (and cost) of lime necessary for neutralization. Common treatment practices can be established if the pH is not significantly different for overburden from above different coal seams. If the spoils piles prove significantly different in pH, this information may be useful to establish costs and priorities for reclaimation.

The table below* gives pH measurements made at a series of test plots over four commerical seams. Test the hypothesis that pH is the same in overburden from all four coals.

Mulky coal	Mineral coal	Weir-Pittsburg coal	Bevier coal
7.170	6.409	4.689	7.029
7.475	4.011	5.317	6.618
6.829	6.825	4.614	6.967
6.765	6.605	5.275	5.869
6.890	5.724	4.382	6.600
7.080	5.879	5.143	7.046
6.136	7.465	5.164	5.325
	6.693	4.910	5.157
	6.836	5.143	6.314
			7.640
			6.620
			7.543

*From Final Report 1971-72 of the Mined Land Redevelopment Project Ozark Regional Commission.

12. GEOLOGIC SAMPLING

"Making sure that the sampling was random presented no difficulties in small "scrapings" and the like, but became a big problem in large exposures. Subconscious selection...is often difficult to detect, yet may be seriously invalidating to the final results. All sampling spots were therefore chosen blindly. In large exposures a hammer or chisel was whirled horizontally round the head five times and then flung towards the rock face with the thrower's eyes closed. If it hit the bed, then the sample was taken from that point; if not, then that part immediately above the spot at which it fell was sampled."

-P. Allen (1948)

Sampling patterns

The objective of sampling is to insure that observations are representative of the population from which they are taken. In the case of samples taken from a geologic body, this may be done in two ways which are completely opposite in philosophy. The first approach is to sample uniformly, taking observations in a regular pattern which completely covers the body being sampled. Examples include grid drilling schemes used in some oil exploration projects and in mine evaluation programs, and point counting of thin sections for petrographic analysis. The opposite approach is random sampling, in which observations are taken in a way that insures that every potential observation has an equal chance of being selected.

The primary disadvantage of uniform sampling is the possibility of bias if the pattern coincides with an unrecognized regular structure in the sampled population. However, there are certain analytical procedures that can be performed only if data are taken in a regular pattern (time series procedures and their two-dimensional extensions, for example, require regularly spaced samples). Random samples are by definition free of bias, but the precision with which local variations that may exist in the sampled body can be detected is not uniform across the body.

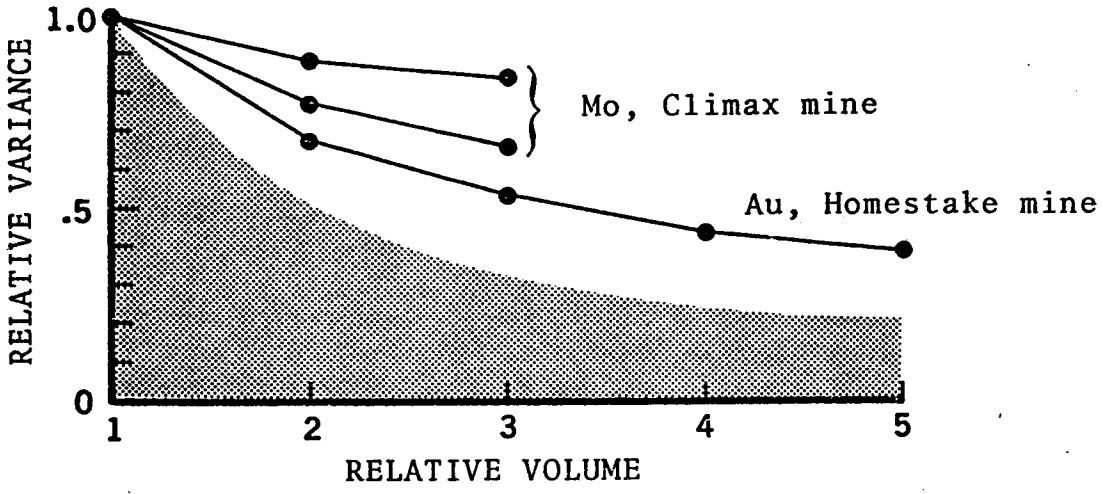
Griffiths and Ondrick (1968) provide an extensive analysis of three different sample patterns suitable for layered sediments, and by extension, for any zoned deposit. These patterns include spot samples in which observations are taken at random locations in the sampled body (in their example, the face of a gravel pit in a glacial till). Stratified samples are those in which the rock body is divided into presumably homogenous zones such as till layers and samples are selected at random from within each zone. Channel samples involve systematically collecting composite samples along a series of randomly placed traverses that cross any zones that might exist.

On the basis of their sampling experiments, Griffiths and Ondrick conclude that it is necessary to use a combination of channel and stratified sampling. Channel samples are self-weighting (i.e., the importance attached to observations from a given layer is proportional to the thickness of the layer), and yield unbiased estimates of the population mean and variance even in the presence of zoning or layering. If stratified samples yield statistics that are significantly different from those of channel samples, this is evidence that layering is present. The stratified samples then yield estimates of the means and variances of the layers. Spot sampling seems to require too many spots and too many observations per randomly located spot to compete with channel samples (i.e., the method is inefficient). Also, a spot sample pattern cannot supply information about the presence or absence of layering and so fails to compete with the stratified sampling design.

In their specific experiment, Griffiths and Ondrick found that approximately 300 or more measurements of pebble lengths were necessary to achieve stability in their statistical estimates, and about 400 measurements to be optimal. The number of observations per channel or per layer was critical. They found five channels with 70 observations per channel to be adequate, or 20-30 observations in each of 15-20 layers for stratified samples.

The importance of sampling has long been recognized in mining, where accurate estimates of ore grade are essential prior to mining. Since large volumes of rock are subsequently mined and the material of interest extracted *in toto*, mining geologists have had excellent

opportunities to check the reliability of various sampling methods. Much of this work is summarized in Koch and Link (1970). They find that diamond-drill samples and well-taken channel samples are superior to chip samples (spot samples taken along a line across a face) which in turn are superior to spot samples. Koch and Link also conclude that the oft-quoted theory that the variance of samples is inversely proportional to the volume of the samples does not hold. In general, larger volumes of rock yield somewhat smaller variances, but the variance due to differences in sample volume is much smaller than differences commonly noted between the samples.



Numbers of samples

The number of samples necessary in a study can be calculated if the variance of the population is known, and the desired precision can be specified. Then, the value of n necessary to reduce the standard error of the statistic to any desired level can be calculated. For example, suppose the population mean must be estimated within ± d, at some level of significance α. The test

$$Z = \frac{x-\mu}{\sigma/\sqrt{n}}$$

can be rewritten to solve for n:

$$n = \left(\frac{Z\,\sigma}{X-\mu}\right)^2$$

where $(X-\mu)$ is the difference d, Z is the standard normal deviate corresponding to the desired $\alpha/2$ level (remember, the interval is two-sided, as the true mean must be within the interval $\mu \pm d$), and σ is the population standard deviation.

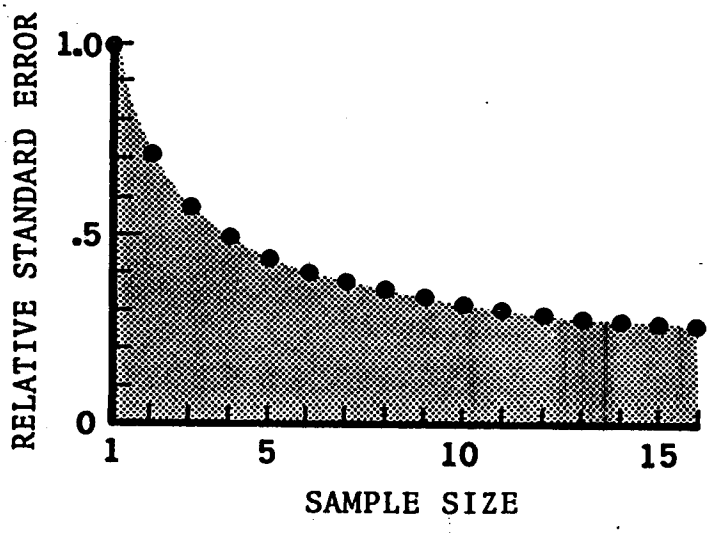
Of course, the variance is seldom known, so it is necessary to use the equivalent t-test:

$$n = \left(\frac{t\ s}{X-\mu}\right)^2$$

where t is the t-value corresponding to the $\alpha/2$ level of significance and m-1 degrees of freedom. The sample variance must be computed from a prior sample of size m. The necessary number of additional observations is n-(m+1).

Therefore, determination of the sample size requires two-stage sampling; first to estimate the variance in order to calculate the standard error, and next to achieve the desired precision. If the number of additional samples is much larger than the size of the initial sample, the estimate of n is probably conservatively large because the estimated variance is inflated. The final sample size may be recalculated part-way through the sampling procedure to see if reduction in the standard error has significantly reduced the necessary number.

This procedure assumes the variable being measured is approximately normally distributed and the samples are taken without bias. It should be noted that past a certain sample size, the reduction of standard error becomes increasingly difficult.



Sampling nets

Many geologic problems involve searching an area for a rare event, usually a mineral deposit. One way in which this can be done is by systematic uniform sampling on a regular grid of points. The probability of "hitting a target" is a function of the abundance of the targets and target size, shape, and orientation relative to the pattern and spacing of the search grid. These probabilities have been extensively investigated by Singer and Wickman (1969), who use methods of geometric probability. For example, if the radius of the target is less than half the grid spacing, the target may either be missed completely or hit only once. The probability of a hit is equal to the area of the target divided by the area enclosed in one cell of the grid. If the radius of the target is larger than half the grid spacing, the possibility of multiple hits exists and the geometrical relationships become more complex. Singer (1972) provides a computer program for generating tables of the probability of a hit on targets of specified size with rectangular or triangular grids of various dimensions.

Minimum spacing required between samples to map a continuous variable:

The density with which control points must be taken to map a continuous variable such as a subsurface structural horizon may be estimated by calculating the semivariance or autocorrelation (q.v.) of the surface. To calculate these functions it is necessary to take an initial sample of measurements at equally spaced points along a transverse across the area to be mapped. If the surface to be mapped is stationary, the semivariance and autocorrelation provide equivalent information. If the surface is not stationary, it is necessary to remove a drift, usually approximated by a low-order polynomial fitted to the observations by least-squares. The semivariance is then calculated for the residuals from the drift.

When the residuals are stationary, the semivariance is found as

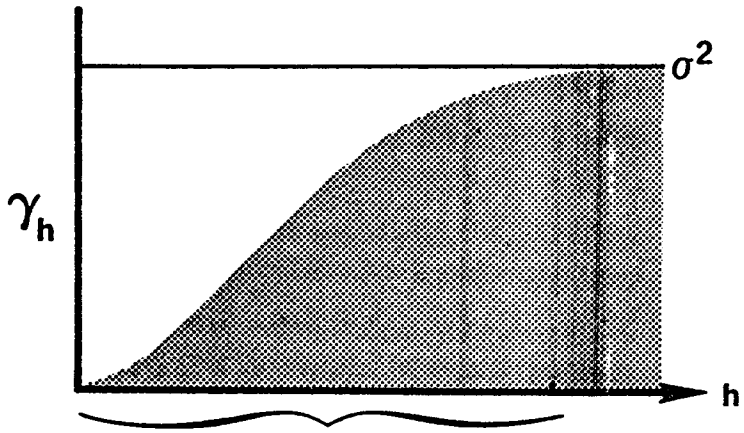
$$\gamma_\tau = 1/2 \; \Sigma \; (Y'_t - Y'_{t-\tau})^2$$

where Y' is the residual at point t. A plot of semivariance versus lag (the semivariogram) has the form shown. From this diagram, the zone of influence can be determined. When the semivariance asymptotically

approaches the variance, there is a limit (called the range) beyond which two samples are statistically independent. The range is that radius L such that

$$\text{var } (Y') - \gamma_L \leq \varepsilon$$

where ε is any small number. The range L divides the samples into two categories. All samples whose distances to a point to be estimated are less than or equal to L provide information about the point. All samples outside the neighborhood defined by L are independent observations with respect to the point to be estimated. Therefore, the minimum distance between samples points on the map must be less than 2L if all areas of the map are to have some degree of sample control.



In practice, the semivariogram may not be the same for all orientations leading to non-circular zones of influence. This indicates the surface is anisotropic, and different degrees of sampling control are required for different orientations across the map.

References

Griffiths, J.C., and C.W. Ondrick, 1968, Sampling a geological population: Kansas Geological Survey Computer Contribution 30, 53 p.

Koch, G.S.Jr. and R.F. Link, 1970, Statistical analysis of geologic data: John Wiley & Sons, Inc., New York, 375 p.

Singer, D.A., and F.E. Wickman, 1969, Probability tables for locating elliptical targets with square, rectangular, and hexagonal point-nets: Penn State Univ., Mineral Sciences Experiment Station, Special Publ. 1-69, 100 p.

Singer, D.A., 1972, ELIPGRID, a FORTRAN IV program for calculating the probability of success in locating elliptical targets with square rectangular and hexagonal grids: Geocom Programs, no. 4, 16 p.

13. CHI-SQUARE DISTRIBUTION AND TEST

The χ^2 distribution was derived by Helmert in 1875 and rediscovered by Pearson in 1900. In its simplest form, the distribution is related to the standard normal distribution by:

$$\chi^2 = Z^2$$

which functions as a curvilinear transformation. This may be extended to a more general statistic by summing several Z^2 scores drawn from the same normal distribution, when:

$$\chi^2 = \sum_{i}^{n} Z_i^2 = \sum_{}^{n} \frac{(X_i-\mu)^2}{\sigma}$$

Chi-square is directly related to the sampling distribution of the variance, since

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

"Student" could not derive this formula mathematically, but showed that it held true for small sample variances calculated for heights of 3000 criminals.

The distribution is uniquely specified by ν (the number of degrees of freedom) which is the (n-1) term in the above expression. As a square power, the variable is always positive. Standard statistical tables for each distribution (corresponding to a different value of ν) list chi-square values for various levels of α (marking areas under the distribution).

The normal distribution is often used as a continuous approximation of the (discrete) binomial distribution when samples are not prohibitively small. In a similar fashion, the chi-square distribution is used as a continuous approximation of the multinomial distribution, which applies to situations where an observed event may take one of several outcomes. This provides the key to the most widely used application of the chi-square test in the comparison of proportions of entire distributions. If measurements in a sample fall into m different and mutually exhaustive classes, then:

$$\chi^2 = \sum_{}^{m} \frac{(O-E)^2}{E}$$

where O = the observed frequency in each class
and E = the expected frequency (as predicted from a null hypothesis distribution).

The number of degrees of freedom are the number of classes (m) minus the number of independent values used in estimating the expected frequencies. (This point is explained in the examples.) Pearson used this relationship as a test of association between categorical attributes and as a goodness-of-fit test between observed and expected distributions. These two applications are illustrated in the examples.

Example: Pearson χ^2 Test of Association

The Pennsylvanian Coal Measures of Ayrshire, Scotland, are comprised of a dominantly nonmarine succession of shales, siltstones and sandstones, with frequent coal seams and rootlet horizons. The relationship between thickness of coal seams and the nature of the succeeding lithology are shown in the contingency table which summarizes frequencies counted from 26 borehole records:

Coal thickness (in.)	"Roof" lithology			
	Shale	Siltstone	Sandstone	Totals
0 (Rootlets)	373	197	111	681
1-6	238	62	26	326
7-18	196	42	23	261
19-42	126	21	7	154
42+	44	2	2	48
Totals	977	324	169	1470

As a motivated hypothesis, it may be postulated that there is an association between coal seam thickness and the overlying lithology. The null hypothesis is then equated with an expectancy of no association or independence, where the joint probabilities are the product of their component marginal probabilities. These joint probabilities multiplied by the grand total of the sample generate the expected frequencies in the contingency table cells under the null hypothesis of independence (as discussed in the section on Probability). The quantity $(O-E)^2/E$ is computed for each cell and summed for the entire table. Then,

$$\chi^2 = 88.72$$

For a contingency table of r rows and c columns, the number of degrees of freedom, $\nu = (r-1)(c-1)$

$$\nu = 8$$

The table for the chi-square distribution with $\nu = 8$ is consulted and the critical value of χ^2 at $\alpha = 0.05$ is found to be 15.51. The computed statistic greatly exceeds this value so that the null hypothesis of independence may be rejected with a probability of considerably less than one in twenty of being wrong. The alternative hypothesis of

association between the categorical attributes of coal thickness and succeeding lithology is therefore accepted.

When a chi-square test of association is applied to a 2 x 2 contingency table, the degrees of freedom are reduced to one. Two conditions should be observed:

(1) A minimum expected frequency of 10 must be observed in order to consider the test adequate.

(2) A modification known as Yates' correction for continuity should be applied to the computational formula:

$$\chi^2 = \Sigma \frac{(|O-E|-0.5)^2}{E}$$

This correction applies only to cases of one degree of freedom. In larger tables, a general rule of a minimum expected frequency of 5 should be observed as a conservative safeguard on the test. An absolute minimum expected frequency of one must be observed to prevent spurious inflation of the computed chi-square. In situations where this basic condition is not satisfied, either a larger sample must be taken to increase the frequencies, or categories must be combined to attain the minimum value. The first course is preferable since the combination of categories results in biasing the sample. However, the selection of categorical boundaries is itself arbitrary, so that the power of the test is influenced to a fair degree by the initial decisions of the investigator.

Example: Pearson Goodness-of-Fit χ^2 Test

In an earlier example, a normal distribution curve was fitted to a sample histogram of skull length measurements of the extinct amphibian, *Trimerorhachis insignis*. The fit was judged to be poor, although the small sample size (33) was not large enough for a conclusive decision. The null hypothesis to be tested is that the sample distribution is drawn from a normal population. Since the sample mean and variance were used as the parameters of the fitted normal curve, these may not be used as test statistics. However, the partition of the distribution histogram into class interval categories enables observed frequencies to be contrasted with normal curve expected frequencies in a chi-square test. The raw data are contained in the table:

Class	O	E
∞ - 50	0	1.09
50 - 60	2	1.68
60 - 70	3	3.04
70 - 80	9	4.72
80 - 90	3	5.84
90 - 100	7	5.74
100 - 110	3	4.82
110 - 120	3	3.20
120 - 130	1	1.72
130 - 140	2	0.76
140 - ∞	0	0.40

where O and E are the observed and normal expected frequencies. In order to maintain a minimum expected frequency of one, the last two classes must be merged. The number of degrees of freedom are the number of classes, 10 (the last two classes counting as one) minus the number of constraints used in computing the expected frequencies. There are three constraints involved, since the normal distribution was computed to have the same total number of skulls, the same mean and the same variance:

$$\nu = 10 - 3 = 7$$

The tabulated critical value of χ^2 at $\alpha = 0.05$ and 7 $\nu = 14.06$. The computed value of χ^2 from the table is 9.33, which is less than the critical value. As a consequence, the null hypothesis that the sample distribution is drawn from a normal distribution is <u>not</u> rejected. This may seem a surprising result. However, it must be remembered that the sample size is fairly limited (33 individuals) and the failure to reject the null hypothesis merely implies that the proportional distribution of the sample is not sufficiently deviant from a normal curve expectation to conclusively reject a normal distribution as the parent population. A larger sample size would reduce the uncertainty involved but, based on

the limited evidence, a normal distribution model cannot be ruled out.

Chi-square tests are normally conducted as one-tailed procedures where the selected significance level is matched with the higher tail of the chi-square distribution. If the computed statistic is higher than the critical value, the null hypothesis is rejected. If the computed statistic is very low and lies in the lower tail of the distribution, then the implications are more esoteric in that the results are "too good to be true." It was mentioned earlier that Mendel's classic genetics experiments appear to have been "faked." In one of these experiments Mendel recorded a sample of second generation seeds resulting from crossing yellow round peas and green wrinkled peas. The results were:

Seed	O	P	E
yellow and round	315	9/16	312.8
yellow and wrinkled	101	3/16	104.3
green and round	108	3/16	104.3
green and wrinkled	32	1/16	34.8
Totals	556	1	556.2

where O is the observed frequency, P the expected probability in Mendel's theory and E, the corresponding expected frequency. The number of degrees of freedom, $\nu = 4 - 1 = 3$ and the computed chi-square value is 0.475. The critical values of chi-square in the lower tail of the distribution for 3 degrees of freedom are 0.352, at $\alpha = 0.05$ and 0.584 at $\alpha = 0.10$.

It follows that for this experiment there was less than a one-in-ten chance of obtaining real-world results which fit so closely to the theoretical model.

The chi-square test is the most widely used example of a non-parametric statistical test. Most tests of hypotheses are directed to ratio statistics from measured samples which estimate parameters such as

means in the case of t-tests and variances for F-tests. As such they are eminently suitable for data measured on the interval and ratio scales. These parameters are not available for nominal and ordinal scales where data are collected in a categorical form such as contingency tables. Non-parametric tests including the chi-square are applied in the elucidation of variable associations against a null hypothesis of independence.

14. DEGREES OF FREEDOM

Degrees of freedom are the number of independent estimates of a population parameter that are contained within a sample. If a random sample of size n is taken from a population, each observation can be regarded as an estimate of the population mean. Because each observation is chosen randomly, they are independent and the number of degrees of freedom associated with the sample mean $\bar{X}$ is n, the number of observations.

Every squared deviation from the population mean is an estimate of the variance. However, if the population mean is not known, it must be estimated by $\bar{X}$. Because $\bar{X}$ is the sum of all the observations in a sample (times $\frac{1}{n}$), only (n-1) of the deviations $(X-\bar{X})^2$ can be independent; the n-th deviation is predetermined. Therefore, the number of degrees of freedom associated with s^2 is n-1. In general, one degree of freedom is lost for every population parameter that must be estimated from the sample in order to calculate another statistic.

In analysis of variance and multiple regression (including trend surface analysis), SS_T has n-1 degrees of freedom because it is an estimate of the variance about the sample mean. The model (which is the m hypothetically different categories in analysis of variance, or the linear relation specified by the fitted equation in regression) is entirely defined by the number of categories or the number of coefficients in the regression equation. Since these m categories or coefficients generate all of the variance of the fitted model about $\bar{X}$, SS_R (or SS_B in analysis of variance) has m-1 degrees of freedom. The sums of squares due to deviation from the model (or error variance) must have degrees of freedom equal to the difference between n-1 and m-1. F-tests, of course, have both ν_1 and ν_2 degrees of freedom because the test statistic is formed as the ratio of two estimated variances, each with its associated degrees of freedom.

In a chi-square test of goodness-of-fit, an attempt is made to estimate the form of a population distribution from a sample frequency distribution in the form of a histogram. Each category of the observed histogram is in effect an estimate of the area of the equivalent portion

of the population distribution. If the histogram has m categories, these constitute m estimates of the form of the parent distribution. However, the histogram must sum to a constant, so only m-1 of these are independent. In addition, a number of population parameters must be estimated from the sample (usually the mean and variance), and one degree of freedom is lost for each of these.

There are rc estimates of the joint frequencies of occurrence in an r X c contingency table. However, in any row, there are only c-1 independent estimates, and only r-1 independent estimates in any column, as the expected frequencies are computed from the row and column totals. As the expected frequencies are found by cross-multiplying row and column totals, as soon as c-1 entries in a row (or r-1 entries in a column) have been found, the remaining element is predetermined. Therefore the degrees of freedom associated with a chi-square contingency table is (r-1)(c-1).

15. LOGNORMAL DISTRIBUTION

Everyone believes in it because the mathematicians think it is a fact of observation and the observers assume it is a mathematical law.

--Poincare, 1898 (on the normal distribution)

The normal distribution was originally applied to the analysis of errors associated with repeated physical measurements such as those made of the location of astronomical bodies. It was later extended to the description of natural variation of attributes measured on biological and social populations. In both case, the "error" is the summation of small arithmetic displacements from the central or expected distribution value. At one time, the normal distribution was thought of as a "universal law" of variation in natural populations. Sample distributions observed to be markedly skewed could be explained as a mixture of differing normal populations. More recently, it has been realized that observed samples may be the product of processes that are multiplicative in character. For example, the formation of particulate matter is, in large part, the result of crushing and breakage. As a result, the distribution of particle size is regulated by a proportioning process and tends to a lognormal distribution.

The lognormal distribution was proposed by Galton and McAlister as a naturally occuring distribution and their theory was extended by Kapteyn (1903). The real-world occurrence of the lognormal distribution was hotly disputed by Pearson (1906) and other critics, but gained acceptance with the development of small particle statistics and econometric theory.

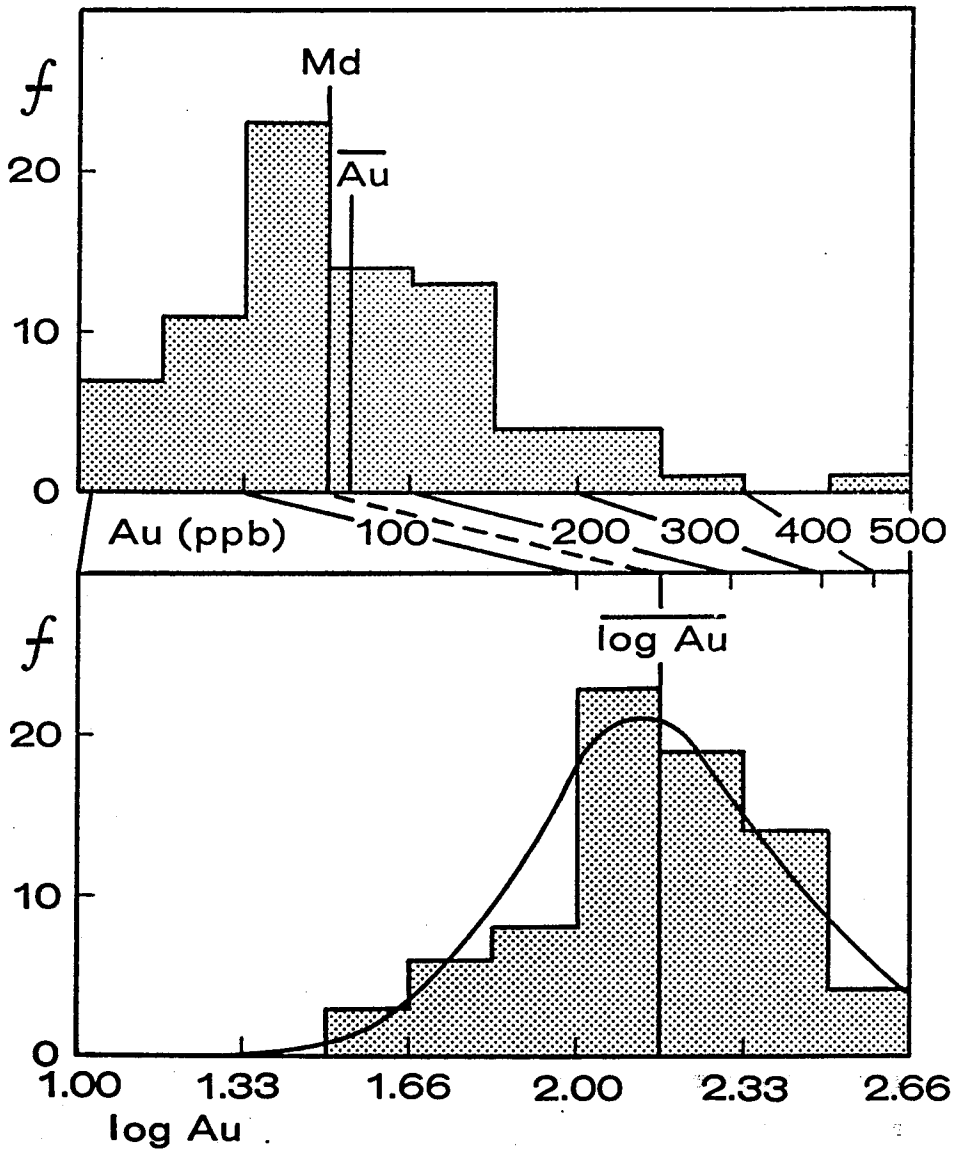
The lognormal distribution is simply the condition where the logarithmic transformation of measurements of a raw variable are normally distributed. The raw variable itself shows a strong positive skew. The mean of the distribution is then at the geometric mean of the original data (the nth root of the product of the n measurements) and equal to the logarithm of the geometric mean on the transformed scale. Computation of the moments of lognormally distributed data follows the conventional formulae after logarithmic transformation of the raw measurements.

Logarithmic transformation in geology was first proposed by Krumbein (1936) as an application to sediment-size distribution analysis. He used a transformation of grain diameters by a negative logarithm to the base two, which he termed phi, ϕ

$$\phi = -\log_2 d$$

The ϕ scale of measurement is a numerical realization of the Wentworth grade scale, expresses the logarithmic properties of grain size distributions, and allows the computation of useful statistical measures. Sand size distributions are clearly the product of a multiplicity of processes reflecting source materials, hydraulic phenomena, etc., but their numerical expression appears to be more closely related to a logarithmic rather than arithmetic scale. It has been shown that alluvial transport of sand results in a lognormal size distribution when the initial bed material is normally distributed (Mahmood, 1973). Moments of ϕ-scaled observations have been widely used as potential indicators of sand depositional environments (see "Moments of a Distribution").

Ahrens (1954) argued "the lognormal distribution of elements" as "a fundamental law of geochemistry" from empirical studies of element distributions in igneous bodies. Element concentrations are commonly highly positively skewed and a logarithmic transformation causes such distributions to be more symmetrical in shape and similar in appearance to a normal distribution. The distribution of the raw data may be judged to be effectively lognormal if the median and geometric mean are approximately the same. While many geochemists would question the dogma of the lognormal distribution as a law, logarithmic transformation of geochemical data is widely used both for data summarization and input for statistical analysis. Since many statistical techniques require normality in data measurements, an approximately normal conformation of transformed data broadly satisfies this condition. The introduction of highly skewed raw data into such techniques largely invalidates their effectiveness as aids in interpretation.



Histograms of gold concentrations (ppb) in 78 soil samples from the Malvern hills, England. (Analyses by D.W. Bullard and P.K. Harvey, University of Nottingham, England). Normal curve fitted to log transformed data.

Sample statistics: Median (Md) = 150 ppb
Arithmetic mean ($\bar{X}$) = 166 ppb
Geometric mean = 146 ppb
Logarithmic mean ($\overline{\log X}$) = 2.165

References

Ahrens, L.H., 1954, The lognormal distribution of the elements, (a fundamental law of geochemistry and its subsidary): Geochimica et Cosmochimica Acta, v. 5, p. 49-73.

Kapteyn, J.C., 1903, Skew frequency curves in biology and statistics: Astronom. Lab., Groningen: Noorhoff.

Krumbein, W.C., 1936, Application of logarithmic moments to size frequency distributions of sediments: J. Sed. Pet., v. 6, no. 1, p. 35-47.

Mahmood, K., 1973, Lognormal size distribution of particulate matter: J. Sed. Pet., v. 43, no. 4, p. 1161-1166.

Pearson, K., 1906, Skew frequency curves. A rejoinder to Professor Kapteyn: Biometrika, v. 5, p. 168.

16. POISSON DISTRIBUTION

There do exist levels where chance is hardly recognized at all... it is music, not without its majesty, this power series

$$Ne^{-m}(1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \ldots\ldots\frac{m^{n-1}}{(n-1)!}),$$

terms numbered according to rocketfalls per square, the Poisson dispensation ruling not only these annihilations no man can run from, but also cavalry accidents, blood counts, radioactive decay, number of wars per year.....

-Thomas Pynchon (Gravity's Rainbow)

The Poisson distribution applies to cases where p, the probability of a success in a trial, is very small ($p < 0.10$) but the number of trials, n, is very large. It is a discrete distribution that is a limiting approximation of the binomial for low values of p, where the distribution becomes highly skewed. The binomial distribution may be applied at this level although the arithmetic becomes very cumbersome. However, a distinct advantage of the Poisson is that it may be applied in situations where the quantity np is known, but neither n (the number of trials) nor p are known independently. This condition occurs in a wide variety of applications where "rare" events are observed to occur within time periods, areas and volumes of fixed size and their non-occurrence ("failure") is not ennumerable.

Poisson model postulates are:

(1) The events are independent.

(2) The probability of an event does not change with time (or area or volume, if spatially measured).

(3) The probability of an event is approximately proportional the size of the interval.

(4) The probability of more than one event in an interval is much smaller than that of a single event.

The Poisson distribution is defined by the Poisson probability, P(X):

$$P(X) = e^{-np}\ (np)^X/X!$$

where e = 2.718

and X = number of events in an interval

The estimate of the mean is np, which is the mean number of events per interval or the average rate of occurrence. The mean is the same as the variance, so that

$$s^2 = np$$

(Note that this is the limiting case of the binomial $s^2 = npq$ as q approaches one.)

The quantity np is usually designated λ.

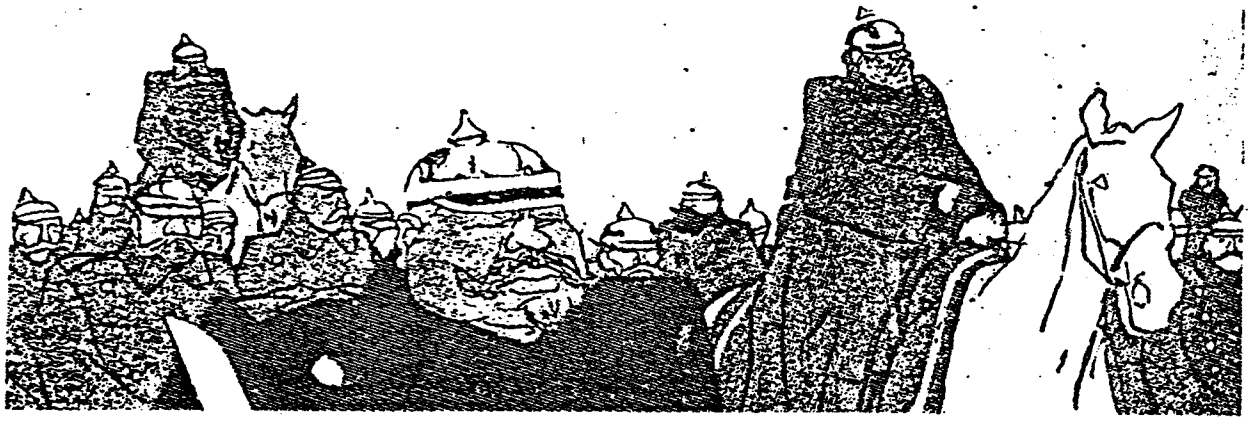
S.D. Poisson (1781-1840) first described the distribution in a páper in 1837 as a theoretical derivation. Practical examples were lacking at first until von Bortkiewicz demonstrated its realization in the real world from bizarre statistics such as the frequency of suicides among children and (the classic) incidence of horse-kick fatalities in the Prussian army. In modern times, the Poisson distribution has been applied to the rate of atomic decay in a radioactive sample, the frequency of malfunctions in mechanical equipment, and the number of arrivals per unit time at service terminals such as check-out counters. The distribution describes phenomena that happen sporadically as "random events" over time or space. As such, it is commonly fitted to experimental data to judge whether observational events appear to be distributed randomly as opposed to alternatives of clustering, ordering, etc. As usual, statistical judgements are phrased negatively. If the Poisson model does not fit, the hypothesis of random distribution is rejected. However, if the Poisson fits, randomness is not rejected but may not necessarily correctly describe the observed situation.

In fitting experimental data, an observation interval of fixed dimension (time, length, area, etc.) is selected that generally contains none or few events. A count is made of the number of intervals (f) that contain zero, one, two...events (X). By conventional computation:

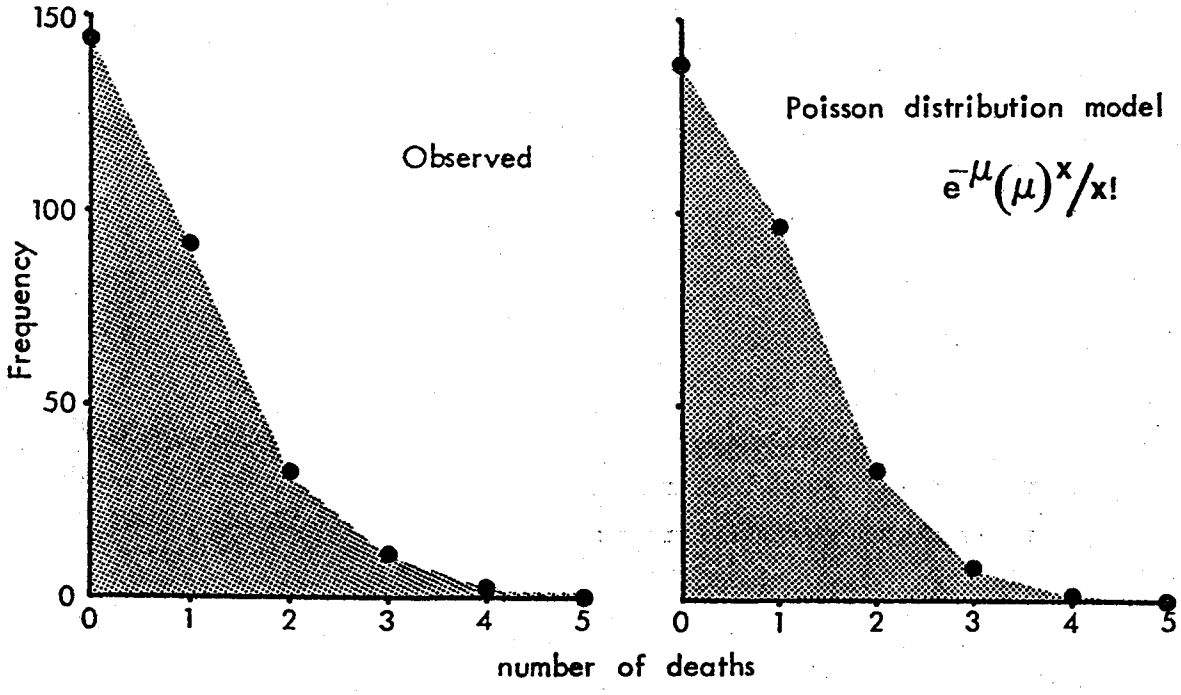
$$\lambda = \bar{X} = \frac{\Sigma fX}{m}$$

Where m is the total number of intervals, and

$$s^2 = (\Sigma fX^2 - m\bar{X}^2)/(m-1)$$



Annual number of deaths from horse kicks in 14 Prussian army corps between 1875 and 1894



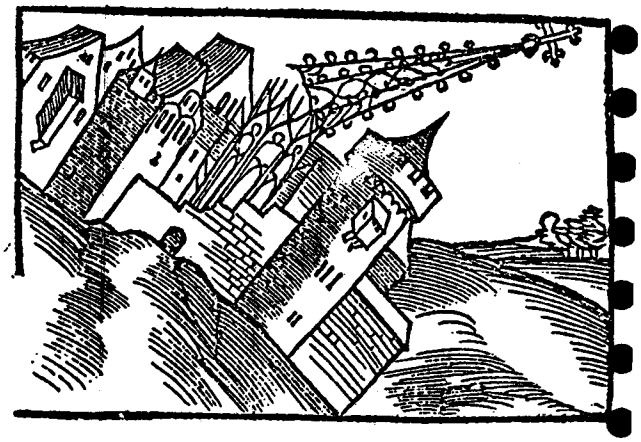
von Bortkiewicz (1898) " Das Gesetz der kleinen Zahlen "

One of the earliest applications of the Poisson distribution to real world data which is often cited as a classic example.

If the observed distribution is the product of a Poisson process, $\bar{X}$ and s^2 should be close in value, subject to the consideration of sample size. For Poisson fit calculations, $\bar{X}$ is used as the maximum likelihood estimator of λ. The procedure is illustrated in the following example.

Example: Earthquakes in Kansas, 1866-1965

Kansas is located in the Central Stable Region of North America and is subject to infrequent, relatively mild earthquake disturbances with epicenters in Kansas and neighboring states. Many of the local shocks are probably related to a major structure, the Nemaha Anticline, and appear to reflect minor adjustments in the basement that occur intermittently (Lee, 1954; Merriam, 1963). The dates of recorded earthquakes experienced in Kansas since 1866 are:



1867	1875	1877	1878	1881	1882
1895	1897	1902	1904	1906	1906
1909	1917	1919	1919	1925	1927
1927	1928	1929	1929	1929	1929
1929	1931	1932	1933	1935	1936
1939	1942	1948	1948	1952	1956
1961	1961	1963	1965		

Total number recorded, n = 40

If numbers of earthquakes (X) are counted for twenty five-year periods (1866-1870,...1961-1965) and summarized as a distribution (f_o), the results are as in the first two columns of the following table. The total number of intervals, m = 20.

Then $\bar{X} = \Sigma x f_o / m = 2.00$

and $s^2 = 3.26$

These two common estimates of λ differ but this may be due either to the small sample size (with poor estimates of their parameters) or a significant deviation from a Poisson model.

Using $\bar{X}$ as the estimate of λ (or np) the Poisson probabilities P(X) for the various levels of X are:

for X = 0: $e^{-\lambda}=0.14$

X = 1: $\lambda e^{-\lambda}=0.27$

X = 2: $\lambda^2 e^{-\lambda}/2 =0.27$

Multiplication of these probabilities by m (the total number of intervals) gives the frequencies (f_e) that would be expected for a Poisson model with the same $\bar{X}$ and m (see the table).

Kansas Earthquake Data, 1866-1965

X	f_o	P(X)	f_e
0	2	0.14	2.71
1	8	0.27	5.41
2	5	0.27	5.41
3	2	0.18	3.61
4	2	0.09	1.80
5	0	0.04	0.72
6	0	0.01	0.24
7	0		0.07
8	1		0.02

The fit of the Poisson model to the observed frequencies may be checked by a chi-square test. Using the chi-square ground rule of a minimum expected tally of one, the last four categories must be combined as a single class with a resulting reduced total of six classes.

The computed total $\chi^2 = 2.20$

The number of degrees of freedom is the number of classes minus the number of constraints used in computing the model ($\bar{X}$ and m). Then,

$$\nu = 6 - 2 = 4$$

The tabulated χ^2 value for 4 ν and $\alpha = 0.05$ is 9.49. So the null hypothesis that the observed and Poisson fit are common estimates of the same distribution is not rejected under the conditions used here.

It may therefore be suggested (in the absence of contrary evidence) that earthquakes experienced in Kansas are basically independent, random occurrences in time. This is contrasted with a possible alternative model of earthquakes clustered in time as multiple reverberations of basement movements. While there is an overall good fit of the Poisson model to the observational data, the strongest discrepancy is provided by a single period (1926-1930) when eight earthquakes were recorded. This is the primary cause of the high value of s^2 relative to $\bar{X}$.

In cases where s^2 is greater than $\bar{X}$ and a Poisson model is rejected, the implications follows that the events are more clustered than would be expected for a random distribution. The descriptive distribution is then likely to be the negative binomial, often known as a "contagious distribution". The negative binomial commonly describes situations where clusters of events are distributed randomly. A geological example is given by Griffiths (1966) who fitted a negative binomial model to the distribution of successes in simulated grid drilling for oil fields in Kansas.

Many random distributions of events appear to be clustered in a non-random fashion when judged by casual observation. This is a reflection of the human tendency to look for patterns and to take special notice of "freak" runs of events while ignoring non-clustered sequences. This has resulted in various superstitions such as the legend that Jesuits commonly die in threes, a belief that is statistically analysed by Solterer (1941). The analysis of distributions of events in space and time by Poisson, negative binomial, and other models provides a means of objectively characterizing their structure.

References

Griffiths, J.C., 1966, Exploration for natural resources: Jour. Operations Res. Soc. Amer., v.14, p. 189-209.

Lee, W., 1954, Earthquake and Nemaha Anticline: Am. Assoc. Petroleum Geologists Bull., v.38, p. 338-340

Merriam, D.F., 1963, The geologic history of Kansas: Kansas Geol. Survey Bull. 162, 317 p.

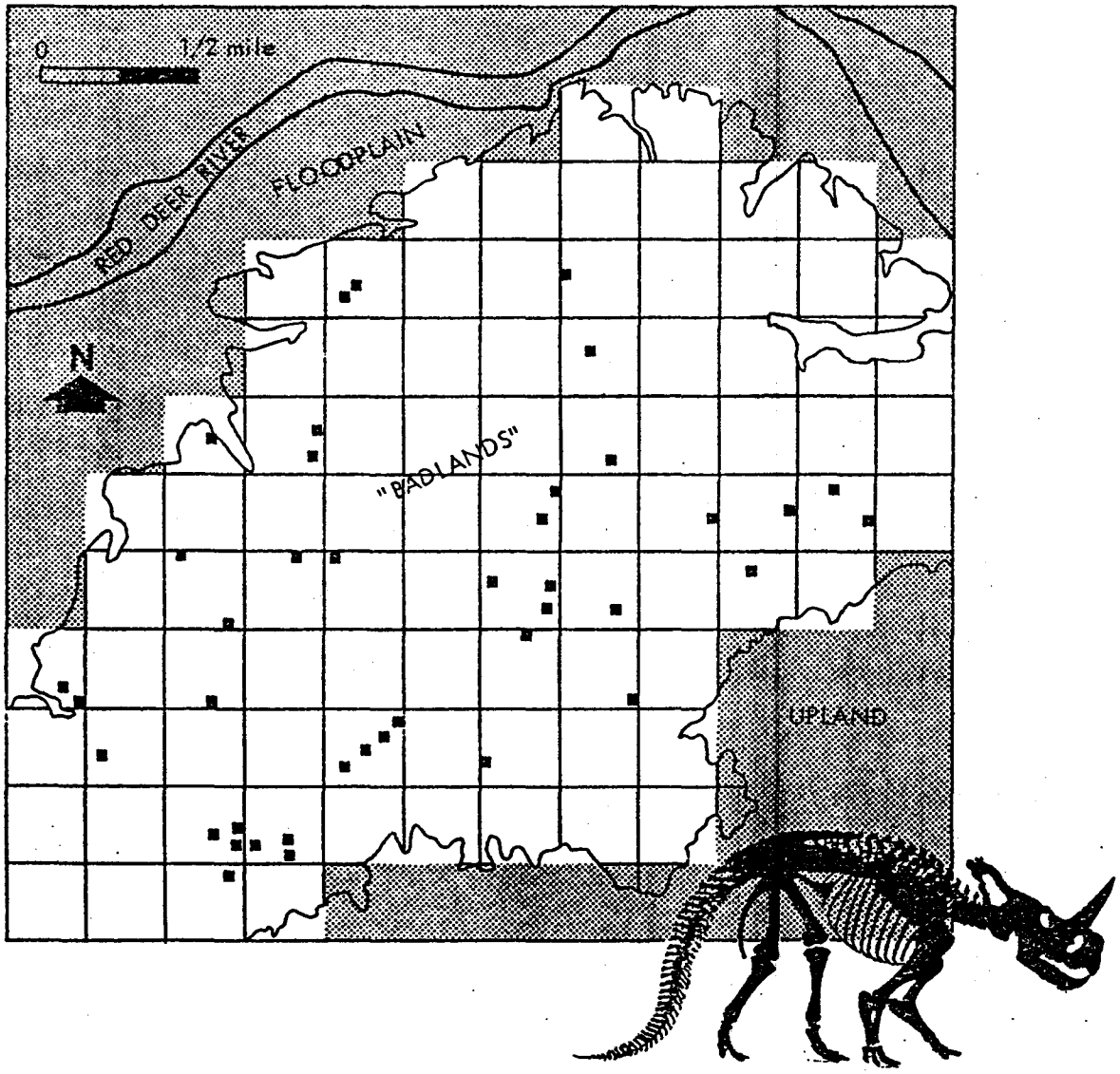
Solterer, J., 1941, A sequence of historical random events: Do Jesuits die in three's?: Jour. Am. Stat. Assoc., v.36, p. 477-484.

Exercise 16.1: Locations of Albertan dinosaurs

The accompanying map shows the locations of dinosaur skeletons found in an area of the outcropping Oldman Formation near Drumheller, Alberta (adapted from Sternberg, 1950). Skulls and other remains are found within the deltaic deposits that make up the formation and are particularly common within channel sandstone units. Using the unshaded squares as observational units of area, compile an observed distribution of skeleton frequencies per square. Fit a Poisson distribution to the data and check the fit with a chi-square test. Do the skeletons appear to be randomly distributed from the available evidence? Can you suggest geological reasons for your observations?

Reference:

Sternberg, C.M., 1950, Steveville west of the 4th meridan, with notes on fossil localities: Geol. Surv. Canada, Map 969 A.



Location of dinosaur remains (marked as black squares) in the Upper Cretaceous Oldman Formation near Drumheller, Alberta (adapted from Sternberg, 1950).

17. THE POISSON PROCESS IN GEOCHRONOLOGY

Geological age-dating is based on measurements of radioactive decay concentrations of various nuclides that have substantially long half-lives. The mathematics of age computation are derived from a Poisson process description of the phenomenon of radioactive decay. The fission of an individual radioactive atom is a rare event and the probability of such an occurrence is proportional to the total number present, R. As decay proceeds, the quantity of radioactive atoms decreases so that the number of atoms that fission per unit time is

$$\frac{-dR}{dt} = \lambda R$$

Where λ is known as the decay constant (and is the λ of the Poisson model). By integrating,

$$R = R_0 e^{-\lambda t}$$

where R_0 is the initial concentration of R at time zero.

$$R_0 = Re^{\lambda t}$$

If D is the number of radiogenic daughter atoms produced in time t, then

$$D = R_0 - R = Re^{\lambda t} - R = R(e^{\lambda t} - 1)$$

Therefore the ratio of daughter atoms to remaining radioactive atoms is

$$D/R = (e^{\lambda t} - 1)$$

and solving for t

$$t = (1/\lambda)\ \log_e\ (1 + D/R)$$

(The half-life corresponds to t when D = R and so is equal to $(1/\lambda)\log_e 2$.) As an example, radiometric ages based on U^{238} decay may be calculated from the equation

$$t = (1/\lambda)\log_e(1 + Pb^{206}/U^{238})$$

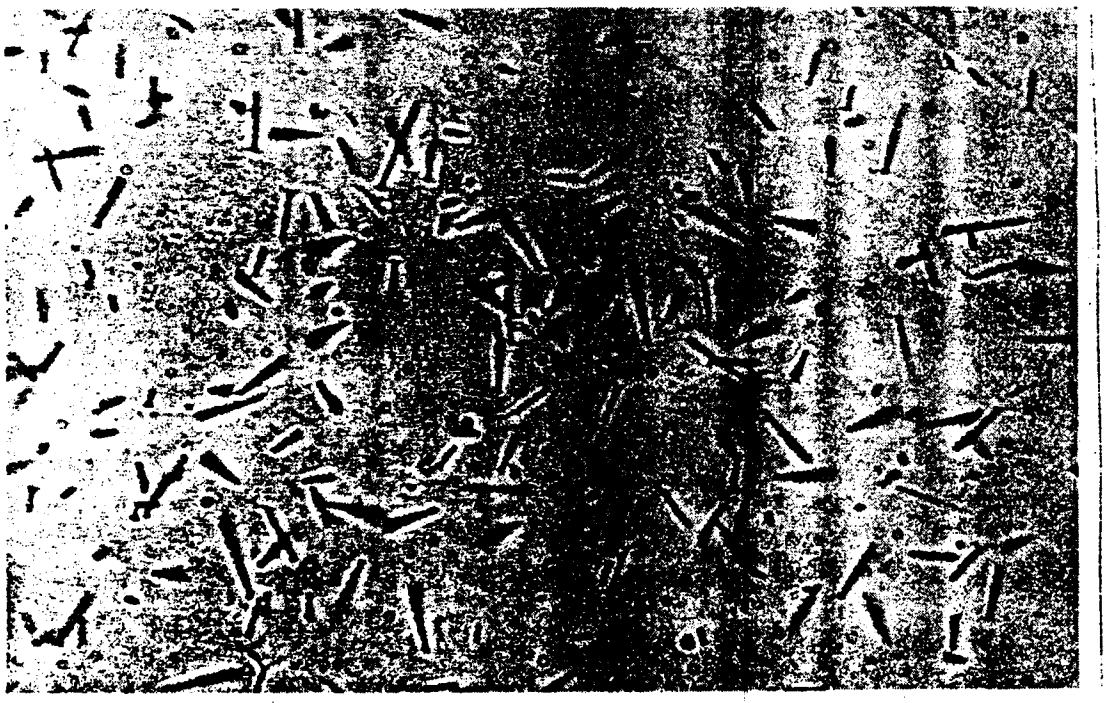
in which the isotope ratio is based on their measured concentrations in the host mineral.

Example: Fission-track dating

A recently developed method is that of fission-track dating. Ages are computed from the density of radiation tracks caused by the spontaneous fission of uranium impurities in minerals such as zircon. The majority of these tracks are produced by U^{238} fission events, by virtue of its markedly greater concentration (over 99%) and its decay rate relative to other uranium isotopes. The mineral specimen is sectioned and etched, and the density of spontaneous fission tracks counted. The mount is then placed in a nuclear reactor and given a measured dose of thermal neutrons which induces fission of the U^{235} in the mineral. The induced fission tracks are counted and the measured density used to compute the relative areal concentration of remaining non-fissioned U^{238} from the equation:

$$R = Nk/\phi$$

where N = The areal density of induced tracks, ϕ = the neutron dose and K is a constant (approximately 1.05×10^{17}) made up of the total and spontaneous fission constants of U^{238}, the ratio of U^{235} to U^{238} and the thermal fission cross-section area of U^{235}. D (the density of spontaneous fission tracks) may then be substituted with R and λ (the decay constant of U^{238}) in the basic age equation.



Spontaneous fission-tracks in a zircon crystal from Nyasaland X 1600 (Fleischer, Price and Walker, 1964).

Naeser and McKee (1970) present data used for fission-track dating of minerals in three Tertiary ash flow tuffs in central Nevada. The numerical quantities for a ziron in the Caetano Tuff No. 11 sample are:

Spontaneous track density, $N = 1.03 \times 10^{6}$ cm^{-2}
Induced track density, $D = 2.28 \times 10^{6}$ cm^{-2}
Neutron dose, $\phi = 1.36 \times 10^{15}$ cm^{-2}

then,

$$R = 176.2 \times 10^{6}$$

and, using the decay constant of U^{238} ($\lambda = 1.54 \times 10^{-10}$ yr^{-1}),

$$t = (1/\lambda)\log_e(1 + D/R)$$
$$= 37.8 \text{ million years.}$$

This figure shows reasonable agreement with a K-Ar date of 32.0 million years for the tuff sample.

References

Fleischer, R.L., Price, P.B. and Walker, R.M., 1964, Fission-track ages of zircons: J. Geophys. Res., v.69, no.22, p. 4885-4888.

Naeser, C.W., and McKee, E.H., 1970, Fission-track and K-Ar ages of Tertiary ash-flow tuffs, north-central Nevada: Geol. Soc. America Bull., V.81, p. 3375-3384.

18. EXPONENTIAL DISTRIBUTION

The exponential distribution is directly related to the Poisson distribution. The Poisson describes the number of rare events within a given time or spatial interval and is a discrete distribution. The exponential expresses the distribution of the periods of time or distances between successive events and so is a continuous function.

If λ is the mean number of events within a unit of time or space (the same convention as in the Poisson) and t is an interval measured in these units, then the density function of the exponential distribution is

$$f(t) = \lambda e^{-\lambda t}$$

The cumulative distribution function is

$$F(t) = 1-e^{-\lambda t}$$

which defines the cumulative area under the density function so that, for example, the area between $t = 0$ and $t = T$ is $(1-e^{-\lambda T})$. This represents the proportion of periods between events that are of length less than T.

The first two moments of the density function are:

$$E(t) = 1/\lambda$$
$$\text{var}\ (t) = 1/\lambda^2$$

The exponential distribution is applied to phenomena described by a Poisson process and provides an alternative to the Poisson distribution. As a continuous distribution, it has the advantage that the distance between adjacent events is an unambiguous measure. By contrast, the application of a Poisson distribution model requires selection of a sampling interval. The choice of different intervals will result in observed distributions of events that may differ in their fit to an expected Poisson distribution (Miller and Kahn, 1962, p. 369). This problem is a function of scale in the sense that a scatter of events sampled at a large scale may show clustering which is not seen if sampled at a fine scale.

The exponential distribution may be applied as a theoretical model to the Kansas earthquake data analysed in the Poisson distribution section. The intervals between successive earthquakes were calculated

from the precise dates listed in Merriam (1963) and accumulated as observed frequencies (f_o) in the histogram table.

Time intervals between Kansas earthquakes, 1866-1965

Intervals (years)	f_o	f_e
0 - 1	11	12.9
1 - 2	9	8.6
2 - 3	8	5.8
3 - 4	3	3.9
4 - 5	3	2.6
5 - 6	1	1.7
6 - 7	1	1.2
7 - 8	1	0.8
8 - 9	1	0.5
9 - 10	0	0.4
10 - 11	0	0.2
11 - 12	0	0.2
12 - 13	0	0.1
13 - 14	1	0.1

From the Poisson example, $\bar{X} = 2.00$ which is used as the estimate of λ, the expected number of earthquakes in a five-year period. Using a time unit of one year,

$$\lambda = 2.00/5 = 0.40$$

Now, the measure of the proportion of periods of length less than t is given by the cumulative distribution function:

$$F(t) = 1 - e^{-\lambda t}$$

Therefore the proportion of periods of length > a but < b is

$$(1-e^{-\lambda b}) - (1-e^{-\lambda a}) = e^{-\lambda a} - e^{-\lambda b}$$

Substituting the histogram table class boundaries as a and b, the expected proportion in each class may be computed under the assumption of an exponential distribution. Multiplication by the total number of intervals (39) gives the appropriate expected frequencies (f_e).

The close similarity of the fitted exponential distribution to the observational data suggests strongly that the Kansas earthquakes are independent events generated by a Poisson process. (The fit may be checked with a chi-square test). The burst of seismic activity in 1929 noted as a possible discrepency in the Poisson distribution exercise appears to be a fortuitous cluster generated by an essentially random process.

The exponential distribution may be used to describe distances between neighboring points that are randomly scattered in space. As an example, the distances between the locations of adjacent Albertan dinosaur skeletons measured on either the east-west or north-south coordinates of the map must conform to an exponential distribuion if the skeletons occur randomly across the area. (see exercise 16.1).

References

Merriam, D.F., 1963, The geologic history of Kansas: Kansas Geol. Survey Bull. 162, 317 p.

Miller, R.L. and Kahn, J.S., 1962, Statistical analysis in the geological sciences: John Wiley and Sons, New York, 483 p.

19. MULTIVARIATE STATISTICS

Introduction

The methods of description and analysis that have been described up to this point have been univariate, where samples of observations have been expressed in terms of a single variable. Most geological investigations generate multivariate data in which multiple measurements are made on each individual for each of a number of variables.

As with many univariate methods, one of the aims of multivariate statistics is to distinguish samples drawn from different populations. However, a most important objective of multivariate studies is the definition of relationships between variables. In situations where there are affinities between measured variables, there is some redundancy of information. Following the scientific dictum of parsimony, the raw variables may often be condensed to a few descriptors of a "fundamental" nature and possible genetic meaning.

If a series of measurements are made on a sample of n individuals in terms of m variables, the observations may be tabulated as an array of the form:

$$\begin{bmatrix} X_{11} & X_{12} \cdots X_{1m} \\ X_{21} & X_{22} \cdots X_{2m} \\ \cdots & \cdots \quad \cdots \\ X_{n1} & X_{n2} \cdots X_{nm} \end{bmatrix}$$

This representation is known as a sample or data matrix, which can be denoted as $\underline{X}$ (conventionally printed in heavy type) and is equivalent to $[x_{ij}]$, where x_{ij} is the value of the jth variable measured on the ith individual.

The matrix can be plotted geometrically as a swarm of n points in m-dimensional space, where each of the reference axes are orthogonal and represent one of the variables. (If there are two variables, this plot can be drawn as a scatter diagram.) Alternatively, each individual can be represented as a vector extending from the origin to the point and denoted as:

$$[X_{i1} \; X_{i2} \cdots X_{im}]$$

Plotted as a swarm of points, the most basic numerical descriptor of the swarm is its centroid or multivariate mean, whose coordinates correspond to the means of the m variables, or

$$[\bar{X}_{i1}\bar{X}_{i2}....\bar{X}_{im}]$$

This is the mean vector of the n sample vectors.

A secondary measure is required to characterize whether the swarm is concentrated or diffused around its centroid. The univariate measure of dispersion is the variance and, for each of the m variables, this statistic specifies the dispersion of the swarm measured along the orthogonal axes. Measures of the spatial dispersion in terms of the joint relationships between all possible pairs of the m variables are given by the covariances. The covariance of the swarm in the X_1,X_2 plane is given by the equation:

$$\text{cov}(X_1,X_2) = {}^{1}/_{n-1} \Sigma (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$$

The complete description of the dispersion of points is given by a matrix of variances and covariances. This variance-covariance matrix is conventially denoted as Σ and has m rows and m columns.

$$\Sigma = \begin{bmatrix} \text{var (1) cov(1,2).....cov(1,m)} \\ \text{cov(2,1) var(2).....cov(2,m)} \\ \text{.....} \\ \text{cov(m,1) cov (m,2)....var(m)} \end{bmatrix}$$

As can be seen, the covariance of a variable with itself is the variance.

Univariate statistical methods are used for analysis of the means and variances which are single quantities and are treated using conventional algebra. The representation of their multivariate analogues by mean vectors and covariance matrices requires that computations be executed by matrix algebra (q.v.). Multivariate analysis is the study of relationships between variables. These methods are known as R-mode and are an extension of classical univariate statistics. In other applications, the varibles are represented as points in a space whose axes are the samples; such analyses are known as Q-mode. Q-mode methods are outside classical statistics and are used by investigators who attach special significance to individuals as distinctive entities rather than as anomyous members of a statistical sample.

20. CORRELATION

Historical

The principle of correlation was first formulated by Sir Francis Galton (1822-1911). Balloonist, African explorer, fingerprint pioneer, meteorologist (he invented the term "anticyclone") and enthusiastic amateur inventor, Galton was fascinated by descriptive statistics of every kind of phenomenon and their apparent relationships with one another. In studying heredity, a method of expressing relationships of multiple causality occurred to him as a single unifying formula. Galton's discovery was the initial point for formal analysis and refinement by Pearson, Edgeworth and Weldon.

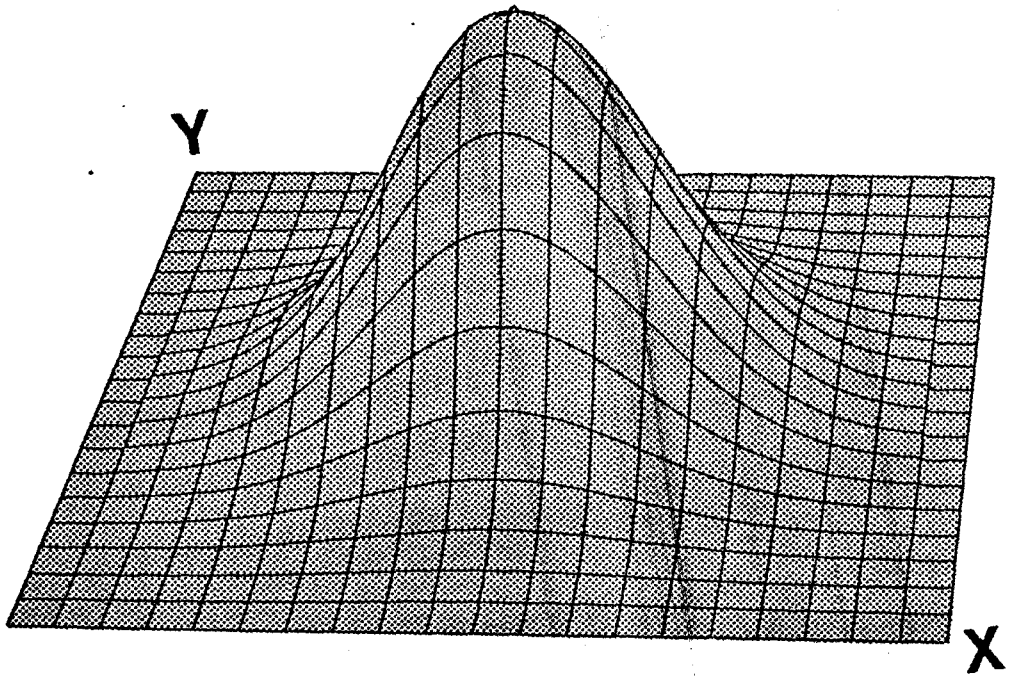
Product-moment correlation coefficient

An index of correlation between two variables is a measure of the intensity of their relationship. The most common measure used is Pearson's product-moment correlation coefficient, which is computed from samples as an estimate, r of the parameter ρ. The correlation between variables X and Y is then:

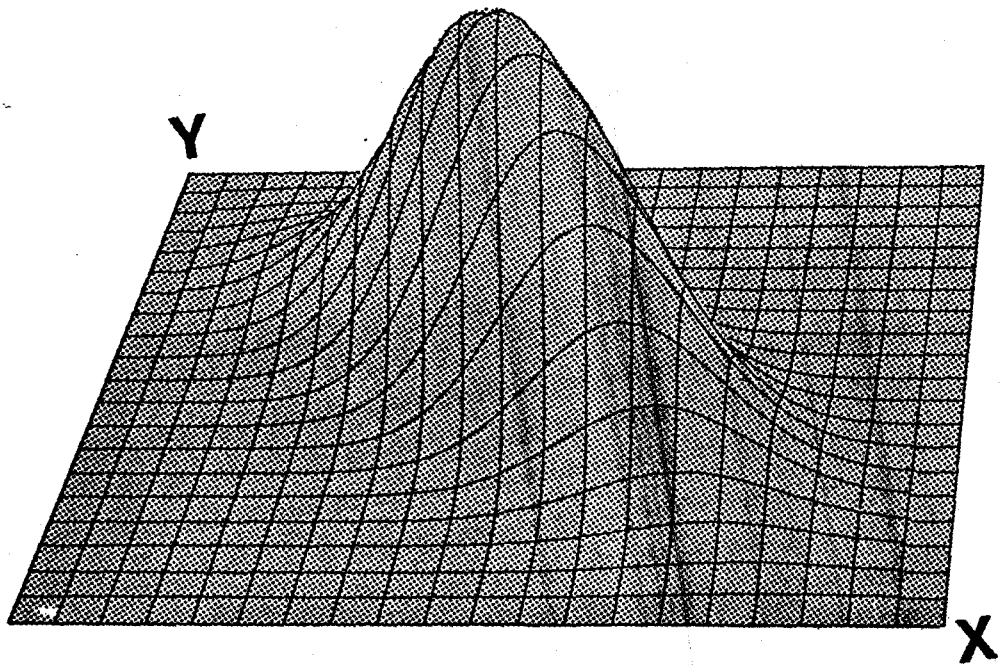
$$r = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(x)\mathrm{var}(y)}}$$

$$= \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\Sigma(Y-\bar{Y})^2}}$$

The value of r can range from +1 (perfect positive relationship) through 0 (no relationship) to -1 (perfect inverse relationship). The geometrical model underlying this equation is the bivariate normal distribution. Normally-distributed bivariate observations from a large sample may be standardized (the measurements are transformed to standard deviation units of the variable from the mean) and plotted as a frequency distribution. If there is no correlation between the variables ($r = 0$) the distribution is radially symmetrical, as shown in the figure, and a horizontal section is circular. If there is correlation, the distribution becomes elongated, the relative constriction reflecting the degree of correlation (see figure) and a horizontal section is elliptical. The correlation coefficient is a standardized measure of

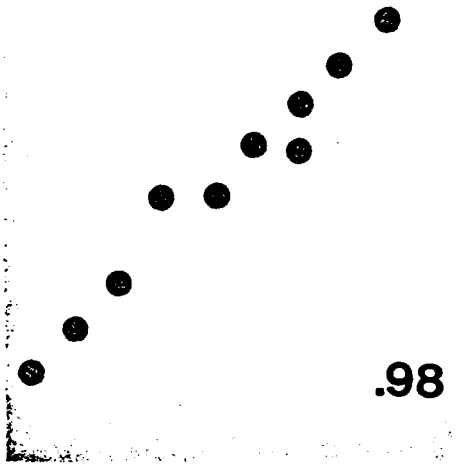


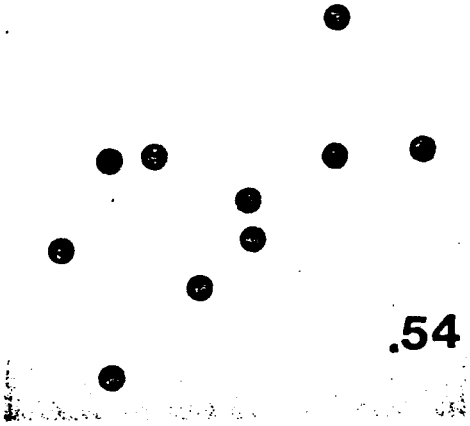
$r = 0$

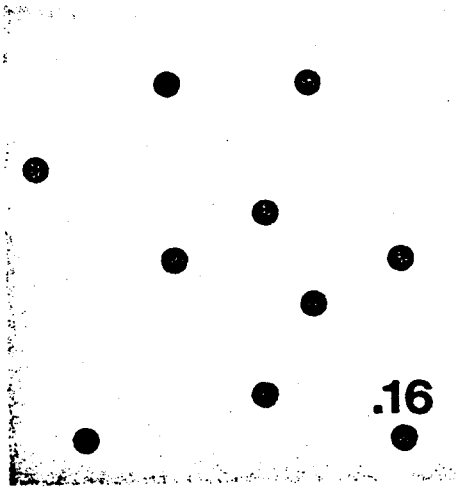


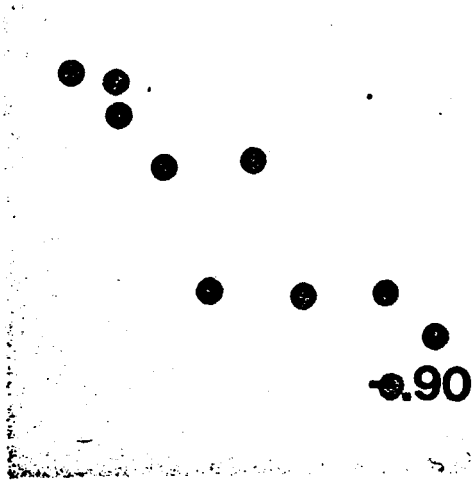
$r = -0{\cdot}8$

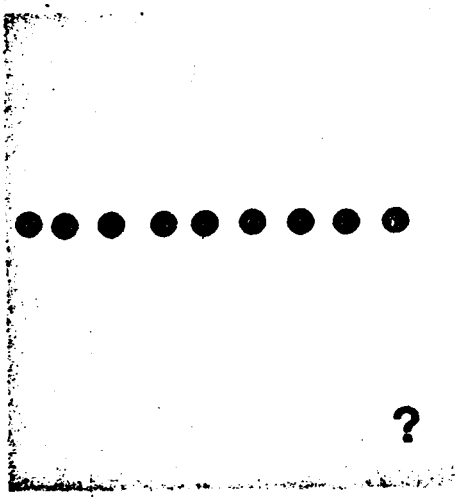
Correlation

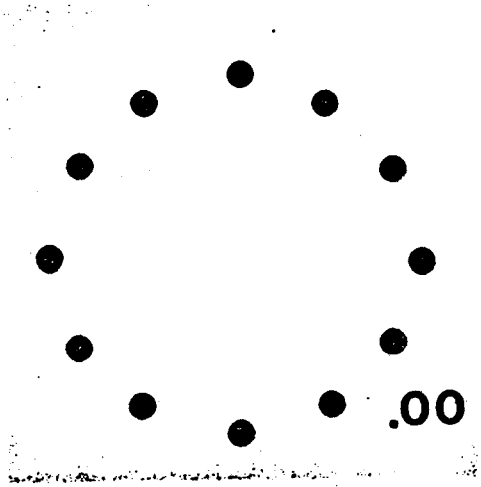




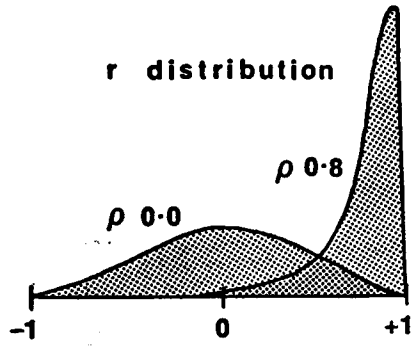








variable interrelationship which is independent of the measurement units, since the denominator is the product of the two variances. If the variables are standardized, the variances both equal one and the correlation coefficient is the standardized covariance.



When r is calculated for a sample from a bivariate population, it is an estimate of the population parameter ρ, and its reliability is a function of sample size. Because the values of r are constrained between the limits of +1 and -1, the sampling distribuiton of r is highly negatively skewed in the vicinity of +1 and highly positively skewed near -1. When = ρ = 0, the distribution of r is symmetrical, though not exactly normal.

(1) As a result, the null hypothesis that ρ = 0 can be tested using a t distribution where:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

with (n-2) degrees of freedom, where r is the correlation coefficient estimate and n is the number in the sample. If t exceeds the tabulated value as a two-tailed test for a selected significance level, the null hypothesis of no correlation is rejected.

(2) The null hypothesis that ρ is some value other than zero requires that a transformation be applied to approximately normalize the sampling distribution. Then,

$$Z_r = \frac{1}{2} \log_e \left(\frac{1+r}{1-r}\right)$$

where Z_r is the transformation of r. The standard error of Z is approximately

$$s_e = \frac{1}{\sqrt{n-3}}$$

To test the null hypothesis that r is an estimate of some hypothesized value ρ, a test score is computed from the normalized correlations:

$$Z = \frac{Z_r - Z_\rho}{s_e}$$

which is approximately normally distributed. The Z-test is applied in the same manner as hypotheses concerning means of normally distributed

variables (see Hypothesis Tests).

(3) Since s_e is the standard error of the correlation coefficient estimate, it may be used in the calculation of confidence limits around r, by computing appropriate Z-values and transforming these back to correlation coefficients.

(4) A Z-test of the significance of the difference between two sample correlation coefficients may be made by computing:

$$Z = \frac{Z_1 - Z_2}{\sqrt{s_{e1}^2 + s_{e2}^2}}$$

The denominator is a pooled estimate of the standard error of Z, where

$$s_{e1} = \frac{1}{\sqrt{n_1 - 3}} \quad \text{and} \quad s_{e2} = \frac{1}{\sqrt{n_2 - 3}}$$

Example: Correlation between boron content of illite and salinity.

Walker and Price (1963) suggested that there was a direct relationship between the boron content of illite of constant grain size and the salinity of the depositional aqueous medium. Since boron concentrations are related to the K_2O content of illite, the measured values are corrected to an "equivalent boron" which is related to a standard K_2O composition.

The equivalent boron content of illite in samples from environments within the Dovey Estuary, Wales were matched with the mean salinities of the environmental aqueous media by Adams, Haynes and Walker (1965). The results are summarized in the table.

Environment	Salinity (ppt), S	Equiv. Boron (ppm), B
marsh	25.5	355
marsh	25.5	345
channel/flats	18.0	260
estuarine	21.5	275
channel/flats	18.0	280
channel/flats	18.0	260
marsh	25.5	355
marsh	25.5	325
marsh	25.5	315
marsh	25.5	310
marsh	25.5	370
estuarine/marsh	31.5	330
tidal	16.0	305
sea	33.0	365
marsh/flats	20.5	270

$n = 15$

$\bar{S} = 23.7 \qquad \bar{B} = 315$

$var(S) = 24.4 \quad var(B) = 1498$

$cov(S,B) = 144.5$

$$r = \frac{cov(S,B)}{\sqrt{var(S)\ var(B)}} = 0.76$$

There appears to be a positive correlation between boron content of illite and environmental salinity. This may be checked by testing a null hypothesis that there is no correlation, i.e., $\rho = 0$. The null hypothesis postulates that the observed correlation is a random sampling fluctuation, where the estimate of the parameter is poor because of the small sample size.

Then,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 6.49$$

With degrees of freedom, $\nu = n-2 = 13$. Selecting a significance level α of 0.05, the tabulated value of t is 2.16. The null hypothesis is therefore rejected and the alternative hypothesis of a significant degree of correlation between the two variables is accepted. The correlation is positive.

Boron analyses of clay materials have been widely used to estimate salinities of ancient environments, but the method drew some fire in the literature of the late sixties. It was variously suggested that illite was primarily detrital with an inheritance of boron from parent rocks and that boron contents were related to organic carbon as well as salinities (see Walker, 1968). The strong positive correlation between the tabulated boron and salinity figures might appear to refute these ideas. However, it must be remembered that correlaton between variables implies nothing concerning cause and effect, it merely demonstrates empirical relationships. So, for example, a tabulation of crime incidence and number of churches in cities would almost certainly result in a strong correlation between these two variables. However, organized religion is unlikely to be a front for organized crime (or vice versa). Instead, both variables are highly correlated with a third variable, the size of the city. Correlations of this type are known as "spurious" correlations.

In the present example, a case might be made for organic carbon as the critical controlling variable. Dilution of illite by organic matter would lead to an inverse correlation between observed boron and carbon. Organic matter would tend to be more highly concentrated in fresh-water and brackish coastal environments relative to the open sea. As a result, the correlation between boron content and salinity could be a function of a third variable, carbon content. This possibility may be evaluated by measuring carbon content in the sample and relating this to boron and salinity. It follows that causal model interpretations are the investigator's responsibility and are not implicit in computed correlations.

Confidence limits for the boron/salinity correlation coefficient may be computed easily after normalization by the transformation:

$$Z_r = \frac{1}{2}\log_e \left(\frac{1+r}{1-r}\right) = 0.996$$

For 90% confidence limits, the boundaries are selected at a standard deviation distance of ±1.645 from the estimate (cutting 5% tails at both ends of the normal distribution).

The standard error of the Z-transformed correlation is

$$s_e = \frac{1}{\sqrt{n-3}} = 0.289$$

The upper 90% confidence limit is then located at a Z-value of

$$0.996 + 1.645\ (0.289) = 1.471$$

Similarly, the lower limit is at

$$0.996 - 1.645\ (0.289) = 0.521$$

The values of r corresponding to those Z transformations are r = 0.90 and r = 0.48. Therefore, the probability is 90% that the population parameter of boron/salinity correlation lies in the confidence range 0.48 to 0.90, or

$$0.48 \leq \rho < 0.90$$

References

Adams, T.D., Haynes, J.R. and Walker, C.T., 1965, Boron in Holocene illites of the Dovey Estuary, Wales, and its relationship to palaeosalinity in cyclothems: Sedimentology, v.4, p.189-195.

Walker, C.T., 1968, Evaluation of boron as a palaeosalinity indicator and its application to offshore prospects: Am. Assoc. Petroleum Geologists Bull., v.52, p.751-766.

Walker, C.T. and Price, N.B., 1963, Departure curves for computing paleosalinity from boron in illites and shales: Am. Assoc. Petroleum Geologists Bull., vol.47, p.833-884.

Assumptions involved in using the product-moment coefficient

The computation of r as a valid measure of association is keyed to:

(1) the trend of the relationship between the two variables is linear, i.e., a straight line relationship where, for any increase in one variable, there tends to be a constant proportional increase in the other, that does not change with the value on either scale.

(2) The variables are measured on an interval or ratio scale, i.e., continuous measurements. Estimation of the significance of r depends on the assumption that the joint distribution of the population is bivariate normal.

Exercise 20.1: Variation of the synodic month through geological time.

Incremental growth patterns of fossil groups such as molluscs and stromatolites have been interpreted to reflect both solar time (formation of daily bands) and synodic time (ridge formation once every lunar month). Growth periodicities related to lunar and solar cycles are common throughout the animal kingdom. Interpretation of growth banding in modern forms is made complex by disturbing factors such as breeding events and generic and specific differences. Analysis of fossil forms is made still more difficult by conditions of poor preservation, pattern ambiguities and lack of equivalent living forms.

The table lists the average "daily" increments per "lunar month" measured for a variety of selected molluscs and stromatolites from different geological periods. From the currently accepted model of the earth-moon system, geophysicists have theorized a decrease in the earth's rotation rate as a result of tidal torque, with consequent "shortening" of the lunar month through geological time. Using the tabulated figures as absolute ages of the fossil samples:

(1) Compute the correlation coefficient of the variables of geological time and growth increment density.

(2) Is the value of r sufficient to reject a null hypothesis of no relatonship between growth increment density and time? Does the value support the geophysical model?

(3) Compute 95% confidence limits for the sample correlation coefficient.

(Assume a bivariate normal parent population as an approximation).

Reference

Pannella, G., Copeland, M., and Thompson, M.N., 1968, Paleotological evidence of variations in length of synodic month since Late Cambrian: Science v. 162, p. 792-796.

SYNODIC MONTH DATA

Growth bands in stromatolite X 20, Sayabec Fmn., Silurian, Quebec

Increments/month	Geological age (MM years)	
29.17	0	Recent
29.40	18	Upper Miocene
29.63	40	Upper Eocene
29.82	46	Middle Eocene
29.92	72	Upper Cretaceous
29.68	205	Middle Triassic
30.05	290	Upper Pennsylvanian
30.22	305	Lower Pennsylvanian
30.37	340	Lower Mississippian
30.53	380	Middle Devonian
31.56	510	Upper Cambrian

Induced correlations

Some correlations between variables do not reflect the relations between the variables, but are induced by an operation or transformation that has been performed on them. The expected correlation between two independent, random variables is zero. However, by performing certain ordering operations or transforming the variables to closed form, the expected correlation may assume some value other than zero even if there is no relation between the variables.

An example of the correlation produced by ordering is the correlation which exists between axial lengths of grains. By definition, the longest axis becomes the A-axis, the shortest becomes the C-axis, and the intermediate axis is B. Therefore, there must always be a positive correlation between any pair of axes, or between the ratios of two axes and the third axis (i.e., between $\frac{b}{a}$ versus c).

The most troublesome induced correlations are spurious negative correlations that appear in closed data sets. A closed data set is one in which all variables measured on an individual add to a fixed total such as 1.00 or 100%, which means the variables are proportions of a whole. Because the sum of the variables is a fixed number, an increase in the proportion of one variable can only occur at the expense of other variables.

In an open data set in which the measurements are not expressed as proportions, two linearly independent variables will have a correlation r_{ij} which is not significantly different from zero. If an open data set is closed by converting the measurements to proportions, apparently significant negative correlations will appear even though the original open data consisted entirely of independent variables. The significance of correlations between variables from a closed array, therefore, should not be tested against $\rho_{ij} = 0$ but against the correlation that would result from closing of independent variables.

Chayes (1971) has devised ways of calculating the correlations that would result fromclosure alone in a data set having more than three variables. This is done by postulating a hypothetical open array of

independent (all $\rho_{ij} = 0$) variables that, when closed, will have the same means and variances as the closed data set being tested. First, the variance of the hypothetical open array must be found by solving the matrix equation (q.v., Matrix algebra):

$$[\bar{X}]\ [\sigma^2] = [s^2]$$

where

$[s^2]$ is the $m \times 1$ vector of variances of the observed closed variables.

$[\sigma^2]$ is the $m \times 1$ vector of unknown variances of the hypothetical "open" variables.

$[\bar{X}]$ is an $m \times m$ matrix of mean proportions of the observed closed variables, equal to

$$[\bar{X}] = \begin{bmatrix} (1 - \bar{X}_1)^2 & \bar{X}_1^2 \ldots . \bar{X}_1^2 \\ \bar{X}_2^2 & (1 - \bar{X}_2)^2 . . \bar{X}_2^2 \\ \bar{X}_m^2 & \bar{X}_m^2 \ldots \ldots \ldots (1 - \bar{X}_m)^2 \end{bmatrix}$$

$\bar{X}_i$ is the mean of observed variable X_i, measured as a proportion (not percentage or ppm).

The total variance of the hypothetical open array is

$$\sigma_t^2 = \sum_{i=1}^{m} \sigma_i^2$$

If all the σ_i^2 variances are positive, the correlations that will result from closure alone can be found by

$$\rho_{ij} \approx (\bar{X}_i \bar{X}_j\ \sigma_t^2 - \bar{X}_j\ \sigma_i^2 - \bar{X}_i\ \sigma_j^2) \ / \ s_i s_j$$

The null hypothesis that the observed correlations are due to closure alone, or that $r_{ij} = \rho_{ij}$, can be tested using the approximate normal transformation given earlier. The difference between a Z-score for the observed correlation (Z_r) and a Z-score of the closure correlation (Z_p) can be tested by

$$Z = (Z_r - Z_p) \sqrt{n-3}$$

The test value Z is approximately normally distributed.

In the special case of a three-variable closed array, as in a ternary diagram, the correlations between the closed variables are determined solely by the observed variances:

$$r_{12} = \frac{s_3^2 - s_1^2 - s_2^2}{2s_1 s_2}$$

No tests are available for the significance of correlations in terniary diagrams.

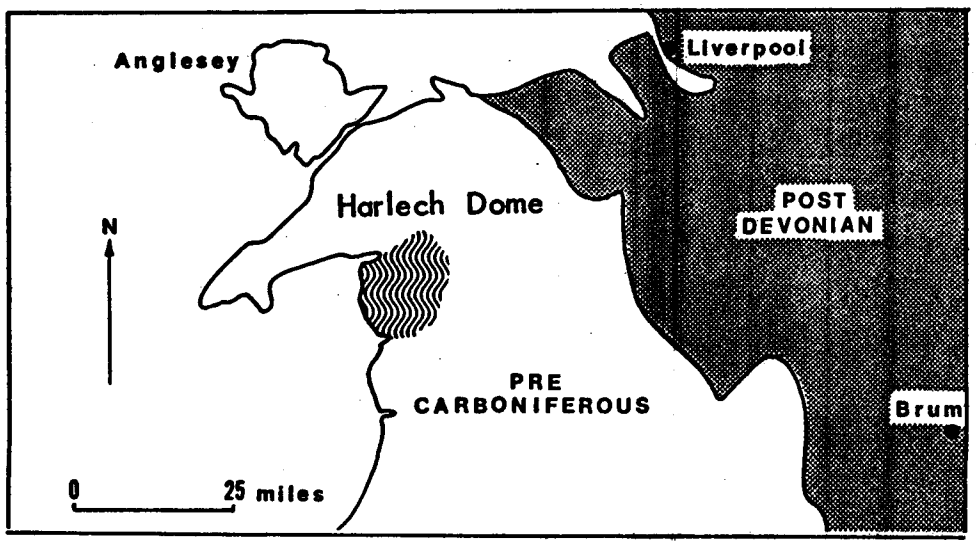
References

Chayes, F., 1971, Ratio correlation: Univ. Chicago Press, Chicago, 99 p.

Okada, H., 1966, Non-greywacke "turbidite" sandstones in the Welsh geosyncline: Sedimentology, v. 7, p. 211-232.

Exercise 20.2: Composition of turbidite sandstones

Okada (1966) studied non-greywake sandstones that are found within turbidite sequences in northern Wales. These rocks are feldspathic and lithic arenites and occur near the bottoms of graded beds. Among other properties, Okada determined their petrographic composition, which is given in the accompanying table. These analyses constitute a five-variable closed array. Are any of the correlations between compositional variables significant, or do they simply result from the effect of closure?



Cambrian turbidites					Silurian turbidites				
Quartz	Feldspar	Rock Fragments	Clay	Cement	Quartz	Feldspar	Rock Fragments	Clay	Cement
57	9	9	25	--	35	14	6	41	3
52	8	7	34	--	43	15	6	32	3
51	13	2	31	3	34	17	26	2	21
8	5	1	83	--	39	19	11	16	14
39	2	2	55	1	47	18	6	22	6
63	12	5	18	2	49	7	14	21	9
41	12	2	44	--	39	16	10	23	11
7	--	--	91	2	37	12	6	35	9
58	19	8	5	10	41	18	10	27	4
50	11	2	37	--	37	9	12	27	14
69	14	1	15	1	27	6	44	11	12
43	12	1	44	--	39	7	8	34	11
69	13	3	14	--	33	5	1	48	11
58	18	2	19	3	61	3	7	25	4
49	11	--	34	6	37	12	34	10	6
69	14	--	9	8	47	6	13	30	4
54	12	1	27	6	33	7	20	37	--
39	5	--	55	1	45	3	6	44	1
					66	9	12	11	2
					57	4	3	35	1
				Means	45.2	10.4	7.9	30.8	5.0
				Variances	207.0	26.7	89.5	356.0	26.2

Spearman's rank correlation coefficient

Since it is impossible to estimate parameters for variables measured on an ordinal scale, some form of non-parametric measure of association must be used for this type of data. If measurments of two ordinal variables are ranked in order for each of the variables, a Spearman's rank correlation coefficient, r' may be calculated:

$$r' = 1 - \frac{6 \Sigma D_i^2}{n(n^2-1)}$$

where n = number of individuals in the sample

and D_i = the difference between the ranks of the ith individual measured on the two variables.

The value is constrained within the range of -1 to +1 in a fashion analogous fashion to the product-moment coefficient.

The r' value computed for a sample is an estimate of the population parameter, ρ'. The sampling distributions of r' are similar to distributions of their equivalent r values, as they are constrained within the same restricted range. However, since the rank numbers are discrete quantities, the actual distributions correspond to the r' values resulting from all the possible combinations of rank orders within the sample (n!). For samples of size $n > 30$, the normal distribution can be used to approximate the sampling distribution of $\rho' = 0$. A two-tailed t-test may then be applied, using the same formula as for r, to test the null hypothesis of zero correlation:

$$t = r'\sqrt{\frac{n-2}{1-r'^2}} \text{ and } \nu = n-2$$

In many cases, there are instances of tied ranks where two or more observations take the same value on a variable and cannot be ordered relative to one another. The rank is averaged between the tied samples of each set and the quantity A computed:

$$A = \frac{(k^3-k)}{12}$$

where k is the number of tied samples. The modified equation of r' for two variables X and Y becomes:

$$r = \frac{\frac{n^3-n}{6} - \Sigma D^2 - \Sigma A_X - \Sigma A_Y}{\sqrt{(\frac{n^3-n}{6} - 2\Sigma A_X)(\frac{n^3-n}{6} - 2\Sigma A_Y)}}$$

When there are no tied ranks, the equation condenses to the original r' formula. It should be obvious that there should be only a few tied ranks relative to the size of the sample for the viability of this statistic to be preserved.

Example: Correlation of textural properties of sandstone

There are many classes of descriptive geological information that are recorded as ordinal data, especially properties such as petrographic textures that are difficult to express succinctly on higher information scales. The evolution of sandstone grain fabrics as a function of source material, tectonic context, climate, hydraulic environment is well entrenched in the literature (e.g. Folk, 1968). Textural pro-

perties of sandstones are primarily attributed to environment of deposition, characterized as either high or low energy in hydraulic terms. Prolonged high energy conditions cause winnowing and abrasion, resulting in coarse well-sorted, well-rounded grains; low energy environments are relatively quiescent and produce finer-grained fabrics that are poorly sorted and often somewhat angular. "Textural maturity" was defined by Folk (1951) as the degree to which a sand is well-sorted and well-rounded. The positive association between sorting and roundness will obviously not be clearcut, since there are situations where well-rounded grains are introduced into a low energy environment or when the time factor in a high energy environment is insufficient for maturation. Sandstone fabrics with "conflicting" textural properties (e.g. grains that are both angular and well-sorted) are said to show "textural inversion".

A suite of twelve reservoir sandstone samples of different ages were examined in thin-section and classified in categories according to degrees of sorting and roundness. These categories were subdivided so the twelve sandstones were ranked in order for the two variables of sorting and roundness. The raw data is summarized in the table.

RESERVOIR SANDSTONE TEXTURAL PROPERTIES

Stratigraphic unit	Age	Sorting		Roundness	Rank	D^2
Lakota Sst.	Cret.	P	4	SR	11	49
Berea Sst.	Miss.	W	10	SA	9	1
Boise Sst.	Plioc.	P	2	A	1	1
Big Clifty Sst.	Miss.	M	8	SA	4	16
Clear Creek Sst.	Penn.	M	6	SA	6	0
Bromide Sst.	Ord.	W	9	SR	12	9
Noxie Sst.	Penn.	P	3	SA	8	25
Green River Sst.	Eoc.	M	7	SA	3	16
Reagan Sst.	Cam.	W	11	SA	7	16
Peru Sst.	Dev.	W	12	SR	10	4
Bartlesville Sst.	Penn.	M	5	A	2	9
Mt. Simon Sst.	Cam.	P	1	SA	5	16

Sorting: P = poor; M = moderate; W = well

Roundness: A = angular; SA = subangular; SR = subrounded

Then,

$$\Sigma D^2 = 162$$

and

$$r' = 1 - \frac{6\ \Sigma D^2}{n(n^2-1)} = 0.43$$

Setting the null hypothesis, $\rho' = 0$, and assuming the sampling distribution to be very approximately normal for this low sample size:

$$t = r\sqrt{\frac{n-2}{1-r'^2}} = 1.52$$

$$\nu = n-2 = 10$$

The critical value of t at $\nu = 10$ and $\alpha = 0.05$ is 2.23. The null hypothesis is therefore <u>not</u> rejected and it cannot be accepted that the correlation between the two variables is necessarily other than zero on the basis of the sample.

There is a weak positive correlation ($r' = 0.43$) whose sense corresponds with the conceptual model. However it can be seen that the sample provides insufficient evidence to support the hypothesized textural interrelationships. If the model is broadly true, a large sample would be needed to inequivocally demolish the null hypothesis that there is no correlation between roundness and sorting. In the meantime, the relationship may be suggested as at best a generalization with many exceptions.

<u>References</u>

Folk, R.L., 1951, Stages of textural maturity in sedimentary rocks: Jour. Sed. Petrology, v. 21, p. 127-130.

_____, 1968, Petrology of sedimentary rocks: Texas Hemphill's Book Store, Austin, 170 p.

21. REGRESSION

Trend lines can be drawn in by eye, but this process is usually sneered at; a mathematical way of doing it is the "least squares" method.

Folk (1968)

While measures of correlation express the intensity of relationship between variables, regression analysis isolates the functional trend relating changes in one variable with changes in the other. The concept of regression derives from Galton, who used this particular term because he noticed that the heights of sons deviated less from the mean than heights of their fathers, and so were "regressing" toward the mean. Galton defined the line which graphed the average relationship between the two variables, the "line of regression." (He formulated the regression concept while sheltering from the rain in a rock crevice -- correlation was conceived when waiting for a train at a country station.) Pearson and others refined regression analysis using the "principle of least squares" which had been devised by Gauss. Gauss had noted that the most probable value of an observation, based on a series of measurements, was that which minimized the quantity

$$\sum^{n} (X-X_i)^2$$

where X is the variable and X_i is a series of i measurements. Then,

$$\frac{d}{dX} \sum^{n} (X-X_i)^2 = 0$$

and so

$$X = \frac{1}{n} \Sigma X_i$$

meaning that the most probably value is the arithmetic average. The principle of least squares can be derived by applying a maximum likelihood criterion if the errors of the measurements follow a normal distribution.

The least-squares principle can be extended to observations on more than one variable. If two variates are related by a linear function, this can be written as:

$$Y = \alpha + \beta X + \varepsilon$$

where X is termed the independent variable (which is assumed to have no error) and Y is the dependent variable that has an associated random error element, ε. The equation of the linear trend Y for a sample of n points is:

$$\hat{Y} = a + bX$$

where a and b are estimates of α and β and are the intercept and slope for a straight-line equation. For any point with value X_i, Y_i the vertical distance between the trend line $\hat{Y}_i$ and the point should be minimized as the error component, ε_i by the least-squares principle. For the n observational points:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \Sigma \varepsilon_i^2 = G(\text{a minimum})$$

As $\hat{Y}_i = a + bX_i$

then $\Sigma(Y_i - a - bX_i)^2 = G$

The quantity G is a function describing the sum of squares of the sample points around any line defined by a and b. If G is plotted against a and b, the function would be curved with a minimum point coinciding with values of a and b that are the descriptors of the best-fit regression line. This point occurs when the slope of the G-function is zero with respect to a and b. From elementary calculus, the minimum is defined when both

$$\frac{dG}{da} = 0 \text{ and } \frac{dG}{db} = 0$$

Partially differentiating the expression for G in terms of a:

$$\frac{dG}{da} = \Sigma -(Y_i - a - bX_i) = 0$$

Or by rewriting,

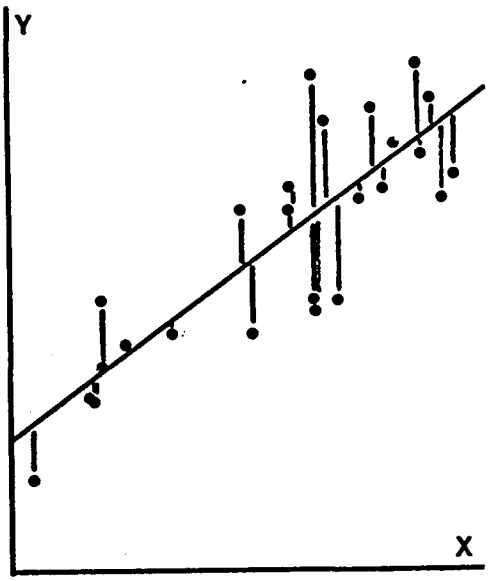
$$na + b\Sigma X_i = \Sigma Y_i$$

Partially differentating with respect to b:

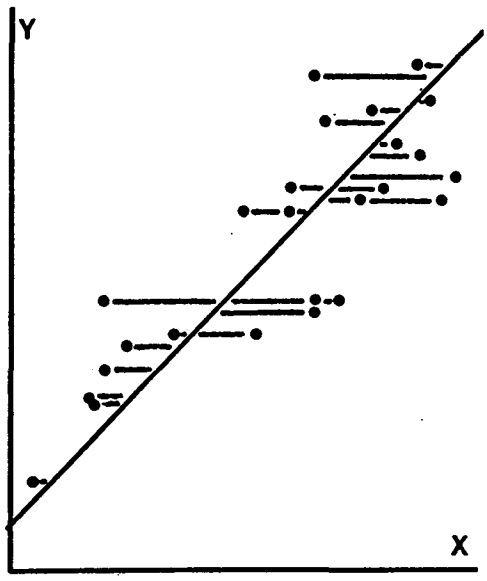
$$\frac{dG}{db} = 2\Sigma - X_i (Y_i - a - bX_i) = 0$$

Collecting terms yields

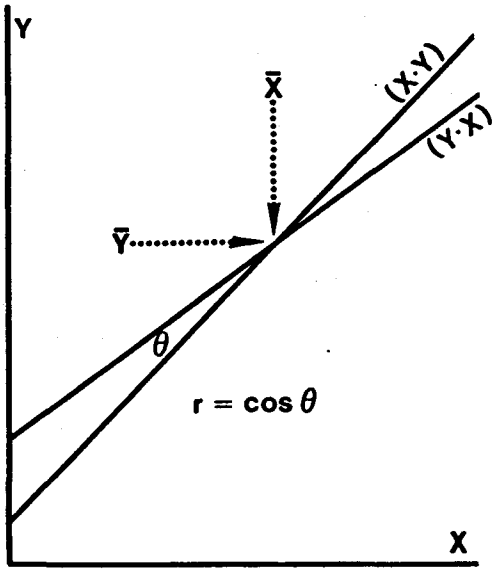
$$\Sigma(X_i Y_i - aX_i - bX_i^2) = 0$$



Regression of Y on X (Y.X)



Regression of X on Y (X.Y)



which can be rearranged

$$a\Sigma X_i + b\Sigma X_i^2 = \Sigma X_i Y_i$$

Combining the two equations gives:

$$b = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

and $$a = \frac{\Sigma Y_i - b\Sigma X_i}{n} = \bar{Y} - b\bar{X}$$

The process employed has minimized the squares of the deviations of points from a linear regression line measured in a vertical direction (or parallel to the Y axis) and constitutes a regression of Y on X. When X is the dependent variable and Y the independent, a regression of X on Y is computed, where the sum of squares are minimized with respect to deviations measured parallel to the X axis. In either case, the conceptual model assumes that the independent variable is measured without error and that the error of the points is contained in the dependent variable and is normally distributed about the regression line. For any scatter of points there exist two possible regression lines corresponding to the two models. They have the properties that they both pass through the means of the variables $(\bar{X},\bar{Y})$ and the cosine of the angle between them is the product-moment correlation coefficient between the variables. For correlations of ±1 the two regression lines are coincident.

Estimation of the error

The variance of the population of all possible observations about the true regression line is the variance of the error, symbolically represented by $\sigma^2_{Y.X}$. An unbiased estimate of $\sigma^2_{Y.X}$ is provided by:

$$s^2_{Y.X} = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{(n-2)}$$

$$= \frac{\Sigma Y_i^2 - a\Sigma Y_i - b\Sigma X_i Y_i}{(n-2)}$$

The square root of this quantity is known as the standard error of the estimate. If the observations are distributed normally about the regression line then about 95% of the actual values would be expected to lie within ±2 standard errors form the regression line (measured parallel to the Y axis).

The slope of the regression, β

The computed value of b is a sample estimate of the true population parameter, β. If the error or residuals, ε, are normally distributed about the regression line then the sampling distribution of b is normal with a mean of β and a variance of σ_b^2. The sample estimate of this variance is:

$$s_b^2 = \frac{s_{Y.X}^2}{\Sigma(X_i - \bar{X})^2}$$

The variance is used in a t-test of the null hypothesis that the true slope of the regression is zero (i.e., that Y is linearly independent of X), or

$$H_0 : \beta = 0$$

The test statistic has (n-2) degrees of freedom and is of the form:

$$t = \frac{b-\beta}{s_b}$$

If the computed t statistic exceeds the critical value of t at the selected significance level, the null hypothesis is rejected with the implication that there is a systematic linear regression of Y on X.

Confidence limits for the regression coefficient β may be set by

$$b - ts_b \leq \beta < b + ts_b$$

The intercept of the regression, α

As with the slope, a is a sample estimate of the parameter, α. Assuming a normal distribution of errors, the variance of the estimated intercept a is:

$$s_a^2 = \frac{s_{Y.X}^2}{(n-2)}$$

Confidence limits for the true intercept α are then

$$a - ts_a < \alpha < a + ts_a$$

Confidence limits for Y_i for a selected value of X_i

Confidence limits that contain the population value of Y_i for any fixed X_i may be calculated for a specific probability level. This estimate involves errors contained in the estimation of α and β by a and b. For this situation the limits are given by:

$$Y_i \pm t s_{Y.X} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{n \text{ var}(X)}}$$

Inspection of the equation shows that for any set probability level, the width of the confidence interval changes with the value of X_i. The interval is narrowest when $X_i = \bar{X}$ because in a bivariate normal distribution of data the concentration of points used in the estimation is the greatest. At increasing distances from the mean, the point density decreases and the interval flares away from the regression line. This "flare" is most exaggerated when n, the sample size, is small.

Fit of the regression line to the sample data.

The total variation in the data to be "explained" is contained in the variance of the dependent variable, or var (Y).
For any single point the raw variation is:

$$(Y_i - \bar{Y})$$

The amount of variation explained by the regression line is:

$$(\hat{Y}_i - \bar{Y})$$

The unexplained or residual variation is then:

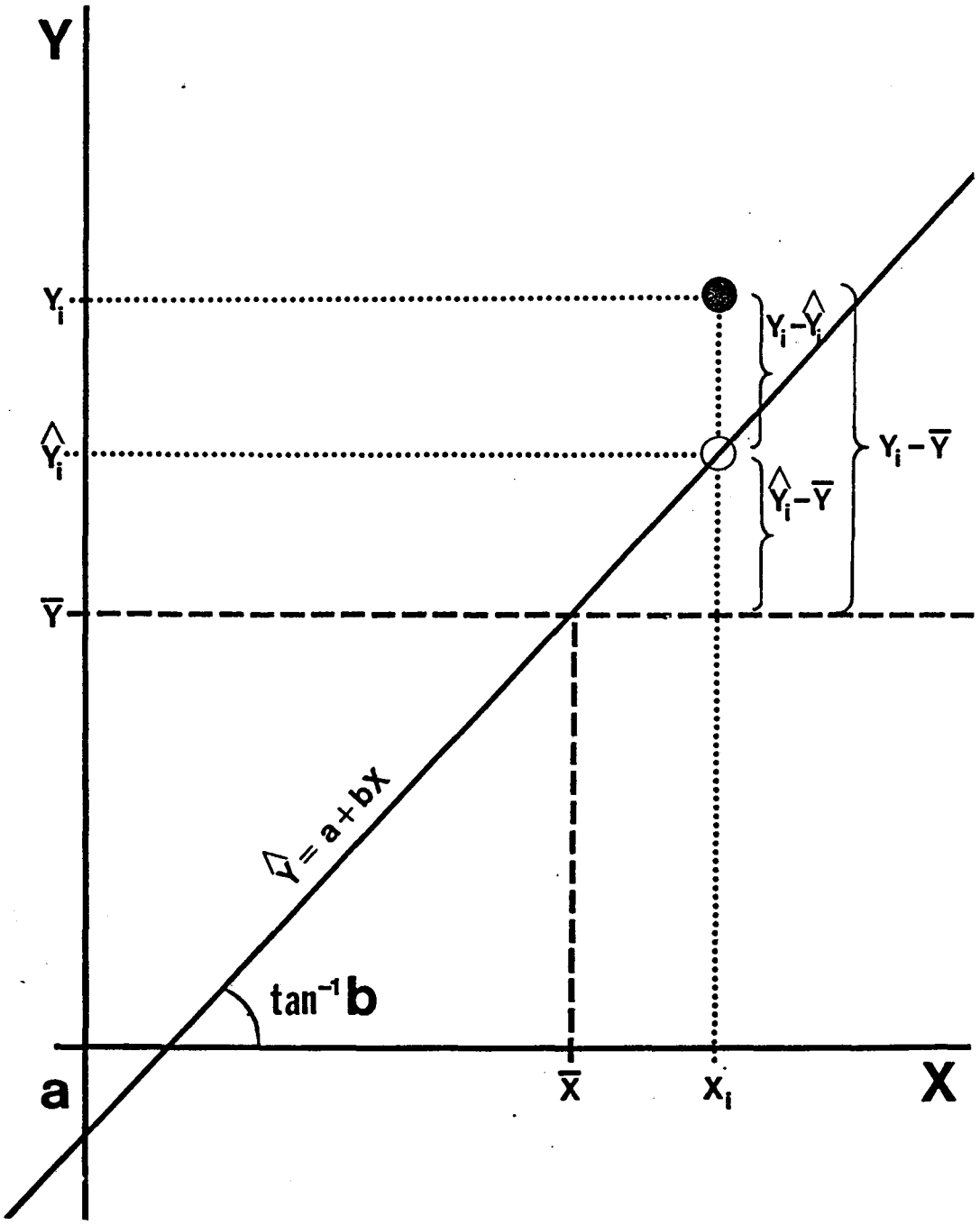
$$(Y_i - \hat{Y}_i)$$

If the total variation is summed for all points, the result is zero because of the cancellation of positive and negative deviations. However, by squaring:

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(\hat{Y}_i - \bar{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2$$

$$\text{or} \quad SS_T = SS_R + SS_D$$

so that the total variation is the sum of the variation explained by the regression plus the residual variation or deviations of the points from the regression.



SOURCES OF VARIATION IN REGRESSION

The ratio $\frac{\text{explained variation}}{\text{total variation}}$ or $\frac{SS_R}{SS_T}$ is known as the coefficient of determination and is equal to r^2 (the square of the product-moment correlation coefficient). The ratio is obviously constrained between the values of zero and one and is often expressed as a percentage figure, which describes the "percentage of sums of squares accounted for" by the regression equation.

An F-test may be used to check whether the regression line explains a significant proportion of the variation in the data in an analysis of variance procedure (q.v.). This is an alternative test of the null hypothesis that the regression coefficient is zero or:

$$H_0: \beta = 0$$

Source of variation	SS	dif.	MS
$\hat{Y}_i - \bar{Y}$ (Regression)	$(\Sigma(X_i-\bar{X})(Y_i-\bar{Y}))^2/\Sigma(X_i-\bar{X})^2 = SS_R$	1	$MS_R = SS_R$
$Y_i - \hat{Y}_i$ (Deviation)	$SS_T - SS_R = SS_D$	n-2	$MS_D = SS_D/(n-2)$
$Y_i - \bar{Y}$ (Total)	$\Sigma(Y_i-\bar{Y})^2 = SS_T$	n-1	$MS_T - SS_T/(n-1)$

In this notation, SS = sums of squares and MS = mean squares. If MS_R and MS_D are the mean squares explained and unexplained by regression, then

$$F = \frac{MS_R}{MS_D}$$

with $\nu_1 = 1$ and $\nu_2 = (n-2)$ degrees of freedom.

Example: Regression of salinity on boron in illite

In the section on correlation, the boron content of illite collected in various environments of the Dovey estuary was related to the average water salinity by a correlation coefficient. Both variables were assumed to have some degree of associated error. The application of a linear regression model to this data requires that one variable be fixed and the other to contain all the error as the random variable. The main error in the boron/salinity data is contained in the salinities which

are estimates of the average figure for each environment. If the analytical error associated with the boron analyses is considered to be relatively low, a linear model of regression of salinity (S) on boron (B) is reasonable.

From the raw data:

$\Sigma B = 4720$ $\Sigma S = 355.0$

$\Sigma B^2 = 1506200$ $\Sigma S^2 = 8743.5$

$\Sigma BS = 113730$

$b = 0.096$

$a = -6.69$

and the regression equation is:

$$S = 0.096B - 6.69$$

the standard error of estimate:

$$s_{S.B} = 3.9$$

If the salinity data were distributed normally about the regression line, approximately 68% of the points would be expected to lie within 3.9 ppt of the line (corresponding to the area under the normal curve between ± one standard deviation).

The variance in the coefficient of the slope, b, is:

$$s_b^2 = 0.000735$$

and $s_b = 0.027$

Setting a null hypothesis that the true slope is zero, or

$$H_0: \beta = 0$$

The critical value for t with $\nu = 13$ and $\alpha = 0.05$ is 2.16

$$t = \frac{b-\beta}{s_b} = \frac{0.096-0}{0.027} = 3.56$$

The null hypothesis is therefore rejected and the alternative hypothesis of a systematic regression of salinity on boron is accepted.

The variance of the estimated intercept is:

$$s_a^2 = 1.19$$

$$s_a = 1.09$$

The intercept of the regression line implies that a zero boron content coincides with a negative salinity. The discrepancy is unlikely to be due to the estimation of the intercept, since the conditon of zero boron/zero salinity would give a "true" intercept which is 6.13 standard deviations from the intercept estimate. It follows that <u>if</u> the relationship is linear over the ranges of boron and salinity, then the illite brought into the estuary has an initial boron concentration. Alternatively, the relationship may not be fully linear and the increase in boron content is more pronounced at higher salinity levels (Adams, Haynes and Walker, 1965).

As a direct application of the salinity/boron regression, estimations of palaeosalinities may be made based on boron concentrations measured in illites in ancient rocks (disregarding such disturbing factors as diagenesis).

The average boron concentration of seven Coal Measure (Pennsylvanian) samples containing non-marine lamellibranchs was 190 ppm. Inserting this figure into the regression equation, the regression estimate of salinity is a brackish 11.6 ppt, which is empirically reasonable since modern seawater contains 30-40 ppt. 95% confidence limits may be computed on this estimate as:

$$11.6 \pm 11.3 \text{ ppt}$$

Assuming the linear regression model is appropriate, it can be stated that the "real" figure for the palaeosalinity of these samples lies between 0.3 and 22.9 ppt, with a probability of 95% that this range is correct.

<u>Reference</u>

Adams, T.D., Haynes, J.R. and Walker, C.T., 1965, Boron in Holocene illites of the Dovey Estuary, Wales, and its relationship to palaeosalinity in cyclothems: Sedimentology, v.4, p.189-195.

<u>Exercise 21.1: Regression of length of synodic month on geologic time</u>

Using the data tabulated in Exercise 20.1 on the number of growth increments (I) per month in fossil shells and their estimated geological age (A), compute a linear regression of I on A (I is the dependent variable and A is the independent variable).

(1) What is the standard error of estimate?

(2) Is the slope of the computed regression significantly different from zero?

(3) What proportion of the total sums of squares are explained by the regression?

(4) What is the regression estimate of the length of the present lunar month? Compute 95% confidence limits for this estimate.

Reference

Pannella, G., Copeland, M., and Thompson, M.N., 1968, Palaeontological evidence of variations in length of synodic month since late Cambrian: Science v. 162, p.792-796.

Reduced Major Axis as a best-fit line

When the correlation between two variables is moderately low, the divergence between the two regression lines of Y on X and X on Y is relatively large. Neither of the two lines appears to be a "best-fit" line to the human eye, which would favor a line approximately bisecting the acute angle between the two regressions. In choosing a best-fit line by eye, the human tends to minimize the sums of squares perpendicular to the line, rather than parallel to either the X or Y axis. The slope of this line may be computed easily as

$$b' = s_Y/s_X$$

As both the standard deviations of X and Y are always positive, the sign of the slope is given by the sign of the correlation or covariance of X and Y. The intercept a' is obtained by

$$\bar{Y} = a' + b' \bar{X}$$

Since, like the regressions, the reduced major axis passes through the sample centroid.

While the reduced major axis appears aesthetically pleasing, it will only function effectively as a prediction equation when the error terms of X and Y are of the same magnitude. Since this is not usually the case, the linear regression model is the conventional choice for analysis and prediction.

Curvilinear regression

The linear regression model may be extended as the general equation:

$$Y = \alpha + \beta X_1 + \ldots\ldots + \omega X_m + \varepsilon$$

where X_1 to X_m may represent different independent variables as a multiple regression, or polynomial powers of one or more independent variables as a curvilinear regression. These models are termed linear since they are linear with respect to the parameters (α, β, etc.) that are estimated. Solutions for these models may be made by more generalized procedures executed as matrix algorithms (q.v.)

In other cases, the theoretical model may be non-linear and take forms such as:

$$Y = e^{\beta X}\varepsilon$$

$$\text{or} \quad Y = \alpha X^{\beta} Z^{\gamma} \varepsilon$$

In many of these examples, the model can be transformed into the equivalent linear form so that

$$\log Y = \beta X + \log \varepsilon$$

$$\text{and} \quad \log Y = \log \alpha + \beta \log X + \gamma \log Z + \log \varepsilon$$

Least-squares procedures are then applied to these models in precisely the same manner as an orthodox linear model. However, the estimated coefficients are least-squares estimates only in the frame of reference of the transformed model, which is not necessarily equivalent to a minimizing of squared deviations in the original model. The problem lies in the transformation of the error term which may or may not conform to its expectancy of independent variable value. Nevertheless, the coefficients from the transformed linear model may provide viable estimates of these quantities.

Example: Sail area and body volume in *Dimetrodon*

The dorsal "sail" of certain Permo-Carboniferous pelycosaurs ("sail-back reptiles") is generally thought to represent a heat-regulating device, whereby body temperature could be raised or lowered by orienting the sail with respect to the sun or by internal control of the blood flow between the body and the sail. The development of the sail is shown particularly well in species of the Permian genus *Dimetrodon*. The body volume increases as the cube of the linear dimensions and body areas as the square. Thus:

$$\text{body volume} = (\text{body area})^{3/2}$$

Dimetrodon data (from Romer, 1948)

	L	S	$\hat{S}$
D. milleri	4.16	390	372
limbatus ♀	5.53	630	621
limbatus ♂	6.08	675	737
grandis	7.61	1190	1104
loomisi	5.53	678	621
gigashomogenes	6.61	920	856

Where L is the orthometric linear unit (OLU)

S is the length of the longest dorsal spine (mm.)

$\hat{S}$ is the regression estimate of S from

$$\hat{S} = 28.6\ L^{1.8}$$

If the hypothesis of heat regulation by the sail is correct as a general relationship,

$$\text{sail area} = (\text{body area})$$

since by the Principle of Similitude, the sail area would increase proportionately with the body volumes (Raup and Stanley, 1971, p. 185).

Romer (1948) checked this relationship from measurements made on several *Dimetrodon* species. As sail areas and body volumes are difficult to measure from fossil specimens, Romer measured the length of the longest neural spine in the sail (the bone support structure) and estimated an "orthometric linear unit" from the cube root of the cross-sectional area of the vertebral centrum. Since the cross-section would be roughly proportional to the weight of the animal (and hence the volume), the "olu" serves as an index of the linear dimension of volume. The spine length would be proportional to the square root of the sail area and the "olu" proportional to the square root of the body area. Therefore, the expected relationship between the two is

$$S = bL^{3/2} + \varepsilon$$

where S is the spine length and L is the "orthometric linear unit." If the exponent is made an unknown quantity, k, the data from fossil skeletons may be used in a non-linear regression model

$$\hat{S} = bL^{k}$$

to estimate the exponent and check it with the figure postulated by the heat regulation hypothesis.

As a linear transformation:

$$\log \hat{S} = \log B + k \log L$$

from which k may be estimated by a least squares procedure, regressing log S on log L.

Data from some of Romer's specimens are shown in the table together with the regression curve estimation of spine length (S) for the measurements of the "orthometric linear unit."

The regression equation is:

$$\hat{S} = 28.6\ L^{1.8}$$

where the exponent of 1.8 computed from the data comes close to the value of 1.5 predicted by the theoretical model.

References

Raup, D.M. and Stanley, S.M., 1971, Principles of Paleontology: W.H. Freeman, San Fransico, 388 p.

Romer, A.S., 1948, Relative growth in pelycosaurian reptiles: Robert Broom Commem. Vol., Special Publ. Roy. Soc. South Africa, p. 45-55.

Exercise 21.2: Relation between porosity and depth of burial

depth, ft.	porosity	
0	54.6	Pierre Shale
1300	52.0	Niobrara Shale
1550	36.0	Carlile Shale
1850	37.6	Greenhorn Fm.
2,100	33.4	Belle Fourche Shale
2,900	32.8	Graneros Shale

Data from Black Hills region (Rubey, 1931).

It has been suggested that differences in the porosity of fine grained sediments should reflect depth of burial. If this is true, it may be possible to estimate the amount of rock section eroded from above a presently-existing outcrop. In order for this to be done, a systematic relationship must be demonstrated between porosity of samples collected at successive depths below an arbitrary point. The Upper Cretaceous section exposed in the Black Hills provides an almost ideal setting to evaluate this hypothesis, because the thick sedimentary section is composed almost entirely of marine shales. The table gives the stratigraphic depth below the Peirre Shale to successive sampled units, and the experimentally determined porosities of the samples. Is there a significant linear or logarithmic relationship between the two variables? What do tests of significance suggest about the reliability of calculations made using Sorby's relationship:

$$\frac{C(100-P)}{P} - D = B$$

where B is thickness of eroded rocks, D is depth below surface, P is porosity, and C is a constant (the value of C = 2200 has been suggested for these rocks)?

References

Rubey, W.W., 1931, Lithologic studies of fine-grained Upper Cretaceous sedimentary rocks of the Black Hills region: U.S.G.S. Prof. Paper 165, p. 1-54.

Sorby, H.C., 1908, On the application of quantitative methods to the study of the structure and history of rocks: Jour. Geol. Soc. London, v. 64, p. 227-231.

Exercise 21.3: Relation between latitude and coiling in forams.

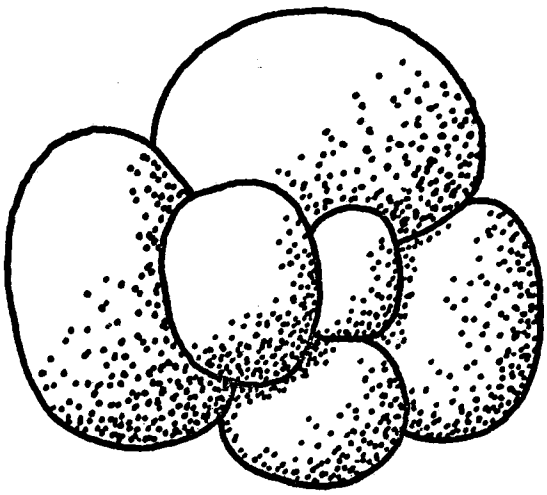
There is evidence that certain modern planktonic forams are sensitive to temperature variations, as the percentages of specimens having dextral coiling seems to increase in the extreme southern latitudes. This is believed to result from the existence of two distinct races or taxa in forms usually classed as single species, with one race preferentially living in a colder environment. The empirically determined difference in coiling ratios (ratio of sinistral to dextral tests) may provide a latitude dependent variable of use as a paleotemperature indicator.

Vella (1974) gives measurements of the coiling ratio for *Neogloboquadrina pachyderma* in the southeastern Indian Ocean. His observations are abstracted in the table below, for two sediment size fractions from samples taken at 17 locations. Determine the relationship between coiling ratio and latitude for the two fractions. Is there a significant difference for the two sizes? If the two regressions are not significantly different, pool the two data sets and compute the regression for the combined groups. In his paper, Vella calculates 3rd order polynomial regressions of coiling ratio versus latitude. Comment on the appropriateness of this.

Reference:

Vella, P., 1974, Coiling rations of *Neogloboquadrina pachyderma* (Elrenberg): Variations in different size fractions: Bull. Geol. Soc. America, vol. 85, p. 1421-1424.

Location Latitude S.	Percent sinistral coiling *N.pachyderma* 0.125-0.175 mm	0.175 mm
28°30.95'	1	6
30°55.81'	9	0
34°29.97'	5	2
36°29.8'	9	5
38°32.93'	8	0
40°36.66'	47	23
41°46.4'	76	28
42°00.81'	63	32
42°35.3'	32	9
43°18.5'	35	27
43°57.2'	24	5
45°00.7'	86	33
47°32.5'	44	20
47°32.0'	75	61
48°01.5'	85	78
51°58.7'	99	98
52°02.3'	94	97



22. MATRIX ALGEBRA

"First of all, say over to yourself ten times "row by column."

- K. Hope on matrix multiplication

Matrix theory in its modern usage originated in the work of Cayley and Sylvester in the 1840's. Matrix algebra is the language of multivariate analysis, making possible the succinct description of techniques that handle cumbersome arrays of data and statistics. From an operational standpoint, matrices are easily stored and manipulated in a computer and most programming languages are structured in terms of matrix representation.

A matrix is a rectangular array of elements such as:

$$\begin{bmatrix} 1 & 4 \\ 7 & 9 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

If a matrix has r rows and c columns, it is said to be or order r $\times$ c (or known as an r $\times$ c matrix). When r = c, the matrix is square. A matrix with only one row is a row vector; a matrix with only one column is a column vector. A matrix having a single row and a single column(i.e., an isolated number or element) is known as a scalar.

In most notations a matrix is denoted by a single capital letter, conventionally printed in heavy type or indicated by an underline as in these notes.

$$\underline{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

a_{ij} is recognized as the element of $\underline{A}$ occurring in the ith row and jth column of the matrix. The leading diagonal of a square matrix corresponds to the a_{ii} elements (i.e. the diagonal, reading from upper left to lower right). A symmetric matrix is a square matrix such that $a_{ij} = a_{ji}$ for all values of i and j. The matrix is symmetrical about the leading diagonal.

$$\begin{bmatrix} 1 & 2 & 5 \\ 2 & 4 & 3 \\ 5 & 3 & 7 \end{bmatrix}$$

The trace of a square matrix is the sum of its leading diagonal elements (12 in the example above). A diagonal matrix has leading diagonal elements which are non-zero values and zero off-diagonal elements. A diagonal matrix is often denoted as $\underline{\Lambda}$.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

An important diagonal matrix is the identity matrix whose diagonal elements are all one. The identity matrix is denoted $\underline{I}$.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The transpose of matrix $\underline{D}$ is written $\underline{D}'$ ($\underline{D}$-hyphen) or sometimes $\underline{D}^t$ and is a matrix whose rows are the columns of the matrix $\underline{D}$ (or equivalently, whose columns are the rows of $\underline{D}$). Then if:

$$\underline{D} = \begin{bmatrix} 3 & 7 \\ 1 & 9 \\ 4 & 2 \end{bmatrix} \qquad \underline{D}' = \begin{bmatrix} 3 & 1 & 4 \\ 7 & 9 & 2 \end{bmatrix}$$

For two matrices to be equal they must be of the same order and their corresponding elements must be identical.

$$\underline{A} = \underline{B} \text{ when } a_{ij} = b_{ij} \text{ for all } i \text{ and } j.$$

Addition of matrices

Only matrices of the same order may be added. The sum of matrices $\underline{A}$ and $\underline{B}$ is the matrix of the sums of their corresponding elements:

$$\underline{A} + \underline{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \\ a_{31} + b_{31} & a_{32} + b_{32} & a_{33} + b_{33} \end{bmatrix}$$

then, $\underline{A} + \underline{B} = \underline{B} + \underline{A}$

Subtraction of matrices operates in a precisely analogous manner, with subtraction rather than addition of individual elements.

Multiplication of matrices

When a matrix is multiplied by a scalar, each element of the matrix is transformed by the scalar:

$$k\underline{A} = \begin{bmatrix} ka_{11} & ka_{12} & ka_{13} \\ ka_{21} & ka_{22} & ka_{23} \\ ka_{31} & ka_{32} & ka_{33} \end{bmatrix}$$

Multiplication of two matrices is more complex and is only possible when the number of columns in the first matrix (the prefactor) is the same as the number of rows in the second matrix (the postfactor). If $\underline{C} = \underline{AB}$, each element c_{ij} in the $\underline{C}$ matrix is the sum of the products obtained by multiplying the ith row of $\underline{A}$ by the jth row of $\underline{B}$ or:

$$c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$$

where m is the number of columns of $\underline{A}$ and rows of $\underline{B}$. For example,

$$\text{if } \underline{A} = \begin{bmatrix} 3 & 1 & 2 \\ 4 & 0 & 2 \end{bmatrix} \text{ and } \underline{B} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 3 & 0 \end{bmatrix}$$

$$\text{then } \underline{C} = \underline{AB} = \begin{bmatrix} 10 & 7 \\ 10 & 8 \end{bmatrix}$$

$$\begin{aligned} \text{where } c_{11} &= 3X1 + 1X1 + 2X3 \\ c_{12} &= 3X2 + 1X1 + 2X0 \\ c_{21} &= 4X1 + 0X1 + 2X3 \\ c_{22} &= 4X2 + 0X1 + 2X0 \end{aligned}$$

The order of notation in a product expression is important since $\underline{AB} \neq \underline{BA}$

$$\underline{BA} \text{ in the above example is } \begin{bmatrix} 11 & 1 & 6 \\ 7 & 1 & 4 \\ 9 & 3 & 6 \end{bmatrix}$$

In many cases, a change in the order of matrices to be multiplied may result in a situation where the product does not exist (i.e., the number of prefactor columns does not equal the number of postfactor rows). So, if $\underline{AB}$ exists if does not necessarily follow that $\underline{BA}$ exists.

The minor product moment = $\underline{A'A}$ and the major product moment = $\underline{A'A}$ and consists of the sums of squares and cross products of the columns and rows of the matrix $\underline{A}$, respectively.

The inverse matrix

Division of one matrix by another in a manner analogous to the scalar operation is not directly possible. However, the operation may be made through multiplication by an inverse matrix.

Now if,

$$\underline{AB} = \underline{C}$$

the value of the matrix B is conceptually equivalent to the division of C by A. However, if a matrix A^{-1} can be found such that

$$\underline{A}\underline{A}^{-1} = \underline{I} \text{ and } \underline{A}^{-1}\underline{A} = \underline{I}$$

then by multiplication:

$$\underline{A}^{-1}\underline{AB} = \underline{A}^{-1}\underline{C}$$
$$\underline{IB} = \underline{A}^{-1}\underline{C}$$
$$\underline{B} = \underline{A}^{-1}\underline{C}$$

$\underline{A}^{-1}$ is known as the inverse matrix of A and is defined by the property that a matrix multiplied by its inverse yields an identity matrix. The inverse is defined only for square matrices and does not exist for all of these. If a square matrix does not have an inverse, it is called a singular matrix.

The widest use of the inverse matrix in practical applications is the solution of unknown quantitites in simultaneous equations. For example, the following equations:

$$3x + 0y + 2z = 19$$
$$x + 4y + 2z = 29$$
$$2x + y + 5z = 35$$

may be written in matrix notation as:

$$\begin{bmatrix} 3 & 0 & 2 \\ 1 & 4 & 2 \\ 2 & 1 & 5 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 19 \\ 29 \\ 35 \end{bmatrix}$$

Solution for the column vector of x, y, z unknowns requires a divisor operation which is achieved by multiplying both sides of the matrix equation by the inverse of the coefficient matrix.

Evaluation of the inverse of low order matrices (< 4) may be made through determinant computation, but becomes prohibitively time-con-

suming for matrices of high order. A variety of iterative algorithms are available for inverse matrix calculation of which one of the best known is the Gauss-Jordan method. The procedure is illustrated in finding the inverse of the simultaneous equations coefficient matrix as follows. First, an identity matrix I is written next to the coefficient matrix:

$$\begin{bmatrix} 3 & 0 & 2 \\ 1 & 4 & 2 \\ 2 & 1 & 5 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The procedure now aims to transform the coefficient matrix into an identity matrix by performing the operations of multiplying any row by a constant and adding or subtracting multiples of any row from any other row (elementary transformations). Each of the operations is simultaneously applied to the corresponding rows of both matrices. The sequence of transformations then runs:

$$\begin{bmatrix} 1 & 0 & 2/3 \\ 1 & 4 & 2 \\ 2 & 1 & 5 \end{bmatrix} \quad \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 2/3 \\ 0 & 4 & 4/3 \\ 2 & 1 & 5 \end{bmatrix} \quad \begin{bmatrix} 1/3 & 0 & 0 \\ -1/3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 2/3 \\ 0 & 4 & 4/3 \\ 0 & 1 & 11/3 \end{bmatrix} \quad \begin{bmatrix} 1/3 & 0 & 0 \\ -1/3 & 1 & 0 \\ -2/3 & 0 & 1 \end{bmatrix}$$

.....

.....

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 9/20 & 1/20 & -1/5 \\ -1/40 & 11/40 & -1/10 \\ -7/40 & -3/40 & 3/10 \end{bmatrix}$$

The original identity matrix is now transformed into the inverse of the coefficient matrix. This may be verified by computing:

$$\underline{A}^{-1}\underline{A} = \underline{I}$$

The column vector of unknowns is solved by premultiplying the equation total vector by the inverse coefficient matrix. The solution is a 3 X 1 vector containing the numerical values of x, y, and z.

Further definitions

A square matrix $\underline{A}$ is said to be orthogonal when $\underline{A}'\underline{A} = \underline{D}$ where $\underline{D}$ is a diagonal matrix. If $\underline{A}'\underline{A} = \underline{A}\underline{A}' = \underline{I}$ then the matrix A is orthonormal. The rank of a matrix is defined as the maximum number of linearly independent rows of $\underline{A}$. Its value is either equal to or less than the order of the matrix. (For a coefficient matrix, the rank states the condition of consistency in the simultaneous equations they describe). A diagonal matrix has one non-zero element in each row on the leading diagonal and so its rank is equal to the number of non-zero elements along the diagonal as seen by inspection. Since the rank of a matrix is invariant during elementary transformations, a matrix may be transformed to its equivalent diagonal form and its rank counted from the number of non-zero elements.

23. MATRIX ALGORITHMS

All God's chillun got algorithm

Reverse II (computational poetics program)

The correlation and linear regression procedures were described in terms of conventional scalar algebra. This is adequate for the task of computing correlations for two variables or estimating two parameters in a regression model. However, if analysis is extended to the consideration of correlation between all possible pairs of a number of variables or a regression model with a large number of parameters, the application of scalar algebra becomes increasingly involved and cumbersome. The number of separate computations increases roughly as the square of the number of unknowns involved. Algebraic manipulations may be expressed in terms of matrices of data and statistical quantities rather than by separate treatment of the component scalars (which are equivalent to the matrix elements). The analytical treatment is made considerably more concise and is in a form that is ideal for programming in digital computers. The matrix format highlights the general case for various statistical models and enhances an appreciation of their basic structure. If scalar algebra can be thought of as a simple mathematical language with a rudimentary syntax, matrix algebra is a higher level language of discourse, enabling a more succinct expression of multivariate analysis. Two examples of matrix algorithms or procedures are examined that translate a set of raw data into the estimates of a statistical model.

Computation of correlation coefficients between m variables for a sample of size n.

The correlation coefficient between two variables X_1 and X_2 is given by:

$$r_{1.2} = \frac{\text{cov}(X_1,X_2)}{[\text{var}(X_1).\text{var}(X_2)]^{1/2}} = \frac{\text{cov}(X_1,X_2^2}{s_{X1} \cdot s_{X2}}$$

If correlations between m variables are required, an m $\times$ m correlation matrix may be computed by the following matrix procedure:

Given: $\underline{X}$ = n $\times$ m raw data matrix

and $\underline{U}$ = 1 $\times$ n unit vector (all elements equal one)

Then:

$$[\text{cov}(X_j, X_k)] = \underline{C} = (\underline{X}'\underline{X} - \frac{1}{n}(\underline{UX})'(\underline{UX}))\frac{1}{n-1}$$

and $[\text{var}(X_j)] = \underline{D}$ (the diagonal matrix form of $\underline{C}$)

so $[s_j] = \underline{D}^{1/2}$

Matrix "division" is achieved by multiplication by the inverse of the appropriate matrix. Now, the inverse of a diagonal matrix is also diagonal, of the same order, and contains the reciprocals of the elements of the original matrix:

$$(\underline{D}^{1/2})^{-1} = \underline{D}^{-1/2}$$

Then the correlation matrix, $[r_{jk}] = \underline{R} = \underline{D}^{-1/2}\underline{C}\underline{D}^{-1/2}$

This concise matrix description may be explained in terms of the computational nuts and bolts (scalar algebra) from the following derivation of the covariance matrix, $\underline{C}$.

$$\text{cov}(X_j, X_k) = \frac{1}{n-1} \sum_{i}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

$$= \frac{1}{n-1} (\Sigma X_{ij}X_{ik} - \frac{1}{n}\Sigma X_{ij}\,\Sigma X_{ik})$$

The expression $\text{cov}(X_j, X_k)$ is an element of the $\underline{C}$ matrix.

Similarly, the quantities on the right-hand side of the equation are elements of matrices and vectors:

(1) $[\Sigma X_{ij}X_{ik}]$ is an m × m matrix of sums of raw squares and cross-products and is produced by premultiplying the raw data matrix, $\underline{X}$ by its transpose = $\underline{X}'\,\underline{X}$.

(2) ΣX_{ij} and ΣX_{ik} are sample totals of raw variables X_j and X_k and are common elements of a vector of sample totals. This vector is a 1 × m row vector computed by multiplying the unit row vector, $\underline{U}$ by the data matrix, $\underline{X} = \underline{UX}$. The m × m matrix of $[\frac{1}{n}\Sigma X_{ij}\Sigma X_{ik}]$ is obtained by premultiplying the sample totals vector by its transpose and multiplying by the scalar (1/n).

(3) The m × m matrix of $[\Sigma X_{ij}X_{ik} - \frac{1}{n}\Sigma X_{ij}\Sigma X_{ik}]$ is the matrix of sums of squares and cross-products of deviations from the variable means. The matrix is the result of subtraction of the matrix of (2) from the matrix of (1).

(4) Multiplication of the "corrected" sums of squares and cross-products by the scalar (n-1) yields the variance-covariance matrix, C.

The matrix equation for computation of R may be easily programmed as an algorithm with relatively few statements in FORTRAN or an equivalent language.

Solution of regression coefficients in the general linear model.

The first order linear regression model (considered in section 21) may be written as the form:

$$\hat{Y} = a + bX$$

Recall that by applying the least-squares principle to the errors and partial differentiation to minimize the squares, two simultaneous equations were obtained for the solution of a and b:

$$na + b\Sigma X_i = \Sigma Y_i$$
$$a\Sigma X_i + b\Sigma X_i^2 = \Sigma X_i Y_i$$

which may be rewritten in matrix form as:

$$\begin{bmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{bmatrix} \quad \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{bmatrix}$$

If this is extended to a multiple regression with Y regressed on m independent variables, the function may be written as

$$\hat{Y} = a_0 X_0 + a_1 X_1 + \ldots\ldots\ldots + a_m X_m$$

where X_0 is a "dummy variable" whose value is always one; $X_1 \ldots . X_m$ are the m independent variables; and $a_0 \ldots . a_m$ are the parameters to be estimated. Then, the sums of squares to be minimized are:

$$G = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma(Y_i - (a_0 X_{i0} + a_1 X_{i1} + \ldots a_m X_{im})^2$$

By partially differentiating G with respect to each of the (m+1) unknown parameters $(\frac{dG}{da_i})$, (m+1) equations are produced for their solution which take the form:

$$a_0\Sigma X_{i0}X_{i0} + a_1 \Sigma X_{i0}X_{i1} + \ldots\ldots\ldots a_m\Sigma X_0X_{im} = \Sigma X_{i0}Y_i$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$a_0\Sigma X_{im}X_{i0} + a_1\Sigma X_{im}X_{i1} + \ldots\ldots\ldots a_m\Sigma X_{im}X_{im} = \Sigma X_{im}Y_i$$

These may be solved as simultaneous equations by conventional algebra or the entire procedure executed as a matrix operation starting from the raw data. The model is written:

$$\hat{Y}_i = a_0X_{i0} + a_1X_{i1} + \ldots + a_mX_{im}$$

where $\hat{Y}_i$ is the regression predicted value of Y for the ith observation of variables $X_1 \ldots . X_m$ in a sample of size n. The vector of coefficients can be premultiplied by the data matrix to yield the matrix of these predicted values:

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} X_{10} & X_{11}\ldots\ldots\ldots\ldots X_{1m} \\ X_{20} & X_{21}\ldots\ldots\ldots\ldots X_{2m} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ X_{n0} & X_{n1}\ldots\ldots\ldots\ldots X_{mn} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ .. \\ .. \\ a_n \end{bmatrix}$$

$$\text{or } \underline{\hat{Y}} = \underline{XA}$$

(Note that the first column of $\underline{X}$ is $[X_{i0}] = [1]$.)

In order to obtain the set of partial differentiation simultaneous equations in matrix form, $\underline{X}$ is first premultiplied by its transpose to give a matrix of sums of squares and cross-products of the raw mariables. Postmultiplication of this matrix by the column vector of unknowns $\underline{A}$ gives

$$\underline{X'X}\ \underline{A}$$

which is the block of left-hand sides of the simultaneous equations. The right-hand sides of the equation are the column vector $[\Sigma X_{ji}Y_i]$ which is the sums of products between the observed Y values and the independent variables. The vector is given by:

$$\underline{X'Y}$$

Then,

$$\underline{X'XA} = \underline{X'Y}$$

Solving for $\underline{A}$:

$$\underline{A} = (\underline{X'X})^{-1}\underline{X'Y}$$

This generalized solution is the basis for programming the linear regression model for the computer, including the three-dimensional polynomial form which is trend surface analysis (q.v.). Polynomial regression models with a single independent variable such as

$$Y = a_0X^0 + a_1X^1 + a_2X^2 + \ldots\ldots + a_mX^m$$

are merely a particular case of the general linear regression model. If a raw data matrix is assembled in the form:

$$\begin{bmatrix} 1 & X_1 & X_1^2 & \ldots\ldots X_1^m \\ 1 & X_2 & X_2^2 & \ldots\ldots X_2^m \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 1 & X_n & X_n^2 & \ldots\ldots X_n^m \end{bmatrix}$$

The solution proceeds as already shown and $\underline{X}'\underline{Y}$ can be written as:

$$\begin{bmatrix} n & S_1 & S_2 & \ldots\ldots S_m \\ S_1 & S_2 & S_3 & \ldots\ldots S_{m+1} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ S_m & S_{m+1} & S_{m+2} & \ldots\ldots S_{2m} \end{bmatrix}$$

Where $S_1 = \Sigma X_i, S_2 = \Sigma X_i^2 \ldots\ldots\ldots\ldots\ldots\ldots S_m = \Sigma X_i^m \ldots S_{2m} = \Sigma X_i^{2m}$

Exercise 23.1: Origin of "stream channels" emanating from craters on Mars

Mars never got big enough to have much geology.

-Sir Edward Bullard

The photographic reconnaisance missions of the Mariner space probe series have substantially changed ideas concerning the nature of the Martian surface. Volcanoes (including the largest known in the solar system, *Nix Olympica),* dune fields and streamlike channels have been discovered. Many of the channels have been suggested to have formed from erosion by water, like their terrestrial equivalents. One type of channel emanates from certain craters and has been hypothesized to be the result of the release of water from a subsurface permafrost layer by meteoritic impact and melting (Maxwell, Otto, Picard and Wilson, 1973). These authors recorded data for seven craters which are foci of channel networks and observed that the majority are in equatorial regions, possibly because equatorial surface temperatures would favor a more prolonged lifetime for released water to carve channels. If these suppositions are true, a simple linear multiple regression model may be proposed in which the average length of crater channel, L, is a function of the crater diameter, D, (as a surrogate variable of meteoric impact heat magnitude) and Martian latitude, T, (reflecting surface temperature) or:

$$\hat{L} = a + bD + cT$$

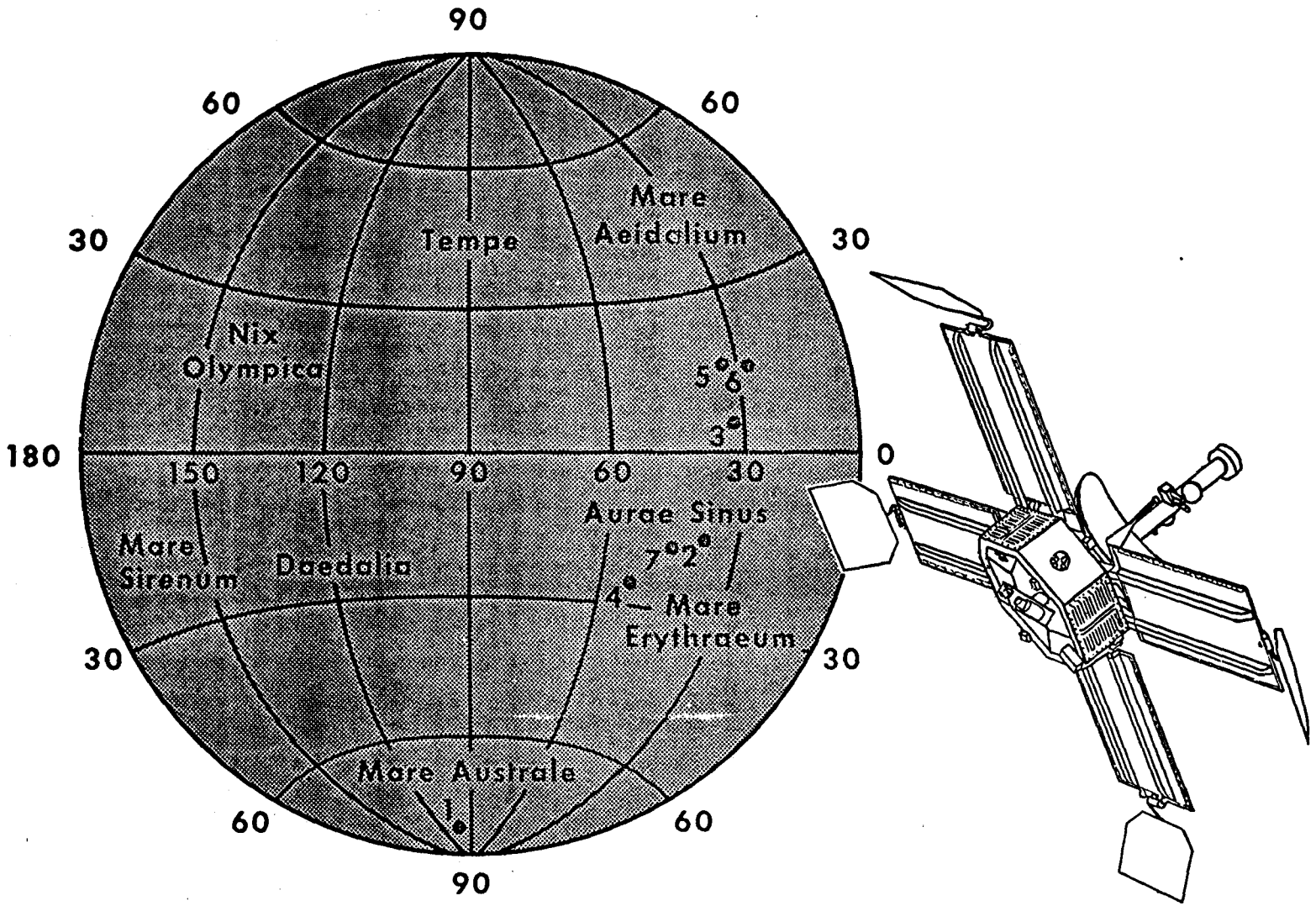
(1) Using the data listed in the table, compute the estimates a, b, and c of the regression parameters by conventional or matrix algebra.

(2) Compute the regression model estimates of the average channel length ($\hat{L}$) for each of the craters.

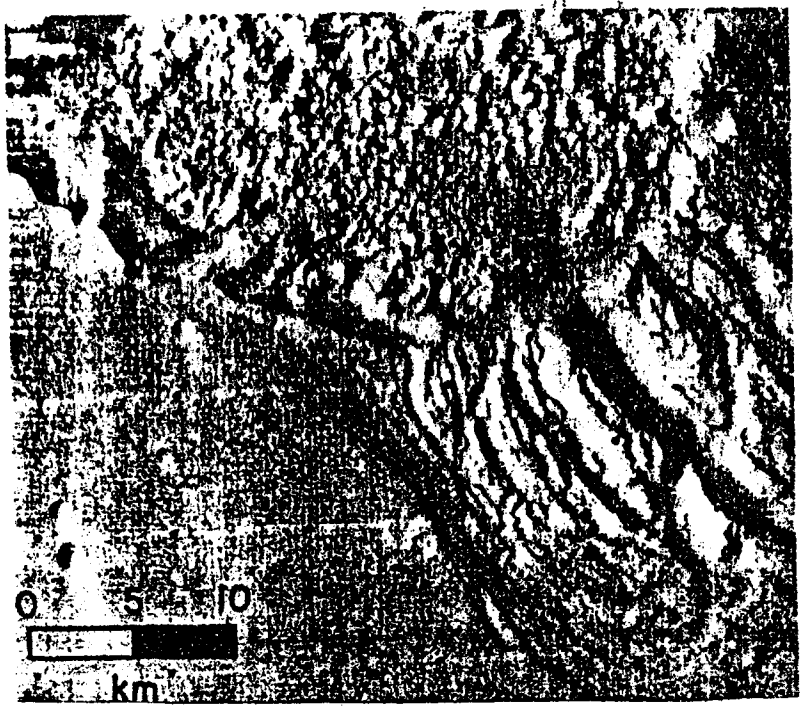
(3) Using the ratio of (explained variation/total variation) in terms of sums of squares, calculate a "goodness-of-fit" for the regression.

(4) Tabulate the explained, unexplained and total sums of squares as an analysis of variance table, together with the associated degrees of freedom. (The ν of the regression is 2.) Compute the mean squares and make an F test of the null hypothesis that the regression of L on D and T is a meaningless relationship for prediction of L.

(5) Compare the goodness-of-fit estimate with the F-test result and comment in terms of the physical nature of the variables and the characteristics of the data set.



Location of Martian craters with associated "stream channels"



Channels originating from crater 3

(NASA photo 4245-048; orbit 211; 2/27/72

CHARACTERISTICS OF CHANNELS ORIGINATING FROM MARTIAN CRATERS

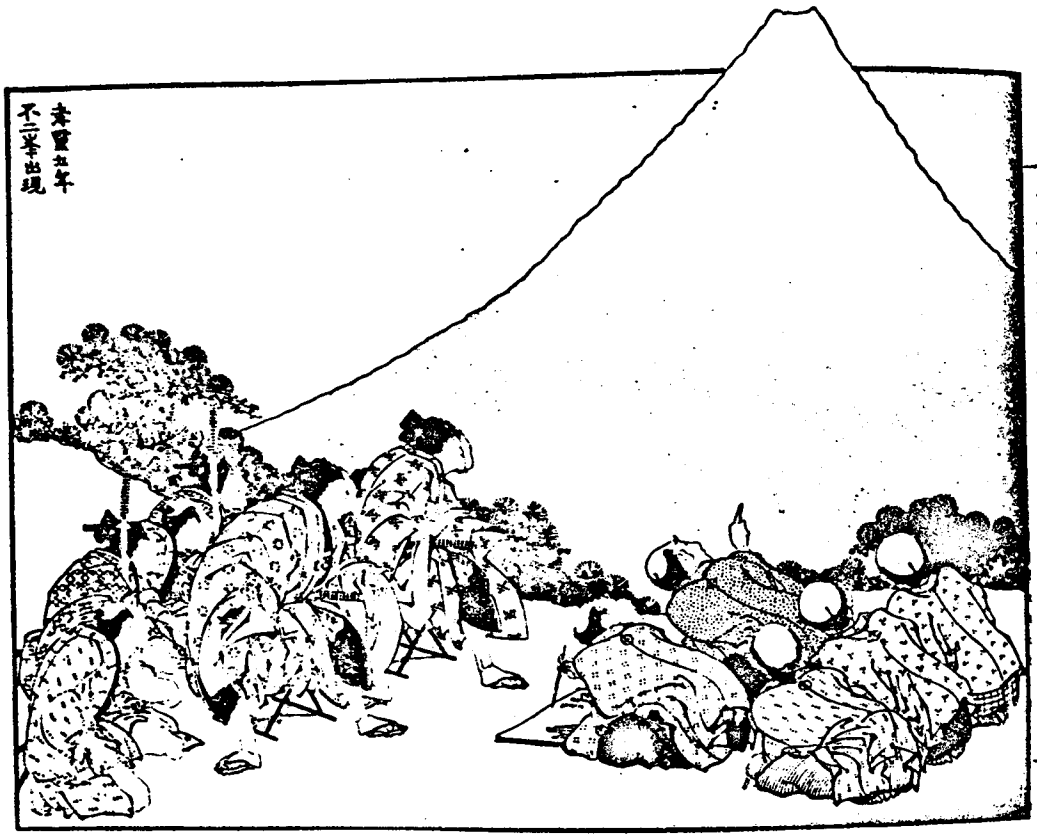
ID	L	D	T
1	88.5	114.6	83.53
2	221.8	46.2	16.07
3	23.8	36.8	4.55
4	37.2	25.4	25.19
5	292.7	114.5	15.97
6	271.0	119.4	15.14
7	31.2	18.2	18.16

ID = Crater identification number
L = Average length of channel (km.)
D = Crater diameter (km.)
T = Latitude of crater (degrees)
(Basic data drawn from Maxwell et al, 1973)

Reference

Maxwell, T.A., Otto, E.P., Picard, M.D., and Wilson, R.C., 1973, Meteoric impact: A suggestion for the origin of some stream channels on mars: Geology, v.1, no. 1, p. 9-10.

24. TIME SERIES - SERIES OF EVENTS



1229	1376	1583	1780	1927
1239	1377	1584	1804	1928
1240	1387	1587	1806	1929
1265	1388	1598	1814	1931
1269	1434	1611	1815	1932
1270	1438	1612	1826	1933
1272	1473	1613	1827	1934
1273	1485	1620	1828	1935
1274	1505	1631	1829	1938
1281	1506	1637	1830	1949
1286	1522	1649	1854	1950
1305	1533	1668	1872	1951
1324	1542	1675	1874	1953
1331	1558	1683	1884	1954
1335	1562	1691	1894	1955
1340	1563	1708	1897	1956
1346	1564	1709	1906	1957
1369	1576	1765	1916	1958
1375	1582	1772	1920	1962

~~A time series, in the widest sense, is the record of a variable which changes through time or with changes in distance along a line. The record may be continuous or discontinuous in time or distance, and the variable may be of any rank.~~ The simplest form of time series is a series of events, which consists of a record of the times (or distances) between "happenings." The events are considered to be: (1) instantaneous points, (2) haphazardly occurring, and (3) distinguishable only by their time (or place) of occurrence.

Geologic examples include the historic record of eruptions of a volcano, the record of first arrival times at a seismic station, spacings between successive fractures along a prospect trench, and distances to successive grain boundaries in a traverse across a thin section.

The historic record of eruptions of the volcano Aso in Kyushu, Japan, dates from 1229 onwards (Kuno, 1962) and is given in the table. Although Aso is a complex strato-volcano, all historic eruptions have been explosive, ejecting ash of andesitic composition. Analysis of such records may shed light on the nature of eruptive mechanisms and can even lead to physical models of the structure of volcanoes (Wickman, 1965).

Series of events studies are designed to

(1) estimate the mean rate of occurrence

(2) estimate the pattern of occurrences, for the purpose of

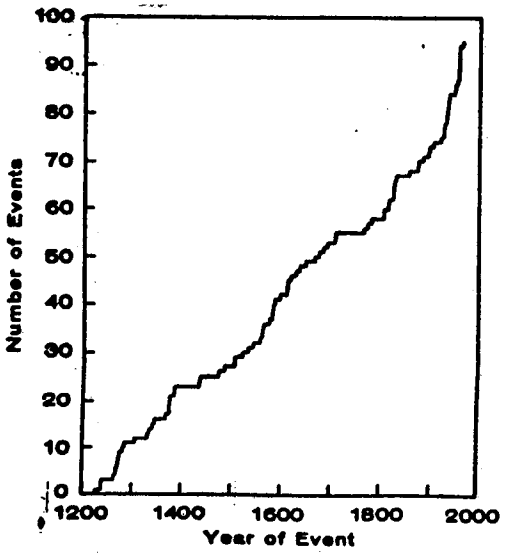
(a) determining the precision of the estimate of the rate of occurrence;

(b) assessing the appropriateness of the sampling scheme;

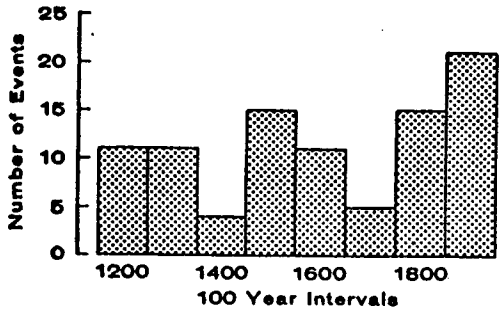
(c) detecting a trend;

(d) detecting some other systematic feature in the rate of occurrence.

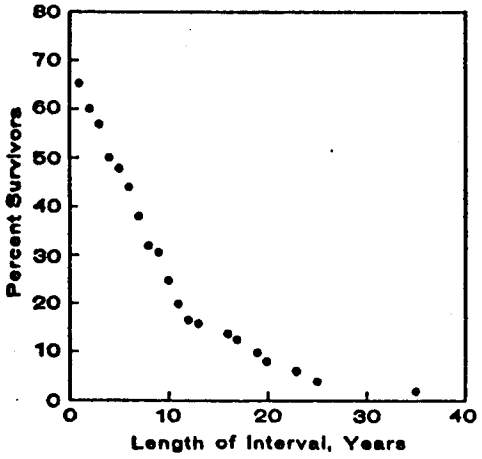
The simplest techniques for studying series of events are graphic. Some of the more useful forms of plots that can be made are shown, using the data for eruptions of Aso.



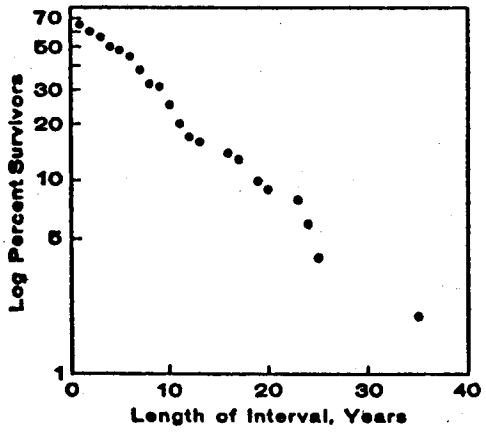
(1) Total number of events (N_t) to have occurred at or before time t, against time t. This is a cumulative plot of the number of events. This plot is especially good for showing changes in the average rate of occurrence. The slope of a line connecting any two points on the curve is the average number of events per unit of time for that period.



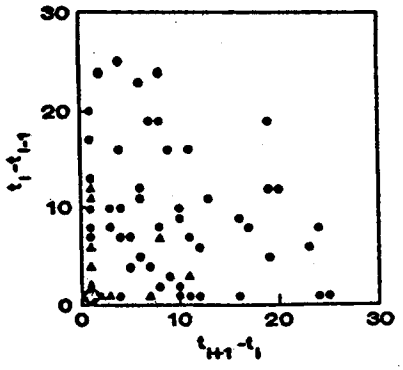
(2) Number of events occurring in successive equally spaced intervals of time. This histogram directly indicates local periods of fluctuation from the average rate of occurrence.



(3) Empirical survivor function, obtained by plotting the percent "survivors," or R(X) = proportion of time intervals longer than X, against X = length of time interval. If events occur randomly in time, this plot will be exponential in form. The survivor function estimates the probability that an has not occurred up to time X.



(4) Log empirical survivor function, a plot of log R(X) against X. Departures from linearity indicate departures from the exponential distribution expected from a Poisson or random process.



(5) Plot of serial correlation between successive intervals between events. The degree of correspondence between the length of an interval and the length of the interval immediately preceding is shown by plotting $X_{i+1} = t_{i+1} - t_i$ against $X_i = t_i - t_{i-1}$.

Here, t_i is the time of occurrence of the ith event. This plot reveals any tendency for intervals to be followed by intervals of similar length; a scatter diagram with large dispersion and relatively high concentration of points near the axes is typical or random series of events.

Tests for trend, or slow changes in the rate of occurrence of events, are usually run first, because any trend present must be removed before other components of the series can be measured. Two possible models for a series of events are the Poisson process (q.v.) in which the rate of occurrence, $\lambda(t)$, is a constant:

$$\lambda(t) = e^{\alpha}$$

and a Poisson-like process in which the rate changes slowly with time:

$$\lambda(t) = e^{\alpha+\beta t}$$

That is, the coefficient β has some value other than zero. This possibility can be tested by linear regression (q.v.).

(1) Transform the series into a new series of the variable X' by dividing the sequence into successive segments which contain ℓ events each. (X'_1 is the elapsed time from the start of the series to event ℓ, X'_2 is the elapsed time from event ℓ to event 2ℓ, and so on.) (2) Create an independent variable Z, defined as the time at the center of each interval of X'. If the events are Poisson,

$$E\ (\log X') = -(\alpha')$$

$$\text{var}\ (\log X') = \frac{1}{\ell - 1/2}$$

If the events are Poisson-like with a trend,

$$E\ (\log X') = -(\alpha' + \beta Z)$$

$$\text{var}\ (\log X') = \frac{1}{\ell - 1/2}$$

In both models, $\alpha' = \log \lambda + e_\ell$.

(3) Estimate the coefficients α' and β by linear regression.

(4) Test the null hypothesis $H_0 : \beta = 0$ by regression methods.

(5) Compare the mean squares for deviation (MS_D) about the regression to var (log X'). If the process is Poisson, they should be the same.

An alternative test for trend in a Poisson-like process is based on the characteristics of an ordered sample from an exponential distribution of durations between events. It consists of a comparison between the centroid of the observed times to successive events, to the mid-point of the total time of observation of the series. If t_i is the time from the start of the series to the ith event, the centroid S is

$$S = \frac{\Sigma t_i}{n}$$

where n is the number of events in the series (if the series begins and ends with an event, use n-1). The mid-point of the series is simply T/2 where T is the total length of the series.

The statistic

$$U = \frac{S - \frac{T}{2}}{\left(\frac{T}{\sqrt{12n}}\right)}$$

has standardized normal form (Cox and Lewis, 1966).

If a series proves to be stationary, successive events may be checked for <u>independence</u>. This can be done by considering the lengths of successive intervals between events as a random variable X. The <u>first order autocorrelation</u> of X is

$$\rho_1 = \frac{\text{cov } (X_i, X_{i+1})}{\text{var } (X)}$$

If the series consists of a large number of observations, the null hypothesis of no autocorrelation, i.e.,

$$H_0 : \rho_1 = 0$$

can be tested by

$$|\rho_1| > \frac{Z}{\sqrt{(n-1)}}$$

where Z is the standard normal variate for a two-tailed test having a signigicance level of α. A series of events which exhibits no trend and no autocorrelation between successive events may be presumed to be random and hence attributable to a Poisson process. If the series exhibits autocorrelation, more complex interrelationships between the lengths of successive intervals can be investigated. These include higher order autocorrelation, special types of Markov processes (q.v.), branching processes, and cyclical changes in rate which can be examined by Fourier analysis (q.v.) of the succession of interval lengths.

References

Cox, D. R., and P.A.W. Lewis, 1966, The statistical analysis of series of events: Methuen & Co., Ltd., London, 285 p.

Kuno, H., 1962, Catalogue of the active volcanoes of the world, part XI, Japan, Taiwan and Marianas: Inter. Volcanological Assoc., Naples.

Wickman, F. E., 1965, Repose period patterns of volcanoes, General discussion and a tentative stochastic model: Arkiv for Mineralogi och Geologi, v.4,p. 351-367.

Exercise 24.1: Occurrence of bentonites in Mowry Shale

The Mowry Shale is a black, siliceous shale of Early Cretaceous age, occurring in Colorado, Wyoming, and Montana. The interval is characterized by numerous bentonite beds, which are mined at several locations in Wyoming and Montana for drilling mud and foundry clay. Bentonite is composed almost entirely of montmorillonite, developed as an alteration produce of rhyolitic or andesitic volcanic ash. The table gives the thickness, in feet, between successive bentonite beds measured in an outcrop of Mowry Shale in Freemont County, Wyoming. These beds are presumed to record the eruptions of volcanoes in western Idaho. If it is assumed that the enclosing black shale was deposited at an approximately uniform rate, it may be possible to analyze this sequence of thicknesses as a series of events analogous to the historical series formed by the eruptions of Aso.

Test these data for trends in the rate of occurrence. If none are observed, test for autocorrelation of successive intervals between events. Comment on the possible effects of (a) unequal sedimentation rates in the sedimentary basin, and (b) presence of more than one volcano in the bentonite source area.

(Bottom)	4	4	14
	26	35	17
	4	2	5
	5	15	10
	4	10	5
	17	23	6
	3	8	11
	6	7	29 (top)
		47	

25. TIME SERIES - AUTOCORRELATION

Time series may exhibit properties which reveal the nature of the process which generated them. These properties may include a trend, periodic fluctuations, or dependency of observations on the value of preceding observations. By transforming a time series into the lag domain through autocorrelation, the nature of the series may be more apparent.

A simple time series consists of a sequence of observations of a variable Y, measured at successive instants in time or points in space. Each observation is separated from the preceding observation by an interval of time or distance which is constant for the series. The position of an observation within the series is indicated by a subscript, as in Y_t. A complete set of observations contains n points and has a total length of $T = n\Delta t$. The interval between points Y_t and $Y_{t+\tau}$ is referred to as a lag of length τ, and is the offset between the series and itself at a previous time or location.

Examples of geologic time series include hydrographic records from stream gauges, electric logs from wells, seismograms, bathymetric traces, and all other manner of continuously-recorded geophysical measurements. It would also be possible to use a time series approach in certain geomorphic problems by considering changes in ground elevation along a line, or to analyze the bed form in a flume tank as a time series.

The autocovariance for lag τ is the covariance between observations Y_t with observations $Y_{t+\tau}$. That is, the covariance is calculated between a series and itself displaced by τ lags.

$$cov_\tau = \frac{1}{n} \Sigma Y_t \, Y_{t+\tau} - \bar{Y}_t \, \bar{Y}_{t+\tau}$$

In dimensionless form this is the autocorrelation or serial correlation for lag τ,

$$r_\tau = \frac{\text{cov}_\tau}{\text{var } Y} = \frac{\Sigma Y_t Y_{t+\tau} - \bar{Y}_t \bar{Y}_{t+\tau}}{\Sigma (Y_t - \bar{Y})^2}$$

For a time series of finite length, the computational equation for autocovariance is

$$\text{cov}_\tau = [n-\tau(\Sigma Y_t Y_{t+\tau}) - \Sigma Y_t \Sigma Y_{t+\tau}]/(n-\tau)(n-\tau-1)$$

and the denominator of the autocorrelation function becomes $\sqrt{\text{var } Y_t \text{ var } Y_{t+\tau}}$. In these equations, the span of summation changes for each value of τ. At $\tau = 0$, summation extends from $t = 1$ to n; at $\tau = 1$, summation extends from $t = 2$ to $(n-1)$ and the calculated value is the first-order autocovariance or autocorrelation. In general, the limits on the summations extend from $t = (1+\tau)$ to $(n-\tau)$. A plot of autocovariance or autocorrelation versus lag is called a covariogram or correlogram.

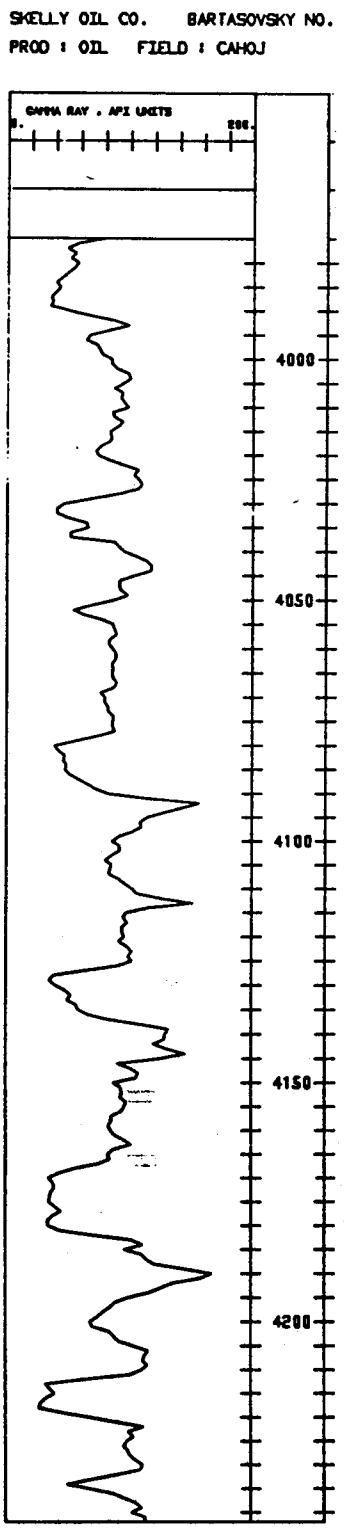
A closely related function is the semivariogram, used in Kriging (q.v.). The semivariance γ is half the variance of the differences between observations Y_t and $Y_{t-\tau}$, or

$$\gamma_\tau = \frac{\text{var } (Y_t - Y_{t-\tau})}{2}$$

If the time series is stationary, this is equivalent to

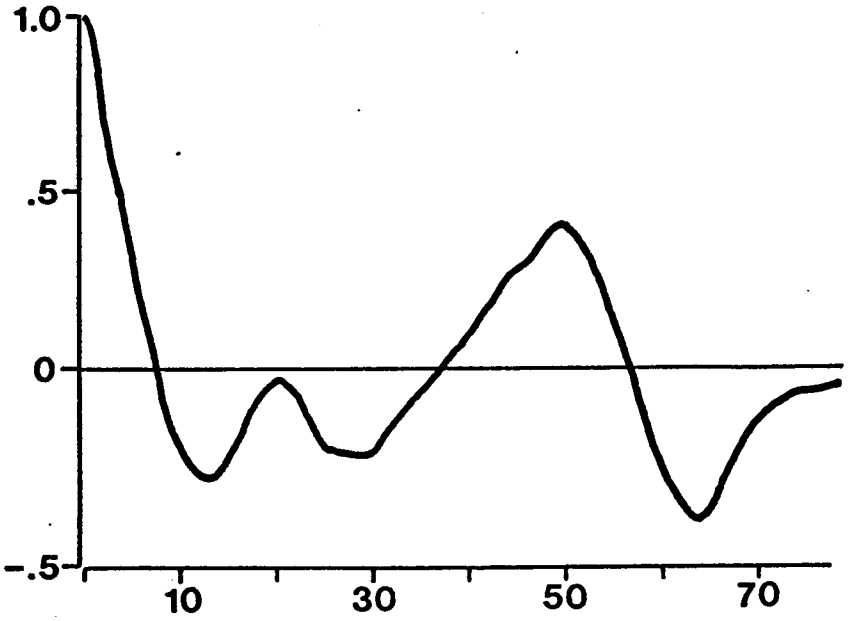
$$\gamma_\tau = \text{var } Y - \text{cov}(Y_t - Y_{t-\tau})$$

Example: Periodicity in gamma-ray logs



The gamma-ray well log at the left is from the Bartasovsky No. 1 well in the Cahoj field in Rawlins Co., Kansas. The segment shown in the Lansing-Kansas City Group (Pennsylvanian), which consists of alternating limestones (curve deflects to the left) and shales (curve deflects to the right). These lithologic differences in the units are reflected in the gamma-ray trace, which responds to the K^{40} content of clay minerals in the shales. In eastern Kansas, the same stratigraphic interval is characterized by classic developments of cyclothems (Moore and Merriam, 1965), and it would be instructive to see if a cyclic pattern is detectable in the log of the Bartasovsky well.

The autocorrelation function for this gamma-ray well log is shown below. If successive gamma-ray measurements were independent, the function would drop to r = 0 at the first lag; instead it is significantly greater than zero up to about lag six. This indicates there is a significant but decreasing relationship between the gamma-ray response at any depth and that measured at points further down the well, reflecting the typical thickness of the stratigraphic units in the interval. However, the most notable feature of the correlogram is the increase in autocorrelation around lag 47, suggesting that there is a pronounced tendency for the stratigraphic succession to be repetitive, and for similar lithologies to recur about every fifty feet.



Analysis of correlograms

1. Select a hypothetical mathematical model for the series, which is characterized by a number of unknown parameters. This provides a family of curves which can represent the correlogram.

2. Estimate the parameters, which is equivalent to selecting a particular curve from the family of curves given by the model.

3. Test the goodness-of-fit of the curve and accept or reject the hypothesis of equivalency between the model and the observed series.

4. Recycle through steps 1 through 3 if the model proves inappropriate.

One of the simplest models that can be proposed for a time series is that successive observations are independent and are normally distributed. That is, there is no relation between an observation at one point and the value at any other point. The expected autocorrelation at any lag is zero,

$$\rho_\tau = 0$$

with an expected variance of

$$\sigma^2 = \frac{1}{n-\tau+3}$$

Approximate confidence bands around the expected autocorrelation for series longer than 75 observations are given by

$$5\% \text{ confidence band} = \frac{-1 \mp 1.64\ n-\tau-1}{n-\tau}$$

$$1\% \text{ confidence band} = \frac{-1 \mp 2.33\ n-\tau-1}{n-\tau}$$

More complex models may assume the series is linearly dependent. A moving average model has the form

$$Y_t = \sum_{\tau=0}^{m} b_\tau Z_{t-\tau}$$

where Z is a standardized independent variable. The expected autocorrelation for a moving averate of span w is

$$\rho_\tau = 1 - \frac{\tau}{w}$$

Autoregressive or Markov models (q.v.) have the general form

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_m Y_{t-m} + \varepsilon_t$$

The correlogram drops asymptotically from one to zero. A first-order autoregressive series is

$$Y_t = \alpha_1 Y_{t-1} + \varepsilon_t$$

and the expected autocorrelation function is

$$\rho_\tau = \rho_1^\tau$$

where ρ_1 is estimated by the first autocorrelation coefficient, r_1. Second-and higher-order models are discussed in Yule and Kendall (1968).

Periodic-component models involve sine and cosine terms and generate series that repeat at regular intervals. Correlograms of such models are cosine functions and also repeat at regular intervals. These models are usually investigated using Fourier analysis (q.v.).

The models discussed above describe a stationary time series. That is, the statistics of the time series converge on the same population parameters regardless of what segment of the series is observed. If the series exhibits a trend, it is nonstationary because different segments will have different means. The model for a linear trend is

$$Y_t = a + bt$$

whose expected correlogram is

$$\rho_\tau = 1$$

for all τ.

Another form of nonstationarity is a jump, or abrupt change in average value at time t', where

$$E(Y_t) = C_1 = \text{constant for } t<t'$$

$$E(Y_t) = C_2 = \text{another constant for } t>t'$$

The exact form of the expected correlogram of a series containing a jump depends upon the difference between C_1 and C_2 and the lengths of the two segments of the series. If the two segments are both of length ℓ and the jump is symmetrical around zero (i.e., $C_1 = -C_2$), the autocorrelation is given by

$$\rho_\tau = \left(1 - \frac{3\tau}{\ell}\right) \frac{\ell}{1-\tau}$$

26. TIME SERIES - SPECTRAL ANALYSIS

Spectral analysis is the partitioning of the variation in a time series into components according to the duration of the intervals over which they occur. If a phenomenon is measured either continuously or at discrete points spaced through time (or along a line in space) the measurements can be regarded as a function of time (or distance).

$$Y = f(x)$$

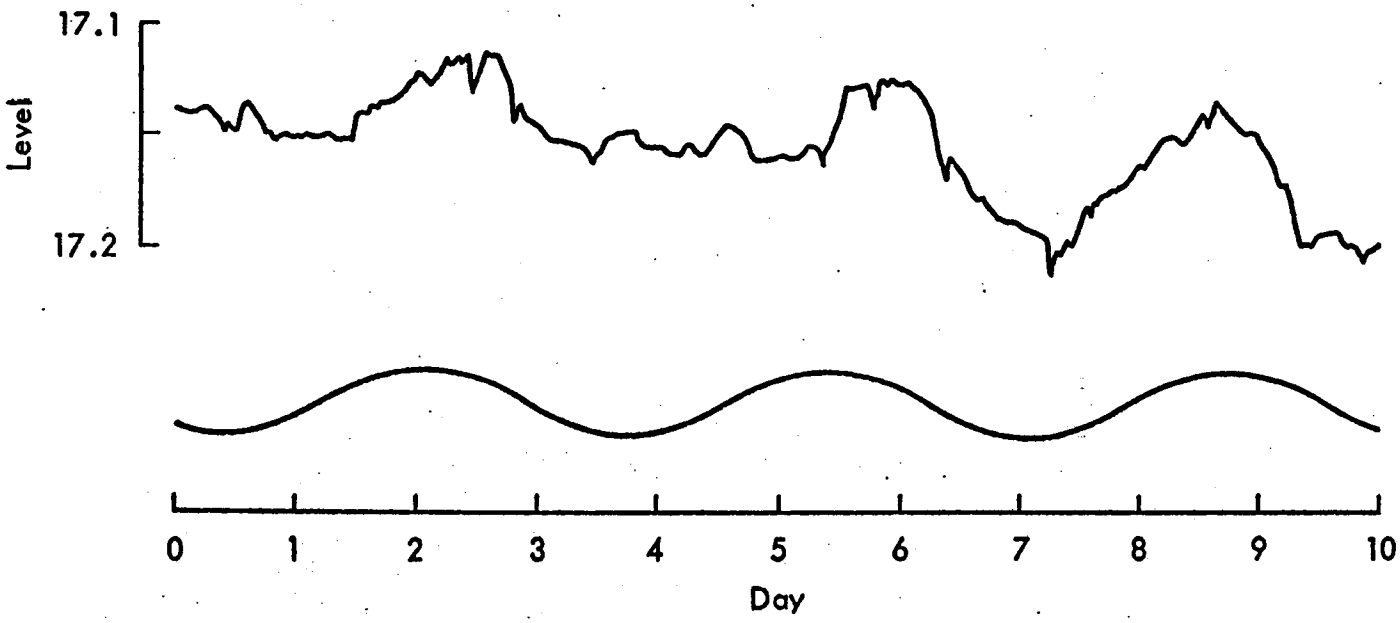
f(x) may also be expressed as the sum of number of sine and cosine terms. Such a series is called a Fourier series and the determination of these coefficients is called spectral analysis, harmonic analysis, Fourier analysis, or frequency analysis.

Historical development

Spectral analysis has roots which are interwoven with those of musical theory, reflecting their common concern with vibratory motion. Pythagoras, in the 6th century B.C., discovered the relation between musical notes produced by a vibrating string and the length of the string. Kepler, applying the harmonic relationships he found in arithmetic, geometry, and music, discovered the laws of planetary motion in the early 17th century. Bernoulli and Euler studied the mathematical functions that described the vibratory motion of musical strings, leading to Bernoulli's publication in 1728 of the relationship that eventually would be known as the Fourier theorem.

Jean Baptiste Fourier, a mathematician, physician, and one-time overlord of Egypt during the Napoleonic era, provided proof that a continuous, single-valued function could be represented by a series of sinusoids, the relationship that now bears Fourier's name. This allows a function such as a time series to be decomposed into a hierarchy of simpler wave-

forms, each of which contains a certain percentage of the variance of the original series.



For example, a 10-day record of groundwater level is an observation well near Wichita, Kansas, is shown above. (Williams and Lohman, 1949). The variance in water level is 6.55×10^{-4} feet. A sinusoid with a wavelength of approximately three days was fitted to this series and is shown beneath the original; this wave has a variance of 1.45×10^{-4} feet, or about 22% of the original variance.

Basic trignometric relationships

The illustration shows a sinusoidal curve. Assume the wave makes one complete fluctuation in a time (or distance) T. The total duration (or length) of the interval can be expressed in radians by the conversion

$$\theta = \frac{2\,t}{T} \text{ radians}$$

(1) The equation for the curve is

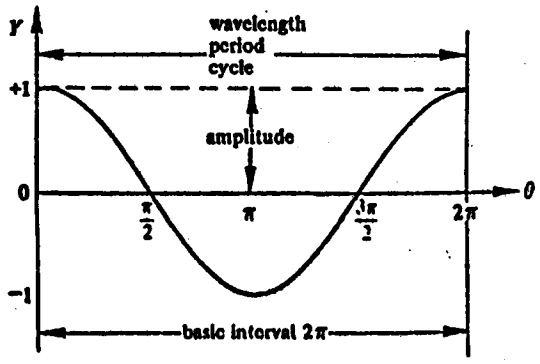
$$Y = \cos \theta$$

(2) The amplitude may be altered to any magnitude by

$$Y = A \cos \theta$$

(3) The number of cycles per fixed interval (the frequency) may be increased or decreased by

$$Y = \cos (k\theta)$$



(4) The crest of the wave shown starts at the origin, but may be moved to any position by subtracting a phase angle

$$Y = \cos(\theta - \phi)$$

(5) In general, any curve of regular sinusoidal form may be written

$$Y_k = A_k \cos(k\theta - \phi_k)$$

Because of the trigonometric relationship cos (R-S) = cos S cos R + sin S sin R, this equation may be rewritten

$$Y = A_k \cos\theta_k \cos(k\theta) + A_k \sin\theta_k \sin(k\theta)$$

(6) Define $\alpha_k = A_k \cos\theta_k$ and $\beta_k = A_k \sin\phi_k$. Then the equation becomes

$$Y_k = \alpha_k \cos(k\theta) + \beta_k \sin(k\theta)$$

(7) A complex time series can be defined as the sum of these sinusoids

$$Y = \sum_{k=0}^{\infty} \alpha_k \cos(k\theta) + \beta_k \sin(k\theta)$$

which is an expression of Fourier's relationship.

Calculation of the Fourier coefficients

If the time series is sampled at n equally spaced points, one of which is j, the computational equations used to find the α and β coefficients are

$$\beta_k = \frac{2}{k} \sum_{j=0}^{n-1} Y_j \sin\left(\frac{2\pi jk}{n}\right)$$

$$\alpha_k = \frac{2}{k} \sum_{j=0}^{n-1} Y_j \cos\left(\frac{2\pi jk}{n}\right)$$

Because of trigonometric relationships, β_0 vanishes and α_0 simplifies to

$$\alpha_0 = \frac{1}{n} \sum_{j=0}^{n-1} Y_j$$

which is simply the mean of the time series.

The Fourier coefficients can be combined to obtain the amplitudes

$$A_k = \alpha_k^2 + \alpha_k^2$$

and phase angles of the constituents sinusoids

$$\phi_k = \tan^{-1} \left(-\frac{\alpha_k}{\beta_k}\right)$$

(The expression $\tan^{-1}$ means "find the angle whose tangent is given.")

The Variance Spectrum

The variance of a sinusoidal wave form sampled at regular intervals is simply half the square of the amplitude of the wave. That is,

$$s_k^2 = A_k^2 / 2$$

Therefore, the contribution of any frequency k of the total variance of a time series is

$$\text{contribution in percent} = \frac{s_k^2}{s^2} \cdot 100\% = \frac{A_k^2}{2s^2} \cdot 100\%$$

The variances of the individual frequencies may be plotted as a variance spectrum or power spectrum (the latter term is favored in engineering applications). The variance spectrum which is calculated for an observed time series is sometimes called the raw spectrum, and is an estimate of the true or population spectrum. If the original time series is continuous, but sampled only at regular intervals, the computed spectrum is not complete as wavelengths shorter than twice the sample spacing (the Nyquist frequency) cannot be estimated. Instead these short wavelengths are confounded in the longer wavelengths; this is called aliasing. In addition, the observed time series is usually only a sample from a much longer (or infinite) series and so the variances calculated are statistics estimating the true population parameters. The standard error of these raw spectral estimates commonly is very high, on the same order of magnitude as the raw estimates themselves. Better estimates can be created by first calculating the autocovariance or autocorrelation of the time series, taking the Fourier transform of this function, and then smoothing the spectrum by averaging adjacent values. Alternative equations have been proposed for the smoothing operation. A commonly used form calculates the raw spectrum as

$$s_k^2 = 2\left[1 + 2 \sum_{\tau=1}^{m} \text{cov}_\tau \cos 2\pi k\tau\right] \text{ where } \tau \text{ is the lag}$$

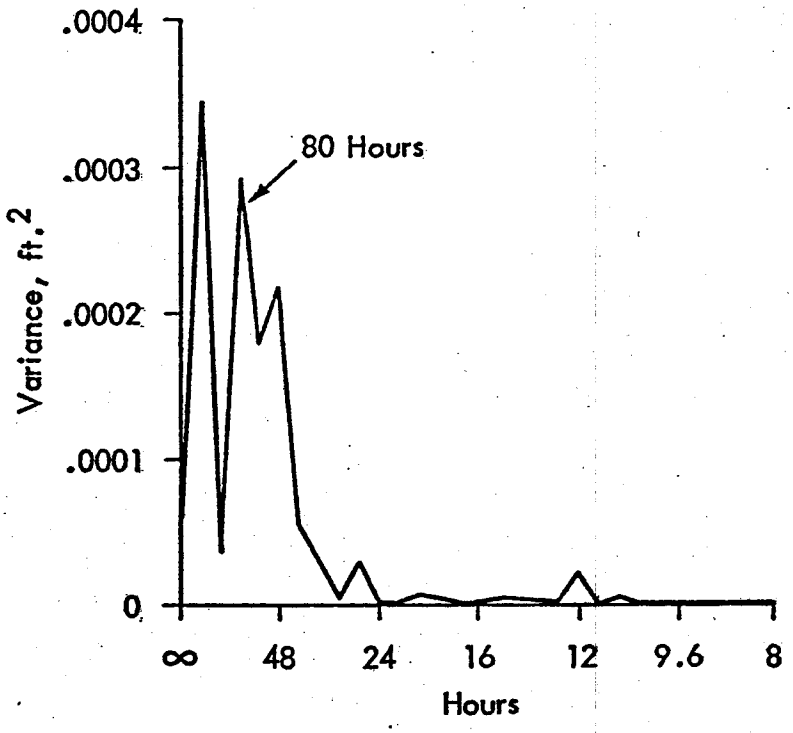
up to a maximum lag of $\tau = m$ (usually $\frac{n}{4}$ or less). The raw spectrum is

smoothed by a filter such as the Hanning filter:

$$s_k^2 = 1/4\ s_{k-1}^2 + 1/2\ s_k^2 + 1/4\ s_{k+1}^2$$

A third approach is to express the Fourier transform in an exponential form involving the imaginary number i. Then, the Fast Fourier Transform (FFT) algorithm can be used to find the complex coefficients for all wavelengths up to the Nyquist frequency, provided the number of points in the time series is some power of 2.

The figure below shows the variance spectrum of the water level data. Note the pronounced peak at 80 hours, corresponding to the periodic three-day component.



Components of time series

Time series may be deterministic or stochastic, or mixtures of the two. Deterministic components are those whose behavior can be predicted exactly. These include periodic changes which repeat at regular intervals, and transients, most commonly trends and jumps. Stochastic components, in contrast, are characterized by their statistical properties. The development of a stochastic time series is governed by probability

functions, and its exact state at any instant cannot be predicted with certainty.

Most geologic time series are either completely stochastic or stochastic with a periodic component. For example, hydrologic series such as stream levels may be stochastic with an annual periodic components related to spring runoff. Other geologic phenomena may exhibit trends or nonstationarity (q.v., autocorrelation).

The presence of a suspected periodic component may be tested by calculating the probability that a spectral value s_k^2 will exceed the value σ_k^2 of an independent stochastic process. The test, devised by Fisher, involves calculation of the ratio

$$\hat{g} = \frac{s^2_{max}}{2s^2}$$

where s^2_{max} is the largest value in the variance spectrum and s^2 is the variance of the time series. The critical value of g for a specified probability P is given by

$$g \approx 1 - e^{\frac{ln\ P - ln\ m}{m-1}}$$

where $m = \frac{n}{2}$ if the series contains an even number of observations and $m = \frac{n-1}{2}$ if n is odd. If the test value of $\hat{g}$ exceeds the critical value $\hat{g}$, the periodic component may be presumed to exist, and may be removed from the time series to isolate the stochastic component. If the test value does not exceed the critical value, the observed variance s_k^2 could have arisen by chance from a purely stochastic process.

References

Box, G.E.P., and A.M. Jenkins, 1970, Time series analysis, forecasting and control: Holden-Day, Inc, San Francisco, 553 p.

Lee, Y.W., 1960, Statistical theory of communication: John Wiley & Sons, Inc., New York, 509 p.

Rayner, J.N., 1971, An introduction to spectral analysis: Pion Ltd., London, 174 p.

Williams, C.C., and S.W. Lohman, 1949, Geology and ground-water Resources of a part of South-central Kansas: Kansas Geological Survey Bull. 79, 455 p.

Yevjevich, V., 1972, Stochastic processes in hydrology: Water Resources Publications, Ft. Collins, Colo., 276 p.

27. MARKOV CHAINS

"I drew one conclusion which I believe to be correct: that is, though there is no system, there really is a sort of order in the sequence of casual chances--and that, or course, is very strange."

-Dostoevsky (The Gambler)

There are many situations in which a sequence of events in either time or space, is observed as a succession of mutually exclusive states. Geological examples include traverses across a thin-section in which each 'event' corresponds to the mineral ('state') recorded at each point and stratigraphic successions, where observations spaced at constant intervals record the occurrence of a rocktype at those points. The relationship between adjacent events may be summarized by a transition tally matrix in which each cell sums the number of times that one state (identified by the row) is succeeded by another (identified by the column). So, for example, a succession consisting of lithologies A,B,C and D may be summarized as a transition tally matrix by taking observations at successive one-foot intervals, accumulating transition totals in the appropriate cells and might appear as:

	A	B	C	D	
A	7	4	3	1	15
B	4	5	2	1	12
C	1	2	4	5	12
D	3	1	3	3	10
	15	12	12	10	

where, for example, the number of times C succeeds A is 3. It can be seen that the ith row total is equal to the ith column total, which is a property of all transition matrices of this type, since every lithology that is entered is also left (with the exception of the initial and terminal events). These row/column totals may be written as the vector:

[15 12 12 10]

which represents the number of times the succession observations are in each of the four states.

Division of the tally matrix by each of the row totals leads to a transition probability matrix, P:

$$\underline{P} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.47 & \text{~~0.37~~ .27} & \text{~~0.30~~ .20} & 0.07 \\ 0.33 & 0.42 & 0.17 & 0.08 \\ 0.08 & 0.17 & 0.33 & 0.42 \\ 0.30 & 0.10 & \text{~~0.20~~ .30} & 0.30 \end{bmatrix} \end{array}$$

Similarly, division of the totals vector by the grand total results in an estimate of the fixed probability vector:

$$[0.31 \quad 0.24 \quad 0.24 \quad 0.20]$$

which expresses the proportions of each lithology in the total sequence. Since A, B, C and D are mutually exclusive events, the probability that one state is followed by another is either a conditional or an unconditional probability. P(B/A) is the notation that A will be followed by B, given that A has occurred as the previous event. In the unconditional case:

$$P(B/A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A)\,P(B)}{P(A)} = P(B)$$

as opposed to the conditional alternative where:

$$P(B/A) \neq P(B)$$

If all the transitions are unconditional, then:

$$\text{e.g. } P(B/A) = P(B/B) = P(B/C) = P(B/D)$$

and the model is one of independent events. The expected transition probability matrix, A, for independent events consists of rows of the fixed probability vector. For the example:

$$\underline{A} = \begin{bmatrix} 0.31 & 0.24 & 0.24 & 0.20 \\ 0.31 & 0.24 & 0.24 & 0.20 \\ 0.31 & 0.24 & 0.24 & 0.20 \\ 0.31 & 0.24 & 0.24 & 0.20 \end{bmatrix}$$

A matrix of expected tallies is computed by multiplying by the row totals of the observed tally matrix. The null hypothesis of independent events may be tested as a chi-square contingency (q.v.) with

$(m-1)^2$ degrees of freedom (where m is the number of states). If the null hypothesis is rejected, the alternative model is accepted of a partial dependency between successive events, and is known as a Markov chain of first order.

Named after its discoverer, A.A. Markov, whose inspiration was the alternation of vowels and consonants in Pushkin's poem 'Onegin', Markov chains are an example of a stochastic process model. Markov chain models occur in the range between the extremes of determinism, where every event is exactly specified by its predecessor, and independent events, where there is no relationship between successive events.

If the matrix $\underline{P}$ is squared,

$$\text{i.e. } \underline{P}^2 = \underline{P}\,\underline{P}$$

the resulting matrix is the expectation of the probability of the (i + 2) event given that of the ith event, as predicted by the first order Markov chain. If this matrix differs significantly from that observed in the sequence (as judged by chi-square test of the appropriate tally matrices), the sequence has second order Markov properties. If the matrix $\underline{P}$ is successively powered to the limit, the matrix approaches a matrix of equilibrium proportions of the states, which corresponds to $\underline{A}$.

There are several problems that must be resolved before analysing typical geological sequences for Markov properties:

(1) The length of interval between successive events must be selected. In analysing a lithological succession, if too small an interval is chosen, the number of transitions of states to themselves becomes extremely large and a Markov property reflects the trivial fact that successive observations tend to be within the same bed. If too large an interval is used, many thin bed 'events' are missed altogether. The problem may be resolved by structuring the model in terms of an embedded Markov chain, where transitions are recorded between successive states. The transition matrices of the embedded case have zero entries on their leading diagonal.

(2) A long sequence of events must be recorded to provide an adequate sample for parameter estimation and the testing of hypotheses, since

the number of transition types to be estimated increases as the square of the number of states. The conventional ground rules of the chi-square test that stipulate a minimum expected tally count in each cell must be obeyed as a conservative safeguard on test validity. (3) Each computed transition probability is a sample estimate of its population parameter. Consequently, the transition probabilities must not change systematically over the sequence i.e. the sequence must be stationary. In geological applications, a sequence of events may be non-stationary as, for example, a succession with long-term facies variation, where the transition properties of rocktypes may change.

Example: Embedded Markov chain analysis of a logged borehole succession from the Ayrshire Coal Measures (Pennsylvanian)

The Ashentree No. 1 Bore penetrates 1600 feet of interbedded shales, siltstones, sandstones and coals of dominantly non-marine aspect and interpreted to be the product of delta-plain depositional environments. Five states may be defined which represent the main lithologies:

A = 'Barren' shale or mudstone
B = Shale containing non-marine bivalves
C = Siltstone
D = Sandstone
E = Rootlet horizon or coal

The bed succession from the base of the bore to Skipsey's Marine Band are listed in the table. Transitions from one bed to another are analyzed so that the transition matrices are in the format of an embedded Markov chain.

The observed one-step transition tally matrix is:

	A	B	C	D	E	
A	0	11	36	21	52	120
B	28	0	4	4	0	36
C	34	2	0	45	13	94
D	29	1	45	0	3	78
E	28	23	9	8	0	68
	119	37	94	78	68	//396

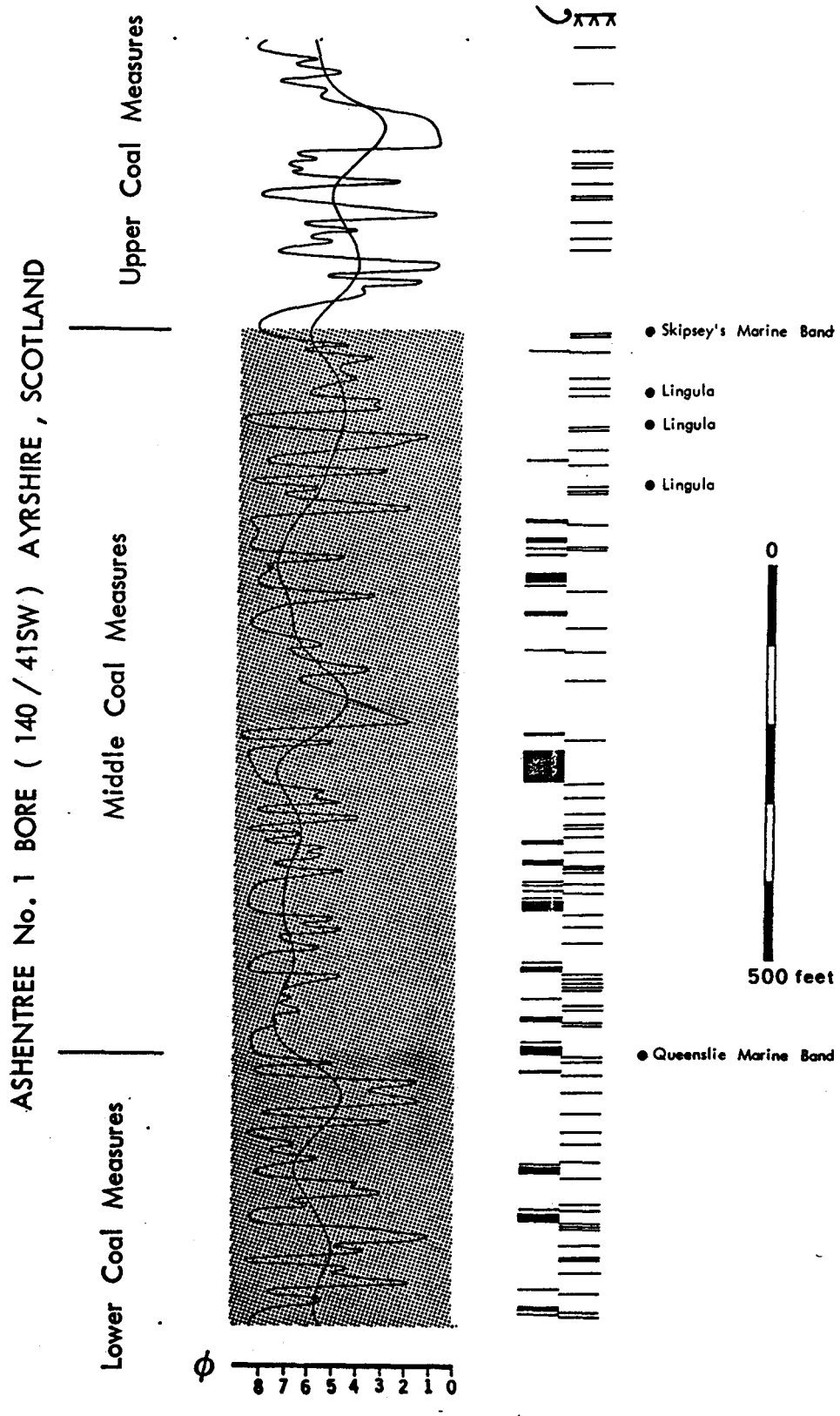
SEQUENCE OF BEDS IN THE ASHENTREE NO. 1 BORE (140/41SW) AYRSHIRE, SCOTLAND RANGING FROM TOTAL DEPTH TO SKIPSEY'S MARINE BAND (PENNSYLVANIAN COAL MEASURES)

1 A = 'barren' shale or mudstone
2 B = shale with non-marine bivalves
3 C = siltstone
4 D = sandstone
5 E = rootlet horizon or coal

READ BY ROWS
Bottom of Succession

A	E	A	B	D	A	E	B	A	B
A	D	A	C	D	C	D	A	E	B
D	C	E	A	D	C	A	E	A	E
A	C	D	C	E	D	A	D	A	E
D	C	A	E	A	E	B	A	E	B
A	D	C	E	A	E	A	D	C	D
C	D	C	E	A	C	A	E	A	B
D	B	A	D	E	C	A	C	D	A
D	A	E	A	D	A	C	D	C	A
E	A	D	C	D	A	E	D	C	D
A	E	B	A	C	D	A	E	D	C
A	E	B	A	D	A	E	A	C	A
E	B	C	A	E	C	D	C	A	E
A	D	C	B	A	C	A	E	A	C
A	E	A	C	A	E	C	D	C	E
D	C	D	C	E	A	D	A	E	B
A	C	D	C	D	C	A	E	C	A
C	A	C	D	C	D	C	A	E	A
D	C	A	C	E	B	A	B	A	E
B	A	B	A	E	B	A	C	A	D
A	E	D	C	D	A	E	B	D	A
C	E	C	B	A	C	E	C	D	C
A	E	A	E	A	D	C	A	E	A
C	D	A	D	A	E	C	A	D	A
E	B	A	B	A	C	D	C	A	E
B	A	D	C	D	C	A	D	C	A
E	C	D	C	A	C	D	A	C	E
B	C	A	C	A	E	B	C	A	C
D	C	D	C	D	C	A	E	B	A
B	A	C	A	C	A	B	A	E	B
A	E	A	B	A	E	B	A	C	D
C	D	C	D	C	A	E	C	D	A
E	D	A	E	A	B	A	C	D	C
E	A	C	D	A	E	B	A	C	A
D	A	E	D	A	C	E	A	C	D
E	A	B	A	C	D	C	D	C	D
A	E	B	C	D	C	E	A	C	D
A	C	D	E	A	C	D	C	D	C
D	A	E	B	A	C	D	C	D	C
A	D	A	E	A	E	B			

Skipsey's Marine Band



Smoothed grain-size profile of the Ashentree Bore succession.

: occurrence of non-marine bivalves
: rootlet horizons and/or coals
Shaded part of the succession is the sequence used for Markov chain analysis.

The vector of column totals is:

[119 37 94 78 68] /396

The observed transition probability matrix is: [.30 .09 .24 .20 .17]

$$\underline{P} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & 0.09 & 0.30 & 0.18 & 0.43 \\ 0.78 & 0 & 0.11 & 0.11 & 0.00 \\ 0.36 & 0.02 & 0 & 0.48 & 0.14 \\ 0.37 & 0.01 & 0.58 & 0 & 0.04 \\ 0.41 & 0.34 & 0.13 & 0.12 & 0 \end{bmatrix} \\ .30 \quad .09 \quad .24 \quad .20 \quad .17 \end{array} \begin{array}{c} \\ .30 \\ .09 \\ .24 \\ .20 \\ .17 \end{array}$$

and the column estimate of the fixed probability vector is:

[0.30 0.09 0.24 0.20 0.17]

Since the one-step transition precludes the succession of a state by itself in the embedded model, the expected transition matrix for the situation of independent events is not a matrix of rows of the fixed probability vector. By taking account of this constraint, the expected transition probability matrix for independent events is:

$$\underline{A} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & 0.12 & 0.36 & 0.28 & 0.24 \\ 0.38 & 0 & 0.25 & 0.20 & 0.17 \\ 0.46 & 0.10 & 0 & 0.24 & 0.20 \\ 0.43 & 0.09 & 0.29 & 0 & 0.19 \\ 0.42 & 0.09 & 0.28 & 0.22 & 0 \end{bmatrix} \end{array}$$

and the corresponding expected tally matrix is:

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & 13.8 & 43.7 & 33.9 & 28.6 \\ 13.8 & 0 & 9.1 & 7.1 & 6.0 \\ 43.7 & 9.1 & 0 & 22.4 & 18.8 \\ 33.9 & 7.1 & 22.4 & 0 & 14.6 \\ 28.6 & 6.0 & 18.8 & 14.6 & 0 \end{bmatrix} \end{array}$$

Applying a chi-square test to the null hypothesis of independent events (using a conventional chi-square contingency table procedure):

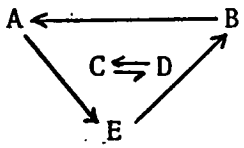
$$\text{Computed } \chi^2 = 178.0$$

$$\nu = (m-1)^2 - m = 11$$

The critical value of χ^2 for $\nu = 11$ and $\alpha = 0.05$ is 19.7. The null hypothesis is rejected and the alternative of a first order Markov property is accepted.

The descriptive structure of the Markov chain may be found by inspection of the corresponding transition tallies observed in the sequence as contrasted with those predicted for independent events. Alternatively, each transition may be tested as an independent chi-square test with one degree of freedom, applying Yates' correction for continuity (q.v.) and observing the minimum expected frequency rule.

The transitions that occur more often than would be expected in an independent events model are:



Passage time statistics may be computed from $\underline{P}$. The matrix of the mean first passage time, $\underline{M}$, contains the mean number of events between the occurrence of one state and another. The matrix, $\underline{W}$, consists of the variances of these passage times. A fundamental matrix, $\underline{Z}$, may be defined as:

$$\underline{Z} = (\underline{I} - \underline{P} + \underline{A})^{-1}$$

Then $$\underline{M} = (\underline{I} - \underline{Z} + \underline{E}\,\underline{Z}_d)\underline{D}$$

and $$\underline{W} = \underline{M}(2\underline{Z}_d\,\underline{D} - \underline{I}) + 2(\underline{ZM} - \underline{E}(\underline{ZM})_d)$$

where $\underline{E}$ is an $m \times m$ matrix of ones;

$\underline{Z}_d$ is the diagonal matrix of Z;

$\underline{D}$ is a diagonal matrix whose elements are the reciprocals of the fixed probability vector;

and m is the number of states.

(Additional details of these computations are given in Kemeny and Snell, 1960).

For the Ashentree bore,

$$\underline{Z} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0.80 & 0.04 & 0.01 & -0.03 & 0.17 \\ 0.32 & 0.92 & -0.10 & -0.09 & -0.05 \\ 0.02 & -0.08 & 0.88 & 0.21 & -0.03 \\ 0.01 & -0.11 & 0.27 & 0.91 & -0.09 \\ 0.14 & 0.21 & -0.12 & -0.10 & 0.87 \end{bmatrix} \end{array}$$

(For an independent events process, the fundamental matrix is an identity matrix).

The mean first passage time matrix,

$$\underline{M} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 3.31 & 9.40 & 3.63 & 4.74 & 4.05 \\ 1.57 & 10.71 & 4.10 & 5.08 & 5.35 \\ 2.57 & 10.68 & 4.20 & 3.53 & 5.26 \\ 2.59 & 10.95 & 2.55 & 5.06 & 5.61 \\ 2.17 & 7.57 & 4.18 & 5.13 & 5.83 \end{bmatrix} \end{array}$$

and the matrix of variances of the first passage time,

$$\underline{W} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 3.18 & 84.15 & 9.85 & 16.86 & 18.24 \\ 1.96 & 84.84 & 9.81 & 16.97 & 18.91 \\ 3.63 & 85.70 & 9.15 & 14.72 & 19.83 \\ 3.73 & 85.66 & 7.49 & 16.03 & 19.65 \\ 2.69 & 78.44 & 10.01 & 17.14 & 19.28 \end{bmatrix} \end{array}$$

Potential Second order Markov properties may be examined by summing transitions between the ith and (i + 2) beds in the sequence as the tally matrix:

	A	B	C	D	E
A	50	21	15	28	6
B	5	6	13	6	6
C	22	3	40	7	22
D	20	1	7	22	28
E	22	6	19	15	5

The corresponding matrix predicted for a first order Markov model is found by squaring $\underline{P}$ and multiplying by the bed row totals:

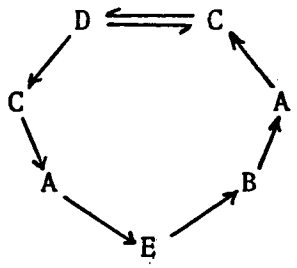
A	50.8	18.6	20.2	24.6	5.8
B	2.9	2.7	10.7	6.8	12.8
C	23.6	8.1	38.1	7.7	16.5
D	18.3	4.6	9.2	27.1	18.8
E	23.8	2.8	15.3	11.6	13.5

A chi-square contingency test of the null hypothesis of a pure first order Markov model gives a calculated χ^2 value of 37.1.

$$\nu = (m-1)^2 = 16$$

The critical value of χ^2 at $\nu = 16$ and $\alpha = 0.05$ is 26.3. The sequence is therefore likely to have additional second order Markov properties.

The description of the composite first and second order Markov model may be made by comparing the totals of transition trios (e.g. ACD) with their expected frequencies for an independent events process, either by inspection, or statistical test where warranted by adequate frequencies. The description of the model is represented as:



which is a pictorial expression of the 'preferred' transitions of lithologies with the Ashentree Bore, taking account of both first and second order Markov properties. There is a strong concordance between this scheme and the conceptual model of the main 'theme' of rocktype ordering in the British Coal Measures proposed by Trueman (1954) and many others.

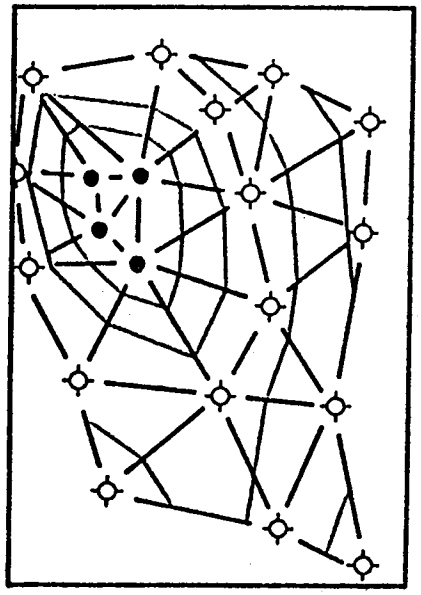
References

Kemeny, J.G., and Snell, J.L., 1960, Finite Markov Chains: D.Van Nostrand, Princeton, 210 p.

Trueman, A.E., 1954, The Coalfields of Great Britain: Edward Arnold, London, 396 p.

28. COMPUTER MAPPING

Programs for drawing contour maps from scattered data points fall into three categories, although these basic procedures are embellished almost endlessly.



(1) Triangulation procedures, which simulate the process of manual contouring. Lines are projected from each data point to the nearest three points, dividing the map area into triangles. The points where contour lines cross the triangles are established by linear interpolation down the sides of the triangles. The final step is to connect points of intersection that have equal value to form the contours. Essentially this process represents the surface as a "geodesic dome" composed of flat triangular plates.

Modifications include fitting curved rather than flat plates to the triangular areas, subdividing the basic triangles into finer subtriangles of similar form, and restricting the manner in which control points are connected so the resulting triangles are as nearly equilateral as possible.

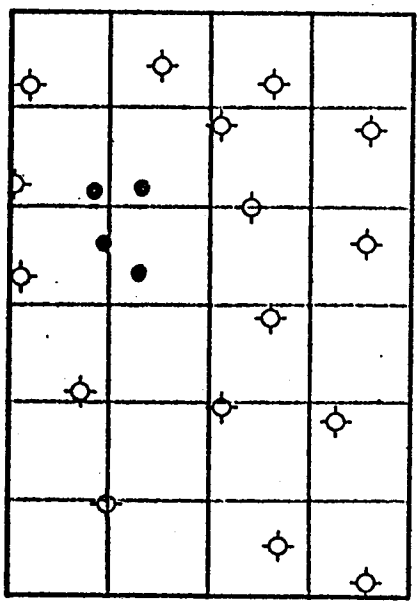
The principal advantages of this procedure are the directness of the methodology and the fact that all control points must lie on the contoured surface. The principal drawbacks are the non-uniqueness of the triangular mesh, which can result in different patterns of contour lines for the same data, and the extreme slowness of the procedure as compared to gridding routines.

(2) Global fit procedures, which fit a complex mathematical function of the geographic coordinates to the control point values. Polynomial trend surfaces and double Fourier surfaces are examples; their computation is discussed in another section.

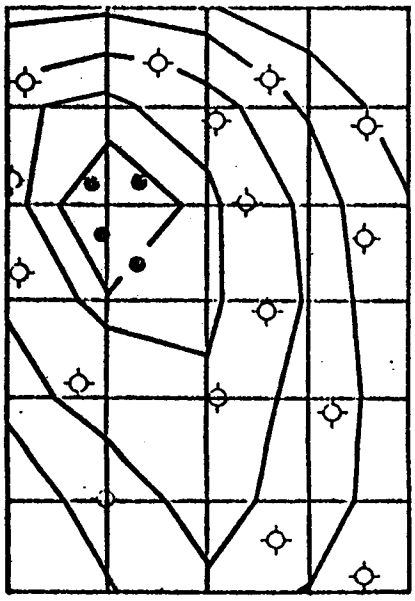
Modifications include two-stage procedures for fitting small "trend

surfaces" to local areas of residuals from the global trend surface.

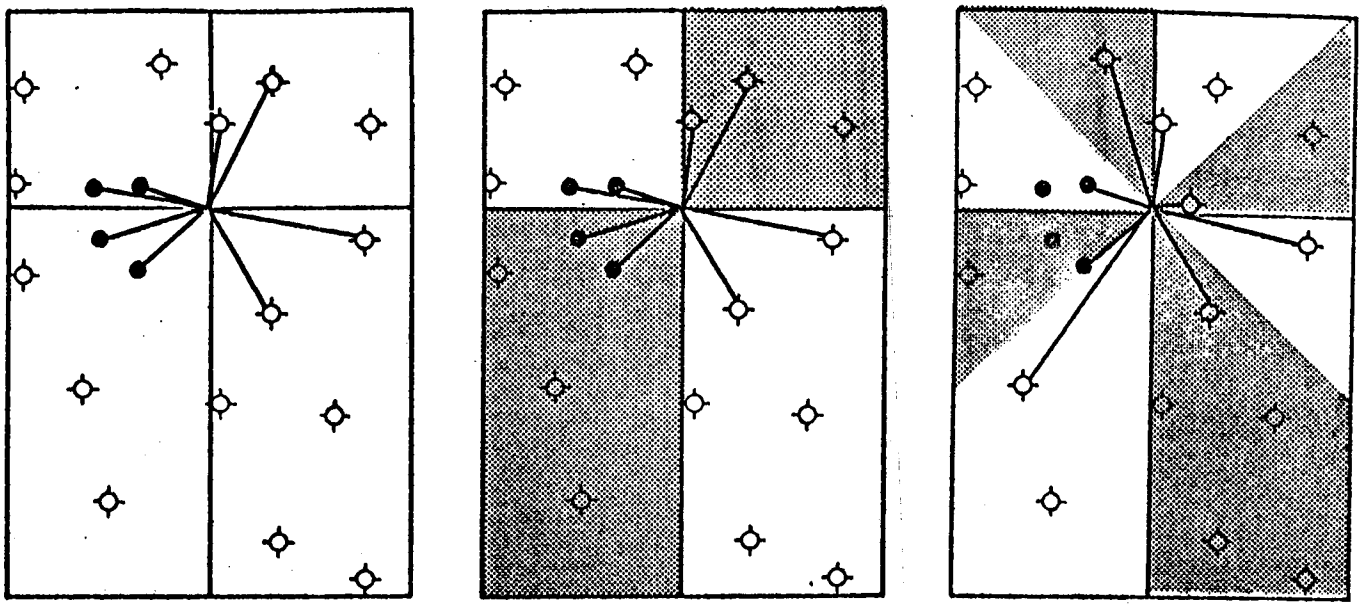
The advantage of global fitting as a contouring procedure is its extreme computational speed. Its disadvantage is that it provides a very poor map of the data, as it is impossible to represent the detail in most mapped variables with any single, tractable equation.



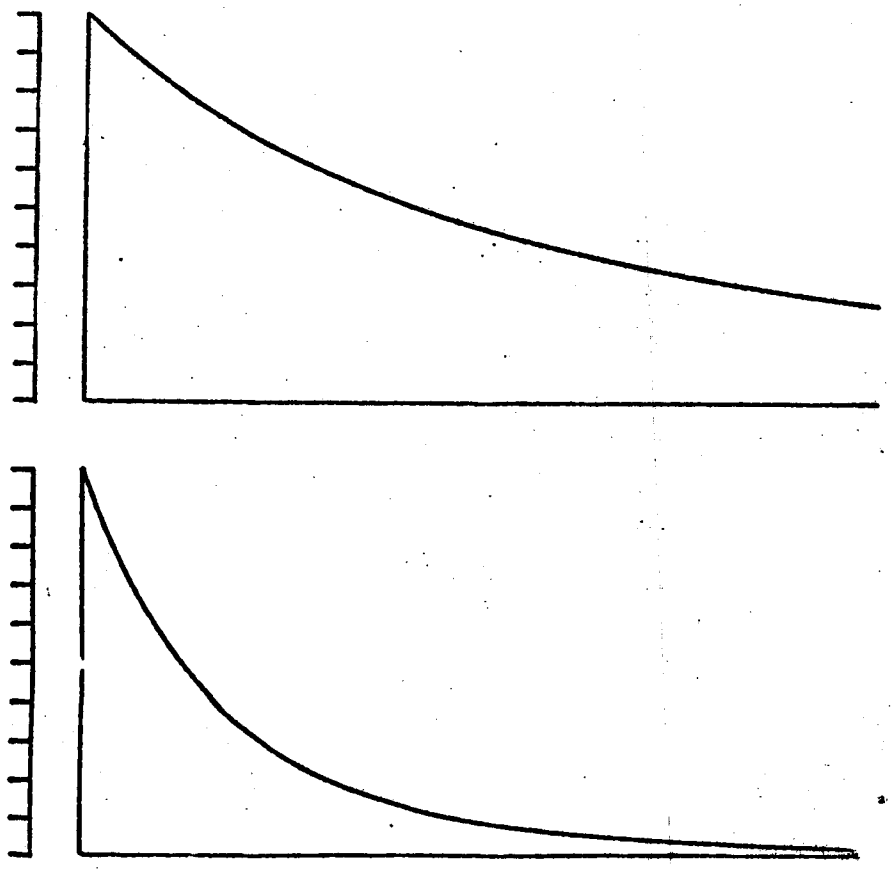
(3) Local fit methods estimate values at the nodes of a regular grid across the map from a weighted average of the control points nearest each grid node. Contours are laced through the gridwork by linear interpolation between the nodes to find the points of intersection of the contour levels with the grid lines. Points of common elevation are then connected to form the contour lines.



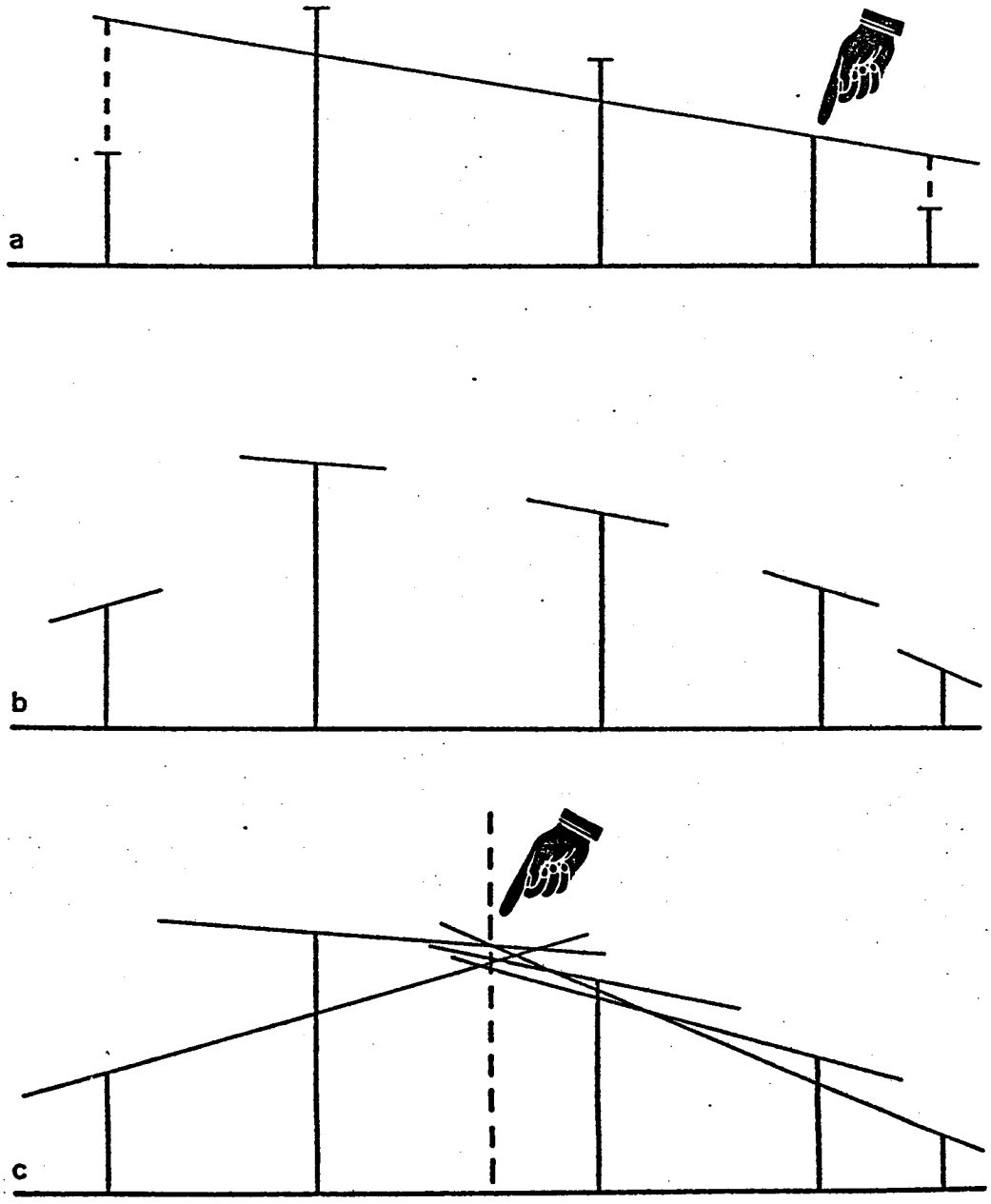
Estimating the regular grid of values is called "gridding" and consists of two steps. First, the nearest neighbors must be found. The simplest procedure is to take the n nearest points to the grid node being estimated. With certain distributions of control points, this may result in unconstrained estimates of the surface, if all the nearest points lie on one side of the node to be estimated. Constraints may be introduced to insure some equitable radial distribution of the nearest neighbors used. These include a quadrant search, where n points must be found in each of four quadrants around the estimated point, and the octant search which carries the concept one step further.



The second step is the estimation of grid values from control points that have been located in the first step. The estimates may simply be weighted averages, where the control points are weighted by a function of their distance from the grid node. The most commonly used functions decline with distance at least as rapidly as $1/D^2$ (top curve) and some of the most effective functions decline at the rate of $1/D^6$. (lower curve).



A more elaborate procedure divides the estimating process into two phases. During the first phase, the dip of the surface at each control point is found by fitting a weighted least-squares plane to the surrounding control points. In the second phase, these dips are projected from the control points that surround a grid node to that location. An estimate of the surface at that node is then made as a weighted average of these projections.



Contouring programs combine different weighting functions, search procedures, and other modifications in great variety. The superiority

of specific combinations is loudly proclaimed by their proponents, but the relative merits of the more elaborate procedures are questionable. Commercial contouring packages usually have the ability to construct block diagrams, isopach maps, and maps of other transforms of the surface.

The primary advantages of the local fit method derive from the intermediate gridding step; this allows storage of the mathematical representation of the surface as an array in the computer. Storage is minimized and the process of drawing contour lines is speeded. Two or more variables can be compared (by isopaching or other methods) even if they are measured at different locations, because the grids, rather than the control points, are compared. However, the gridding step also is the cause of most of the drawbacks of the method, especially the distressing tendency for contour lines to sometimes pass on the wrong side of control points in areas of low dip.

References

Davis, J.C., and McCullagh, M.J., eds., 1975, Display and analysis of spatial data: John Wiley & Sons Inc., New York, 383 p.

Walters, R.F., 1969, Contouring by machine: Bull. Am. Assoc. Pet. Geologists, v. 53, p. 2324-2340.

29. TREND SURFACE ANALYSIS

Trend surface analysis is a mpaping procedure in which a surface is fitted to map observations by least-squares. Trend surfaces are used for two contrasting purposes: (1) to map a variable which is too erratic or too poorly sampled (or both) to be mapped by conventional means. (2) As a method of high-pass spatial filtering to isolate local residuals or "anomolies" which deviate from the fitted surface. Conventional trend surfaces are mathematical equations in which Z, the variable to be mapped, is expressed as a polynomial function of the geographic coordinates of the sample points. The coefficients of the trend equation are found by solution of a series of simultaneous equations so the sum of the squared deviations of the function from the observations is a minimum.

The trend equation can be evaluated at any location to yield Z, the trend, which is a smoothly dipping or undulating surface which represents the average tendency of the observations. The trend is closely related in concept to drift in universal Kriging (q.v.) but is a global function calculated over the entire map rather than over a local neighborhood. The trend will not, in general, coincide with the original $\hat{Z}$ values at the control points leaving a residual or deviation, $Z-\hat{Z} = \varepsilon$.

If the objective of a trend surface analysis is to map the general form of highly variable data, such as geochemical or assay values, the residuals from the fitted trend should be independent of one another and approximately normally distributed. Then, trend surface analysis can be regarded as a form of two-dimensional curvilinear regression (q.v.), and the significance of the fitted model can be tested by analysis of variance.

Trend surfaces are also widely used to separate a geologic variable (usually elevation of a stratigraphic horizon) into two components, the regional trend and local deviations which may be identified with small structural features. The purpose of the analysis is to locate anomolous small patches where the residuals from the trend are autocorrelated, or all of the same sign. Mapping these residuals in effect subtracts the trend component from a map of the original data, thus acting as a high-pass spatial filter. The trend itself usually is of no interest. As the

residuals are not independent (indeed, the purpose of the analysis has been defeated if they are), statistical tests are not appropriate. The correct form of the trend surface equation must be selected on empirical grounds.

Fitting the trend equation:

The general form of a polynomial trend surface equation is

$$\hat{Z} = \alpha_0 + \beta_1 X + \beta_2 Y + \beta_3 X^2 + \beta_4 Y^2 + \beta_5 X Y + \ldots + \beta_p X^m Y^m$$

The first three coefficient define a first degree trend surface, the first six coefficients define a second degree surface, and so on. The first degree surface is a dipping plane, the second degree surface has a dome, basin, or saddle-like shape in which the direction of curvature is constant along any cross-section. Higher orders are increasingly complex because the addition of terms adds degrees of freedom to the equation of the surface and allows more flexibility. The coefficients can be found by solving a series of simultaneous linear equations.

In matrix form, the equations are

$$[X,Y]\ [\beta] = [Z]$$

and the unknown vector of coefficients is equal to

$$[\beta] = [X,Y]^{-1}\ [Z]$$

For a second degree polynomial, the [X,Y] matrix is

$$[X,Y] = \begin{bmatrix} n & \Sigma X & \Sigma Y & \Sigma X^2 & \Sigma Y^2 & \Sigma XY \\ \Sigma X & \Sigma X^2 & \Sigma XY & \Sigma X^3 & \Sigma XY^2 & \Sigma X^2Y \\ \Sigma Y & \Sigma XY & \Sigma Y^2 & \Sigma X^2Y & \Sigma Y^3 & \Sigma XY^2 \\ \Sigma X^2 & \Sigma X^3 & \Sigma X^2Y & \Sigma X^4 & \Sigma X^2Y^2 & \Sigma X^3Y \\ \Sigma Y^2 & \Sigma XY^2 & \Sigma Y^3 & \Sigma X^2Y^2 & \Sigma Y^4 & \Sigma XY^3 \\ \Sigma XY & \Sigma X^2Y & \Sigma XY^2 & \Sigma X^3Y & \Sigma X^3Y & \Sigma X^2Y^2 \end{bmatrix}$$

Note that elements in the first row and in the first column are sums corresponding to the right-hand side of the second-degree trend equation, which is

$$\hat{Z} = \alpha_0 + \beta_1 X_1 + \beta_2 Y + \beta_3 X^2 + \beta_4 Y^2 + \beta_5 XY$$

The remaining elements are sums corresponding to cross-products of the first row and column. In general, if the trend-equation contains p

coefficients, this matrix will be $(P+1) \times (p+1)$. The vector $[\beta]$ is the $(p+1) \times 1$ array of unknown coefficients, and $[Z]$ is a $(p+1) \times 1$ vector of sums of Z and its cross-products with X and Y. For a second degree polynomial these two vectors are

$$\begin{bmatrix} \alpha_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} \qquad \begin{bmatrix} \Sigma Z \\ \Sigma XZ \\ \Sigma YZ \\ \Sigma X^2 Z \\ \Sigma Y^2 Z \\ \Sigma XYZ \end{bmatrix}$$

The matrix equation is solved by inverting [X,Y] and post-multiplying by [Z] to yield $[\beta]$. The lower right elements of [X,Y] are high powers of the geographic coordinates; if the coordinate values are themselves large numbers, these elements may become enormous. Because these elements may be used as divisors during the inversion process, other elements may be reduced to the vanishing level. The result is a succession of exponential underflows and overflows referred to as "matrix blow-up." This situation usually can be avoided by scaling the geographic coordinates.

Evaluating a fitted trend

Once the coefficients of the trend equation have been determined, the equation can be evaluated at the location of each data point to yield the trend values $\hat{Z}$. The differences between the trend value and the Z value at each point can be squared and summed for all the observations to give the sum of squares of deviations about the trend (or sum of squares of the residuals):

$$SS_D = \sum_{i=1}^{n} (Z_i - \hat{Z}_i)^2 = \sum_{i=1}^{n} \varepsilon_1^2$$

(In this notation Z_i refers to the value of Z for the ith data point, located at coordinates X_i and Y_i). Variation in the trend surface itself is given by the sum of squares due to regression:

$$SS_R = \sum_{i=1}^{n} \hat{Z}^2 - \frac{(\sum_{i=1}^{n} Z)^2}{n}$$

The total variation in the original data may be expressed as a sum of squares:

$$SS_T = \sum_{i=1}^{n} Z^2 - \frac{(\sum_{i=1}^{n} Z)^2}{n}$$

The three sums of squares are related by

$$SS_T = SS_R + SS_D$$

The goodness-of-fit of the trend surface, designated R^2, is the proportion of the total sum of squares accounted for by the sum of squares due to regression, or

$$R^2 = \frac{SS_R}{SS_T}$$

The square root of R^2 is the multiple correlation coefficient, R. The total number of degrees of freedom in a trend surface problem is n-1. Division of SS_T ny n-1 gives the variance of the original data. The number of degrees of freedom associated with the trend is a function of the number of coefficients in the trend surface equation, as these coefficients alone determine the shape of the fitted surface. If these are p coefficients there will be p-1 degrees of freedom for the trend, and the variance of the trend surface is SS_R divided by p-1. The residuals have n-p degrees of freedom, so the variance around the trend can be found by dividing SS_D by n-p. The statistical significance of a trend surface may be tested by computing the F-ratio between the variance of the trend and the variance around the trend or

$$F = \frac{SS_R/p-1}{SS_D/n-p} \quad \text{with p-1 and n-p degrees of freedom.}$$

Such tests, however, are valid only under specific assumptions of independence and normality of the deviations. In many instances, the appropriate order of trend surface must be determined empirically.

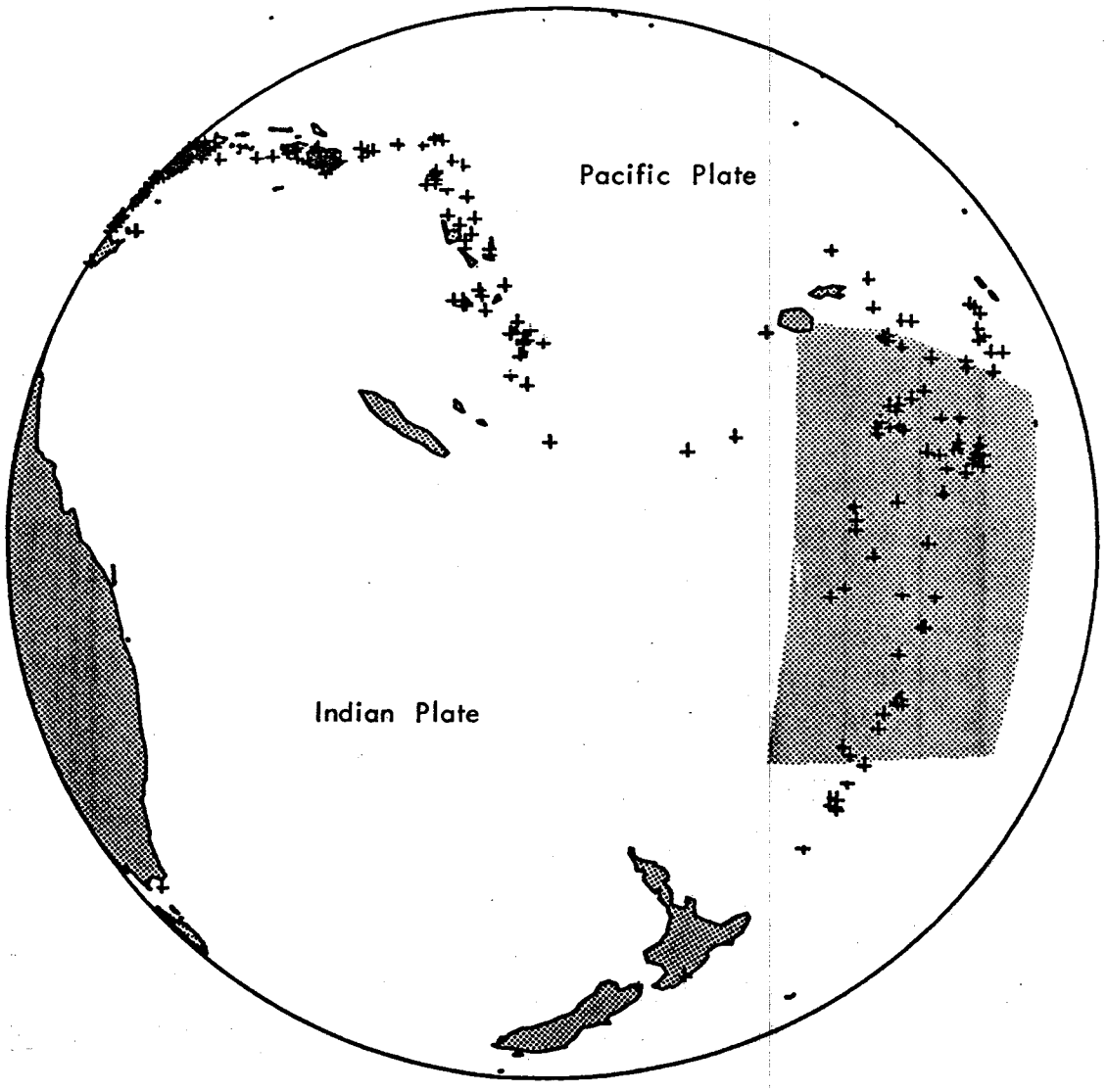
Example: Mapping a Benioff zone

Within the last decade it has been realized that the majority of the world's seismic activity is concentrated around plate margins and is caused by differential movements between adjacent crustal plates. A plot of the locations of earthquakes epicenters over even a brief time-span is sufficient to indicate the surface trace of plate boundaries. The illustration shows epicenters in the Pacific region for the period June-August, 1972. The epicenters loosely define the boundary between two major plates which form a "destructive" margin, where the Pacific plate is moving against the Indian plate and slipping underneath it at an angle of about 45°. The subduction of the Pacific plate marked physiographically by the deep ocean trenches which flank island areas (which are themselves largely the product of volcanic activity associated with subduction).

Earthquake foci along the plate margin define a surface of subductive movement, called a "benioff zone" (named for the pioneer researcher, Hugo Benioff). Along relatively linear parts of the plate margin, the Benioff zone is a simple dipping plane. One such linear area includes the Tonga and Kermadec trenches, north of New Zealand (shaded on the globe), and is marked by intense seismic activity.

Benioff zones were discovered by study of the location of destructive margin earthquake foci and are among the most convincing pieces of evidence for the new global tectonics. In the Tonga-Kermadec region, the reality of a Benioff zone may be checked by a mathematical model of the zone as a linear trend surface fitted to earthquake hypocenters. Those shown on the illustration are listed in the table, drawn from the National Earthquake Information Center's monthly publication "Preliminary Determination of Epicenters."

Since degrees of latitude and longitude are approximately equal at these low latitudes, a trend surface may be computed directly for the data without transforming the geographic scales. (A degree latitude or longitude in this area is approximately 105 km). The fitted linear trend surface is



Plot of earthquake epicenters (shown by crosses) in the Pacific hemisphere during June, July, and August, 1972. Shaded inset area is the Tonga-Kermadec trench region.

EARTHQUAKE HYPOCENTERS IN THE TONGA-KERMADEC

TRENCH AREA (17.5°S-32.5°S; 177.5°E-170°W)

JUNE-AUGUST, 1972

Day	Latitude S	Longitude W	Depth (km.)
June			
1	21.1	174.4	N*
5	20.8	178.3	551
7	31.9	178.2	N
8	19.4	177.1	290
9	21.1	174.0	N
10	21.2	174.3	62
11	20.0	175.3	N
11	20.9	175.4	75
13	29.7	177.3	46
15	17.7	174.7	N
15	29.8	177.0	N
16	17.7	178.6	570
17	21.4	174.4	41
17	31.5	179.0	72
21	18.0	174.7	132
22	30.1	177.8	62
23	20.3	178.4	595
25	24.0	180.0	515
25	24.3	180.0	464
26	26.1	180.3	487
29	20.1	178.3	608
30	17.9	178.7	600
July			
1	21.9	175.0	41
2	20.8	174.2	N
3	21.0	179.3	579
4	31.2	179.4	229
7	32.5	178.8	N
9	20.2	176.3	278
10	22.8	176.2	N
16	20.9	178.7	600
16	21.2	175.7	N
21	21.9	176.0	118
24	20.8	179.1	643
25	27.2	176.5	62
25	27.2	176.7	60
25	21.5	176.4	180
26	23.6	180.1	552
27	20.9	178.1	496
30	18.0	178.0	525
31	20.2	178.7	633

Day	Latitude S	Longitude W	Depth (km.)
August			
2	18.9	173.3	60
2	26.2	177.8	224
3	28.1	177.6	232
5	17.9	178.6	619
5	29.8	177.2	63
8	26.4	180.8	533
9	23.3	178.3	323
10	21.5	174.6	N
11	25.1	179.2	390
12	24.5	176.8	128
13	29.6	177.1	39
14	17.9	178.6	622
15	17.8	178.8	573
15	17.7	173.0	N
19	21.5	177.0	318
26	30.6	177.9	118
26	21.5	174.0	N
26	26.2	176.2	48
28	18.1	176.6	N
28	21.3	174.3	N
29	20.0	175.3	154
30	19.8	177.7	562
31	21.2	179.2	625

*N = probable shallow focus earthquakes where depths are not precisely defined by the data.

Source of data

Preliminary Determination of Epicenters (Monthly listings)
U.S. Govt. Printing Office, Washington D.C.

shown in the following figure and has the equation:

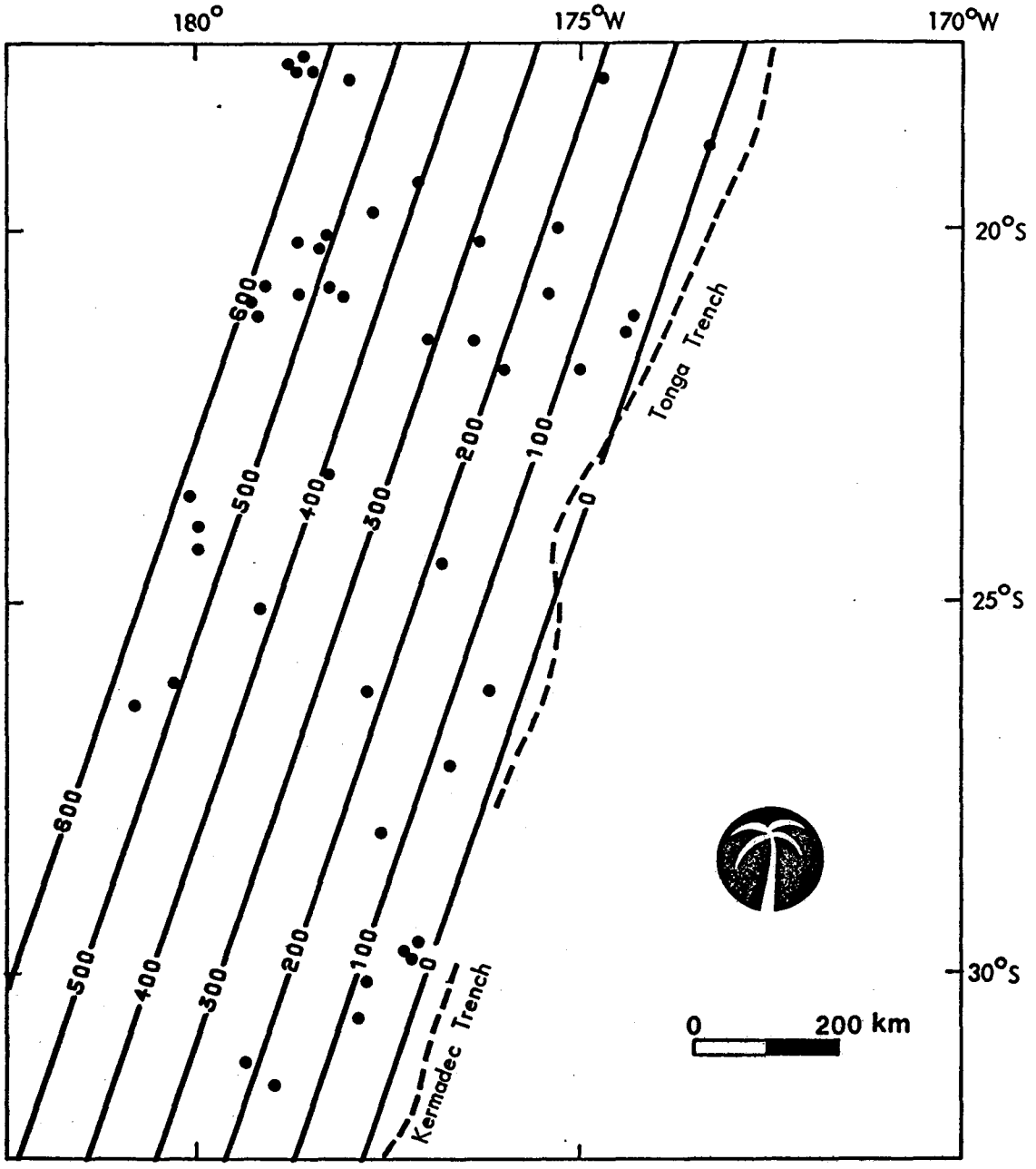
$$H = -18552 + 37.0T - 111.1G$$

where H = depth of earthquake focus (km.)
T = latitude of hypocenter (degrees south)
G = longitude of hypocenter (degrees west)

(Latitude and longitude locations were entered in the regression as negative values).

The goodness-of-fit of the trend surface is

$$R^2 = \frac{SS_R}{SS_R} = 0.94$$



Linear trend surface of earthquake hypocenters for June, July and August, 1972, in the Tonga-Kermadec trench region (depth shown in kilometers).

and the multiple correlation coefficient, R = 0.97.

It seems reasonable to assume that earthquakes observed in the region for the arbitrary three-month period are a random sample. Therefore the significance of the fit of the trend surface to the epicenters can be tested by an analysis of variance procedure. The ANOVA table below contains statistics calculated from the epicenter data.

Source of Variation	Sum of Squares	Degrees of freedom	Mean Squares	F
Trend (Regression)	234.5×10^4	2	117.3×10^4	360
Deviation from trend	14.7×10^4	45	3257	
Total variation	249.2×10^4	47		

The column labeled "mean squares' contains the sums of squares divided by their degrees of freedom. Mean squares are simply estimates of variances; the mean square of the regression (MS_R) is an estimate of the variance in the trend, and the mean square deviation (MS_D) is an estimate of the variance around the trend. If there is no regression (that is, if both slopes β_1 and β_2 are zero), these two variances will approximately be the same, and the F-ratio will be near 1. In this example, the F-ratio is highly significant, so the model of the Benioff zone as a dipping plane seems reasonable.

A confidence band around the estimate of the Benioff zone can be found by calculating the standard errors of the trend coefficients β_1 and β_2. These are

$$s_{e1} = \sqrt{\frac{MS_D}{SS_X}} \qquad s_{e2} = \sqrt{\frac{MS_D}{SS_Y}}$$

Here, SS_X and SS_Y are the sums of squares of the independent variables:

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 801.5$$

$$SS_Y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 137.6$$

The standard errors of the coefficients are

$$s_{e1} = \sqrt{\frac{3257}{801.5}} = 2.0 \quad , \quad s_{e2} = \sqrt{\frac{3247}{137.6}} = 4.9$$

Approximate 95% confidence limits about the coefficients can be calculated as $\mp$ 2 standard errors.

$$\beta_1 \mp 2\, s_{e1} = 37 \mp 4.0 = 41;\ 33$$

$$\beta_2 \mp 2\, s_{e2} = 111 + 9.8 = 120.8;\ 101.2$$

The two trend surface coefficients give the apparent dip of the Benioff zone in the north-sourth and east-west directions. The true dip is the vector resultant of these two apparent dips, or

$$\text{true dip} = \sqrt{\beta_1^2 + \beta_2^2} = \sqrt{(37)^2 + (-111)^2} = 117$$

which is expressed in km/degree, the original units of measurement. As degree of latitude or longitude in this region is approximately 105 km, the dip in degrees is found as

$$\text{dip}^\circ = \arctan\ (\text{true dip}/105) = 48^\circ$$

Similarily, the upper and lower confidence limits on the coefficients can be included to find the range of probable dips for the zone:

steepest dip = $\sqrt{(37 + 4)^2 + (111 + 9.8)^2} = 127.6$ km/degrees = 50.5°

Shallowest dip = $\sqrt{(37-4)^2 + (111 - 9.8)^2} = 106.4$ km/degree - 45.4°

This example is adapted directly from a student exercise described by Shea (1973), who also discusses general assumptions of the model and includes a bibliography of relevant geological literature.

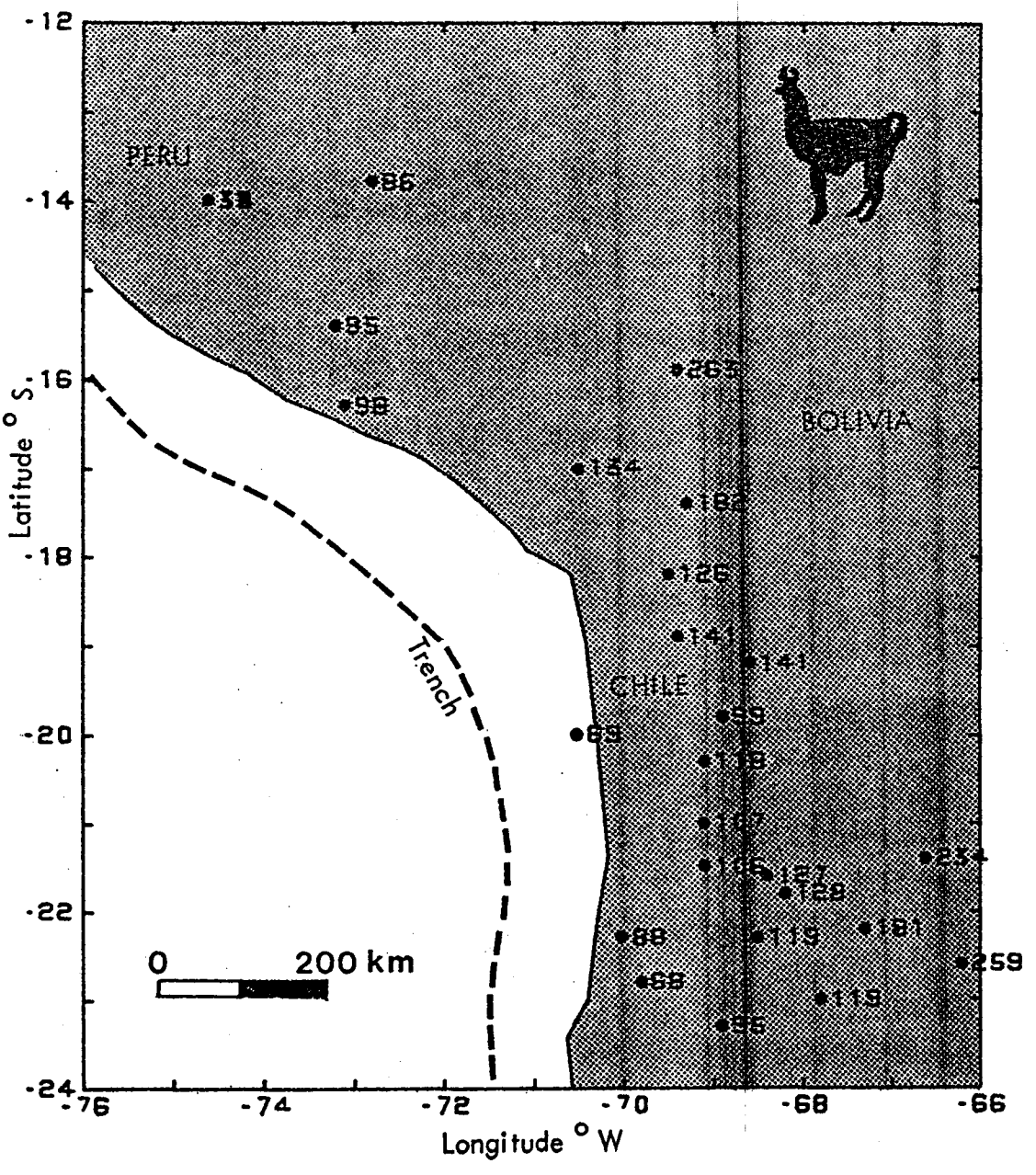
References

Preliminary Determination of Epicenters (Monthly listings for June, July, and August, 1972) U.S. Govt. Printing Office, Washington D.C.

Shea, J.H., 1973, Treatment of earthquake hypocenter data with a programmable calculator: Jour. Geol. Education, v. 21, p.29-34.

Exercise 29.1: Plate tectonics in the central Andes

On the eastern side of the Pacific, the Pacific plate underthrusts the South American plate and there is frequent earthquake activity in the Andes range of Chile, Bolivia and Peru. Earthquake hypocenters for the period June-August, 1972, are tabulated for the region bounded by longitude parallels of 66°W and 76°W and latitude parallels 12°S and 24°S.



EARTHQUAKE HYPOCENTERS IN THE CENTRAL ANDES

June - August 1972

Latitude ^{o}S	Longitude ^{o}W	Depth (km.)
-17.4	-69.3	182
-22.6	-66.2	259
-13.8	-72.8	86
-22.3	-70.0	88
-25.4	-69.2	89
-18.2	-69.5	126
-22.2	-67.3	181
-21.8	-68.2	128
=23.3	-68.9	95
-21.5	-69.1	106
-21.4	-66.6	234
-20.3	-69.1	118
-20.0	-70.5	69
-22.3	-68.5	119
-14.0	-74.6	39
-19.8	-68.9	99
-16.3	-73.1	98
-18.9	-69.4	141
-22.8	-69.8	68
-17.0	-70.5	134
-19.2	-68.6	141
-21.0	-69.1	107
-15.4	-73.2	95
-14.0	-74.6	14
-23.0	-67.8	119
-21.6	-68.4	127
-15.9	-69.4	263

(1) Using latitude and longitude as X and Y coordinates, compute and map first, second and third order trend surfaces of the earthquake hypocenters.

(2) What are the goodness-of-fit and multiple correlation coefficients of these surfaces?

(3) Compile an analysis of variance table for the first order regression surface. Make an F-test of the null hypothesis that the plane has no statistical significance.

(4) Assuming one degree latitude or longitude is 105 km., what is the true dip of the linear surface? Compare the value with that of the Tonga-Kermadec Benioff zone and speculate on any difference in terms of geological models.

Reference

James, D.E., 1971, Plate tectonic model for the evolution of the central Andes: Geol. Soc. American Bull., v. 82, p.3325-3346.

31. EIGENVECTORS AND EIGENVALUES

Consider a sample of observations in which measurements of variable X and Y have been made and which are bivariate normally distributed. The observations may be standardized to units of standard deviation of X and Y, referenced with respect to their means by the transformation:

$$x_i = \frac{X_i - X}{s_X} \quad \text{and} \quad y_i = \frac{Y_i - Y}{s_Y}$$

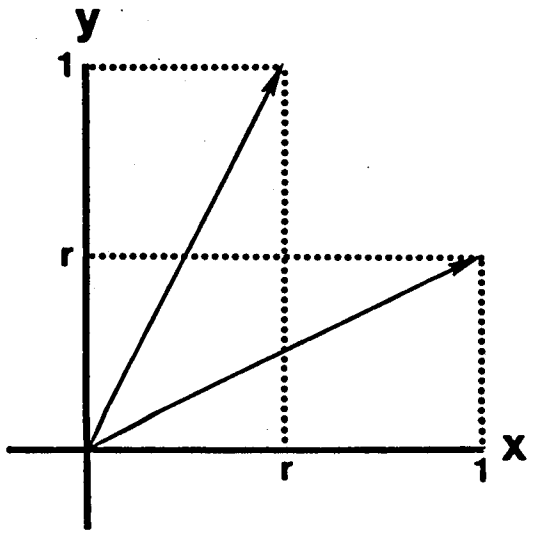
If the sample is plotted as points on an x-y orthogonal coordinate system, it will take the appearance of a cloud whose centroid is the bivariate mean, which has been transformed to the origin. The 'spread' of the cloud is measured by the variances of X and Y (which have both been standardized to one) and the covariance of X and Y, which has been transformed to the correlation coefficient, r, since:

$$r = \frac{\text{cov}(X,Y)}{s_X\, s_Y} = \text{cov}(x,y)$$

The standardized variance-covariance matrix is:

$$\underline{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

which is a correlation matrix. The matrix consists of two row vectors which can be geometrically represented as:



If r is zero, the cloud appears circular in shape and the two vectors are at right angles. If r is one, the cloud is compressed to a diagonal line and the two vectors are coincident. At an intermediate value of r, the cloud takes on an elliptical shape. The variance of x measures the x^2 component of the cloud since $s_x = (\Sigma x^2)/n-1$. Similarly, the variance

of y measures the y^2 component, and the covariance of x and y (r), and xy component. If these sources of variation are summed as a quantity of total variation, s^2:

$$\text{var}(x)\ x^2 + 2\ \text{cov}(x,y)\ xy + \text{var}(y)\ y^2 = s^2$$

which simplifies to

$$x^2 + 2r\ xy + y^2 = s^2$$

This is the equation of an ellipse. If r is zero, the equation becomes:

$$x^2 + y^2 = s^2$$

which is the equation of a circle.

The equation may be written in matrix form as

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = s^2$$

or $\underline{X}'\underline{R}\,\underline{X} = s^2$

An ellipse is symmetrical about two principal axes which are orthogonal and can be specified by vectors. The major axis of the ellipse expresses the major component of variation in the cloud, while the remaining variation is contained in the minor axis. Since the two axes are orthogonal, these two components are uncorrelated. If the reference system could be rotated so that the major and minor axes form the new reference axes, the ellipse could be rewritten as:

$$\lambda_1 u^2 + \lambda_2\ v^2 = s^2$$

where λ_1 and λ_2 are the lengths of the semi-major and semi-minor axes, and u and v are units expressed in terms of the new reference axes. (When the ellipse principal axes coincide with the reference axes, the cross-product term vanishes). In matrix form the equation is

$$\begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = s^2$$

The geometric rotation of axes has been accompanied by a change in the descriptive matrix. The correlation matrix has been transformed to a diagonal matrix, which is known as the canonical form of the original

matrix. The values of λ are known as the eigenvalues of the original matrix and are a measure of the principal components of variation. The vectors that specify the principal axes are called eigenvectors.

It should be noted that when r = 0, the original matrix is already in canonical form. The principal components of variation are then the variances of x and y which comprise the total variation. The eigenvectors coincide with the variable axes which are uncorrelated, by definition.

Eigenvectors and eigenvalues may be computed for any square matrix, although these quantities sometimes have imaginary values. In almost all multivariate statistical applications, analysis is directed to the variance-covariance matrices (or their standardized form) in order to systematically partition components of variation. These matrices are symmetrical and can be represented geometrically by an elliptical function, which has tangible principal axes and, consequently, real-number eigenvectors and eigenvalues.

The process of deriving the eigenvectors and eigenvalues from a variance-covariance matrix may be achieved by application of the following consideration:

(1) Geometric representation of the matrix, R

The multiplication of a vector by a matrix represents the transformation of the vector to a new vector

$$\text{e.g. } \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 7 \\ 4 \end{bmatrix}$$

which, geometrically, is a change in magnitude and a rotation. However, if a matrix is applied to the transformation of one of its eigenvectors, the resultant vector maintains the same direction, but its magnitude is increased by a proportion which is the eigenvalue associated with that eigenvector. So, for example:

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

where [1 1] is an eigenvector of the matrix with an eigenvalue of 3. In geometrical terms, a principal axis vector may be 'stretched' but not rotated. The relationship may be expressed as:

$$\underline{R}\,\underline{X} = \lambda\,\underline{X}$$

and used for solution of eigenvalues (λ) and eigenvectors ($\underline{X}$) of the matrix $\underline{R}$.

(2) Solutions of eigenvalues and eigenvectors by use of the determinant

$$\text{If } \underline{R}\,\underline{X} = \lambda\underline{X}$$
$$\text{then } (\underline{R} - \lambda\underline{I})\underline{X} = 0$$

which represents m simultaneous equations of the form:

$$(r_{11} - \lambda)\, x_1 + \ldots\ldots\ldots r_{1m}x_m = 0$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$r_{m1}x_1 + \ldots\ldots\ldots\ldots (r_{mm}-\lambda)x_m = 0$$

This is a system of homogeneous equations as there are no isolated scalars for a non-trivial unique solution of $x_1 \ldots z_m$. One immediate solution for X is a vector of zeroes which is termed a trivial solution. In order for a non-trivial solution to exist, (R - I) must be a singular matrix which by definition has a determinant of zero. This condition can be demonstrated by an example where:

$$\begin{bmatrix} 4 & 2 \\ 2 & 8 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

In this case, the only solution for x_1 and x_2 is trivial --they are zero. However, if

$$\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

non-trivial solutions for x_1 and x_2 are possible. As a general case, if the matrix takes the form

$$\underline{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

a non-trivial solution exists if

$$ad - bc = 0$$

The quantity (ad-bc) is the determinant of $\underline{A}$ which is written as $|A|$. The singular matrix condition applies to higher order matrices representing homogeneous equations, although computation of the determinant involves more complex expressions.

Returning to the original matrix expression, it follows that:

$$|\underline{R} - \lambda \underline{I}| = 0$$

If R is a 2 × 2 matrix then

$$\begin{vmatrix} r_{11} - \lambda & r_{12} \\ r_{21} & r_{22} - \lambda \end{vmatrix} = 0$$

and so

$$\lambda^2 + (-r_{22} - r_{11})\,\lambda + r_{11}r_{22} - r_{12}r_{21} = 0$$

This is a quadratic expression which may be easily solved by the high school algebra solution:

$$\lambda = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

where $a\lambda^2 + b\lambda + c = 0$

The two values λ_1 and λ_2 are the eigenvalues of the matrix $\underline{R}$.

If λ_1 is substituted in the matrix equation:

$$(\underline{R} - \lambda \underline{I})\underline{X} = 0$$

two simultaneous equations result:

$$(r_{11} - \lambda_1)x_1 + r_{12}\,x_2 = 0$$

$$r_{21}\,x_1 + (r_{22}-\lambda_1)x_2 = 0$$

which may be solved for the vector $\underline{X}_1$ which is the eigenvector associated with the eigenvalue, λ_1. Since the equations are homogeneous, the eigenvector will have direction, but no magnitude i.e. there are a variety of solutions for x_1 and x_2, but they all lie on the same axis in $x_1 - x_2$ space. (x_1 and x_2 represent the original variables which form the reference axes of the matrix $\underline{R}$.)

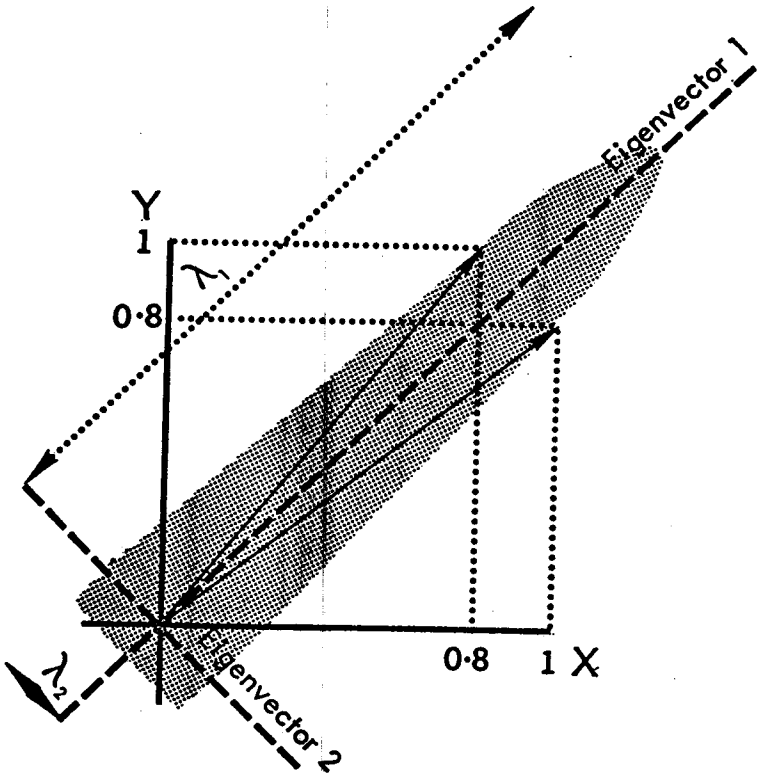
Similarly, substitution of λ_2 in the matrix equation leads to the computation of a second eigenvector which is orthogonal to the first.

The determinant procedure may be applied to higher order matrices for the computation of eigenvalues and eigenvectors in a similar method to that outlined, although the determinant computations are more involved. Computer programs normally make use of an alternative procedure such as the Jacobi method (q.v.).

Numerical example of computations for a two variable data set

A correlation coefficient between two variables, X and Y, is computed to be 0.8. The correlation matrix, $\underline{R}$ is then:

$$\underline{R} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



Then

$$|\underline{R} - \lambda\underline{I}| = 0$$

and so

$$\lambda^2 - 2\lambda + 0.36 = 0$$

$$\lambda_1 = 1.8 \text{ and}$$

$$\lambda_2 = 0.2$$

$$(\underline{R} - \lambda\underline{I})\underline{X} = 0$$

Using λ_1:

$$-0.8X' + 0.8Y' = 0$$

$$0.8X' + 0.8Y' = 0$$

The first eigenvector is:

$$g\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

where g is any scalar (i.e. the solution specifies direction, but not magnitude).

Using λ_2, the second eigenvector is:

$$g\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Eigenvalues and eigenvectors of variance-covariance matrices-the general case

A variance-covariance (or correlation matrix) of m variables describes an m-dimensional ellipsoid in a space where the measurement variables form orthogonal reference axes. If the eigenvalues of the matrix are arranged in order from greatest to smallest, the corresponding eigenvectors specify the principal axes ordered in terms of their lengths. The first eigenvector

therefore defines the direction in m-dimensional space along which the composite variance is at a maximum, while the remaining eigenvectors account for remaining orthogonal components of variance. Since an m X m matrix has m eigenvectors, a new frame of reference has been created where sample data points may be related to m orthogonal axes of maximum variance which are uncorrelated, rather than the original axes of the m variables.

Although there will be solutions for m eigenvalues, some of these may be zero. If there are m' zero eigenvalues, the total variation is expressed in $(m-m')$ dimensions, with a lower dimensionality than that implied by the raw variables. This is the case, for example, in closed variable systems such as percentage data where each set of m measurements sums to 100%. If the sample is plotted in m space, the cloud of points is inevitably constrained to an (m-1) dimensional form.

The inherent dimensionality of variation that is implied by a variance-covariance matrix can be analysed by use of the fact that eigenvalues collectively sum to the trace of the original matrix (the sum of the leading diagonal values). If the matrix is a correlation matrix, the sum of the eigenvalues is m, the number of variables. This relation is used in an accounting procedure to specify the proportion of the total variance accounted for by each eigenvector. In the two-variable numerical example:

$$\lambda_1 = 1.8 \qquad \lambda_2 = 0.2$$

and so the first eigenvector accounts for 90% of the total variation. It follows that if the original data is projected onto the axis of the first eigenvector as a series of scores, these values will contain 90% of the total information on variation expressed in the original two dimensions. This property is used extensively in the techniques of principal components (q.v.) and factor analysis (q.v.) as a means of condensing a large number of variables to linear scales of scores and two-dimensional mappings which contain the highest proportions of variance in the original m dimensions. This approach is true to the basic scientific philosophy of parsimony which directs that analysis should strive to reduce a model to the smallest number of explanatory variables and minimize redundancy in observational data.

The Jacobi method: A computer algorithm for eigenvalue solution

The determinant solution for eigenvalues is simple to apply to low order matrices but becomes increasingly cumbersome for matrices of higher order. The most widely used algorithm employed in computer programs for solutions of symmetric matrices is the Jacobi method.

It will be recalled that if the original axes of a symmetric matrix $\underline{R}$ can be rotated to the principal axes, $\underline{R}$ will have been transformed to a diagonal matrix, $\underline{L}$, which is its canonical form. The elements of $\underline{L}$ are the eigenvalues. The process of rotation is achieved by the operation of a suitable matrix which, for a 2 X 2 matrix, is:

$$\underline{Q} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

where θ is the angle of rotation, then:

$$\underline{Q}'\underline{R}\,\underline{Q} = \underline{L}$$

from which the zero elements of $\underline{L}$ are equivalent to:

$$(r_{11}-r_{22})\ \sin\theta\ \cos\theta + r_{12}\ (\cos^2\theta-\sin^2\theta)$$

Solving for θ,

$$\tan 2\theta = \frac{2r_{12}}{r_{22}-r_{11}}$$

Substitution of θ in the matrix equation leads to a solution of the eigenvalues, from which the eigenvectors are computed.

For higher order matrices, this approach is applied in an iterative manner, to successively transform the off-diagonal elements of $\underline{R}$ to zero values. On the first pass, the pair of largest off-diagonal elements of $\underline{R}$ are located, r_{jk} and r_{kj}. A matrix $\underline{Q}$ is assembled where all the off-diagonal elements are zeroes with the exception of q_{jk} and q_{kj}, which are made $\sin\theta$ and $-\sin\theta$; all the diagonal elements are ones with the exception of q_{jj} and q_{kk} which are both made $\cos\theta$. θ is selected to satisfy the equation:

$$\tan 2\theta = \frac{2r_{jk}}{r_{kk} - r_{jj}}$$

The transformation $\underline{Q}'\underline{R}\,\underline{Q}$ leaves all the rows and columns of $\underline{R}$ unaltered, with the exception of those of j and k. However, r_{jk} and r_{kj} are transformed to zero.

The procedure is repeated, this time applied to the transformed matrix. The iterative process is continued until all of the off-diagonal elements of the transformed matrix are zero (or approximately so). The Jacobi method may be implemented as a computer algorithm and has a 'bootstrap' style of operation that is remniscent of matrix inversion algorithms such as the Guass-Jordan method (q.v.).

32. PRINCIPAL COMPONENTS ANALYSIS

"Geometry, it has been said, is not true (or false), but convenient. Considered as a geometrical method, principal component analysis is not true or false, meaningful or meaningless, but simply a possible, and often convenient, transformation."

-K. Hope

Principal components are linear transformations of a set of m original variables into m new variables which are orthogonal or uncorrelated. The total variance in the original set of variables is defined as the sum of their m variances; all of this variation is contained in the m new variables defined by the principal components. However, one or more of the new variables may exhibit so little variation that they become constants; such variables in effect vanish. Other transformed variables may have such restricted variation that they are judged to make a negligible contribution to the total variance and may be discarded. Thus, principal components analysis may be used to reduce a set of m intercorrelated raw variables into $p \leq m$ independent variables.

The coefficients for each of the linear transformations are of the form:

$$(\underline{S} - \lambda_p \underline{I})a_p = 0$$

where $\underline{S}$ is the $m \times m$ matrix of variances and covariances of the original variables, λ_p is the p-th eigenvalue of this matrix, $\underline{I}$ is an identity matrix of size m, and a_p is the $1 \times m$ vector of coefficients of the p-th principal component. The coefficients, therefore, are the successive eigenvectors of $\underline{S}$.

An original observation X_{ij} can be transformed in Y_{ip}, a score on the p-th principal component, by

$$Y_{ip} = a_{1p} X_{i1} + a_{2p} X_{i2} + \ldots + a_{mp} X_{im}$$

Here, a_{jp} is the j coefficient of the p-th principal component, and X_{ij} is the i-th observation of variable j. The variance of the p-th principal component is equal to the p-th eigenvalue, as the sum of

the eigenvalues is equal to the total variance of the original variables. The contribution of a specific original variable to a principal component is given by

$$r_{jp} = a_{jp}\sqrt{\lambda_p / s_j}$$

which is the correlation of variable j with principal component p.

Objectives of Principal Components Analysis

Principal components are used for two different purposes; to achieve independent variables for use in multiple regression and discriminant analysis, and for "factor analysis." Most geologic studies labeled as "factor analyses" are actually principal components analyses, and much confusion exists as to the distinction between the two. Factor analysis is a statistical hypothesis testing procedure, requiring *a priori* assumptions about the number and probability distributions of the latent factors. Principal components analysis is simply the mathematical transformation of a set of variables into orthogonal form. Certain components may vanish because of mathematical relationships among the original variables (correlation matrices from closed data sets, for example, will be reduced to their true rank which is equal to the number of non-zero eigenvalues), but the discarding of "insignificant" components is done on an empirical basis.

Where principal components are computed as an end unto themselves, the results are often presented in two types of diagrams. These include plots of the scores of observations on successive pairs of principal components, and plots of the correlations of the original variables on the components. The first is useful for identifying clusters or inhomogeneities in the data. The second is useful for *a posteriori* interpretation of the "meaning" of the components.

"Reading" components

Most algorithms for finding the eigenvalues and eigenvectors of symmetrical matrices are variants of the Jacobi method (see matrix algorithms). These return the eigenvectors in order of the magnitude of their associated eigenvalues. It commonly happens that all coefficients of the first eigenvector in a principal components analysis will be positive in sign; thus the first principal component may be a measure of magnitude.

If the raw variables consist of measurements on fossils, for example, scores on the first principal component will be ranked in order of the overall size of the organisms. If the raw measurements are trace elements, the first component may reflect variation in total concentration.

The remaining eigenvectors must contain elements of both positive and negative signs if the first component is entirely composed of positive elements. In many studies, including those of fossils, drainage basins, and sedimentary particles, the second principal component represents length-width ratios and can be interpreted as a generalized "shape" measure. Succeeding principal components may be difficult to interpret, especially if they represent only a minor portion of the total variation. Sometimes, however, they will be composed almost exclusively of a single original variable, indicating that variable is essentially independent of the other raw variables. If the raw variables consist of successive measurements along a distribution (such as a grain size frequency histogram), the principal components may be interpreted as the successive moments of the distribution.

Example: Principal components of turbidite sandstones.

The data given in Exercise 20.2 on the composition of non-greywacke sandstones from Welsh turbidites yields a 5 X 5 covariance matrix:

$$\begin{bmatrix} 207.0 & & & & \\ 23.5 & 26.7 & & & \\ -32.1 & 1.6 & 89.5 & & \\ -177.9 & -59.5 & -81.7 & 356.0 & \\ -14.5 & 8.2 & 22.4 & -42.4 & 26.2 \end{bmatrix}$$

Because all variables are expressed in common units (percent), it is not necessary to standardize the observations. Therefore, each variable is weighted according to the magnitude of its variance.

The successive eigenvalues of this matrix are given below. Each represents the variance contained within its corresponding eigenvectors or principal component. Also given is the percent of the total variance contained within the principal components.

Principal component	Eigenvalue	Percent of total variance	Cummulative percent of total variance
1	491.8	69.7	69.7
2	167.4	23.7	93.4
3	33.4	4.7	98.2
4	12.9	1.8	100.0
5	0.1	0.0	100.0

The first principal component has more than twice the variance of the most variable of the original measurements. The second component contains a greater percentage of the total variance than do three of the original variables. This suggests that a limited number of the orthogonal principal components may be highly efficient in characterizing the differences between the original samples. Note that the last principal component has vanished, because the closed data array from which the covariances were calculated contains only four truly independent variables.

The eigenvectors are coordinates of the principal component axes given in the 5 X 5 matrix below. Columns are the individual eigenvectors, rows are the original variables:

	Principal Component				
	1	2	3	4	5
Quartz	-.51	-.66	-.31	.09	.44
Feldspar	-.13	.03	.61	-.63	.46
Rock fragments	-.13	.64	-.48	-.18	.45
Clay	.83	-.28	.15	-.01	.45
Cement	-.07	.26	.41	.75	.44

The first component, accounting for about 70% of the variance, represents variation in the proportion of clay matrix. The second accounts for an additional 24% of the variance and is produced by variation in the ratio (quartz + clay)/(rock fragments + cement + feldspar). The third component accounts for about 5% of the total variance and represents the ratio between (feldspar + cement) and the other constituents of the rock. These three components may be roughly equated to the sandstone classification parameters of Pettijohn (1957). The first component corresponds to Pettijohn's fluidity index ("ratio of sand detritus to the interstitial detrial matrix."), the second to his maturity index ("ratio of quartz...to feldspar plus rock fragments") and the third to his source rock index ("ratio of feldspar to rock fragments").

Components of a correlation matrix

If the matrix S is replaced by R, the matrix of correlations between the m original variables, the effect is the same as if the original raw variables were standarized. The magnitudes of the variables are no longer different, so the generally observed"size" component may be much less pronounced. In general, the first p components from a correlation matrix will account for less of the total variablility than will the first p components of a covariance matrix computed from the same data and may be more difficult to "read." If the raw variables are compatible in terms of their measurement units (all measurements of lengths, for example), standardization may be unwarranted. If they include a variety of different measures, however, the analysis may be extremely sensitive to these differences unless the corelation matrix is used.

Principal coordinates analysis

Eigenvalue techniques can be used in the so-called "Q-mode" to investigate interrelations between individual observations. First, it is necessary to calculate a matrix of similarities between all objects in a study. If m variables have been measured on n items, the matrix of similarities will be n X n. The similarity measure may be any of a number proposed in the literature of numerical taxonomy (q.v.), but the most widely used includes the cosine theta coefficient:

$$\text{Cos } \theta_{ik} = \frac{\sum_{j=1}^{m} X_{ij} X_{kj}}{\sqrt{\sum_{j=1}^{m} X_{ij}^2 X_{kj}^2}}$$

where X_{ij} represents the j variable measured on object i, and the cosine is measured between objects i and k. If the variables measured on objects are regarded as vectors in m-dimensional space, cos θ is the angle between two vectors. The measure will range from 1 when the two vectors coincide to 0 when the vectors are orthogonal.

Alternative measures include the average difference between two observations expressed as a proportion of the range of variables in the data set:

$$D_{ik} = 1 - \frac{\sum_{j=1}^{m} |X_{ij} - X_{kj}| / Rj}{m}$$

where R_j is the range of variable j. The quantity is subtracted from 1 so

D_{ik} ranges from 1 for complete similarity of the two observations to 0 for complete dissimilarity. This is essentially equivalent to a standardized Euclidean distance measure such as

$$D_{ik} = 1 - \frac{\sum_{j=1}^{m} \sqrt{(X_{ij}-X_{kj})^2/R_j^2}}{m}$$

Once the n X n matrix of similarities has been determined, it is necessary to find its n eigenvalues and the n X n matrix of eigenvectors or principal coordinates. Each of the eigenvectors contains n elements, one associated with each of the original n observations. Each eigenvector or principal coordinate is standardized so the sum of the squares of its elements is equal to its associated eigenvalue:

$$\sum_{j=1}^{n} a_{jp}^2 = \lambda_p$$

Observation X_i can now be represented as a point in principal coordinate space, whose location is defined by each of the ith elements of the n principal coordinates. The distance between any two points X_i and X_k is given by

$$d_{ik}^2 = \sum_{p=1}^{n} a_{ip}^2 + \sum_{p=1}^{n} a_{kp}^2 - 2 \sum_{p=1}^{n} a_{ip}a_{kp}$$

If the eigenvalue λ_p associated with a particular principal coordinate is small, then the contribution $(a_{ip}-a_{jp})^2$ to the distance between X_i and X_j will be small and that principal coordinate may be discarded. If λ_p is large, but the elements a_{ip} of its associated principal coordinate are all very much alike, then $(a_{ip}-a_{jp})^2$ will also make little contribution to the distances between X_i and X_j, and this coordinate may be discarded as well. The only significant coordinates are those having a wide range of elements and whose eigenvalues are large. Usually, only two or three principal coordinates are necessary to adequately display the set of n observations. It should be noted that if the original n X n matrix of similarities was calculated from m variables where $m < n$, all but the first m eigenvalues will of necessity be zero. This insures

an immediate reduction in the dimensionality of the problem, but usually it is possible to arbitrarily eliminate additional coordinates by the criteria given above until the samples can be plotted in ordinary space along axes defined by the remaining coordinates.

References

Blackith, R.E. and R.A. Reyment, 1971, Multivariate morphometrics; Academic Press, London, 412 p.

Hope, Keith, 1968, Methods of multivariate analysis: University of London Press Ltd., London, 165 p.

33. FACTOR ANALYSIS

"The name of the song is called 'Haddocks' Eyes.'"
"Oh, that's the name of the song, is it?" Alice said, trying to feel interested.
"No, you don't understand," the Knight said, looking a little vexed. "That's what the name is called. The name really is 'The Aged Aged Man.'"
"Then I ought to have said that's what the song is called?" Alice corrected herself.
"No you oughtn't: That's quite another thing! The song is called 'Ways and Means': But that's only what it's called, you know!"
"Well, what is the song then?" said Alice, who was by this time completely bewildered.
"I was coming to that," The Knight said. "The song really is 'A-sitting On A Gate': and the tune's my own invention."
--Lewis Carroll (Through the Looking-Glass)

Factor analysis is probably the most well-known (but controversial) multivariate technique used for data analysis. The theoretical model postulates that observed variables are correlated with a lesser number of "hidden" variables or factors which explain the systematic variation in the measured sample. The method had its origins in psychological studies, where factors were equated with personality traits and aptitudes and thought to account for observed variation in data such as questionaire responses and examination scores. In a geological context, two causal (but unknown) variables of paleotemperature and pressure might be postulated to "explain" the variation in an m-variable data set of mineral suite measurements sampled across an igneous body.

Before computers were available, factor analyses were calculated by the centroid solution. However, the initial phase of factor analysis is now programmed as a principal components solution (q.v.) of the correlation matrix of the observational variables. The computation of principal components is a simple algebraic reduction of the data and does not involve implicit hypotheses. A new frame of reference is created for the m variables, where the m axes are uncorrelated and located along directions of maximum variance. For the purposes of factor analysis, there are two drawbacks in principal components as a final solution: (1) Unless there are some zero eigenvalues, the m variables are still expressed in terms of m principal components.
(2) The position of the principal component axes is a generalized "average" of the variances contained in the original variables.

Theory of factor analysis

Suppose that the systematic variation contained in m variables can be expressed by k ($k < m$) factors that are mutually uncorrelated (and therefore geometrically orthogonal). If Z_{ij} is the standardized score of the ith individual in terms of the jth observed variable, a factor model equation can be written:

$$Z_{ij} = a_{1j}F_{i1} + a_{2j}F_{i2} + \ldots + a_{kj}F_{ik} + a_jE_{ij}$$

This equation states that the observed value of the jth variable is composed of a linear combination if k factor scores (the F's) multiplied by factor loadings (the a's) plus a residual "error" term (E) associated with the jth variable. Measured over n observations of the jth variable, the general equation can be written as:

$$Z_j = a_{1j}F_1 + a_{2j}F_2 + \ldots a_{kj}F_k + a_jE_j$$

which can be recognized as very similar to a multiple regression equation with the important distinction that the independent variables are unknown. A complete factor model can be written with m of these equations, one for each observed variable.

There are three types of unknowns to be resolved:

(1) k, the number of factors that account for the systematic variation in the m variables (as opposed to "error").

(2) The factor loadings a, for each of the k factors with respect to each of the m variables.

(3) The factor scores F, for each sample individual with respect to each of the k factors.

Determination of k, the number of factors, is a crucial problem in factor analysis. In its early psychological applications, the researchers "knew" the underlying factors and could identify them by name. This is rarely the situation in geological systems and the purpose of most research is to identify the causative mechanisms that account for measureable properties as an end product of the analysis. The identification of these causal factors (and so their number) as a preface to analysis is alien to conventional geological thinking which would regard this step as prejudging the issue. Instead, the basic properties of the data set are examined for clues regarding its basic structure. It will

be recalled that if the eigenvalues of an m-variable data set are computed and g of these are zero, the true dimensionality of the data is (m-g) and can be described by (m-g) orthogonal (and uncorrelated) axes. This situation will only occur where there is a strict redundancy of description, such as in closed data sets. As open data is measured on m random variates, there will almost always be m non-zero eigenvalues. However, some of these eigenvalues may be sufficiently small to be disregarded as trivial and equated with "error". If the eigenvalues of five variables account for

54% 41% 2% 2% 1%

of the total variance, 95% of the variaton is contained in a two-dimensional plane in the original five-dimensional space. It may be theorized that two orthogonal "causal" factors account for the five variables, with a residual error term contained in the remaining dimensions. Selection of k in less clear-cut cases is a matter of judgement, although various "rules" have been suggested such as counting the number of eigenvalues that each account for more than 1/m of the total variance. The initial phase of factor analysis (computation of principal components) may therefore be used as an aid in deciding an appropriate value for k.

The matrix of m eigenvectors $\underline{U}$, is standardized by multiplication by a diagonal matrix of the square roots of the eigenvalues $\underline{D}$ to the matrix of factor loadings $\underline{A}$:

$$\underline{A} = \underline{U}\,\underline{D}$$

The effect of this is a standardization of the transformed variables in an analogous manner to the standardization of the original raw data.

If k is the number of causal factors, then k factors are contained in the dimensions implied by the first k columns of $\underline{U}$. The remaining (m-k) eigenvectors express the dimensions in which the "error" terms are contained. The first k eigenvectors are located along the axes of maximum variance which generally correspond to a composite mixture of loadings on the original variables. The aim of factor analysis is to look for a "simple" structure in which each of the variables is expressed either as a strong loading (ideally, +1 or -1) or a weak loading (ideally, zero) with respect to each of the factors. The orthogonal axes are therefore rotated from the eigenvectors in search of a solution which best matches a simple structure.

Factor rotation

The most commonly used method for axis rotation is that devised by Kaiser (1958) according to the varimax criterion. The original eigenvectors are rigidly rotated and their orthogonality maintained. The axes are moved to a position such that the sum of the variances of the factor loadings is a maximum. A rotation operation is achieved in matrix algebra by multiplication by a matrix of the form:

$$\underline{Q} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

where θ is the required angle of rotation. If the matrix of the first k eigenvectors is $\underline{K}$ and $\underline{V}$ is the matrix of rotated factors then:

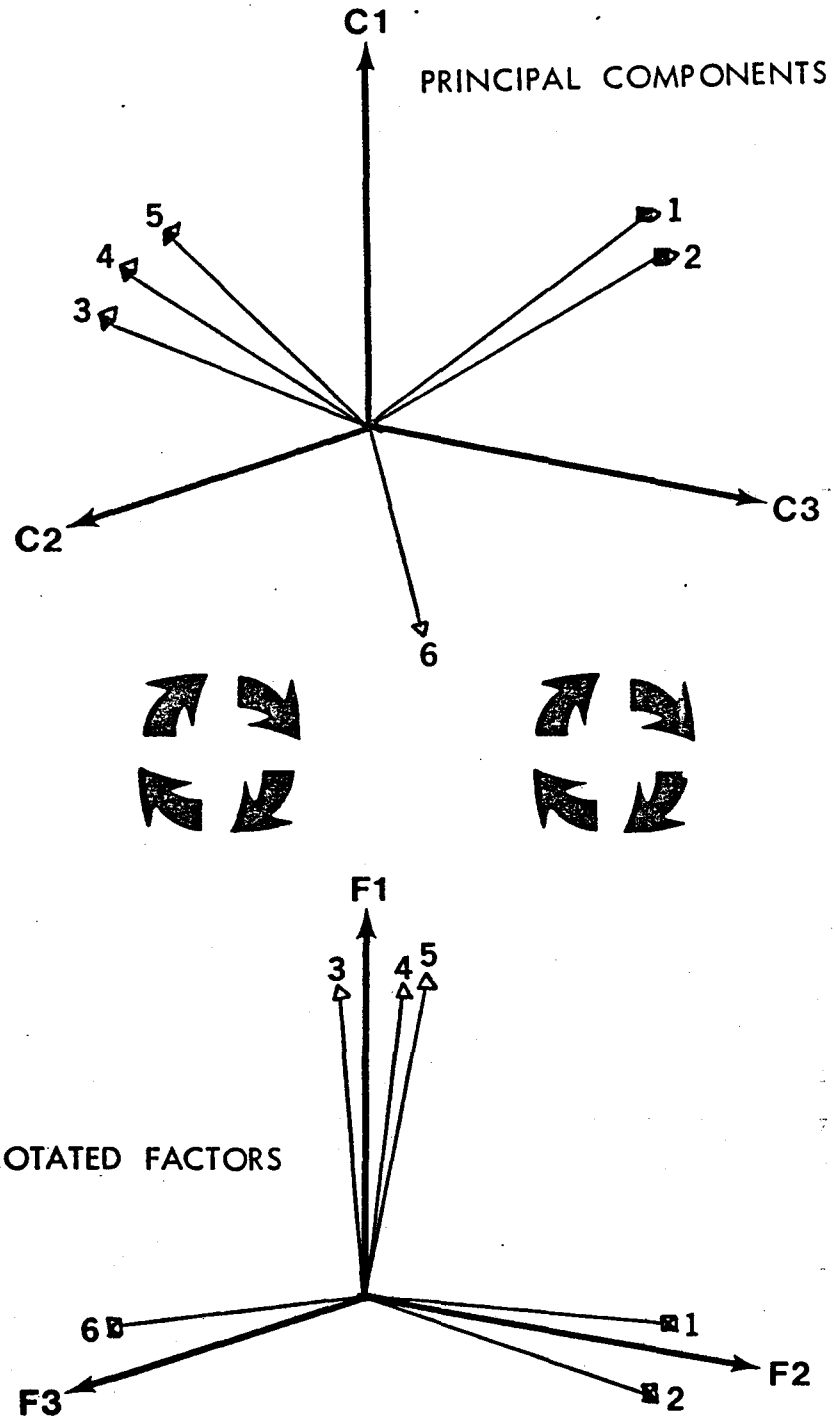
$$\underline{V} = \underline{K}\,\underline{Q}$$

The values of θ are determined by a programmed algorithm as an iterative process that rotates two axes at a time and is geared to the maximization of the total variance of the factor loadings. (For computational details, see Harman, 1960).

Interpretation of the factor matrix

The k standarized eigenvectors may be interpreted as factors (the so-called principal component solution), but the rotated factors (the varimax or other rotational solution) are more commonly used. Since the factor matrix has been standardized, the sum of the squared factor loadings for each variable is the amount of variance in the variable accounted for by the k factors and is called the communality. If all the variance of variables is contained, the communality is 1.0. Similarly the sum of the squared factor loadings within a factor is the proportion of the total variance contained in the factor.

The factor loadings of any factor are examined to interpret the physical nature of the causal variable that the factor represents. Obviously, this step requires a basic understanding of the data together with the application of some external criteria. The interpretation process is extremely delicate and open to ridicule by critics of factor analysis. Interpretation may often appear to confirm innate prejudices concerning causal mechanisms or suggest surrealist models to the naive investigator.


C1
PRINCIPAL COMPONENTS
1
2
3
4
5
6
C2
C3
F1
3
4
5
ROTATED FACTORS
6
1
2
F3
F2

The factor score matrix

The factor score matrix $\underline{F}$ may be obtained by matrix manipulation. Since the original factor model is:

$$\underline{Z} = \underline{F}\,\underline{A}'$$

then

$$\underline{ZA} = \underline{F}\,\underline{A}'\,\underline{A} = \underline{F}\,\underline{\Lambda}$$

$$\underline{F} = \underline{Z}\,\underline{A}\,\underline{\Lambda}^{-1}$$

where $\underline{\Lambda}$ is the diagonal matrix of eigenvalues.

The scores of individuals may be mapped as scatter plots using any two of the factors as references axes and interpretation of their spatial relationships made in terms of the "meaning" of these factors.

References

Harman, H.H., 1960, Modern Factor Analysis: Univ. of Chicago Press, Chicago, 471 p.

Kaiser, H.F., 1958, The varimax criterion for analytic rotation in factor analysis: Psychometrika, v. 23, p. 187-200.

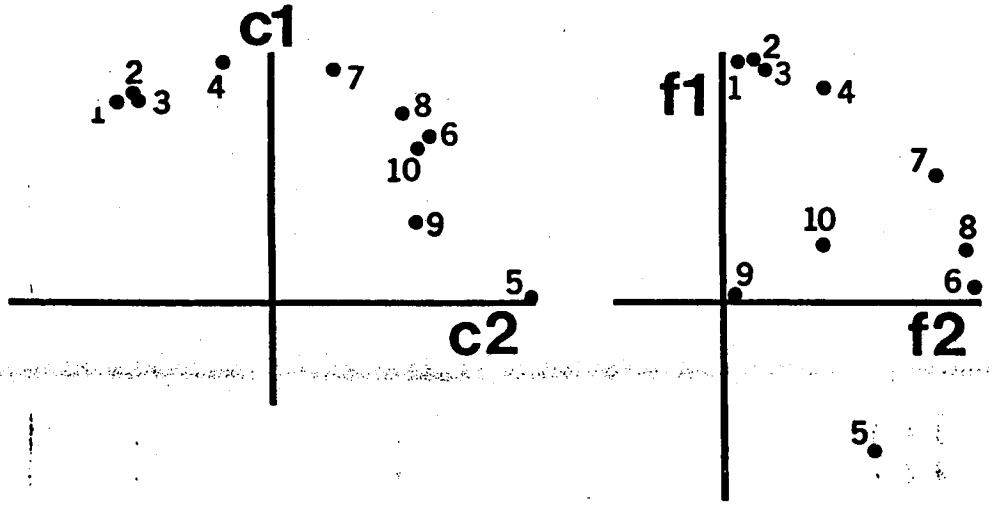
Example: The Pickle Crow Mines exploration program

Many of the terms and concepts outlined may be more readily understood by consideration of an example devised by Klovan (1968). The following treatment is drawn directly from Klovan's paper (with the exception of the name of the hypothetical mining company). While the example is artificial, it illustrates well how factor analysis ought to work under ideal conditions.

Pickle Crow Mines Inc. considered acquiring a one-hundred square mile area surrounding a major lead-zinc ore deposit that they had mined. A reconnaissance study was undertaken using data collected from twenty stations located on a square sampling grid across the area. Company geologists believed that a variety of controls influenced the location of mineralization such as paleotemperature, tectonic style and the transmission properties of the rocks with respect to ore-bearing solutions. Since no direct measurements could be made of these controls, data were collected on a series of rock properties that were believed to reflect them:

1. Magnesium content in calcite
2. Iron content in sphalerite
3. Sodium content in muscovite
4. Sulphide content
5. Crystal size of carbonates
6. Spacing of the rock cleavage
7. Elongation of oolites
8. Tightness of folds
9. Veins per square meter
10. Number of fractures per square meter

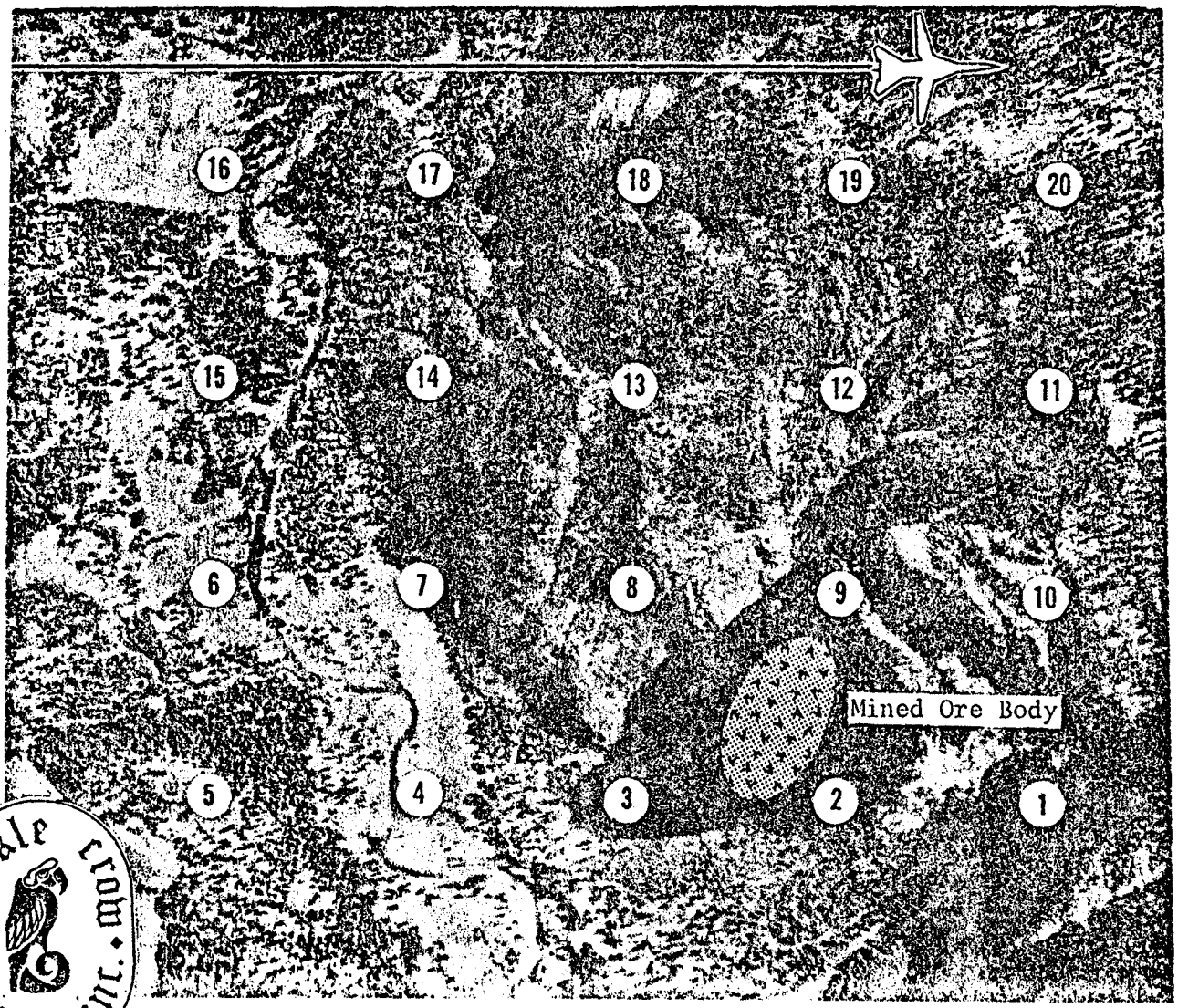
One of the aims of the study was to interpret these observational variables as products of a few, simple causative controls (corresponding to factors). Another was to locate a site where ore deposits might occur similar to the mined ore body. Computation of the eigenvalues of the 10 X 10 correlation matrix of variables resulted in only three non-zero eigenvalues. (This is obviously a reflection of the artificial nature of the data but this situation might be closely paralleled by real random variate data where the first three principal components account for almost all of the data variance.) The geometrical implications are that the data points are constrained within an ellipsoid rather than a ten-dimensional hyperellipsoid. The matrix of standardized eigenvectors could be used as the so-called principal component factor matrix. However, many of the factor loadings are of intermediate strength (notable on the second factor) making interpretation rather ambiguous. The three principal component factors were rotated by the varimax method and the new factor axes located to conform more closely with the separate clusters of variables rather than their composite average.



GEOLOGICAL PROPERTIES

	Mg In Calcite	Fe In Sphalerite	Na In Muscovite	Sulphide	Crystal Size of Carbonates	Spacing of Cleavage	Elongation of Ooliths	Tightness of Folds	Veins/Meter^2	Fractures/Meter^2
LOCALITY 1	1175	999	975	625	158	262	437	324	431	433
LOCALITY 2	936	820	813	575	267	379	478	413	411	428
LOCALITY 3	765	711	716	599	457	548	579	558	491	513
LOCALITY 4	624	598	600	542	471	515	531	520	490	500
LOCALITY 5	417	422	422	432	444	441	437	439	437	437
LOCALITY 6	401	403	375	401	405	270	317	290	515	465
LOCALITY 7	520	504	488	469	427	370	410	386	507	482
LOCALITY 8	661	626	618	553	462	466	506	480	529	523
LOCALITY 9	877	787	773	594	354	401	493	434	500	498
LOCALITY 10	1060	932	898	656	315	312	468	370	580	552
LOCALITY 11	1090	960	935	681	334	375	518	427	567	555
LOCALITY 12	896	811	790	629	403	411	511	448	570	555
LOCALITY 13	748	688	672	560	401	399	472	426	525	512
LOCALITY 14	617	573	553	477	360	315	385	342	487	462
LOCALITY 15	436	424	389	393	361	207	277	236	514	455
LOCALITY 16	664	587	560	419	212	182	287	221	397	369
LOCALITY 17	750	665	651	484	259	299	387	331	399	396
LOCALITY 18	903	797	791	573	291	396	486	427	421	437
LOCALITY 19	998	888	887	657	366	499	583	527	480	506
LOCALITY 20	1162	999	994	671	252	404	539	450	449	471

DATA MATRIX



pickle crow mines inc.

CORRELATION COEFFICIENTS

VARIABLE	1	2	3	4	5	6	7	8	9	10
1	1.000	0.998	0.994	0.908	-0.576	0.130	0.581	0.282	0.012	0.258
2	0.998	1.000	0.998	0.933	-0.523	0.183	0.625	0.334	0.057	0.313
3	0.994	0.998	1.000	0.942	-0.497	0.235	0.664	0.383	0.035	0.312
4	0.908	0.933	0.942	1.000	-0.180	0.477	0.834	0.610	0.286	0.590
5	-0.576	-0.523	-0.497	-0.180	1.000	0.616	0.258	0.519	0.539	0.550
6	0.130	0.183	0.235	0.477	0.616	1.000	0.880	0.987	0.181	0.524
7	0.581	0.625	0.664	0.834	0.258	0.880	1.000	0.944	0.216	0.604
8	0.282	0.334	0.383	0.610	0.519	0.987	0.944	1.000	0.208	0.573
9	0.012	0.057	0.035	0.286	0.539	0.181	0.216	0.208	1.000	0.909
10	0.258	0.313	0.312	0.590	0.550	0.524	0.604	0.573	0.909	1.000

Correlation Coefficients Between the Ten Geological Properties

		EIGENVALUE	PERCENT VARIANCE EXPLAINED
FACTOR	I	5.46	54.61
FACTOR	II	3.19	86.54
FACTOR	III	1.35	100.00

EIGENVALUES OF CORRELATION MATRIX

PRINCIPAL COMPONENT FACTOR MATRIX

		FACTORS		
	COMM.	1	2	3
1	1.0000	0.8029	-0.5894	0.0886
2	1.0000	0.8385	-0.5367	0.0940
3	1.0000	0.8579	-0.5122	0.0407
4	1.0000	0.9760	-0.1961	0.0943
5	1.0000	0.0176	0.9998	-0.0098
6	1.0000	0.6538	0.5999	-0.4611
7	1.0000	0.9297	0.2393	-0.2799
8	1.0000	0.7647	0.5018	-0 4042
9	1.0000	0.3268	0.5407	0.7751
10	1.0000	0.6641	0.5437	0.5132
VARIANCE		54.614	31.928	13.459
CUM. VAR		54.614	86.542	100.000

VARIMAX FACTOR MATRIX

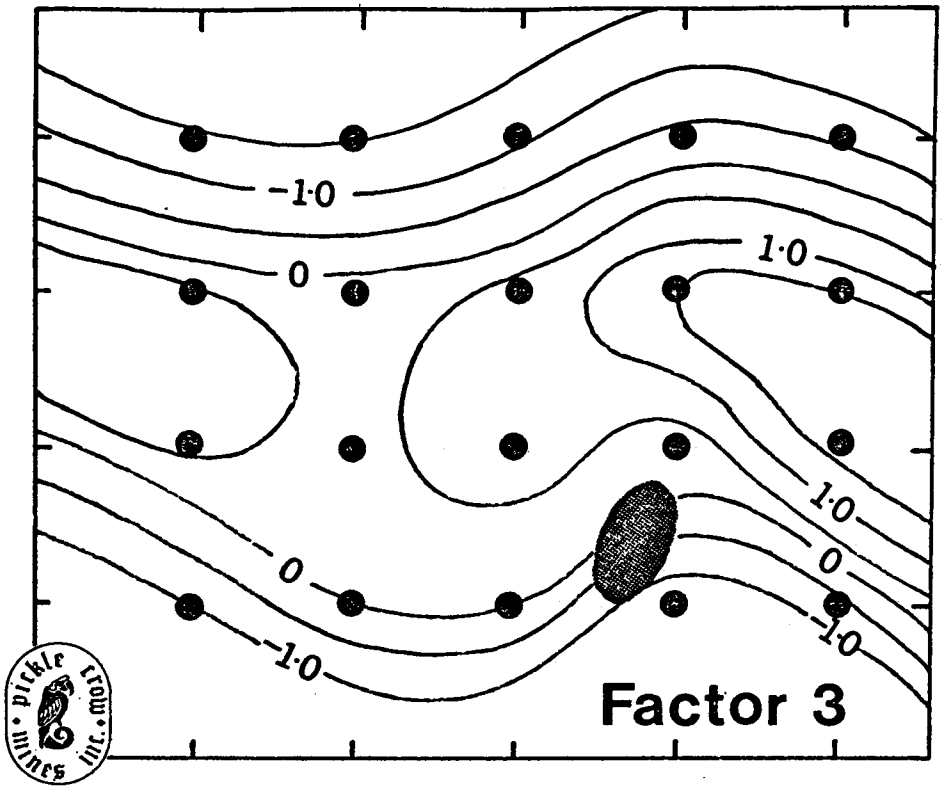
VAR	COMM.	FACTORS 1	2	3
1	1.0000	0.9971	0.0765	-0.0060
2	1.0000	0.9916	0.1241	0.0362
3	1.0000	0.9835	0.1804	0.0117
4	1.0000	0.8813	0.3985	0.2540
5	1.0000	-0.6197	0.5880	0.5198
6	1.0000	0.0558	0.9897	0.1317
7	1.0000	0.5191	0.8380	0.1680
8	1.0000	0.2102	0.9648	0.1580
9	1.0000	0.0146	0.0488	0.9987
10	1.0000	0.2338	0.3979	0.8872
VARIANCE		44.771	33.318	21.912
CUM. VAR		44.771	78.089	100.000

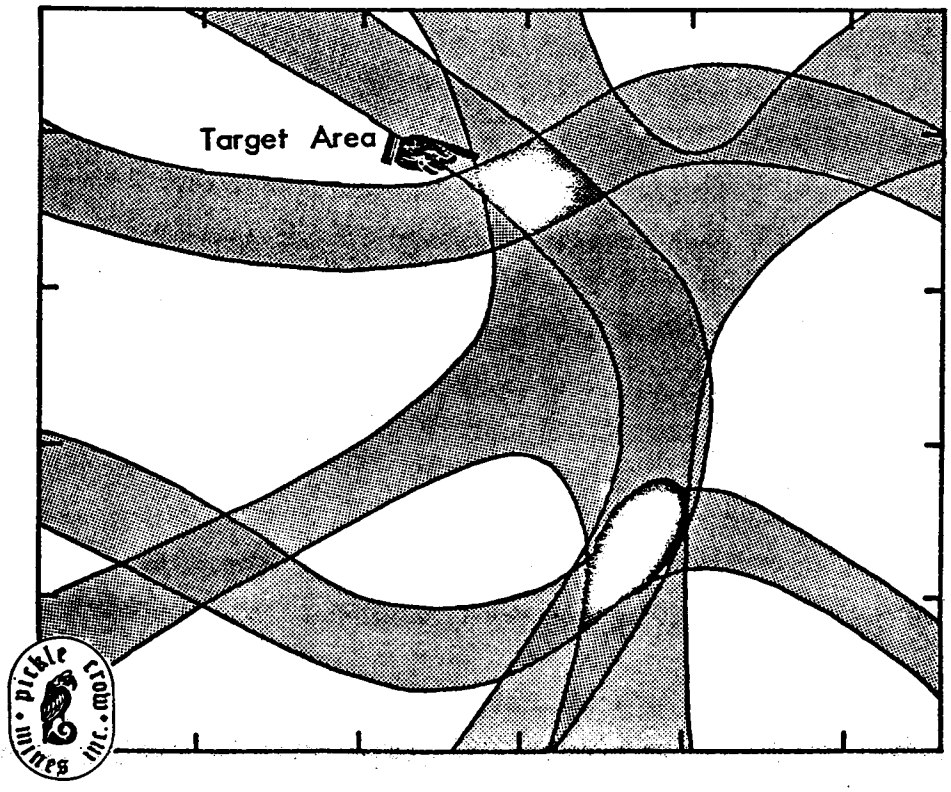
VARIMAX FACTOR SCORE MATRIX

		FACTORS		
LOCALITY	1	1.7191	-1.1345	-0.9480
LOCALITY	2	0.6130	0.2141	-1.3682
LOCALITY	3	-0.2291	1.8610	0.0226
LOCALITY	4	-0.7962	1.5403	0.0196
LOCALITY	5	-1.6301	0.9332	-0.8890
LOCALITY	6	-1.5345	-1.0766	0.6182
LOCALITY	7	-1.1157	-0.0212	0.4175
LOCALITY	8	-0.5908	0.9143	0.7663
LOCALITY	9	0.3711	0.2439	0.2588
LOCALITY	10	1.2434	-0.9398	1.7595
LOCALITY	11	1.3197	-0.2480	1.4876
LOCALITY	12	0.4639	0.1764	1.5324
LOCALITY	13	-0.1684	0.1884	0.7211
LOCALITY	14	-0.6634	-0.5747	0.0782
LOCALITY	15	-1.3364	-1.7591	0.6394
LOCALITY	16	-0.3829	-1.7847	-1.5171
LOCALITY	17	-0.1144	-0.5587	-1.5440
LOCALITY	18	0.4618	0.3838	-1.1944
LOCALITY	19	0.7982	1.3086	-0.1605
LOCALITY	20	1.5620	0.3336	-0.6997

-1·0
0
1·0
Factor 1

-2·0
-1·0
0
0
1·0
1·0
1·0
Factor 2


-1·0
0
1·0
1·0
0
-1·0
Factor 3


Target Area

The first rotated factor may be interpreted as "paleotemperature." Loadings on variables 1,2,3 are very high and all three have been considered as paleothermometers, while loadings on the other variables are very low. In the same way, Factor 2 may be equated with "deformation" and factor 3 with "permeability." Scores for each of the sample stations were computed with respect to each of the factors. The independent geographic mappings of factor scores were contoured as hypothetical surfaces of indices of paleotemperature, deformation and permeability. In order to locate new economic ore bodies, inspection was first made of the set of factor scores at the site of the mine that collectively described the optimal "mix" of physical controls required for major lead-zinc mineralization. An overlap of the three factor score maps was then examined for locations with similar sets of scores favorable to mineralization. One such area was located approximately four miles to the north of the mine and was made a target area for more detailed evaluation.

As Klovan points out, the artificial origin of the data limits the example as a truly effective demonstration of the power of factor analysis. The factor analysis has merely succeeded in recovering the model that was used to generate the data. However, situations in the real world may exist where causal processes are dominated by a few distinctive and independent controls that imprint their operation on measurable geological variables. Under these conditions, factor analysis may have some potential in unravelling simple patterns from complex arrays of observational data. Numerous qualifications must be attached to this statement since errors are involved in measurement, variables of no importance may be used and interpretation may be completely misjudged for a variety of reasons. In addition, this basic factor model of orthogonal linear factors may be totally inappropriate and "hidden variables" (if they exist) may be partially dependent and/or non-linear.

Reference

Klovan, J.E., 1968, Selection of target areas by factor analysis: Western Miner, Feb. 1968, p. 44-54.

Other factor models

Factor analysis may be conducted in either R-mode or Q-mode and interpretation directed either to factor loadings, factor scores or both. The classic factor model stipulates that the factors be orthogonal (and therefore uncorrelated). It can be argued that many real phenomena are interrelated and therefore that the factors that represent them will often be correlated. This possibility may be accomodated by postulating oblique factor axes. Rotation of the principal component axis is made without the constraint of orthogonality and axes are selected to fit clusters of observational variables.

Factor analysis: The Great Debate

In the words of Blackith and Reyment (1971), "It is very hard to discuss factor analyses (for there are many different kinds) without generating more heat than light." The philosophies of factor analysis are a strange hybrid of scientific rationalization and subjective criteria. As a way of analysing data and testing hypotheses it is clearly beyond the pale of orthodox statistical analysis, which is structured in terms of probability statements and decision procedures. This contrast in character is discussed by Hope (1967):

"Factor analysis is akin to medieval scientific method in that it allows Nature to dictate the explanations, whereas analysis of variance has the modern, Kantian, spirit which 'puts Nature to the Inquisition', imposes unnatural conditions upon her in order to wrest from her the answer to preconceived questions.....It is best regarded as a sophistication of the kettle-watching which led to experiments with steam engines."

This characteristic is widely recognized and factor analysis is frequently used as a "fishing expedition" method to generate instant interpretations from mind-congealing masses of data. Hope comments that "factor analyses has proved an excellent way to get a research degree without the need for sound hypotheses and good experimental design." Geologists who make thoughtful use of factor analysis, such as Klovan, counter this observation by stating that factor analysis should be used as an initial exploratory method to generate hypotheses to be tested by more rigorous statistical techniques. In Klovan's words:

"I regard the technique of factor analysis as a tool and I use it much as I do my geological hammer--to reveal hidden facts. As such it is an exploratory tool. Once patterns are revealed by it, more sophisticated methods of data analyses can, and should, be used to evaluate the data more fully."

Some of the philosophical problems that are introduced in applying factor analysis have been touched on in these pages. Questions have to be resolved concerning the number of factors, the type of rotation criteria to be used, whether to compute orthogonal or oblique factors, etc., etc. Their resolution is almost always a matter for the investigator's judgement. A more thorough treatment of these problems is given in Blackith and Reyment (1971).

All the preceding discussion may be dismissed by the applied geologist as academic wrangling. The question uppermost in the mind of the mining or petroleum geologist can be simply stated, "Does it work?" This is difficult to assess in geological studies because the "real" answer is rarely ever known. However, the causes of phenomena in the physical sciences are often precisely understood. Armstrong (1967) applied factor analysis in the standard procedure to data derived from a known physical model. Without an almost exhaustive knowledge of the "real" factors, the interpretation of results was often both misleading and ludicrous. Armstrong concluded:

> "In these studies where the data stand alone and speak for themselves, my impression is that it would be better had the studies never been published. The conclusion that this factor analytic study has provided a useful framework for further research may not only be unsupported - it may also be misleading."

References

Armstrong, J.S., 1967, Derivation of theory by means of factor analysis or Tom Swift and His electric factor analysis machine: The American Statistican, Dec. 1967, p.17-21.

Blackith, R.E., and Reyment, R.A., 1971, Multivariate Morphometrics: Academic Press, London, 412 p.

Hope, K., 1967, Elementary Statistics: Pergamon Press, Oxford, 101 p.

Klovan, J.E., 1968, Selection of traget areas by factor analysis: Western Miner, Feb. 1968, p. 44-54.

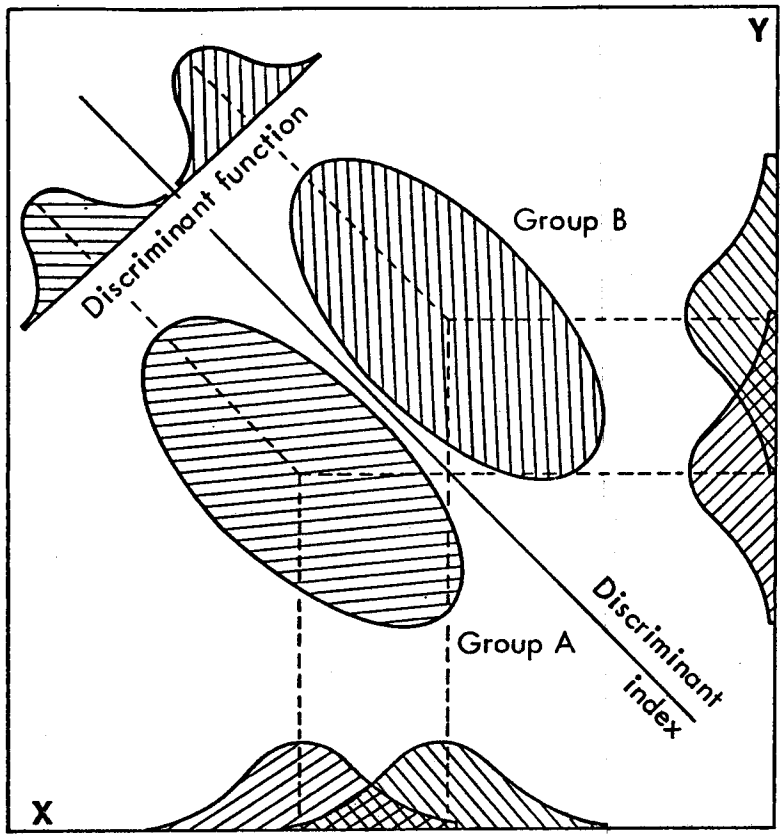
34. DISCRIMINANT FUNCTIONS

Discriminant function analysis is one of the most popular mathematical techniques used in geological studies because if offers solutions to many of the central problems of geology. The practical application of the principle of uniformitarianism requires that a selective combination of variables observed in modern environments be used in the interpretation of the genesis of ancient environmental products. Suites of key variables such as facies associations, diagnostic mineral assemblages, and grain-size statistics have traditionally been used in a qualitative approach to this problem. However, the definition of the critical variables and their separate weightings is equivocal and eminently debatable. The central problem boils down to finding the most effective criterion compounded from obserable variables, which best discriminates samples from known environments. Then, applying this criterion to the classification of samples from unknown ancient environments. Naturally, the structure of this optimal criterion contains inherent information on genetic processes.

Fisher (1936) first derived the linear discriminant function as a statistical method for separating two populations on the basis of their properties by means of a linear weighted function of their variables. A test sample of size n is assigned to two groups (each corresponding to a distinct population) on the basis of prior knowledge. The two groups (sizes n_1 and n_2) may be plotted as clouds of data points in m-dimensional space (for m variables) and are either distinctively separated or have some degree of overlap. It is possible to locate an axis on which the distance between each cloud is maximized while, at the same time the dispersion within each cloud is minimized. This axis defines a <u>linear discriminant function</u> and is derived from the multivariate means, variances and covariances of the two groups. The data-points of the two groups may be projected onto this axis; their relative distances along the axis are conventionally denoted as z-scores (not to be confused with standardized Z-scores).

If z_i is the score of the ith individual, the projection corresponds to the equation:

$$z_i = \lambda_1 X_{i1} + \lambda_2 X_{i2} + \ldots \lambda_m X_{im}$$



Pictorial representation of discriminant analysis applied to a simple bivariate example. (The example is idealized, in the sense that a perfect separation is illustrated, unlike realistic situations where there is commonly an overlap of group distribution clouds across the discriminant index.)

---

where the lambdas are a set of weighting coefficients which translate the $X_{i1}....X_{im}$ coordinates to a scalar value on the disciminant function axis. (λ is the conventional symbol for these coefficients but they are <u>not</u> eigenvalues).

If the centroids of the two groups are entered in the equation as $\overline{X1}_1$, $\overline{X1}_2...\overline{X1}_m$ and $\overline{X2}_1$, $\overline{X2}_2...\overline{X2}_m$, the z-scores are denoted $\overline{z1}$ and $\overline{z2}$. The ratio,

$$F = \frac{(\overline{z1}-\overline{z2})^2}{var(z)}$$

is a measure of the squared deviation distance between the group centroids divided by a pooled estimate of the variance within the two groups. Solving

for the values of λ that produce the best discrimination results in a maximum value of F. This condition occurs when the following equations are satisfied:

$$s_{11}\lambda_1 + s_{12}\lambda_2 + \ldots\ldots\ldots\ldots s_{1m}\lambda_m = d_1$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$s_{m1}\lambda_1 + s_{m2}\lambda_2 + \ldots\ldots\ldots\ldots s_{mm}\lambda_m = d_m$$

where $d_j = \overline{X1}_j - \overline{X2}_j$

and s_{jk} is the sum of the corrected cross products of the two groups, divided by the total degrees of freedom, or

$$s_{jk} = \frac{(n_1-1)\ \mathrm{cov}\ (X1_j, X1_k) + (n_2-1)\ \mathrm{cov}\ (X2_j, X2_k)}{(n_1 + n_2 - 2)}$$

The set of equations describes a regression of d (the dependent variable) on $X_1 \ldots X_m$ as independent variables. Written in matrix form:

$$[s_{ik}]\ [\lambda_k] = [d_i]$$

or $$\underline{S}\ \underline{L} = \underline{D}$$

$$\underline{L} = \underline{S}^{-1}\underline{D}$$

giving a column vector $\underline{L}$ of λ coefficients.

Since a pooled variance-covariance matrix is used, the solution assumes that the variances and covariances of each group are approximately equal. Under this condition, a discriminant index, z_c may be computed from:

$$z_c = \frac{\overline{z1} + \overline{z2}}{2}$$

The discriminant index is a boundary z value between the two groups, selected for the classification of a new individual under the stipulation that the chances of misclassification are divided equally between the two groups. Even if the covariance matrices of the two groups are fairly dissimilar, the linear discriminant function and index provide a viable discrimination tool. If the covariance matrices are markedly dissimilar and the precision of the observation variables warrants more sophisticated discrimination, computation of a quadratic discriminant may be made, which

is also robust for non-normal, but unimodal symmetric, distributions (Burnaby, 1966).

Significance test of discrimination

The linear discriminant function algorithm will necessarily provide a numerical solution but this does not constitute implicit evidence that there is a statistically distinct difference between the two groups in terms of the measured variables. A test of the significance of group separability may be made if the following conditions are approximately met:

(1) The individuals of each group are randomly selected.
(2) The measurement variables in each group are normally distributed.
(3) The variance-covariance matrices of each group are equal.
(4) The individuals used in the function computation were all correctly classified.
(5) A new individual has an equal probability of belonging to either group.

Only rarely will all these rules be met with geological data, but the discriminant function is robust to limited deviations from these conditions.

A generalized distance, D^2, called Mahalanobis' distance, may be computed from:

$$D^2 = \lambda_1(\overline{X1}_1 - \overline{X2}_1) + \ldots\ldots \lambda_m(\overline{X1}_m - \overline{X2}_m)$$

An F ratio is calculated from:

$$F = \left[\frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2)m}\right]\left[\frac{n_1 n_2}{n_1 + n_2}\right] D^2$$

and the null hypothesis that the two group centroids are equal is tested as an F test with m and $(n_1 + n_2 - m - 1)$ degrees of freedom at a selected significance level.

Contribution of observational variables to discrimination

The proportional component of the variable X_j in the distance between the two group means is given by:

$$E_j = \frac{\lambda_j (\overline{X1}_j - \overline{X2}_j)}{D}$$

The ratios may be converted to percentage contributions by multiplication by 100. These contributions are measures of the variables as independent factors and do not take account of possible interactions. Alteratively, a computation of the potency of each variable may be made from:

$$D_i = \frac{(\overline{X1}_i - \overline{X2}_i)^2}{(n-2)\text{var}(X_i)}$$

Variables with the highest potency make the greatest contribution to discrimination. The values of D_i may be used in an F-test procedure to test the significance of each variable within the total discimination as described by Schultz and Goggans (1961).

Canonical analysis: The general case

The linear discriminant function forms a special limiting case of the generalized procedure of canonical analysis in which there are only two groups and where a single discriminant axis is computed, If an optimum discrimination is desired between g groups in terms of m variables, an approach similar to that of principal components may be used. The aim of canonical analysis is to locate a series of orthogonal axes in m-dimensional space which successively account for ordered proportions of the greatest variability between the two g groups as opposed to variability within them. The linear discriminant function represents the first canonical variate. Computation of an additional second major canonical variate in the g group case allows a mapping of the groups in a plane of maximum discrimination. Canonical analysis is made by operation on the g group mean vectors and variance-covariance matrices. Though widely used in biological studies (Seal, 1964), canonical analysis has not been applied in geology to the extent of the linear discriminant function.

References

Burnaby, T.P., 1966, Distribution-free quadratic discriminant functions in paleontology: Kansas Computer Contribution 7, Kans. Geol. Survey, p. 70-77.

Fisher, R., 1936, The use of multiple measurements in taxonomic problems: Ann. of Eugenics, v. 7, p. 179-188.

Schultz, E.F., and Goggans, J.F., 1961, Procedure for determining potent independent variables in multiple regression and discriminant analysis: Agric. Expt. Stat., Auburn Univ., Bull. 336, 75 p.

Example: Discrimination of dry and producing wells in the Kansas "Mississippian chat"

Small oil and gas fields are found in the Mississippian 'B' chert interval in South-West Stafford C unty, Kansas. Conventional exploration for these targets relies mainly on the location of structural highs by reflection seismic methods, although it is recognized that chances of finding a new field are enhanced in locations where the interval is relatively thick and has a low shale content. Data for three variables were assembled for a sample of 124 wells in the area which measure the following characteristics:

(1) The average gamma ray log deflection in the interval, standardized as a proportional measure of shale content, ranging between the extremes of zero (no shale) and one (total shale).

(2) The thickness of the interval (feet).

(3) Deviations from a linear trend surface (q.v.) of the structural elevation of the top of the interval as a measure of the local structure with simple regional dip removed (positive and negative elevations in feet).

The sample was subdivided into a producing group P (33 wells) and a dry group D (91 dry holes). Histograms of the two groups for each variable are shown in the accompanying figure and indicate the degree of distinction between the groups in terms of each variable considered separately. Computation of a linear discriminant function led to the equation:

$$z = -6.34\,G + 0.07\,T - 0.01\,S$$

where G = shale ratio

T = thickness

S = structural residual

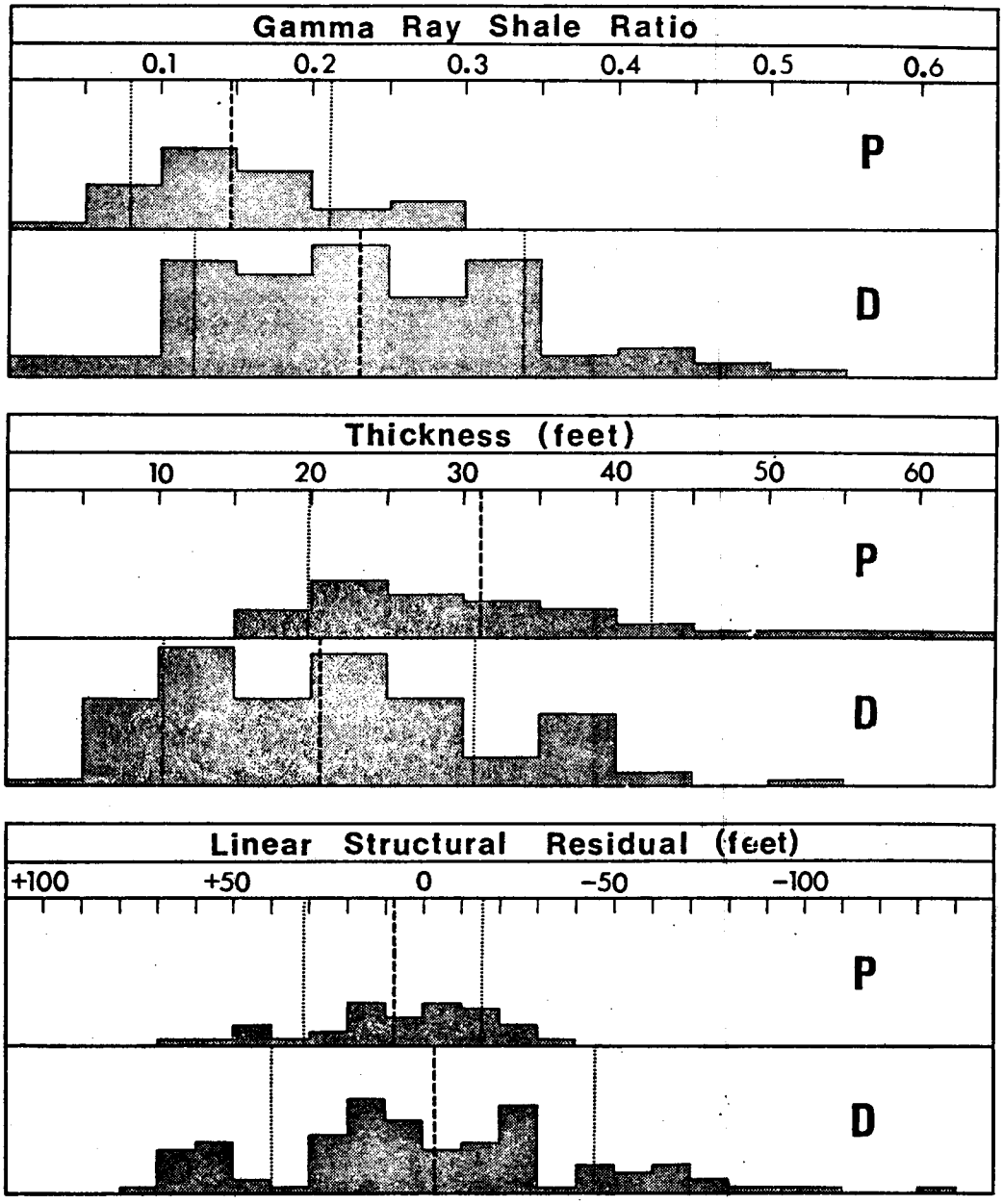
The scores of the group centroids are:

$$\overline{z_P} = 1.30 \text{ and } \overline{z_D} = -0.08$$

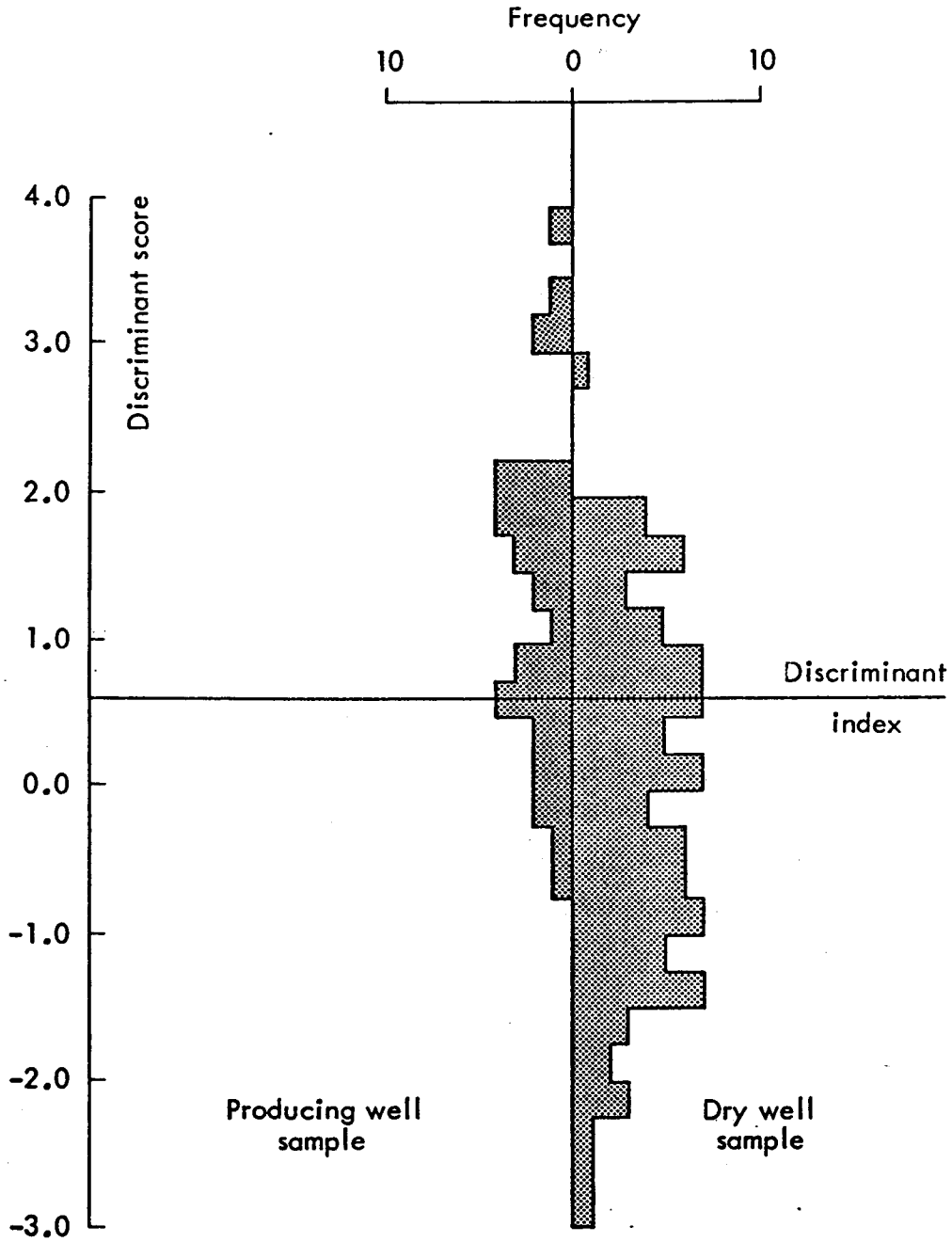
where the discriminant index,

$$z_c = 0.61$$

Mahalanobis' $D^2 = 1.38$



Frequency histograms of Mississippian 'B' shale ratio, thickness and structural residuals for producing (P) and dry (D) wells. Mean values are indexed by a heavy dashed line; boundaries about the mean of one standard deviation are indicated by dotted lines.



Discriminant score frequency histograms for Mississippian 'B' dry and producing wells.

Testing for significance of discrimination,

$$F = 10.98$$

$$\nu_1 = 3 \text{ and } \nu_2 = 120$$

The critical value of F at $\nu_1 = 3$, $\nu_2 = 120$ and $\alpha = 0.05$ is 2.68

The null hypothesis that the two group multivariate means are equal is rejected provided the conditions of the test are accepted as approximately satisfied by the analytical data.

Contribution to discrimination by each variable as computed from their fractional component of Mahalanobis' distance are:

shale ratio, G:	39.6%
thickness, T:	52.6%
structural residual, S:	7.8%

The discriminant function analysis therefore suggests that the differentiation between dry and producing locales is made primarily in terms of thickness and shale content with relatively minor contribution by local structure. Producing wells are to be found in thick, "clean" chert sections and trapping contexts are stratigraphically, rather than structurally, controlled. An empirical measure of the success of the discriminant index as a critical classification boundary value between dry and producing wells may be made by tabulating the frequencies of correct classifications and misclassification.

DISCRIMINANT INDEX CLASSIFICATION

		Producer	Dry
Actual status	PRODUCER	23	10
	DRY	29	62

The discriminant function may be used to classify new wells that are drilled. Suppose a well has been drilled into the Mississippian 'B' and a gamma ray log has been run, but a decision has yet to be made whether to drill-stem test the interval. If the observations at the well are:

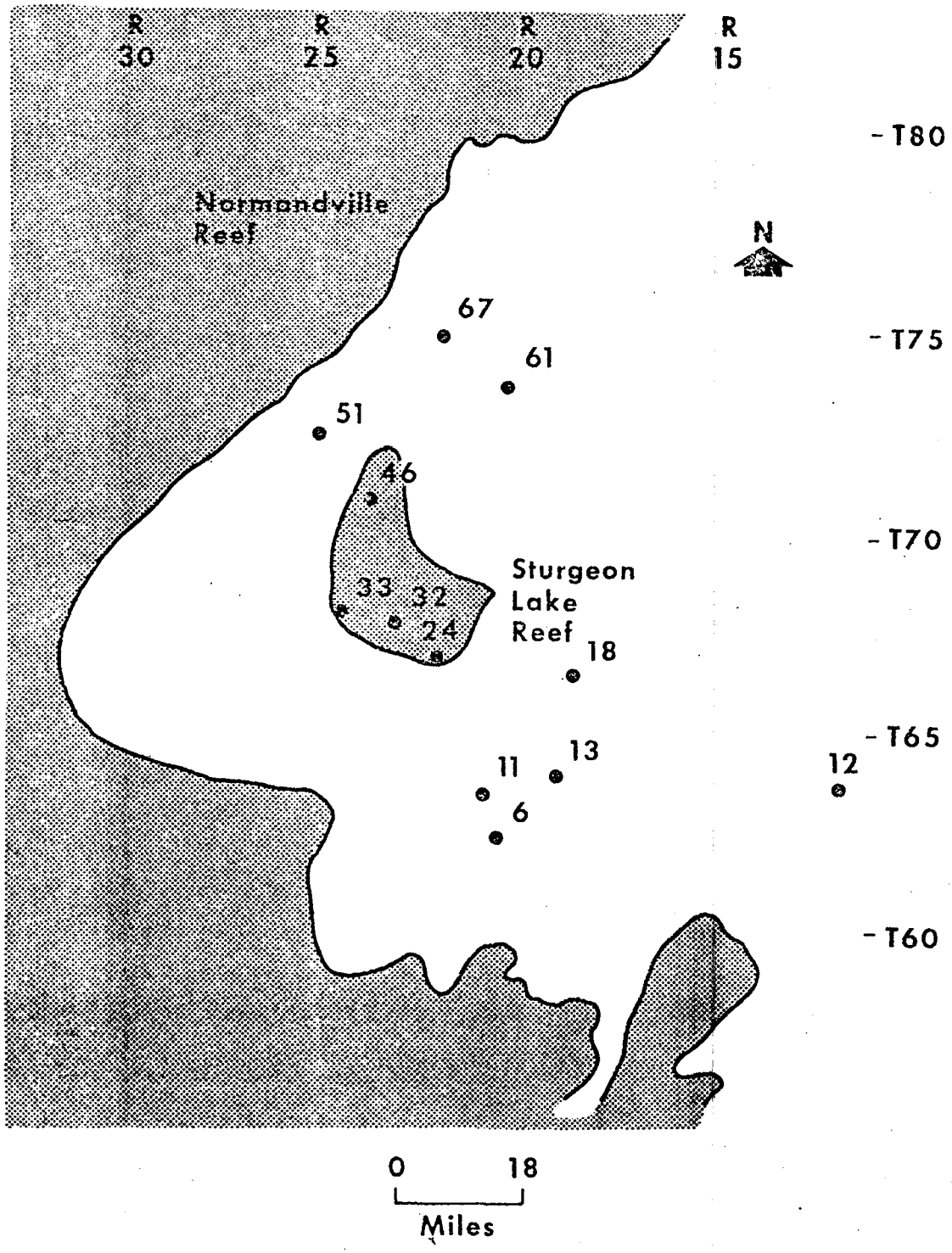
G	0.15
T	34
S	+20

The computed discriminant score is z = 1.23 and exceeds the discriminant index, classifying the new well with the group of producers. Naturally, the decision on the appropriate action to take is phrased in economics since the "cost" of misclassification of either group is not equal as is the assumption of the classical linear discriminant model. If a producing well is missed by misclassification as a dry well, the cost may be the loss of a major new field; if a dry well is tested due to a misleading classification as producer, the cost is that of a drill-stem test.

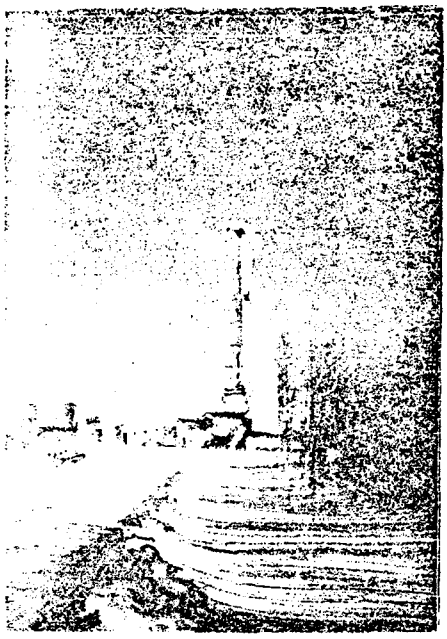
Exercise: The King Kong well crisis - reef or non-reef?

There is uproar at the exploration offices of King Kong Oil Inc. A prospect in Alberta, Canada, has just been drilled, hinged on the prognosis of an updip margin of a Leduc reef indicated on a seismic line. The well penetrated a Woodbend Group section that had anomalous characteristics. (The Woodbend Group contains the Leduc reefs and their lateral equivalent, the Ireton Shale). While clearly not a typical Leduc reef-core dolomite, the section is not a standard Ireton basinal facies of shales and thin limestones. Thick dolomites and limestones dominate the well section although there is a major content of shale. The well is circulating in the Woodbend, waiting on a decision from King Kong. Meanwhile, at King Kong offices, three highly vocal factions are debating an appropriate course of action at a series of hurriedly held meetings. Group A claims the section is reefoid in character but has penetrated a zone high on the talus slope fronting the reef and immediately adjacent to the reef core. They recommend that drilling be continued with a whipstock operation in a down-dip direction towards the reef (directional drilling from a position higher in the hole). Group B suggests that the well is farther removed from the reef core but has penetrated marginal reef facies, and recommends the drilling of an offset well in a down-dip direction. Group C believes that the well has penetrated an atypical basinal facies and is at some distance from the nearest reef. Their recommended action is to both abandon the well and drop the drilling lease in the area.

Woodbend Group cuttings from the well have been geochemically analyzed for trace element content of the non-detrital (acid soluble) fraction. The



Location of wells in Sturgeon Lake area with trace element analyses of the Woodbend Group (after Chester, 1965).



Well	ID	Ni	Cr	Cu	Pb	Ga
REEF						
Gulf Little Smoky	24	23	3.6	18	29.0	16.0
Amerada Crown XF	33	13	3.4	24	8.6	18.0
Amerada Crown UF	46	8	0.9	18	11.0	4.7
Amerada Crown UJ	32	51	6.6	19	4.7	9.8
NON-REEF						
Pan Am Tony Creek A1	11	82	6.1	26	34.0	20.0
Phillips Tony A	6	45	4.9	18	30.0	16.0
Phillips Kaybob A	13	19	3.8	15	17.0	7.0
Imperial Burntwood	12	16	4.0	14	19.0	13.0
B.A. Crooked Creek No. 1	51	47	7.8	15	21.0	12.0
Imperial Little Smoky No. 1	67	27	6.5	20	19.0	17.0
B.A. Little Smoky	61	31	6.4	27	32.0	15.0
Gulf Goose River	18	32	8.1	23	17.0	11.0

Average non-detrital (acid-soluble) trace element content (ppm) of well sections in the Woodbend Group (Devonian) of the Sturgen Lake area, Alberta, Canada (from Chester, 1965)

average ppm content of analyzed elements within the well section are:

Nickel	32
Chromium	42
Copper	20
Lead	14
Gallium	15

A set of analyses of averages of equivalent non-detrital trace element contents is available for known reef and non-reef wells in the Woodbend Group in the Sturgeon Lake area (Chester, 1965) which is at no great distance from the King Kong well. This data is listed in the table.

(1) Compute a linear discriminant function, using the trace element concentrations, that best distinguishes reef from non-reef wells in the area.

(2) Calculate Mahalanobis' distance and make an F test of the null hypothesis that there is no discrimination between reef and non-reef wells.

(3) Compute and plot the z-scores of the Sturgeon Lake sample, together with the discriminant index.

(4) Calculate the 'potency' of each element in the total discrimination. Which are the most potent elements?

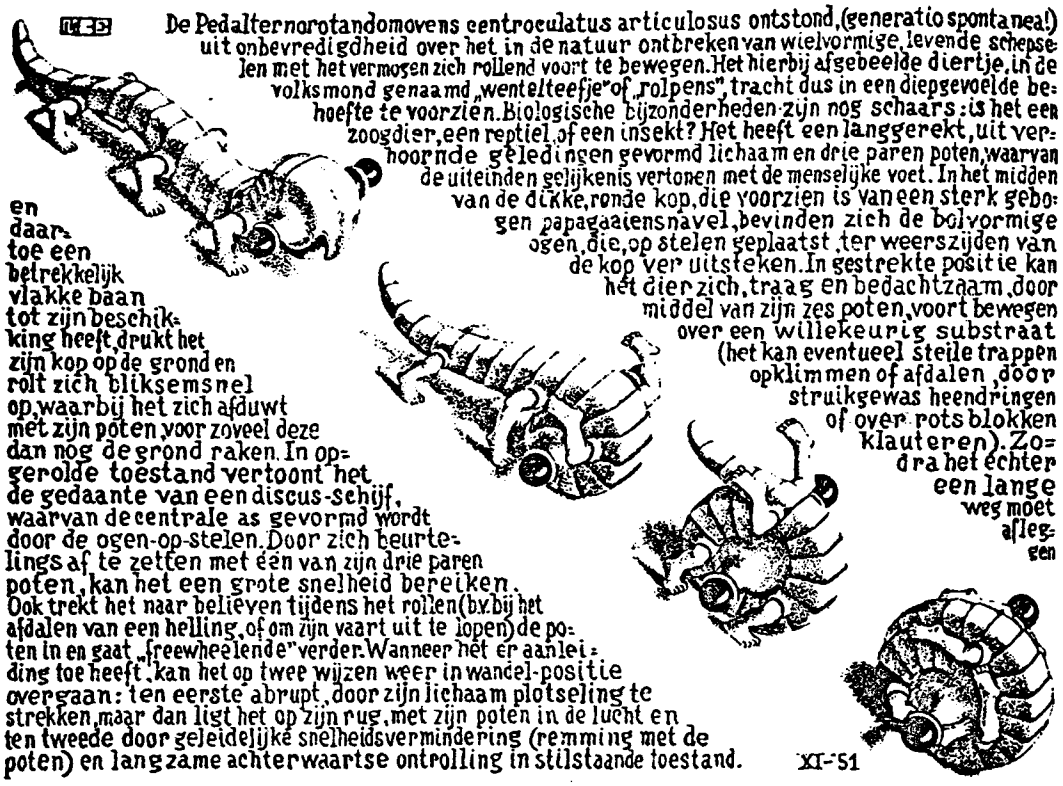
(5) Calculate a discriminant score for the King Kong well analysis and classify the well.

(6) If you were a consultant for King Kong what would be your interpretation of the well sample based on its geochemistry, and what course of action would you recommend?

Reference

Chester, R., 1965, Geochemical criteria for differentiating reef from non-reef facies in carbonate rocks: Am. Assoc. Petroleum Geologists Bull., v. 49, no. 3, p. 258-276.

35. NUMERICAL TAXONOMY



The Pedalternorotandomovens centroculatus articulosus came into being (generatio spontanea!) as a result of dissatisfaction concerning nature's lack of any wheelshaped living creatures endowed with a power of propulsion by means of rolling themselves up.....

- M.C. Escher (1951)

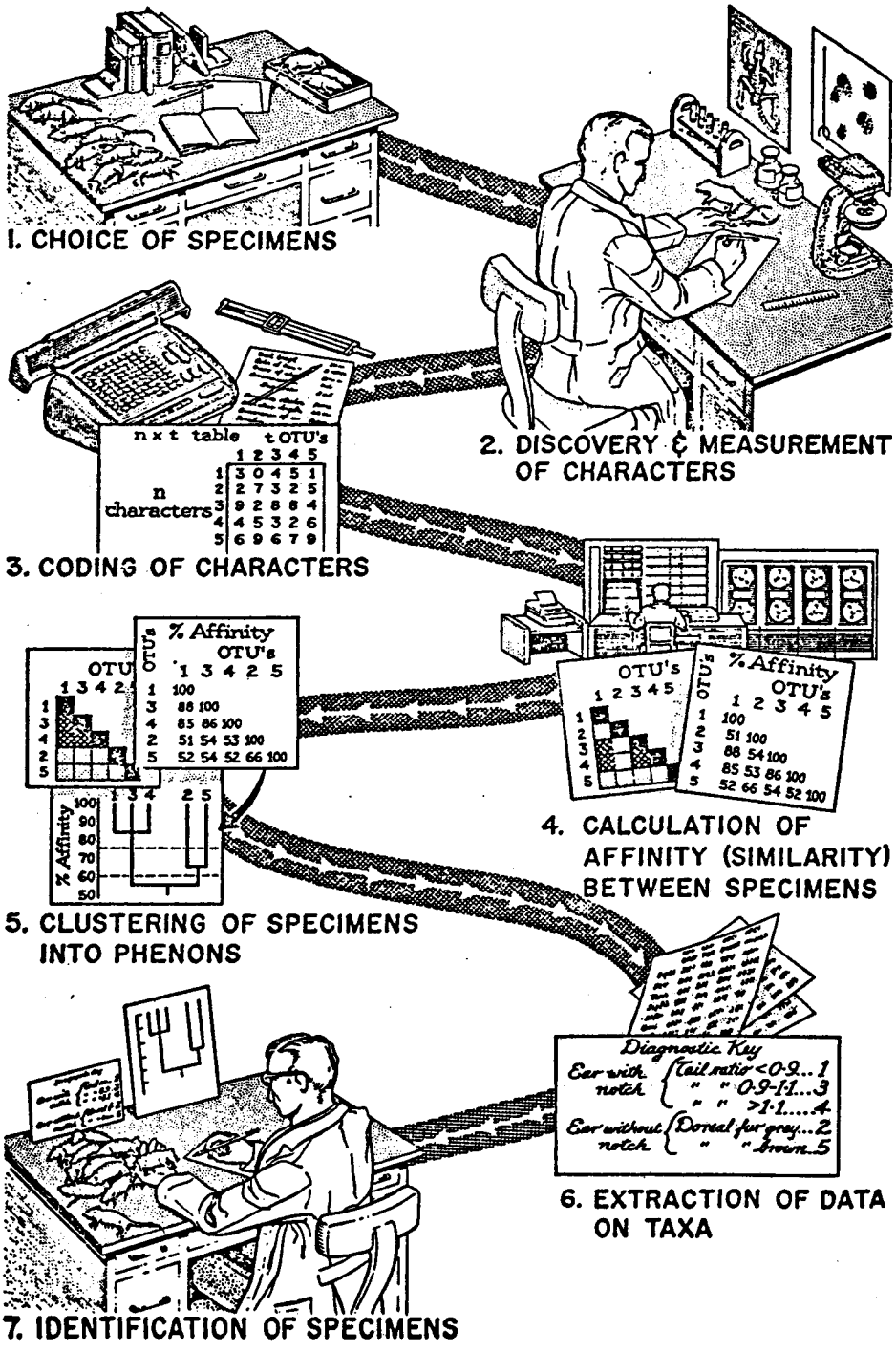
Classification is a central problem in both paleontology and paleoecology. Taxonomic studies of morphological characters of fossils or living forms are directed to finding an arrangement of specimens that describe a "natural" system as a reflection of the evolutionary process. Similarly, observations of the joint occurrences of fossil life-forms are the subject of bioassociation studies where the aim is to define groups that correspond to ancient communities of cohabiting species.

The operation of classification can only be undertaken following a formal definition of criteria of similarity or dissimilarity that govern the objects to be classified. Most "realistic" classifications are

polythetic (based on several observable properties) and so the definition of taxa (classification groups) is usually based on similarity measures drawn from a number of observational attributes. The classification problem is therefore multivariate by nature but is distinct from conventional multivariate statistics. Attention is focused on each measured individual rather than regarding the individual as an anomymous member of a population. Consequently, classification procedures are inherently Q-mode techniques which lack the analytical rigor of the R-mode approach.

The physical end-product of a classification is a spatial representation of the objects arranged according to their mutual relationships, conventionally either as a dendrogram ('tree structure') or as a mapping in two-dimensional space. In either case, the conceptual space of the classification is usually specified as Euclidean with primary orthogonal axes and "standard" trigonometric properties. The distance between each object must therefore be condensed to a scalar that reflects the spatial distance either as a taxonomic distance measure or (more commonly) as a similarity coefficient. Distance measures are generally used in situations where the observation attributes are continuous measurements (interval or ratio scaled) and so the objects already have an implicit representation in continuous m-dimensional space. However, many classification problems involve objects where distinctions are made in terms of the presence or absence of attributes. The fundamental is therefore discrete but may be transformed into an approximately continuous variable by some arithmetic combination of the binary codings of a suite of attributes. Computation is usually made of a ratio constrained between the limits of zero (no similarity) and one (exact similarity). As the ratio is a Q-mode statistic there are no formal theoretical guidelines to its definition. Cheetham and Hazel (1969) report 26 commonly used coefficients of similarity all of which "appear to have been developed intuitively and tested empirically." The reason for this apparent anarchy is that, unlike the R-mode correlation coefficient, similarity coefficients are not estimates of population parameters. Instead, they are dictated by the conceptual model of object interrelation-

A FLOW CHART OF NUMERICAL TAXONOMY



-from Sokal and Sneath (1963)

ships rather than their abstract mathematical properties. A selection of five of the more popular coefficients are listed in the table.

Some simple coefficients of similarity or association

NAME	SYMBOL	FORMULA
Jaccard	S_J	C/N
Simple Matching	S_{SM}	$(C+A)/(N+A)$
Dice	S_D	$2C/(N_1+N_2)$
Otsuka		$C/\sqrt{N_1N_2}$
Simpson		C/N_1

Key to symbols:

- C = Number of attributes matched as present in both objects
- A = Number of attributes matched as absent in both objects
- N_1 = Number of attributes coded as present in the first object
- N_2 = Number of attributes coded as present in the second object
- N = Total number of attributes coded as present in both objects (N_1+N_2-C)

The objects to be classified are conventionally referred to as operational taxonomic units (or OTUs). Tbe computation of similarity coefficients between n OTUs results in a symmetric n X n matrix with ones on the leading diagonal and fractional quantities in the off-diagonal cells. The obvious parallel between this matrix and an R-mode correlation matrix allows the use of ersatz R-mode techniques for the purposes of data condensation and spatial representation (particularly Q-mode factor analysis). This approach is purely an algebraic and geometrical exercise outside the

decision procedures and probability structure of R-mode. A basic computational difficulty arises in that most classifications involve a large number of OTUs and, consequently, a large coefficient matrix. The calculation of the appropriate inverse matrices and eigen properties may become dauntingly prohibitive.

An alternative approach is the use of one of the many methods of cluster analysis (q.v.) which partition the structure of the similarity coefficient matrix as an hierarchical scheme of linkages, most commonly represented by a dendrogram. A flow diagram illustrating a hypothetical study of this type is shown in the figure. The various types of cluster analysis are the most widely used techniques in numerical taxonomy and are described at length in texts such as Sneath and Sokal (1973) and Jardine and Sibson (1971).

More recently, the collection of methods known as non-metric multidimensional scaling (MDS) have been applied with great success to classification problems. MDS attempts to rank objects in multi-dimensional space in terms of the ordering implied by their collective attribute characters rather than to strictly honor their metric coefficient distances. Since the metric properties of similarity coefficients are more apparent than real, the application of MDS is truer to their ordinal nature and is particularly robust since it involves the use of a monotonic, rather than metric, function.

References:

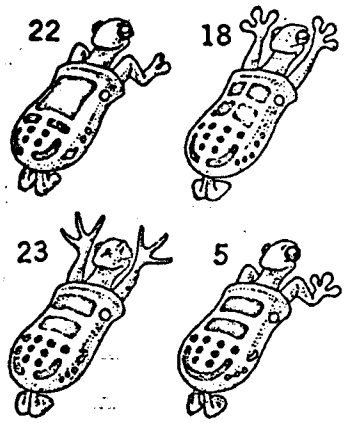
Cheetam, A.H., and Hazel, J.E., 1969, Binary (presence-absence) similarity coefficients: Jour. Paleont., v. 43, no. 5, p. 1130-1136.

Jardine, N., and Sibson, R., 1971, Mathematical Taxonomy: Wiley & Sons, Inc., New York, 286 p.

Sneath, P.H.A., and Sokal, R.R., 1973, Numerical Taxonomy: W.H. Freeman, San Francisco, 573 p.

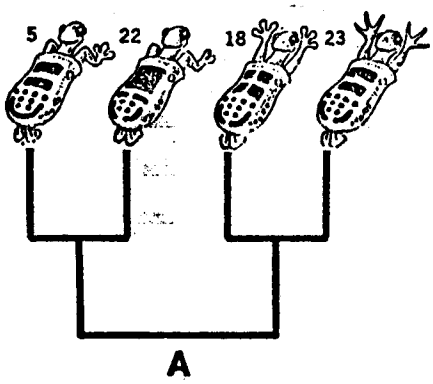
Sokal, R.R., and Sneath, P.H.A., 1963, Principles of Numerical Taxonomy: W.Y. Freeman, San Francisco, 359 p.

Exercise 35.1: Classification of Caminalcules

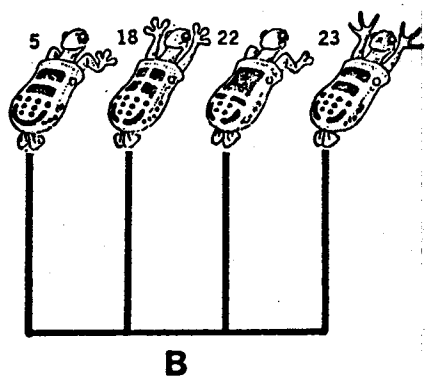


Caminalcules are imaginary animals that were devised by J.H. Camin.

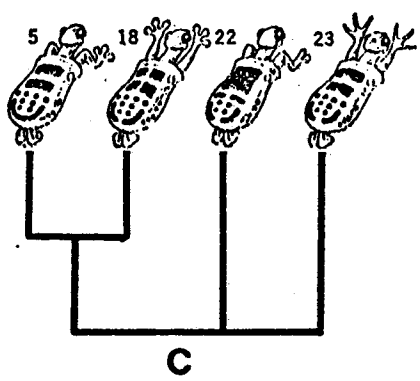
They provide a fundamental data set that enables both the study of criteria used by classical taxonomists and the characteristics of numerical taxonomic methods. The artifical nature of the data largely precludes the intrusion of genetic preconceptions. Sokal (1974) reports the use of twenty-nine Caminalcules in a study of individual differences in taxonomic judgement. Four of the organisms were classified by a distinguished systematic entomologist (A), an invertebrate peleotologist (B) and a graduate student in paleotology (C) in the manner shown. On the basis of the presence or absence of various features, compute similarity coefficients (either Jaccard, Dice or other coefficient) between the four Caminalcules. Draw a simple dendogram that is an approximate representation of the similarity coefficient matrix structure. Apply a simple cluster analysis technique (q.v.) to partition the matrix and generate a numerical classification of the Caminalcules.



A



B



C

References

Camin, J.H. and Sokal, R.R., 1965, A method for deducing branching sequences in phylogeny: Evolution, v. 19, p. 311-326.

Sokal, R.R., 1974, Classification: Purposes, Principles, Progress, Prospects: Science, v. 185, p. 1115-1123.

36. CLUSTER ANALYSIS

Cluster analysis is the name given to a bewildering assortment of techniques designed to assign observations to groups so each group is more-or-less homogeneous and distinct from other groups. There is no analytical solution to this problem, which is general to all areas of classification. Although there are alternative classifications of classification procedures (see Sneath and Sokal, 1973), most may be grouped into four general types.

(1) Partitioning methods operate on the multivariate observations themselves, or on projections of these observations. Basically, these cluster by finding regions in m-dimensional space which are poorly populated with observations, and which separate densely populated regions. Although the analysis is done in m-dimensional variable space, it proceeds by trial-and-error and may be extremely expensive (Switzer, 1970).

(2) Arbitrary origin methods operate on the similarity between the observations and a set of arbitrary points. If n points are to be classified into k groups, it is necessary to compute an asymmetric n $\times$ k matrix of similarity between the k arbitrary points which serve as group centroids and n samples. Samples are iteratively placed in the nearest group, whose centroid is then recalculated for the expanded cluster.

(3) Mutual similarity procedures group together observations which have a common similarity to other observations. First an n $\times$ n matrix of similarities between all pairs of observations is calculated. Then the similarity between columns of this matrix is iteratively recomputed. Members of a cluster will tend to show intercorrelations near +1, while having much lower correlations with nonmembers.

(4) Hierarchic clustering joins the most similar observations, then successively connects the next most similar observations to these. First an n $\times$ n matrix of similarities between all pairs of observations is calculated. Those pairs having the highest similarities are then merged, and the matrix recomputed.

Hierarchic techniques are the most widely applied in the earth sciences, probably because the development of these methods has been closely linked with the numerical taxonomy of fossil organisms. A

large number of variations exist on the basic procedure. First, an arbitrary similarity measure (q.v.) must be calculated between all pairs of observations, yielding an n X n symmetric matrix. Next, observations are "joined" on the basis of close similarity, and their corresponding rows and columns in the similarity matrix are combined.

Note is made of the similarity level at which rows and columns are merged. The similarity matrix is then recomputed by averaging the correlations which the combined rows and columns have with the other rows and columns in the matrix. The process iterates until the matrix is reduced to 2 X 2. The table of levels at which joins occur is used to construct a dendrogram, or tree diagram in which the branches represent observations. The "forks" of the tree diagram represent levels of similarity at which the observations are fused.

Variations in hierarchic clustering

There are different criteria for joining an object to a cluster, or for joining two clusters together. The most common variant (all of which are arbitrary procedures) include:

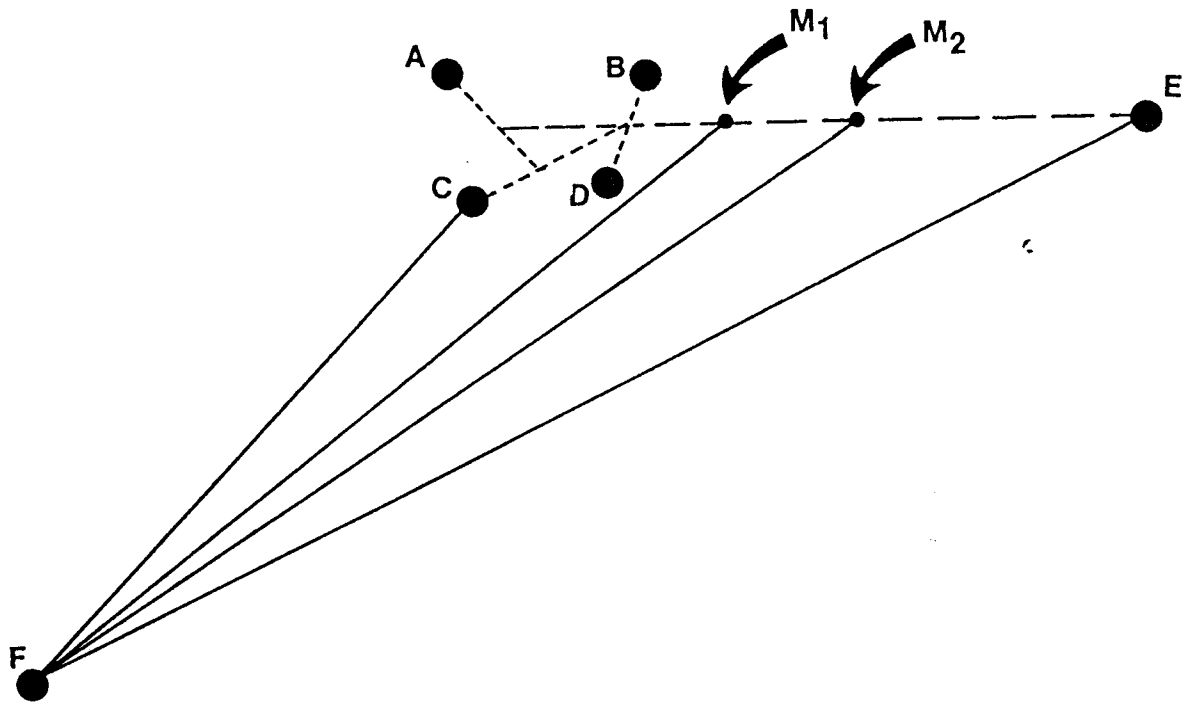
(1) Single linkage. The candidate object or cluster has its greatest similarity with the "nearest" member of the cluster which it will join. Linkage occurs at this level of similarity. This is also called "nearest neighbor" linkage.

(2) Complete linkage. The candidate object or cluster has a greater similarity with the most dissimilar member of the cluster which it will join than with the most dissimilar members of other clusters. Linkage occurs at the level of lowest similarity with the cluster. This criterion is also called "farthest neighbor" linkage.

(3) Average linkage. The candidate object or cluster joins the cluster with which it has the greatest average similarity with the members of the cluster. Linkage occurs at the level of average similarity, without regard to the number of objects in the groups which are contributing to the average.

(4) Centroid linkage. The candidate object or cluster is joined to the cluster whose centroid is "nearest."

The relative effect of the four different linkage strategies is shown below. Here, a cluster is composed of tightly groups objects A, B, C, and D. A more distant object E has just been admitted to the cluster. F is a candidate to join the group. In single linkage, it will join if the distance CF is smaller than the distance to any other object in another cluster. In centroid linkage, it will join if the distance M_1F is smaller than the distance to the centroid of any other group. In average linkage, the candidate will join if the distance M_2F is smaller than the distance to the average of any other cluster. (Note that M_2 is half-way between the average of the cluster ABCD and object E, admitted in the previous cycle. Finally, with complete linkage, object F will join the cluster if EF is less than the distance to the most distant point in other clusters.



A fundamental drawback to hierarchic techniques is the inordinate size of the matrix which must be manipulated if the number of objects is large. Clustering procedures using arbitrary cluster centers were devised to offset this computational difficulty. Probably the most widely used of these is the k-means procedure of McQueen (1967).

Here, k points in m-dimensional space are designated (either by the operator or arbitrarily by the program) as centroids of clusters. A matrix of similarities between the k "centroids" and the n observations is calculated, and the closest points are clustered with the nearest centroid. A new centroid is then calculated and the process iterates exactly like a hierarchial procedure. In principal, the centroid will rapidly move toward the true center of the local group. The advantage is that only a k X n matrix is necessary, and if k is small (5 to 10) and n is large (1000 or more), the process may be faster than a hierarchical method by two orders of magnitude or more. The disadvantage is that a suboptimal clustering may result if the arbitrary starting points do not fall within divergent clusters, leading to premature merger of the centroids and failure to detect outlying clusters.

References

McQueen, J., 1967, Some method for classification and analysis of multivariate observations: *in* LeCam, L.M., and Neyman, J. (eds.), Proc. 5th Berkeley Symp. on Math. Statistics and Probability, vol. 1, p. 281-297, Univ. California Press, Berkeley, Calif.

Sneath, P.H.A., and Sokal, R.R., 1973, Numerical Taxonomy: W.H. Freeman & Co., San Francisco, 573 p.

TABLE 3.8. Cumulative Probabilities for the Standardized Normal Distribution

Standard deviations from the mean	Cumulative probability	Standard deviations from the mean	Cumulative probability
−3.0	0.0014	+0.0	0.5000
−2.9	0.0019	+0.1	0.5398
−2.8	0.0026	+0.2	0.5793
−2.7	0.0035	+0.3	0.6179
−2.6	0.0047	+0.4	0.6554
−2.5	0.0062	+0.5	0.6915
−2.4	0.0082	+0.6	0.7257
−2.3	0.0107	+0.7	0.7580
−2.2	0.0139	+0.8	0.7881
−2.1	0.0179	+0.9	0.8159
−2.0	0.0228	+1.0	0.8413
−1.9	0.0287	+1.1	0.8643
−1.8	0.0359	+1.2	0.8849
−1.7	0.0446	+1.3	0.9032
−1.6	0.0548	+1.4	0.9192
−1.5	0.0668	+1.5	0.9332
−1.4	0.0808	+1.6	0.9452
−1.3	0.0968	+1.7	0.9554
−1.2	0.1151	+1.8	0.9641
−1.1	0.1357	+1.9	0.9713
−1.0	0.1587	+2.0	0.9773
−0.9	0.1841	+2.1	0.9821
−0.8	0.2119	+2.2	0.9861
−0.7	0.2420	+2.3	0.9893
−0.6	0.2743	+2.4	0.9918
−0.5	0.3085	+2.5	0.9938
−0.4	0.3446	+2.6	0.9953
−0.3	0.3821	+2.7	0.9965
−0.2	0.4207	+2.8	0.9974
−0.1	0.4602	+2.9	0.9981
−0.0	0.5000	+3.0	0.9987

Source: Abridged from Table II, A. Hald, *Statistical Tables and Formulas*, John Wiley & Sons, Inc., New York, 1952.

TABLE 3.9. Critical Values of t for ν Degrees of Freedom and Selected Levels of Significance

Number of degrees of freedom, ν	Significance level, α (%)					
	10	5	2.5	1	0.5	0.1
1	3·078	6·314	12·706	31·821	63·657	318·310
2	1·886	2·920	4·303	6·965	9·925	22·327
3	1·638	2·353	3·182	4·541	5·841	10·215
4	1·533	2·132	2·776	3·747	4·604	7·173
5	1·476	2·015	2·571	3·365	4·032	5·893
6	1·440	1·943	2·447	3·143	3·707	5·208
7	1·415	1·895	2·365	2·998	3·499	4·785
8	1·397	1·860	2·306	2·896	3·355	4·501
9	1·383	1·833	2·262	2·821	3·250	4·297
10	1·372	1·812	2·228	2·764	3·169	4·144
11	1·363	1·796	2·201	2·718	3·106	4·025
12	1·356	1·782	2·179	2·681	3·055	3·930
13	1·350	1·771	2·160	2·650	3·012	3·852
14	1·345	1·761	2·145	2·624	2·977	3·787
15	1·341	1·753	2·131	2·602	2·947	3·733
16	1·337	1·746	2·120	2·583	2·921	3·686
17	1·333	1·740	2·110	2·567	2·898	3·646
18	1·330	1·734	2·101	2·552	2·878	3·610
19	1·328	1·729	2·093	2·539	2·861	3·579
20	1·325	1·725	2·086	2·528	2·845	3·552
21	1·323	1·721	2·080	2·518	2·831	3·527
22	1·321	1·717	2·074	2·508	2·819	3·505
23	1·319	1·714	2·069	2·500	2·807	3·485
24	1·318	1·711	2·064	2·492	2·797	3·467
25	1·316	1·708	2·060	2·485	2·787	3·450
26	1·315	1·706	2·056	2·479	2·779	3·435
27	1·314	1·703	2·052	2·473	2·771	3·421
28	1·313	1·701	2·048	2·467	2·763	3·408
29	1·311	1·699	2·045	2·462	2·756	3·396
30	1·310	1·697	2·042	2·457	2·750	3·385
40	1·303	1·684	2·021	2·423	2·704	3·307
60	1·296	1·671	2·000	2·390	2·660	3·232
120	1·289	1·658	1·980	2·358	2·617	3·160
∞	1·282	1·645	1·960	2·326	2·576	3·090

Source: From Table 21, *The Penguin-Honeywell Book of Tables*, copyright F. W. Kellaway (ed.) and Honeywell Controls Ltd. (E.D.P. Division), 1968.

TABLE 3.12a. Critical Values of F for ν_1 and ν_2 Degrees of Freedom and 5% ($\alpha = 0.05$) Level of Significance

Degrees of freedom for numerator, ν_1

Degrees of freedom for denominator, ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	∞
1	161·45	199·50	215·71	224·58	230·16	233·99	236·77	238·88	240·54	241·88	243·91	245·95	248·01	249·05	250·10
2	18·51	19·00	19·16	19·25	19·30	19·33	19·35	19·37	19·38	19·40	19·41	19·43	19·45	19·45	19·46
3	10·13	9·55	9·28	9·12	9·01	8·94	8·89	8·85	8·81	8·79	8·74	8·70	8·66	8·64	8·62
4	7·71	6·94	6·59	6·39	6·26	6·16	6·09	6·04	6·00	5·96	5·91	5·86	5·80	5·77	5·75
5	6·61	5·79	5·41	5·19	5·05	4·95	4·88	4·82	4·77	4·74	4·68	4·62	4·56	4·53	4·50
6	5·99	5·14	4·76	4·53	4·39	4·28	4·21	4·15	4·10	4·06	4·00	3·94	3·87	3·84	3·81
7	5·59	4·74	4·35	4·12	3·97	3·87	3·79	3·73	3·68	3·64	3·57	3·51	3·44	3·41	3·38
8	5·32	4·46	4·07	3·84	3·69	3·58	3·50	3·44	3·39	3·35	3·28	3·22	3·15	3·12	3·08
9	5·12	4·26	3·86	3·63	3·48	3·37	3·29	3·23	3·18	3·14	3·07	3·01	2·94	2·90	2·86
10	4·96	4·10	3·71	3·48	3·33	3·22	3·14	3·07	3·02	2·98	2·91	2·84	2·77	2·74	2·70
11	4·84	3·98	3·59	3·36	3·20	3·09	3·01	2·95	2·90	2·85	2·79	2·72	2·65	2·61	2·57
12	4·75	3·89	3·49	3·26	3·11	3·00	2·91	2·85	2·80	2·75	2·69	2·62	2·54	2·51	2·47
13	4·67	3·81	3·41	3·18	3·03	2·92	2·83	2·77	2·71	2·67	2·60	2·53	2·46	2·42	2·38
14	4·60	3·74	3·34	3·11	2·96	2·85	2·76	2·70	2·65	2·60	2·53	2·46	2·39	2·35	2·31
15	4·54	3·68	3·29	3·06	2·90	2·79	2·71	2·64	2·59	2·54	2·48	2·40	2·33	2·29	2·25
16	4·49	3·63	3·24	3·01	2·85	2·74	2·66	2·59	2·54	2·49	2·42	2·35	2·28	2·24	2·19
17	4·45	3·59	3·20	2·96	2·81	2·70	2·61	2·55	2·49	2·45	2·38	2·31	2·23	2·19	2·15
18	4·41	3·55	3·16	2·93	2·77	2·66	2·58	2·51	2·46	2·41	2·34	2·27	2·19	2·15	2·11
19	4·38	3·52	3·13	2·90	2·74	2·63	2·54	2·48	2·42	2·38	2·31	2·23	2·16	2·11	2·07
20	4·35	3·49	3·10	2·87	2·71	2·60	2·51	2·45	2·39	2·35	2·28	2·20	2·12	2·08	2·04
21	4·32	3·47	3·07	2·84	2·68	2·57	2·49	2·42	2·37	2·32	2·25	2·18	2·10	2·05	2·01
22	4·30	3·44	3·05	2·82	2·66	2·55	2·46	2·40	2·34	2·30	2·23	2·15	2·07	2·03	1·98
23	4·28	3·42	3·03	2·80	2·64	2·53	2·44	2·37	2·32	2·27	2·20	2·13	2·05	2·01	1·96
24	4·26	3·40	3·01	2·78	2·62	2·51	2·42	2·36	2·30	2·25	2·18	2·11	2·03	1·98	1·94
25	4·24	3·39	2·99	2·76	2·60	2·49	2·40	2·34	2·28	2·24	2·16	2·09	2·01	1·96	1·92
26	4·23	3·37	2·98	2·74	2·59	2·47	2·39	2·32	2·27	2·22	2·15	2·07	1·99	1·95	1·90
27	4·21	3·35	2·96	2·73	2·57	2·46	2·37	2·31	2·25	2·20	2·13	2·06	1·97	1·93	1·88
28	4·20	3·34	2·95	2·71	2·56	2·45	2·36	2·29	2·24	2·19	2·12	2·04	1·96	1·91	1·87
29	4·18	3·33	2·93	2·70	2·55	2·43	2·35	2·28	2·22	2·18	2·10	2·03	1·94	1·90	1·85
30	4·17	3·32	2·92	2·69	2·53	2·42	2·33	2·27	2·21	2·16	2·09	2·01	1·93	1·89	1·84
40	4·08	3·23	2·84	2·61	2·45	2·34	2·25	2·18	2·12	2·08	2·00	1·92	1·84	1·79	1·74
60	4·00	3·15	2·76	2·53	2·37	2·25	2·17	2·10	2·04	1·99	1·92	1·84	1·75	1·70	1·65
120	3·92	3·07	2·68	2·45	2·29	2·18	2·09	2·02	1·96	1·91	1·83	1·75	1·66	1·61	1·55
∞	3·84	3·00	2·60	2·37	2·21	2·10	2·01	1·94	1·88	1·83	1·75	1·67	1·57	1·52	1·46

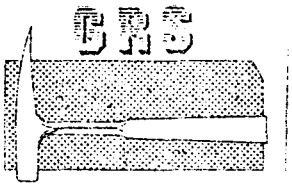
Source: From Table 22, *The Penguin-Honeywell Book of Tables*, copyright F. W. Kellaway (ed.) and Honeywell Controls Ltd. (E.D.P. Division), 1968.

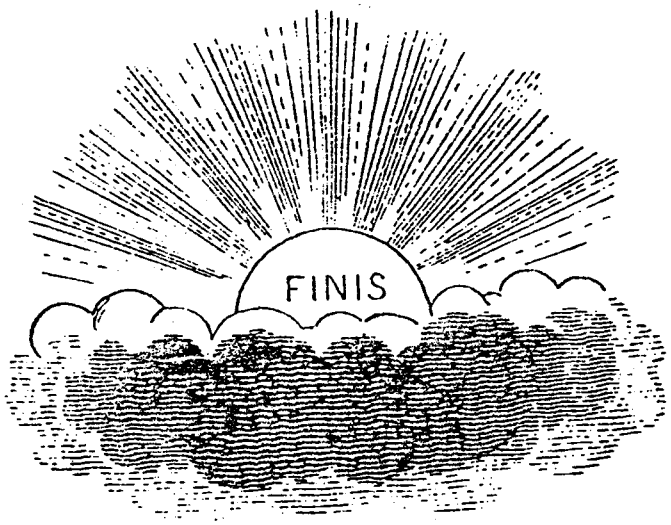
TABLE 3.16. Critical Values of χ^2 for ν Degrees of Freedom and Selected Levels of Significance

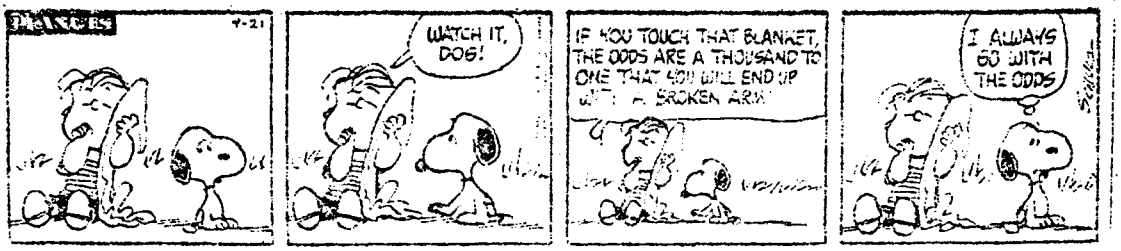
Significance level, α (%)

Number of degrees of freedom, ν	20	10	5	2.5	1
1	1·64	2·71	3·84	5·02	6·63
2	3·22	4·61	5·99	7·38	9·21
3	4·64	6·25	7·81	9·35	11·34
4	5·99	7·78	9·49	11·14	13·28
5	7·29	9·24	11·07	12·83	15·09
6	8·56	10·64	12·59	14·45	16·81
7	9·80	12·02	14·07	16·01	18·48
8	11·03	13·36	15·51	17·53	20·09
9	12·24	14·68	16·92	19·02	21·67
10	13·44	15·99	18·31	20·48	23·21
11	14·63	17·28	19·68	21·92	24·72
12	15·81	18·55	21·03	23·34	26·22
13	16·98	19·81	22·36	24·74	27·69
14	18·15	21·06	23·68	26·12	29·14
15	19·31	22·31	25·00	27·49	30·58
16	20·47	23·54	26·30	28·85	32·00
17	21·61	24·77	27·59	30·19	33·41
18	22·76	25·99	28·87	31·53	34·81
19	23·90	27·20	30·14	32·85	36·19
20	25·04	28·41	31·41	34·17	37·57
21	26·17	29·62	32·67	35·48	38·93
22	27·30	30·81	33·92	36·78	40·29
23	28·43	32·01	35·17	38·08	41·64
24	29·55	33·20	36·42	39·36	42·98
25	30·68	34·38	37·65	40·65	44·31
26	31·79	35·56	38·89	41·92	45·64
27	32·91	36·74	40·11	43·19	46·96
28	34·03	37·92	41·34	44·46	48·28
29	35·14	39·09	42·56	45·72	49·59
30	36·25	40·26	43·77	46·98	50·89
40	47·27	51·81	55·76	59·34	63·69
50	58·16	63·17	67·50	71·42	76·15
60	68·97	74·40	79·08	83·30	88·38
70	79·71	85·53	90·53	95·02	100·43
80	90·41	96·58	101·88	106·63	112·33
90	101·05	107·57	113·15	118·14	124·12
100	111·67	118·50	124·34	129·56	135·81

Source: Abridged from Table 24, *The Penguin-Honeywell Book of Tables*, copyright F. W. Kellaway (ed.) and Honeywell Controls Ltd. (E.D.P. Division), 1968.







THESE NOTES WERE PREPARED WITH THE FINANCIAL ASSISTANCE OF THE KANSAS GEOLOGICAL SURVEY