

R102

GEOSTATISTICS

Lecture notes

W. Nemec

Revised Third Edition



Department of Earth Science
University of Bergen
2011

CONTENTS

1. Introduction
1.1. Why do we need statistics in geology?
1.2. Geostatistics and the scope of this course
2. Basic terminology 1
2.1. Variables - their types and scales 1
2.2. Sample population and general population; population size and dimension 1
3. What is a 'representative' sample population? 2
4. Estimation of an optimal size of sample population 3
5. Frequency distribution 5
5.1. Non-grouped and grouped data; frequency plots 5
5.2. Some properties of normal distribution 8
5.3. Common distribution types 9
5.4. Exceedence frequency plots 12
5.5. The significance of power-law distribution (fractal statistics) 14
5.6. Strategy for using parametric and non-parametric methods 16
6. Basic parameters and their estimators (statistics) 17
6.1. Arithmetic mean, variance and standard deviation 17
6.2. Other basic statistics: median, mode, coefficient of skewness, geometric mean, harmonic mean, standard deviation of the mean, and coefficient of variation 18
6.3. Simple solutions for simple problems: examples 21
7. Confidence intervals for basic parameters 23
8. Confidence interval for mean proportion 24
9. Analysis of univariate populations 25
9.1. Parametric comparison of two populations 25
9.1.1. The Snedecor F-test for the difference between two variances 26
9.1.2. The Student t-test for the difference between two means 26
9.1.3. The t-test for conformity between population mean and assumed mean 27
9.1.4. The chi-square test for the difference between population variance and assumed variance 28
9.1.5. Confidence interval for the difference between two means 28
9.1.6. The Tukey test for the difference between two means 29
9.1.7. Confidence interval for the difference between two variances 29
9.1.8. Confidence interval for the difference between two proportions 29
9.2. Normality tests 30
9.2.1. The Cramer-von Mises normality test for non-grouped data 30
9.2.2. The Pearson normality test for grouped data 31
9.3. Nonparametric comparison of two populations 32
9.3.1. Runs test (U-test) 32
9.3.2. The Mann-Whitney test (rank sum U-test) 33
9.3.3. The Kolmogorov-Smirnov test (D-test) for grouped data 34
9.3.4. The chi-square test for contingency table of qualitative variable 34
9.4. Simultaneous comparison of several populations 35
9.4.1. The Bartlett chi-square test for homogeneity of variances 35
9.4.2. The Tukey-Cramer test for homogeneity of means 36
9.4.3. One-way ANOVA as a parametric test for homogeneity of means 37
9.4.4. Two-way and three-way ANOVA 38
9.4.5. The Kruskal-Wallis test (nonparametric rank sum H-test) 41
10. Analysis of vectorial populations 42
10.1. The construction of circular frequency histogram (rose diagram) 42
10.2. Analysis of two-dimensional (azimuthal) populations 45
10.2.1. The Kuiper test of distribution uniformity 45
10.2.3. The Watson test of distribution uniformity 46
10.2.4. The chi-square test of distribution uniformity for grouped data 46

10.2.5. Parameters and statistics of population with circular-normal distribution: the mean vector direction and length, vectorial concentration, and the Batschelet cricular variance and standard deviation 49
10.2.6. The Rayleigh test of distribution uniformity (parametric) 50
10.2.7. Evaluation of sample population and choice of analytical strategy 50
10.2.8. Confidence interval (sector) for the mean vector direction 51
10.3. Comparison of two vectorial populations 52
10.3.1. The t-test for the difference between two mean vectors 52
10.3.2. The F-test for the difference between two mean vectors 52
10.3.3. Test for the difference between two vectorial concentrations 52
10.3.4. Nonparametric test for the difference between two vectorial populations 53
10.3.5. The F-test for simultaneous comparison of several vectorial populations 54
10.4. Optimization of sampling strategy for a vectorial population 54
10.5. Analysis of three-dimensional (spherical) vectorial data 56
10.5.1. Numerical representation of geological strike/dip measurements 56
10.5.2. The chi-square test for spherical distribution uniformity 57
10.5.3. Parameters and statistics: mean vector direction and length, and vectorial concentration 59
10.5.4. Confidence interval (cone) for the mean vector 60
10.6. Comparison of two spherical vectorial populations 60
10.6.1. The F-test for the differences between two mean vectors 60
10.6.2. The F-test for the difference between two vectorial concentrations 61
11. Analysis of bivariate populations 62
11.1. Relationship between two quantitative variables: linear correlation and regression 62
11.1.1. Covariance and Pearson's correlation coefficient 62
11.1.2. The Fisher t-test for the significance of linear correlation 62
11.1.3. Solution to the problem of data closure 63
11.1.4. Straight-linear regression 64
11.1.5. The Fisher t-test for the significance of regression 64
11.1.6. The Fisher t-test for the difference between two regression coefficients 65
11.1.7. Evaluation of the regression line's goodness-of-fit 65
11.1.8. Confidence interval and confidence belt for regression line 66
11.1.9. Curvilinear regression: polynomial, exponential, logarithmic and power models 67
11.1.10. Nonparametric alternative for linear correlation and regression 69
11.2. Relationship between a quantitative and a qualitative variable 71
11.3. Relationship between two qualitative variables (association analysis) 72
11.3.1. Association of two ordinal variables 72
11.3.2. Association of two nominal variables 73
12. Analysis of multivariate populations 74
12.1. Relationship between a quantitative variable and several other quantitative variables (linear multicorrelation and multiregression) 74
12.2. Classification (grouping) of samples or variables 75
12.2.1. The concept of R-mode and Q-mode analysis 75
12.2.2. Cluster analysis and dendrograms 75
12.2.3. Principal component (factor) analysis 87
12.3. Discrimination analysis of two populations 93
13. Analysis of sequential (serial) data populations 94
13.1. Tests for randomness of data series 94
13.1.1. Runs test for binomial (binary) data series 94
13.1.2. The RUD and RAM tests for binarized quantitative or ordinal data series 94
13.1.3. The Kendall turning-point test binarized data series 96
13.1.4. The Fisz median-crossing test 96
13.1.5. The Gold runs-length test 97
13.1.6. The Meacham rank-difference test 97
13.1.7. The Hurst statistic (rescaled range analysis) 98
13.2. Tests for asymmetric trends in data series 99

13.2.1. The Spearman rank correlation test 99
13.2.2. The Kendall rank correlation test..... 100
13.2.3. The Moore-Wallis signs test..... 101
13.2.4. The mean-square successive difference test 102
13.2.5. The Wald-Bertram test for two equal-length data series..... 102
13.3. Statistical filtering of quantitative data series..... 103
13.4. Autocorrelation and cross-correlation methods 104
13.5. Autoassociation and cross-association methods..... 106
13.6. Semivariogram..... 108
13.7. Markov-chain analysis 111
13.7.1. Embedded and regular chain models..... 111
13.7.2. Test for the bulk significance of Markovian difference matrix..... 114
13.7.3. Test for the significance of the individual elements of difference matrix..... 114
13.7.4. Test for stationarity of Markov chain..... 115
13.8. Fourier (spectral) analysis..... 116
14. Analysis of a continuous 'regionalized' variable (data with map co-ordinates)..... 128
14.1. Trend surface analysis..... 128
14.2. Other surface-fitting methods..... 141

Statistical tables 151

Appendix: Matrix algebra 169

Statistical literature for geoscientists

Textbooks

Agterberg, F.P. (1974). *Geomathematics.* Developments in Geomathematics Vol. 1, Elsevier, Amsterdam.

Anderberg, M.R. (1973). *Cluster Analysis for Applications.* Academic Press, New York.

Andersen, E.B. (1989). *The Statistical Analysis of Categorical Data.* Springer-Verlag, New York.

Bendat, J.S. & Piersol, A.G. (1971). *Random Data: Analysis and Measurement Procedures.* Wiley Interscience, New York.

Blais, R.A. & Carlier, A. (1968). *Application of Geostatistics in Ore Valuation, Ore Reserve Estimation and Grade Control.* CIMM Special Volume 9, Canadian Institute of Mining and Metallurgy, Calgary.

Bock, H.H. (Ed.) (1988). *Classification and Related Methods of Data Analysis.* North Holland, Rotterdam.

Cattel, R.B. (1965). *The Meaning and Strategic Use of Factor Analysis.* Rand, McNally & Co., Chicago.

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

Cooley, W.W. & Lohnes, P.R. (1971). *Multivariate Data Analysis.* John Wiley & Sons, New York.

Cressie, R.F. (1991). *Statistics for Spatial Data.* John Wiley & Sons, New York.

Dautray, R. & Lions, J.-L. (1990). *Mathematical Analysis and Numerical Methods for Science and Technology.* Springer-Verlag, Berlin.

David, M. (1977). *Geostatistical Ore Reserve Estimation.* Developments in Geostatistics Vol. 2, Elsevier, Amsterdam. [A book on semivariograms and the methods of kriging.]

Davis, J.C. (2002). *Statistics and Data Analysis in Geology.* 3rd Edition, John Wiley & Sons, New York. [The 1st edition from 1973 and the 2nd edition from 1986 can also be used.]

Fenner, P. (Ed.) (1972). *Quantitative Geology.* GSA Special Paper 146, Geological Society of America, Boulder, Colorado.

Ferguson, G.A. (1965). *Nonparametric Trend Analysis.* McGill University Press, Montreal.

Fisher, N.I., Lewis, T. & Embleton, B.J.J. (1987). *Statistical Analysis of Spherical Data.* Cambridge University Press, Cambridge.

Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reinhold, New York.

Guillaume, A. (1977). *Introduction a la Géologie Quantitative.* Masson, Paris.

Gradstein, F.M., Agterberg, F.P., Brower, J.C. & Schwarzacher, W.S. (Eds.) (1985). *Quantitative Stratigraphy.* UNESCO Publication Series, Reidel, Dordrecht.

Griffith, J.C. (1967). *Scientific Methods in Analysis of Sediment.* McGraw-Hill, New York.

Haberman, J.W. (1978). *Analysis of Qualitative Data, Vol. 1 & 2.* Academic Press, New York.

Harbaugh, J.W. & Bonham-Carter, G. (1970). *Computer Simulations in Geology.* John Wiley & Sons, New York.

Harbaugh, J.W. & Merriam, D.F. (1968). *Computer Applications in Stratigraphic Analysis.* John Wiley & Sons, New York.

Henley, S. (1981). *Nonparametric Statistics.* Elsevier Applied Science, London.

Hohn, M.E. (1988). *Geostatistics and Petroleum Geology.* Van Nostrand Reinhold, New York.

Jolliffe, I.T. (1986). *Principal Component Analysis.* Springer-Verlag, New York.

Kendall, M.A. (1965). *A Course in Multivariate Analysis.* Monograph Series Vol. 2, Charles Griffin, London.

Koch, G.S., Jr. & Link, R.F. (1970). *Statistical Analysis of Geological Data.* John Wiley & Sons, New York.

Krumbein, W.C. & Graybill, F.A. (1965). *An Introduction to Statistical Models in Geology.* McGraw-Hill, New York.

Lawley, D.N. & Maxwell, A.E. (1971). *Factor Analysis as a Statistical Method.* 2nd Edition, Butterworths, London.

Lehman, E.L. & D'Abrera, H.J.M. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

Loudon, T.V. (1979). *Computer Methods in Geology.* Academic Press, London.

Matheron, G. & Armstrong, M. (Eds.) (1988). *Geostatistical Case Studies.* Pergamon Press, New York.

Mardia, K.V. (1972). *Statistics of Directional Data.* Reidel/Kluver, Dordrecht.

Marsal, D. (1987). *Statistics for Geoscientists.* Pergamon Press, Oxford.

McCommon, R.B. (Ed.) (1975). *Concepts in Geostatistics.* Springer Study Edition, Springer-Verlag, Berlin.

Merriam, D.F. (Ed.) (1976). *Quantitative Techniques for the Analysis of Sediments.* Pergamon Press, Oxford.

Merriam, D.F. (Ed.) (1978). *Recent Advances in Geomathematics.* Pergamon Press, Oxford.

Miller, R.L. & Kahn, J.S. (1965). *Statistical Analysis in the Geological Sciences.* 2nd Edition, John Wiley & Sons, New York.

Neave, H.R. & Worthington, P.L. (1992). *Distribution-free Tests.* Routledge, London.

Ostle, B. (1963). *Statistics in Research.* Iowa State University Press, Ames, Iowa.

Pakes, A.G. & Maller, R.A. (1990). *Mathematical Ecology of Plant Species Competition: A Class of Deterministic Models for Binary Mixtures of Plant Genotypes.* Cambridge University Press, Cambridge.

Priest, S.D. (1993). *Discontinuity Analysis for Rock Engineering.* Chapman & Hall, London. [A book on the statistical analysis of the orientation and spatial distribution of rock fractures.]

Reyment, R.A. (1971). *Introduction to Quantitative Paleoecology.* Elsevier, Amsterdam.

Reyment, R.A. (1980). *Morphometric Methods in Biostratigraphy.* Academic Press, London.
Ripley, B.D. (1988). *Statistical Inference for Spatial Processes.* Cambridge University Press, Cambridge.
Rock, N.M.S. (1988). *Numerical Geology: A Source Guide, Glossary and Selective Bibliography to Geological Uses of Computers and Statistics.* Springer-Verlag, New York.
Schuenmeyer, J. & Drew, L. (2010) *Statistics for Earth and Environmental Scientists*. Wiley, New York.
Schwarzacher, W.S. (1976). *Sedimentation Models and Quantitative Stratigraphy.* Developments in Sedimentology Vol. 19, Elsevier, Amsterdam.
Sen, A. & Srivastava, M. (1990). *Regression Analysis.* Springer-Verlag, New York.
Sherman, M. & Li, B. (2010) *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Wiley, New York.
Size, W.E. (Ed.) (1987). *Use and Abuse of Statistical Methods in the Earth Sciences.* Oxford University Press, Oxford.
Stanley, L.T. (1973). *Practical Statistics for Petroleum Geologists.* Petroleum Publishing Company, Tulsa, Oklahoma.
Swan, A.R.H. & Sandilands, M. (1985). *Introduction to Geological Data Analysis.* Blackwell Science, Oxford.
Walden, A.T. & Guttorp, P. (Eds.) (1992). *Statistics in the Environmental and Earth Sciences.* Edward Arnold, London.
Whitten, E.H.T. (Eds.) (1975). *Quantitative Studies in the Geological Sciences.* GSA Memoir 142, Geological Society of America, Boulder, Colorado.

Journals

A casual supportive use of basic statistics in research can be found in papers published in virtually all geoscience journals, but the following three are fully devoted to geomathematics and more advanced geostatistical studies: ***Mathematical Geology*** (Journal of the International Association of Mathematical Geology), ***Computers & Geosciences*** (journal published by Pergamon Press) and ***Computer Contributions to Geology*** (series published jointly by the Kansas State Geological Survey and the University of Kansas in co-operation with the American Association of Petroleum Geologists).

What is statistics?

STATISTICS is a branch of mathematics concerned with the investigation of variability in natural phenomena, particularly with the recognition and analysis of regularities in natural phenomena and relationships between different natural phenomena. The principal aim of statistics is to provide a basis for drawing objective, justifiable and verifiable conclusions from limited observational material (or sample datasets).

GEOSTATISTICS is a branch of applied statistics concerned with the research issues specific to geosciences.

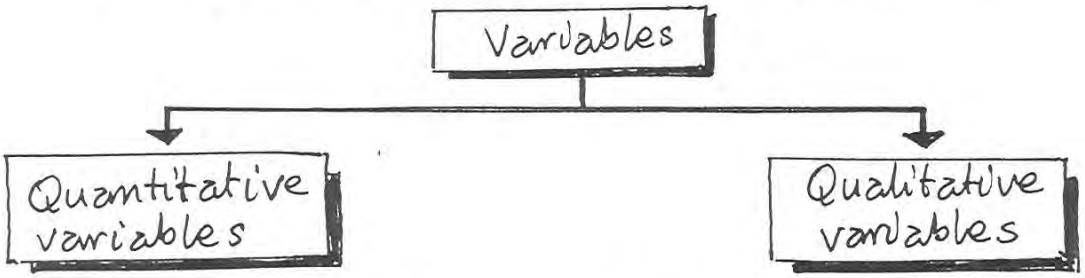
Why do we need statistics in geosciences?

We do need statistics in geosciences because:

- **Our geological data are inherently statistical, derived as some very limited sample datasets.**
- **Our datasets always need to be summarized in an easy and objective way – for an overview, conclusion and comparisons.**
- **Statistics prevents us from drawing subjective inferences and conclusions from our (generally objective) data.**
- **Statistics allows us to evaluate other researchers' data and conclusions and in this way re-assess many existing theories.**

Basic terminology

- Variable – any physical, chemical or biological property of the geological object studied, which is the subject of our systematic measurements or observations.



Quantitative variables

These are measureable and have values (expressed objectively as numbers).
Quantitative variables can be:

- continuous (e.g., the content of a mineral or chemical component in a rock; or bed thickness) or
- discrete (e.g., the number of coal beds in a rock succession or the number of bivalve specimens in a limestone unit studied in an outcrop).

The values of continuous quantitative variables are expressed in:

- ratio scale – which has an absolute zero and equal intervals (e.g., bed thickness, mineral content, length, etc.) or
- interval scale – which has an arbitrary zero and equal intervals (e.g., temperature, pH, altitude, map co-ordinates).

The values of discrete quantitative variables are expressed in the ratio scale.

Qualitative variables

These are non-measurable and assume discrete "states" (descriptive categories).
For example: lithology, lithofacies, stratification type, rock texture, rock colour, fauna species, mineral types and others.
Qualitative variables are expressed in either

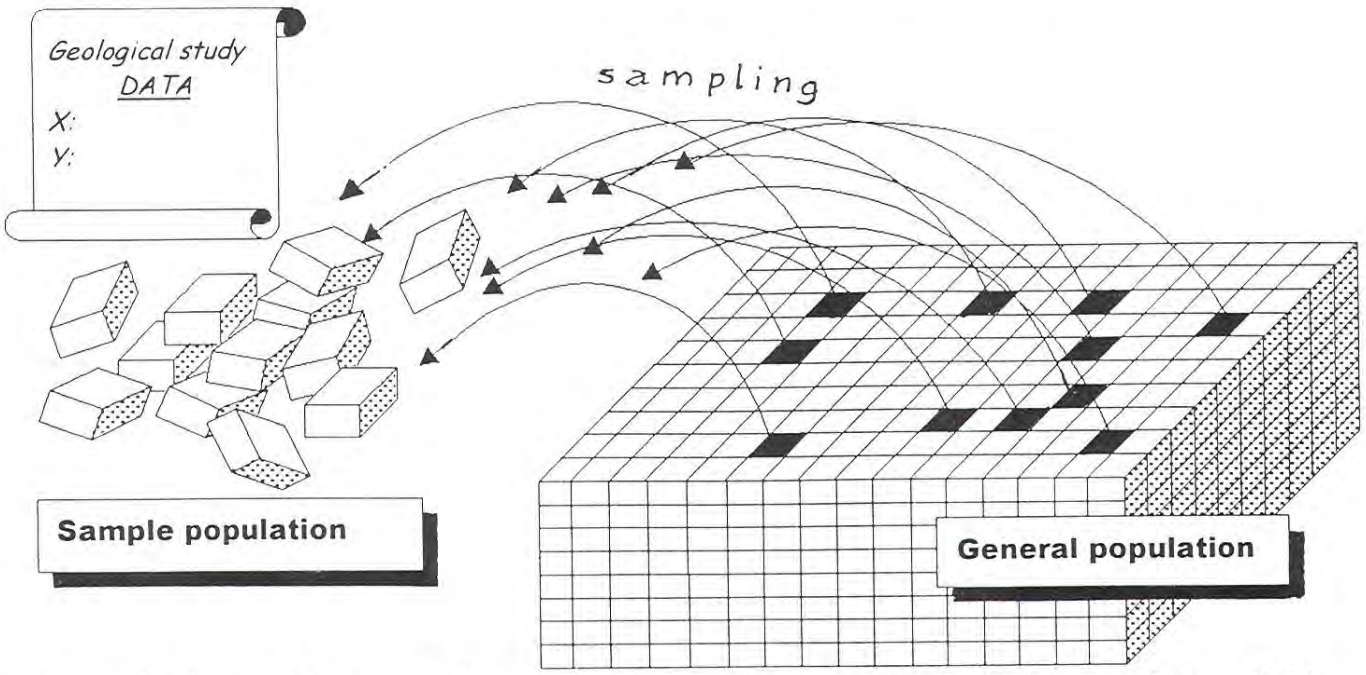
- nominal scale, which is a nonhierarchical multistate scale with all states equal (e.g., lithology, colour, fauna species, etc.)

or

- ordinal scale, which is a hierarchical multistate scale with all states ranked relative to one another (e.g., the Mohs scale of mineral hardness, bioturbation intensity, colour intensity, sandstone texture, metamorphic degree, relative age of rock units in lithostratigraphic section, etc.).

A special case of nominal qualitative variable is binary (binomial) variable, which has two states only (e.g., present/absent, 1 or 0).

• Sample population and general population



Sample population is the collection of samples (data set) available for a particular study. The number of samples (data) is finite, usually very limited. The samples are observations on one or more variables.

General population is the geological object studied as a whole (i.e., all possible "samples" that theoretically might have been taken from the object studied). It may be considered as an infinite population of samples, of which only a finite, small number are available (sample population).

A sample population has its size (number of samples, n) and dimension (number of variables, p, measured in each sample). Sample populations with $n \leq 30$ are generally regarded as "small". When analyzed, a multidimensional ($p > 1$) population can be considered as univariate, bivariate or multivariate, depending on the number of variables analyzed at one time. A one-dimensional population ($p = 1$) is necessarily univariate.

A sample population to be representative, with respect to the general (parental) population, should be:

(1) randomized (i.e., taken in a random, non-preferential manner, such that every part of the object studied has an equal chance to be sampled);

(2) of appropriate size (i.e., not too small, sufficient to reveal full variation of the variable or variables studied).

In geological research, these requirements are seldom fulfilled and a "sample population of convenience" is most often used: we collect data where we can.

● Estimation of an "optimal size" of sample population

Several different formulae can be used, depending on how the "accuracy" or the "economy" of the sampling is defined.

The Avashti-Kumar formula:

$$n_{opt} = \frac{(t_{1/2\alpha} \cdot S_o/S_a)^2}{1 + \left(\frac{1}{n_o} \cdot t_{1/2\alpha} \cdot S_o/S_a\right)^2}$$

where: n_{opt} = optimal number of samples;
n_o = the size of preliminary sample population;
S_o = the standard deviation of the preliminary population;
S_a = the desirable accuracy, defined as the acceptable standard error of the mean;
$t_{1/2\alpha}$ = the value of t-function (from TABLE 2 for DF = $n_o - 1$ and selected α, read for $\frac{1}{2}\alpha$-value).

The confidence level of the estimate is $(1-\alpha)100\%$.

Example: {x} 23.0, 30.3, 26.4, 36.8, 33.5

$n_o = 5$, $\bar{x}_o = 30.0$, $S_o = 5.49$

desirable standard error: $S_a \leq 1.0$

acceptable confidence level: 90% ($\alpha = 0.10$)

from t-table (TABLE 2): $t_{1/2} = 2.13$

$$n_{opt} = \frac{(2.13 \times 5.49/1.0)^2}{1 + \left(\frac{1}{5}\right)^2 (2.13)^2 (5.49/1.0)^2} = \frac{136.53}{6.47} = 21$$

$n = n_{opt} - n_o = 21 - 5 = 16$ additional samples needed

Formulae based on confidence interval for the true mean:

$$n_{opt} = \left(\frac{S_o \cdot t_{1/2\alpha}}{L_\alpha}\right)^2 \quad \text{or} \quad n_{opt} = n_o \left(\frac{L_o}{L}\right)^2$$

where: S_o and $t_{1/2\alpha}$ - as above;
L_α = confidence half-interval for the true mean (see p. 23).

where: n_o, L_o = for pilot dataset;
L = required for true mean (μ), for the same α as used to calculate L_o.

The desirable accuracy here is to have a confidence of $(1-\alpha)100\%$ that the true mean is within the range of $\pm L$ around the calculated mean of sample population (true mean = calc. mean $\pm L$).

Note carefully that the significance (risk-of-error) level of the estimate in each case is α, but the t-value is taken for $\frac{1}{2}\alpha$.

The Cochran formula:

$$n_{opt} = \sqrt{(s_s^2 / s_e^2)(c_e / c_s)}$$

where: s_s^2 and s_e^2 – variances calculated from preliminary sample population (a few samples have been taken from a few outcrops, and the estimates of within-outcrop and between-outcrop variation have been obtained by one-way ANOVA);

c_e = cost or time of study per exposure (outcrop);

c_s = cost or time per sample (measurement or lab analysis).

This formula derives from the cost/time function:

$$C = c_e \cdot n_e + c_s \cdot n_e \cdot n_s$$

where: C = total cost or time of the study;

c_e = cost or time per exposure (outcrop);

c_s = cost or time per sample/measurement;

n_e = number of outcrops to be studied;

n_s = number of samples per outcrop.

Example:

Preliminary sample population has given $s_s^2 = 0.6695$ and $s_e^2 = 0.0238$, and the study's economic conditions are:

$$\begin{cases} c_e = 60 \text{ minutes per outcrop} \\ c_s = 90 \text{ minutes per sample} \end{cases}$$

The use of the Cochran formula gives:

$$n_{opt} = \sqrt{(0.6695/0.0238)(60/90)} = 4.3188 \simeq 5$$

Estimate: 5 samples per outcrop will be an optimal number.

A method for the optimization of sampling in the case of directional (vectorial) data is given later in the text (p. 54-55).

Frequency distribution

- Non-grouped and grouped data

The table below shows an example of a sample population of non-grouped data (original measurements). The data population is one-dimensional (univariate) and has a size of $n = 356$.

Table 1. Measurements of the thickness of a coal bed (cm); the number of data is $n=356$.

0	49	60	70	78	84	92	103	115
0	50	60	70	78	84	92	103	115
5	50	61	70	78	85	93	104	115
16	50	61	70	78	85	93	104	115
27	50	61	70	79	86	93	104	115
28	51	61	71	79	86	93	104	116
29	51	62	71	79	86	94	104	116
32	51	62	71	79	86	94	104	116
36	52	63	71	79	86	94	104	117
36	53	63	71	79	86	94	104	118
37	53	63	71	79	86	95	104	118
37	53	64	72	80	87	95	104	118
38	53	64	72	80	87	95	105	118
39	53	64	72	80	87	96	105	119
39	54	64	72	80	88	97	105	119
39	54	64	72	80	88	97	105	119
40	55	64	72	81	88	97	105	120
40	55	65	73	81	88	97	106	120
41	55	65	73	81	88	97	107	120
41	55	66	73	82	88	97	107	121
42	55	66	73	82	89	98	107	121
42	56	66	73	82	89	98	107	121
42	56	66	73	82	89	99	107	121
43	56	66	74	82	89	99	107	121
43	56	67	74	83	90	99	108	123
43	57	67	74	83	90	99	108	123
44	57	67	75	83	90	99	109	123
44	57	67	75	83	90	99	109	124
44	58	67	75	83	90	99	109	125
44	58	67	76	83	90	100	110	126
45	58	68	76	83	90	101	110	127
46	58	68	76	83	90	101	110	128
46	58	68	76	83	90	101	111	132
46	58	69	76	83	90	101	111	138
47	59	69	77	84	91	102	111	139
47	59	69	77	84	91	102	112	146
47	59	69	77	84	91	102	112	
48	59	69	77	84	91	102	112	
48	60	70	77	84	92	102	113	
49	60	70	77	84	92	103	114	

For an easier review, the data have been ordered from the smallest to the largest value.

The frequency distribution of this data set is shown in Figs. 1-3 (next page).

class limits [cm]	*class midpoint* [cm]	*number frequency*	*relative frequency in proportion*	*cumulative frequency*	*cumulative proportion*
0–10	5	3	0.0084	3	0.0084
10–20	15	1	0.0028	4	0.0112
20–30	25	3	0.0084	7	0.0196
30–40	35	11	0.0309	18	0.0505
40–50	45	27	0.0758	45	0.1263
50–60	55	37	0.1039	82	0.2302
60–70	65	43	0.1208	125	0.3510
70–80	75	51	0.1433	176	0.4943
80–90	85	58	0.1629	234	0.6572
90–100	95	36	0.1011	270	0.7583
100–110	105	42	0.1180	312	0.8763
110–120	115	27	0.0758	339	0.9521
120–130	125	13	0.0365	352	0.9886
130–140	135	3	0.0084	355	0.9970
140–150	145	1	0.0028	356	0.9998
Totals		356	1.0000		

Fig. 1. The data set from Table 1 (previous page) shown as grouped data. The relative frequency (in per cent) is plotted in Figs. 2 & 3. A cumulative frequency plot is shown in Fig. 5.

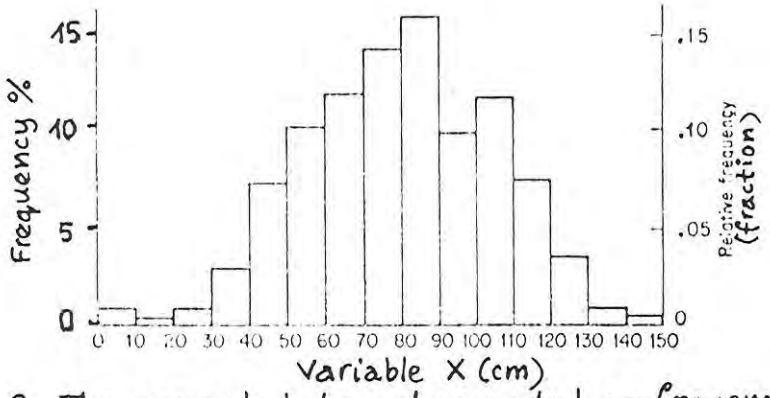


Fig. 2. The grouped data set presented as frequency histogram.

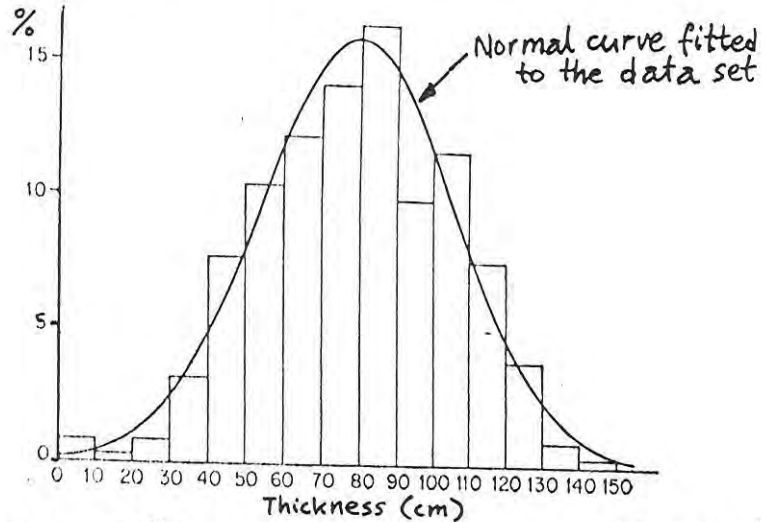


Fig. 3. Normal (Gaussian) curve approximating the histogram; vertical scale somewhat enlarged for visual purpose. The data set is the same as in Table 1 and Figs. 1–2.

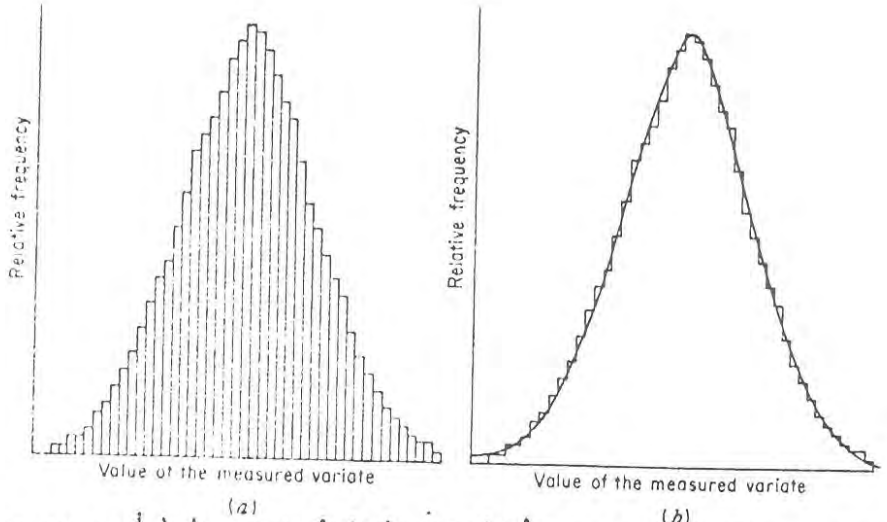


Fig. 4. Frequency histogram (a) turns into a smooth frequency curve (b) when the number of data is very large ($n \to \infty$) and the frequency classes are made very narrow.

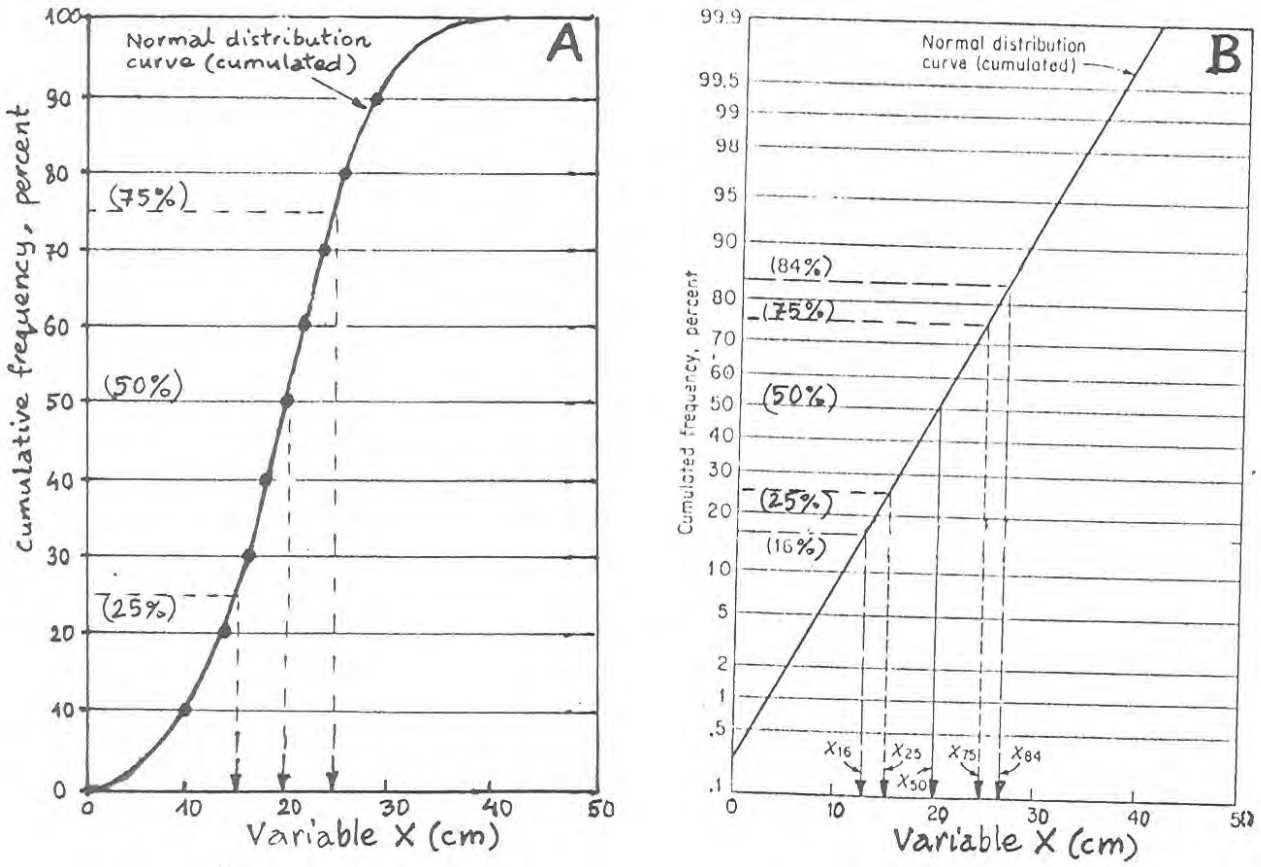


Fig. 5. Cumulative frequency of grouped data plotted in arithmetic (A) and probability scale (B) on the percentage axis. Normal frequency distribution is a straight line when plotted in probability scale (B). Note that a whole range of percentiles can be read from the curve to characterize the shape of the frequency distribution; the 50th percentile, X_{50}, is called "median".

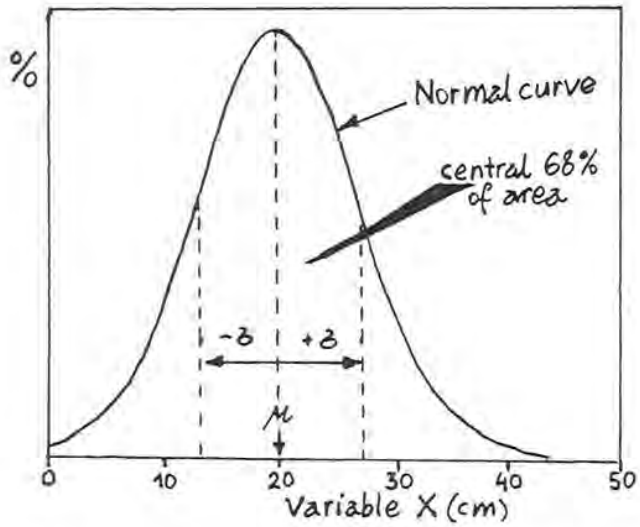


Fig. 6. The area under the frequency curve is called the distribution density; ca. 68% of the density is within the range of one standard deviation ($\pm\sigma$) around the mean (μ), and nearly whole density is within the range of $\pm 3\sigma$ around the mean. The mean value (μ) defines the distribution central tendency, or the bell-shaped curve's central point.

Some properties of normal distribution

- The non-cumulative frequency curve is symmetrical and bell-shaped, with the maximum ordinate (frequency peak) at μ and with inflection points at $\mu-\sigma$ and $\mu+\sigma$
- The cumulative frequency curve is sigmoidal (S-shaped) in arithmetic percentage scale, but is a straight line when plotted in the so-called probability scale (Fig. 5)
- The frequency distribution is completely described by two parameters: the mean value (μ) and the standard deviation (σ)
- The mean, the median (50th percentile) and the mode (x-value corresponding to the curve's highest point) are identical if the distribution is perfectly symmetrical.
- The area under the curve (Fig. 6) is distributed as follows:

central part of the area under normal curve	Interval around the mean
50%	$\mu \pm 0.6745\sigma$
68.3%	$\mu \pm \sigma$
95,4%	$\mu \pm 2\sigma$
99.7%	$\mu \pm 3\sigma$

Distribution type	formula (density function)	parameters	point estimators	illustration (diagram next page)
Normal	$n(X;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-[(X-\mu)^2/2\sigma^2]}$	μ (mean) σ^2 (variance)	$\hat{\mu} = \bar{X}$ $\hat{\sigma}^2 = s^2$	A
Lognormal	$\mathcal{L}(X;\mu_L,\sigma_L) = \frac{1}{X\sigma_L\sqrt{2\pi}} e^{-[(\log X-\mu_L)^2/2\sigma_L^2]}$ $n(L;\mu_L,\sigma_L) = \frac{1}{\sigma_L\sqrt{2\pi}} e^{-[(L-\mu_L)^2/2\sigma_L^2]}$ where $L = \log X$	μ_X σ_X^2 μ_L σ_L^2	$\hat{\mu}_X$ $\hat{\sigma}_X^2$ $\hat{\mu}_L = \bar{L}$ $\hat{\sigma}_L^2 = s_L^2$	B
Gamma	$G(X;r,\beta) = \frac{X^{r-1}e^{-X/\beta}}{\Gamma(r)\beta^r}$ mean = $r\beta$ variance = $r\beta^2$	r,β (r known)	$\hat{r}, \hat{\beta}$ $\hat{\beta} = \frac{1}{rn}\sum_{i=1}^{n} X_i$	C
Circular normal (von Mises)	$C(X;\gamma,\kappa) = Ke^{\theta\cos(X-\gamma)}$ (K is a constant that gives unit area to the density)	γ κ	$\hat{\gamma}$ = vector mean = $= \arctan\frac{\Sigma f_i \sin X_i}{\Sigma f_i \cos X_i}$ $\hat{\kappa}$ = vector concentration	D
Point binomial	$B(X;p) = p^X(1-p)^{1-X}$ $X = 0, 1$	p	$\hat{p} = \bar{X}$	E
Poisson	$P(X;\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$ $X = 0, 1, 2, \ldots$	λ (mean) λ (variance)	$\hat{\lambda} = \bar{X}$ $= \bar{x}$	F

Fig. 7. Common distribution types (density functions, parameters and their estimators); example curves are shown in Fig. 8.

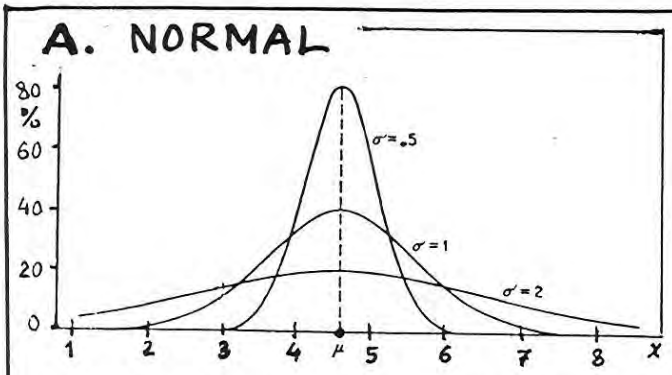
Normal distribution: $f(x) = a \cdot e^{-bx^2}$ (generalized equation) is a special type of negative exponential distribution, with $a = 1/\sqrt{2\pi}$ and $b = 1/2$; $e = 2.71828$.
Standard normal distribution: $f(z) = a \cdot e^{-bz^2}$, where $z = \frac{x-\mu}{\sigma}$; z is a standardized variable (non-dimensional), and this distribution is used as a reference for comparative purpose in some tests.
Lognormal distribution: $f(x) = a \cdot e^{-(\ln x)^2}$ is a normal distribution of variable $x_L = \ln x$; the mean $\bar{x}_L$ and variance s_L^2 are calculated for x_L, but they are geometric; to obtain arithmetic $\bar{x}$ and s_x^2, the following transformations are used:

$$\bar{x} = e^{\left[\bar{x}_L + \frac{1}{2}s_L^2\right]} \quad \text{and} \quad s_x^2 = \left(e^{s_L^2} - 1\right) e^{\left[2\bar{x}_L + s_L^2\right]}$$

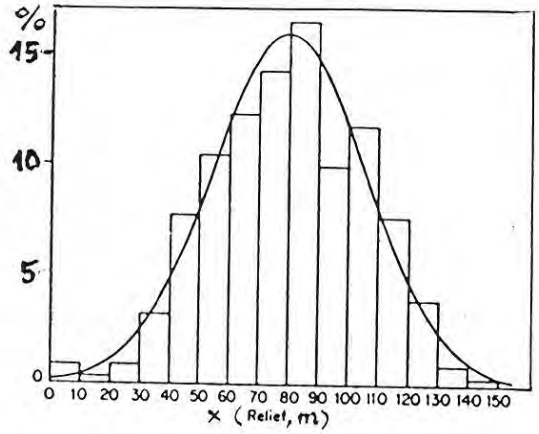
If $x_L = \log_{10} x$ is used, the constant e is replaced with 10 in these formulae.

Exponential (Poissonian) distribution: $f(x) = a \cdot x^b$, and its mean and variance are equal.
Gamma distribution: $f(x) = \left(x^{a-1} \cdot e^{-x/b}\right) / G(a)\, b^a$, where $x \geq 0$, $a > 0$, $b > 0$; its mean and variance are obtained as follows: the ordinary arithmetic mean, $\bar{x}$, and geometric mean, $\bar{x}_G$, are calculated; these are transformed into $z = \ln\bar{x} - \ln\bar{x}_G$ and a corresponding value η is derived from special table; values $a = \eta/z$ and $b = \bar{x}/a$ are calculated, which give mean $= a \cdot b$ and variance $= a \cdot b^2$.
Hyperbolic (power) distribution: $f(x) = a \cdot x^b$; it implies a fractal variable and is discussed later in the text (p. 14-15).

A. NORMAL

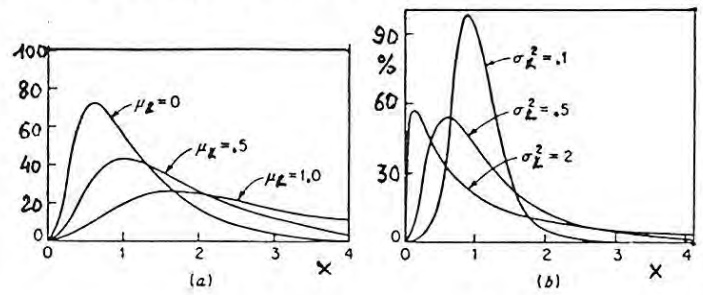


Examples of normal curves with the same mean and differing values of σ. (From Mood and Graybill, 1963.)



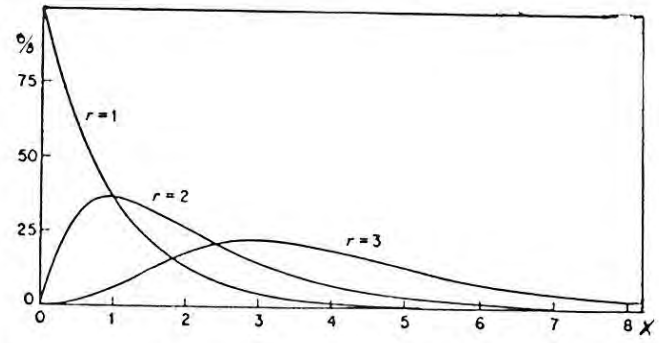
Normal curve that approximates a histogram.

B. LOGNORMAL



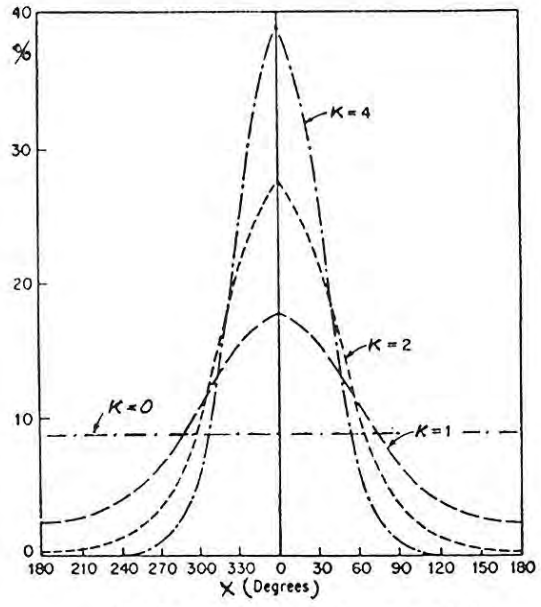
Examples of lognormal densities. (a) Frequency curves of lognormal distribution for three values of μ; (b) frequency curves of the lognormal distribution for three values of σ^2.

C. GAMMA



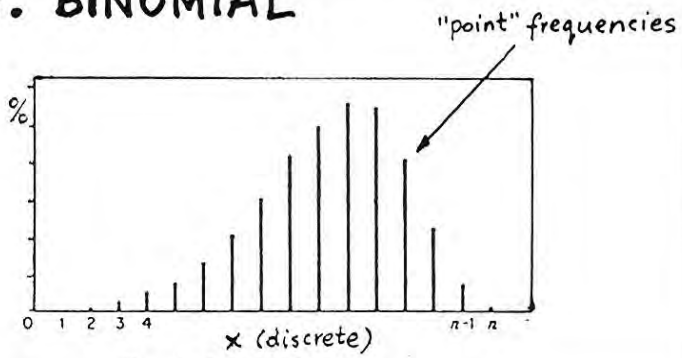
Examples of gamma density for various values of r; $\beta = 1$. (Adapted from Mood and Graybill, 1963.)

D. CIRCULAR NORMAL



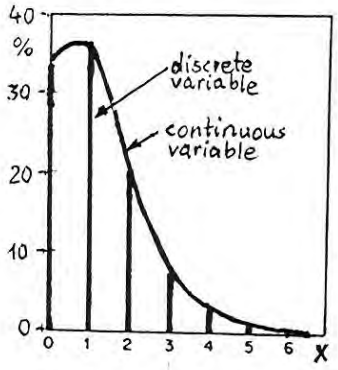
Circular normal densities in linear scale for vector concentration $\kappa = 0, 1, 2, 4$.

E. BINOMIAL



Example of a binomial density.

F. POISSON



Poisson distribution for $\lambda = 1.1$.

Fig. 8. Example curves of common frequency distribution types (from Krumbein & Graybill, 1965).

Examples of geological population distributions

1. Normal distributions
   - Topographic relief
   - Firmness of beach sand
   - Pebble sphericity for fixed particle size
   - Pebble roundness for fixed pebble size
   - Water levels in well through time
   - Drainage density (miles of stream per square mile of drainage-basin area)
   - Specific gravity of rock specimens from a granite pluton
   - Packing density of grains in sandstone
   - Various dimensions (hinge length, etc.) of invertebrate fossils
   - Angle of slope on beach foreshores†
   - Angle of slope in valley walls†
   - Angle of dip of sandstone cross-beds†
   - Porosity of sandstones (expressed as percent void space)
   - Percentages of abundant minerals in rocks
   - Percentage of frosted grains in some dune sand
   - Percentage of moisture in sediments
   - Percentage of some chemical elements or oxides in rocks
   - Mean values based on n observations from normal or nonnormal densities
2. Lognormal distributions
   - Particle-size distributions (by weight or number frequency) of some sediments
   - Thickness of sedimentary beds
   - Length of first-order streams in drainage basins of given order
   - Permeability of sedimentary rocks (sandstone, limestone)
   - Concentrations of trace elements in rocks
   - Lengths of beach segments on some cliffed coasts
   - Areas of river placer deposits
3. Gamma distributions
   - Thickness of sedimentary beds
   - Sand-shale ratio and clastic ratio of some stratigraphic units
   - Percentage of rare components in some rocks (organic matter, trace elements, moisture, heavy minerals)
   - Particle sphericity over a large particle-size range
   - Particle roundness over a large particle-size range
4. Circular normal distributions
   - Orientation of rock joints and fractures
   - Orientation of particle axes in sediments
   - Direction of dip in sandstone cross-beds
5. Binomial distributions
   - Abundant minerals in rocks expressed as number of grains in subsamples of fixed size
   - Abundant fossils in rocks in subsamples of fixed size
   - Occurrence of cross-beds in sandstone (0 = not present, 1 = present)
6. Poisson distributions
   - Rare minerals in rocks expressed as number of grains in subsamples of fixed size
   - Number of alpha particles emitted per unit time from radioactive sediments
   - Sizes of invertebrate fossils in a "death" population

† These items are normal approximations of circular distributions.

Fig. 9. Natural populations and their common types of frequency distribution (from Krumbein & Graybill, 1965).

Note carefully that the list above is a summary of empirical observations, not rules. A particular geological variable may have one type of distribution, or may have different types under different conditions. Note also that only populations with normal distribution can be treated directly by parametric methods; populations with other distributions need to be transformed (e.g., a lognormal distribution becomes normal if log x values are used instead of the raw x values).

● Exceedence frequency plot

This is a rapid graphical method for the recognition of distribution type and for distinction between Gaussian (normal), lognormal, hyperbolic (power-law) and Poissonian (exponential) distributions.

Exceedence frequency, $EF(x_i)$, of a particular value x_i of the measured variable (X) is defined as the number of data (n_i) with values greater than x_i, divided by the total number of data (n):

$$EF(x_i) = \frac{n_i\,(X > x_i)}{n}$$

and this relative frequency is plotted against x_i (see example in Fig. 10).

i	x_i	$EF(x_i)$
1	0.8	0.9
2	1.7	0.8
3	2.0	0.7
4	2.9	0.6
5	3.4	0.5
6	3.7	0.4
7	4.8	0.3
8	6.5	0.2
9	9.0	0.1
10	15.0	0.0

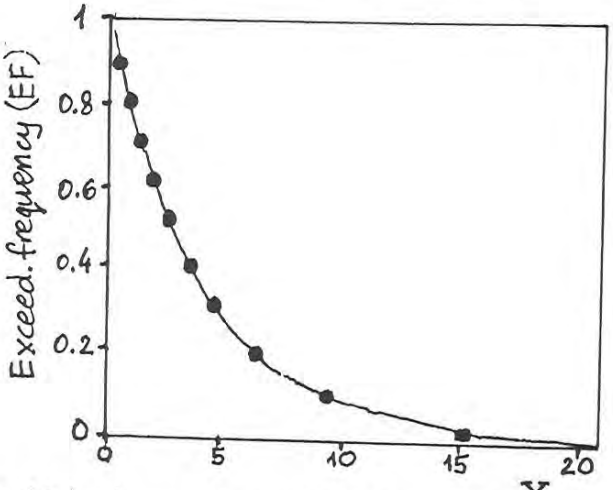


Fig. 10. A sample population of 10 beds with thicknesses, x_i (in cm), listed from the thinnest to the thickest (table). Nine of the 10 beds are thicker than the thinnest bed (EF = 0.9), eight of 10 are thicker than the next thinnest (EF = 0.8), seven of 10 are thicker than the next (EF = 0.7), etc.

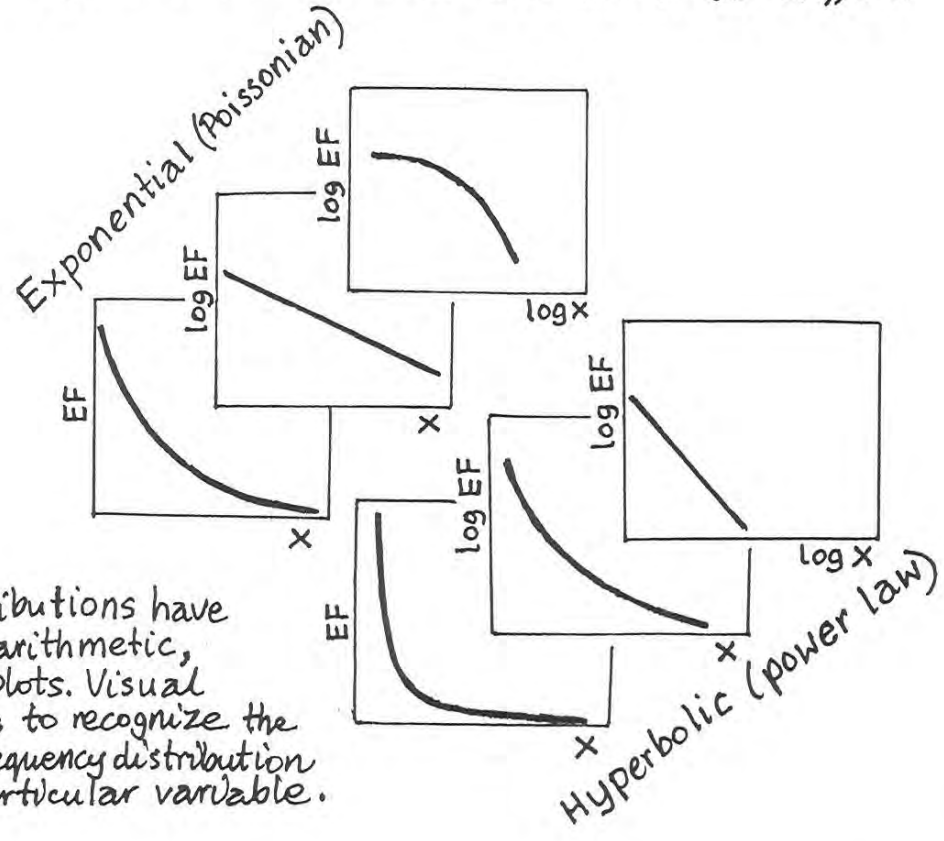


Fig. 11. Different distributions have different shapes in arithmetic, semi-log and log-log plots. Visual inspection thus helps to recognize the underlying form of frequency distribution that characterizes particular variable.

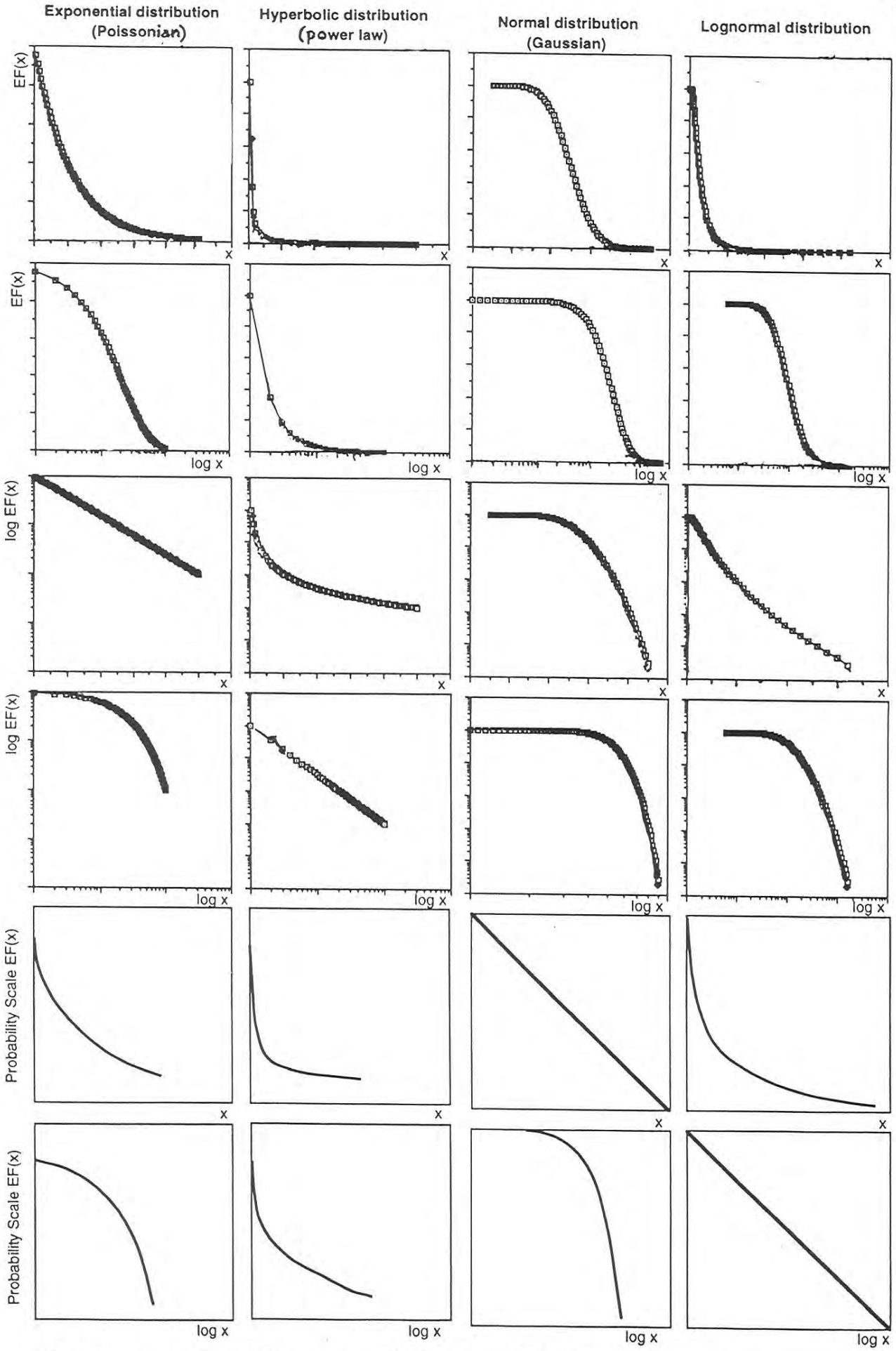


Fig. 12. Exceedence frequency (EF) plots of four common types of data distribution in different scales – a guide to the recognition of distribution type.

- Possible causes of distribution lognormality

There may be at least three reasons why a particular geological variable has a lognormal frequency distribution:

(1) the system of measurement units (scale) used;

(2) the variable measured is the cumulative effect, or sum, of a large but finite number of other, independent random variables (X_i) and $X = \sum^{n} X_i$ will approach normal distribution only when $n \to \infty$;

(3) the variable measured is equal to the product of the effects of several random variables: $X = \prod^{n} X_i$ or $\log X = \sum^{n} \log X_i$, for which it follows that $\log X$ is normally distributed and that X is lognormally distributed.

- The significance of power-law distribution (fractal statistics)

Variables characterized by hyperbolic (power-law) distribution are scale-invariant, or fractal. [The term derives from the Latin word "fractus", meaning broken, and has been applied by Mandelbrot (1982) to objects that were too irregular in shape/form to be described by ordinary Euclidean geometry. Similar scale-indifference may characterize other, non-geometrical geological variables.]

The plot of $\log EF(x)$ vs. $\log X$ (Fig. 12) is a straight line in the case of hyperbolic distribution, and the fractal dimension (D) of the variable x is the line's gradient:

$$D = \frac{\log (EF_i / EF_{i+1})}{\log (X_{i+1} / X_i)} = \frac{\log EF_i - \log EF_{i+1}}{\log X_{i+1} - \log X_i}$$

The notion of dimensionality is abstract, like the fractal concept itself, but can be understood by noting how the "size" of an object changes as its linear dimension increases:
If the linear dimension of a line segment is doubled, the segment's length (one-dimensional characteristic measure of size) is increased by two times. If the linear dimensions of a rectangle are doubled, its area (two-dimensional characteristic measure of size) is increased by four times. If the linear dimensions of a box are doubled, the box's volume (three-dimensional characteristic measure of size) increases by eight times. The relationship between dimension (D), linear scaling (L) and the resulting increase in size (S) is:

$$S = L^{D}$$

Hence, if we upscale a two-dimensional object ($D=2$), its size (area) increases by the square of the scaling (L). If we upscale a three-dimensional object ($D=3$), its size (volume) increases by cube

of the scaling. The equation can be rearranged to give an expression for the topological dimension as a function of the object's characteristic change in size with a change in its linear scaling: $D = \log S / \log L$.

In Euclidean geometry, $D=0$ means a point; $D=1$ means a line; $D=2$ means a surface (two-dimensional object); $D=3$ means a three-dimensional object. The dimension (D) of fractal objects is not necessarily an integer, but a fractional value (fractional dimension); hence the term "fractal."

A variable, X, to be considered fractal, must have a fractal dimension greater than its topological dimension and be either self-similar (i.e., with the changes in its higher values resembling changes in the small values) or self-affine (i.e., with the higher values needed to be rescaled – by scale compression or expansion – before they are self-similar). With reference to the EF plot (Figs. 10-12), the self-similarity of a fractal can be defined by the relationship:

$$EF_i = C \cdot X_i^{-D}$$

where: EF_i = number of data exceeding the value X_i;
C = proportionality coefficient (constant);
D = fractal dimension.

This relationship is manifested by the straight-line plot:

$$\log EF_i = \log C - D \cdot \log X_i$$

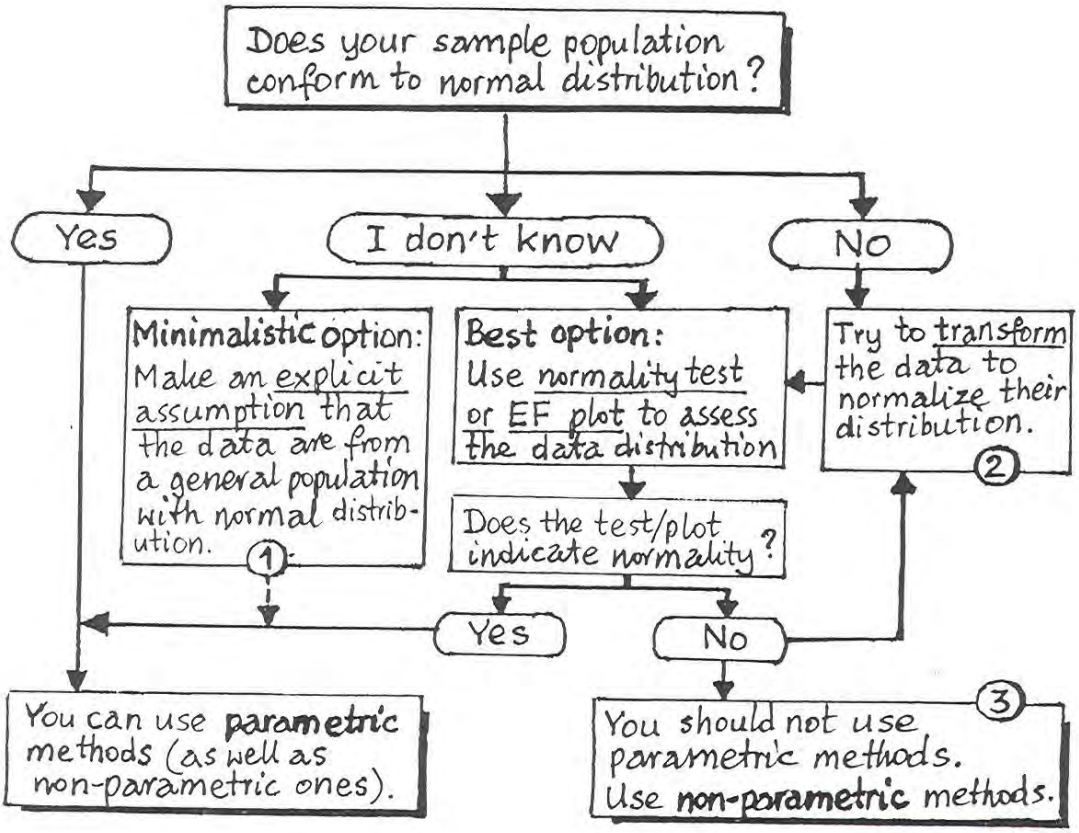
The self-affinity of a fractal can be defined by the relationship:

$$EF_i = C \cdot X_i^{-m \cdot D}$$

where: m = transformation (rescaling) coefficient, on which self-similarity of variable X relies.

Some geological variables show fractal properties, but the causes of this are still little understood. The fact is that most natural variables are results of some other controlling variables, mutually independent or partly dependent; the values, or "size" (S), of a particular variable may thus be subject to linear "scaling" (L) with a fractal dimension ($D > 1$, greater than the variable's one-dimensionality). The notion of self-similarity underlies the concept of self-organized criticality: a natural variable in a marginal stable state, when perturbed from this state by upscaling, will evolve back to the state of marginal stability; in this critical state there is no natural size scale for the variable, so that fractal statistic (D) is applicable.

STATISTICAL METHODS	PARAMETRIC	These refer to the distribution's parameters (such as mean, stand. deviation, variance) and generally assume distribution normality.
	NON-PARAMETRIC	These use the raw data sets as they are, and make no assumption about the distribution



Comments:

(1) You should be aware that, if you use parametric method, your whole statistical analysis will rely on this assumption. If the assumption appears to be invalid, so may be your analysis. However, some parametric method lean more heavily on the distribution normality than do others, and the results in some case may still be valid, especially if the distribution's departure from normality is not too strong; but this you may never know..

(2) The data can be transformed in various ways: instead of the raw measurements, x, use $\log x$, $\ln x$, x^2, x^k, $\sqrt{x}$, x^{-k} $(x+a)^k$, $\log(x+a)$, e^x and other simple functions (with selected k and a constants).

(3) If the distribution is non-normal and you are unable to normalize it, try to identify the actual type of the distribution by using the EF plots. A representative mean and variance can be calculated for every known distribution type (Fig. 7), and some further analysis can also be done, even if limited. Anyway, most parametric methods have their non-parametric equivalents, and use these.

Fig. 13. The analytical strategy for using parametric and non-parametric statistical methods.

Basic parameters and statistics

These are descriptors of distribution, used for populations with normal distribution and in parametric methods.

- Basic parameters (general population characteristics)

μ = mean value of the variable (x)
σ_x^2 = variance of the variable
σ_x = standard deviation of the variable

} unknown "true" values

- Their estimators, or statistics (sample popul. characteristics)

$\bar{x}$ = sample population mean
s_x^2 = sample population variance
s_x = sample population standard deviation

} calculated from observational data of x (see formulae below)

- Calculation for non-grouped data

MEAN

$$\bar{x} = \frac{\sum^{n} x_i}{n}$$

This is the arithmetic average value of x;
n = number of data (or the size of the sample population)

VARIANCE

$$s_x^2 = \frac{\sum^{n} (x_i - \bar{x})^2}{n}$$ "biased" estimator (definition)

$$s_x^2 = \frac{\sum^{n} (x_i - \bar{x})^2}{n-1}$$ "unbiased" estimator (used in practice)

$$s_x^2 = \frac{\sum^{n} x_i^2 - \frac{1}{n}\left(\sum^{n} x_i\right)^2}{n-1} = \frac{n\sum^{n} x_i^2 - \left(\sum^{n} x_i\right)^2}{n(n-1)}$$ formula used for calculation

STANDARD DEVIATION

$$s_x = \sqrt{s_x^2}$$

- Calculation for grouped data

MEAN

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$$

k = number of classes of x-values
x_i = class midpoint value
f_i = class frequency (if in per cent, then n = 100 %)

VARIANCE

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2$$ "unbiased" estimator

$$s_x^2 = \frac{1}{n-1}\left[\sum^{k} f_i x_i^2 - \left(\sum^{k} f_i x_i\right)^2 / n\right]$$ formula used for calculation

STANDARD DEVIATION

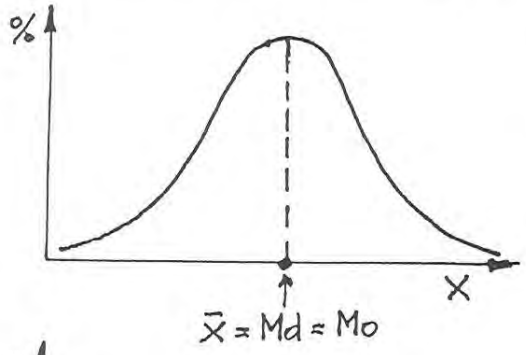
$$s_x = \sqrt{s_x^2}$$

Other useful basic statistics

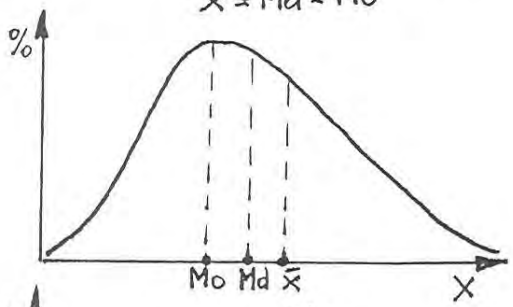
- MEDIAN (Md) = the x-value that divides the data population into two equal parts; i.e., which corresponds to 50% cumulative frequency. Median can be derived directly from the cumulative frequency plot, as the 50th percentile, or find in the following way: the data are ordered from the smallest to the largest and the middle value is taken

$$Md = X_{(n+1)/2} \quad \text{if } n \text{ is an odd number}$$
$$\text{or } Md = \frac{1}{2}(X_{n/2} + X_{n/2+1}) \quad \text{if } n \text{ is an even number}$$

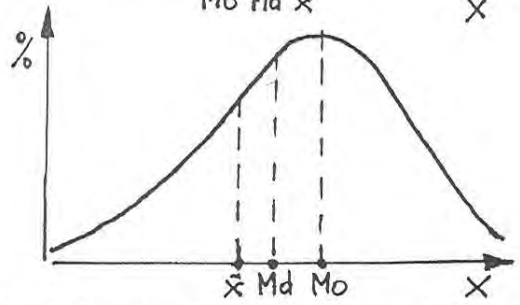
- MODE (Mo) = the x-value with the highest frequency in the data population. Mode is derived from non-cumulative frequency curve/histogram, as the x-value corresponding to the curve's peak or the midpoint value of the histogram class with the highest frequency.



In an ideal, perfectly symmetrical distribution, the arithmetic mean ($\bar{x}$), median (Md) and mode (Mo) are equal.
If these are not equal, the distribution is asymmetrical (or skewed).



The distribution is "positively skewed" when $\bar{x} > Md > Mo$



The distribution is "negatively skewed" when $\bar{x} < Md < Mo$

- COEFFICIENT OF DISTRIBUTION SKEWNESS (Sk)

$$Sk = \frac{\frac{1}{n}\sum^{n}(x_i - \bar{x})^3}{S_x^3}$$

where: S_x^3 is the cube of standard deviation.

A slightly skewed bell-shaped distribution can still be regarded as normal, but a stronger skewed distribution will fail the normality test.

■ GEOMETRIC MEAN ($\bar{X}_G$)

$$\bar{X}_G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

definition formula *

$$\boxed{\bar{X}_G = e^{\frac{1}{n}\sum^{n} \ln x_i}}$$

practical formula used for calculation

* The product may be a very large value and exceed the capacity of calculator or PC if the data set is large and the x-values are high. Therefore, the formula in frame is used in practice.

Note: All data must be $x_i \neq 0$. A zero value will render $\bar{X}_G$ indeterminable. If zero data values occur, these should be replaced with a very low non-zero "dummy" value.

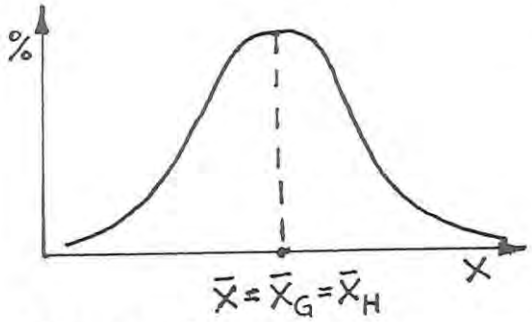
■ HARMONIC MEAN ($\bar{X}_H$)

$$\boxed{\bar{X}_H = n\sqrt{\sum^{n} \frac{1}{x_i}}}$$

definition and practical formula for calculation *

Note: All data must be $x_i \neq 0$. If zero data values occur, these should be replaced with a very low non-zero "dummy" value.

* The value of $\bar{X}_H$ is sensitive to low x_i values, which may result in a biased (too high) estimate. Therefore, uncertain low data values should preferably be excluded.



For an ideal, symmetrical normal distribution:

$$\bar{X} = \bar{X}_G = \bar{X}_H$$

For positively skewed distribution:

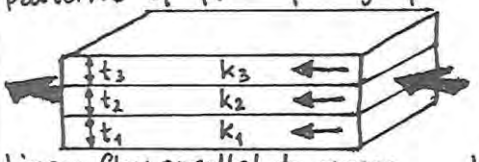
$$\bar{X} > \bar{X}_G > \bar{X}_H$$

For negatively skewed distribution:

$$\bar{X} < \bar{X}_G < \bar{X}_H$$

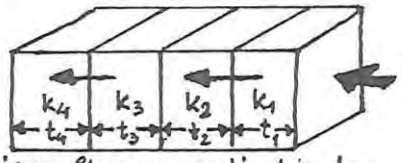
Example

The mean permeability (k) of porous medium is calculated differently for different patterns of flow of single-phase fluid.



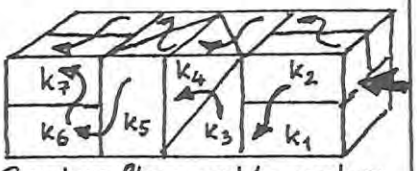
Linear flow parallel to zones of k-layering.
Arithmetic mean is used:

$$\bar{k} = \sum_{i=1}^{n} k_i t_i \Big/ \sum_{i=1}^{n} t_i$$



Linear flow perpendicular to zones of k-layering.
Harmonic mean is used:

$$\bar{k} = \sum_{i=1}^{n} t_i \Big/ \sum_{i=1}^{n} \frac{t_i}{k_i}$$



Random flow and/or random variation in k-values.
Geometric mean is used:

$$\bar{k} = \exp\left[\sum_{i=1}^{n} \ln(k_i t_i) \Big/ \sum_{i=1}^{n} t_i\right]$$

■ STANDARD DEVIATION OF THE MEAN ($S_{\bar{X}}$)

$$\boxed{S_{\bar{X}} = \sqrt{\frac{S_X^2}{n}}}$$

where: S_X^2 = variance
n = number of data

■ COEFFICIENT OF VARIATION (CV)

$$CV = S_x / \bar{x}$$

The ratio of standard deviation and the arithmetic mean

CV is useful for comparing the variation of populations with different variables (different measurement units) or different range of x-values. (Note that variables with high values have greater variances, in numerical terms, than do variables with low values, which renders their variances difficult to compare in absolute terms.)

CV not much greater than 1 indicates a narrow distribution, whereas a high CV, in the range of 2 to 5, indicates a long-tailed, wide distribution.

Perfect normal distributions have $CV \leq 0.5$, whereas $CV > 0.5$ indicates considerable skewness, which may mean a non-normal distribution.

CV provides also some information on the population's pattern of variation:

$CV > 1$	indicates a population with clustred x-values;
$CV \simeq 1$	indicates a population with random x-values;
$CV < 1$	indicates a population with anticlustered (regularly spaced) x-values.

For comparative purposes, the CV-value is sometimes given as a percentage: $CV\% = 100\% \cdot S_x / \bar{x}$.

Example: In the permeability studies of petroleum reservoirs, the CV of permeability ($= S_k / \bar{k}$) is used to assess the reservoir's heterogeneity:

$0 \leq CV < 0.5$	homogeneous reservoir
$0.5 \leq CV < 1$	heterogeneous reservoir
$1 \leq CV$	very heterogeneous reservoir

● Simple solutions for simple problems (examples)

■ EXAMPLE 1: Comparison of isotopic dates

We have two radiocarbon dates, 10300 (±120) yrs and 10100 (±90) yrs, and we wish to assess if the difference between them is significant or insignificant in statistical terms. How to do it?

Calculate: $\Delta t = 10300 - 10100 = 200.$

Calculate: $S_{\bar{t}} = \sqrt{(120)^2 + (90)^2} = \sqrt{14400+8100} = \sqrt{22500} = 150$

($S_{\bar{t}}$ is the combined standard deviation of the mean corresponding to the two dates, with $n=1$ for each, since we are dealing with <u>single</u> dates.)

If Δt is larger than $S_{\bar{t}}$, the difference in dates is significant. (If $\Delta t \leq S_{\bar{t}}$, the difference is insignificant.) In our example $\Delta t = 200$ is larger than $S_{\bar{t}} = 150$, which indicates a significant difference between the two dates.

Now, how significant is "significant"? This issue can be defined in terms of the confidence level of the statistical solution.

If criterion $\Delta t > S_{\bar{t}}$ is used (as above), then we can be no more than 68.3% sure that the difference is significant. (Many researchers in chronostratigraphy take this level of confidence as a reasonable minimum for comparative purposes.)

If criterion $\Delta t > 2S_{\bar{t}}$ is used to say "significant", then we are 95.4% sure that the difference is statistically significant. (Criterion $\Delta t > 1.96 S_{\bar{t}}$ would give us exactly 95% confidence.)

If criterion $\Delta t > 3S_{\bar{t}}$ is used to say "significant", then we are 99.7% sure that the difference is statistically significant. (This practically means "absolutely sure", because probability theory precludes 100% confidence.)

The calculation procedure will be different when we are dealing with <u>sets</u> of dates that are to be compared. Let's say we have the following two data sets:

	Set 1	Set 2
	28080 ± 1500	28700 ± 1200
	27300 ± 1200	30900 ± 1200
	26720 ± 1500	29280 ± 1500
	25980 ± 1500	
Number of data:	$n_1 = 4$	$n_2 = 3$
Mean:	$\bar{t}_1 = 26500$	$\bar{t}_2 = 29626$

The respective standard errors of the means are:

$$S_{\bar{t}(1)} = \sqrt{[(1500)^2 + (1200)^2 + (1500)^2 + (1500)^2]/4} = 1430.9088 \simeq 1431$$

$$S_{\bar{t}(2)} = \sqrt{[(1200)^2 + (1200)^2 + (1500)^2]/3} = 1307.6697 \simeq 1308$$

In this case, $\Delta t = \bar{t}_2 - \bar{t}_1 = 29626 - 26500 = 3126$ and the combined ("pooled") standard error of the means is:

$$S_{\bar{t}} = \sqrt{[(1500)^2 + (1200)^2 + (1500)^2 + (1500)^2 + (1200)^2 + (1200)^2 + (1500)^2]/7} = 1379$$

The rest of the assessment is based on the comparison of Δt with $S_{\bar{t}}$, as in the previous case (see above).

EXAMPLE 2: Sampling zircons for provenance studies

The U-Pb dating of detrital zircon grains is commonly used to determine the provenance of sediment in situations when the ages of hinterland crystalline rocks are known (e.g., sediment derivation from the Scandinavian margin and the Shetland margin in the North Sea rift system; or from the Scandinavian margin and the Greenland margin in the Mesozoic Norwegian Sea rift).

In such studies, it is always difficult to know as to how many zircons should be taken from a sandstone sample to be "sufficient" as a representation of the sandstone's zircon population. Some researchers would take 80-100 grains, others claim that 50-75 may be quite sufficient, and yet other argue that some 30 grains may be enough in many cases. How to make an assessment in such cases?

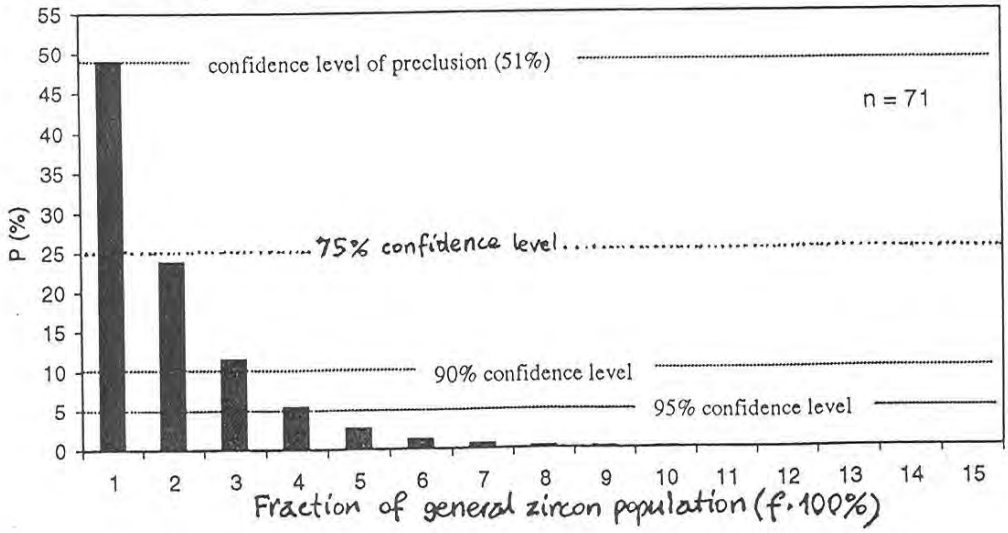
Let's say that we have collected $n = 71$ zircon grains from a sandstone sample. Is this number sufficient for us to be sure that we have not missed some "other" zircons, derived from other source? If there are such "other" zircons in the sediment, what fraction (f) of the total zircon population they may constitute at most? How important can they be, relative to the zircons in our sample population of 71 grains?

This can be estimate by the following formula:

$$P = (1 - f)^n$$

where: P = the probability of occurrence of "other" zircons, not present in the sample population of n grains;
f = the relative frequency of such "other" zircons.

The results of calculation for a range of f-values (0.01 to 0.15) are shown in the diagram below.



We can thus be 95% sure that if any "other" zircons are present in the sandstone, they constitute little more than 4% of the whole zircon population; and 75% sure that they constitute less than 2% of the whole population. But we can be only 51% sure that there is less than 1% of "other" zircons in their population in the rock.

Confidence intervals for parameter values

This is a parametric approach, with the aim to estimate the confidence intervals for the general population's parameters.

Confidence interval for the mean (μ)

$$(\bar{x} - L) < \mu < (\bar{x} + L)$$

with $(1-\alpha)\,100\%$ confidence level

where: $L = t_{\frac{1}{2}\alpha} \cdot S_{\bar{x}}$

- $t_{\frac{1}{2}\alpha}$: value from TABLE 2 for $\frac{1}{2}\alpha$ and $DF = n-1$
- $S_{\bar{x}}$: Stand. deviation of the mean: $S_{\bar{x}} = \sqrt{\frac{S_x^2}{n}}$

When $n \geq 60$, the calculation can be simplified:

$$L = Z_{\frac{1}{2}\alpha} \cdot S_{\bar{x}}$$

where: $Z_{\frac{1}{2}\alpha}$ = value from TABLE 1B for $\frac{1}{2}\alpha$.

Confidence intervals for the variance (σ^2) and standard deviation (σ)

$$\frac{\Sigma(x-\bar{x})^2}{\chi^2_{\frac{1}{2}\alpha}} < \sigma^2 < \frac{\Sigma(x-\bar{x})^2}{\chi^2_{1-\frac{1}{2}\alpha}}$$

For variance

with $(1-\alpha)\,100\%$ confidence level

where: χ^2 = values of chi-square function taken from TABLE 4 for $\frac{1}{2}\alpha$ and $1-\frac{1}{2}\alpha$, respectively.

Note that $\Sigma(x-\bar{x})^2 = (n-1)\,S_x^2$, which may be more convenient for calculation when S_x^2-value has already been obtained.

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{\chi^2_{\frac{1}{2}\alpha}}} < \sigma < \sqrt{\frac{\Sigma(x-\bar{x})^2}{\chi^2_{1-\frac{1}{2}\alpha}}}$$

For standard deviation

with $(1-\alpha)\,100\%$ confidence level

Note that the interval limits for σ are simply square roots of the interval limits obtained for σ^2 (see above).

Confidence interval for mean proportion

This is a non-parametric method for the estimation of a confidence interval for the mean proportion (relative frequency) in a binomial population.

Consider the following parameter and its estimator:

P = unknown true proportion of particular element in the general population of elements studied;

$\hat{P}$ = proportion estimated from the sample population, $\hat{P} = n/N$, where:

n = number of samples with the particular element;

N = total number of samples collected.

Procedure:

Use TABLE 5 (graph), chosing $\alpha = 0.01$ or $\alpha = 0.05$. Enter the graph with $\hat{P}$ (estimated proportion) and read off confidence intervals (CI) via the appropriate pair of curves (selected to approximate your N).

The confidence interval is:

$$\boxed{(\hat{P} - CI_{\alpha}) < P < (\hat{P} + CI_{\alpha})}$$

↑ The minus sign means that the negative CI-value comes here.

↑ Positive CI-value comes here.

and the interval's confidence level is $(1-\alpha)100\%$.

Example: From a limestone, 100 specimens of fauna have been collected, 4 of which are echinoids ($N=100$, $n=4$). The estimated proportion of echinoids in the fossilised fauna population is thus:

$$P = 4/100 = 0.04 \quad (\text{or } 4\%).$$

To find a 95% confidence interval for 'true' P, we read off the L values from the graph (TABLE 5) by using $N = 100$ curves:

$$\begin{cases} L_{0.05} = -0.025 & \text{(negative)} \\ L_{0.05} = +0.035 & \text{(positive)} \end{cases}$$

The confidence interval is:

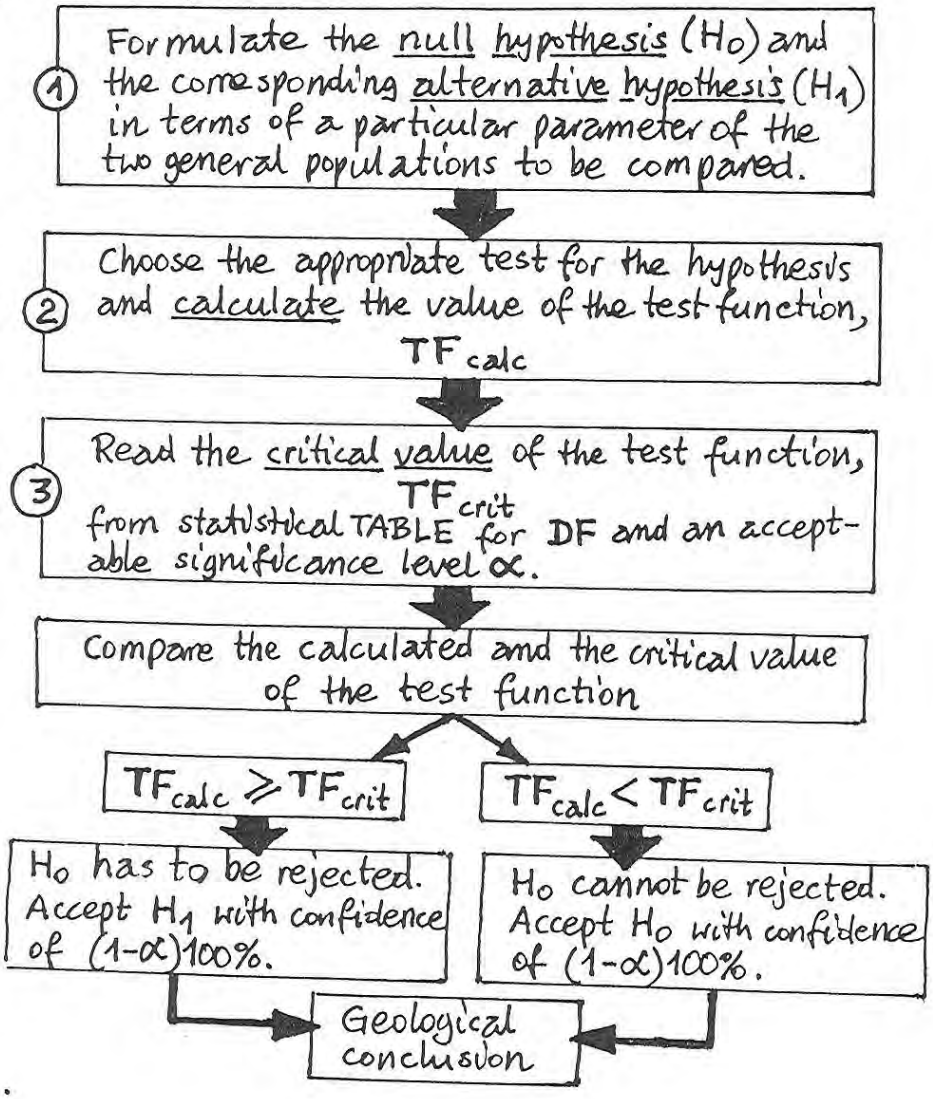
$$(0.040 - 0.025) < \pi < (0.040 + 0.035)$$

$$0.015 < \pi < 0.075 \quad \text{(at 95\% confidence level)}$$

The proportion of echinoids is between 1.5 % and 7.5%, with 95% confidence.

PARAMETRIC COMPARISON OF TWO POPULATIONS

The general scheme for a test of statistical parametric hypothesis is shown in the diagram below.



Comments:

① For example: $\begin{cases} H_0: \mu_1 = \mu_2 & \text{(which means "null difference")} \\ H_1: \mu_1 \neq \mu_2 & \text{(which means a significant difference, either way)} \end{cases}$

This kind of "either-way" formulation is called a two-tail hypothesis, and is used in statistical theory. In research practice, the calculation of statistics normally indicates as to which way the general difference seems to be (i.e., $\bar{x}_1 > \bar{x}_2$ or $\bar{x}_1 < \bar{x}_2$); this is the basis of geological hypothesis. It is more beneficial, therefore, to formulate one of the following one-tail hypotheses: $\begin{cases} H_0: \mu_1 \leq \mu_2 \\ H_1: \mu_1 > \mu_2 \end{cases}$ or $\begin{cases} H_0: \mu_1 \geq \mu_2 \\ H_1: \mu_1 < \mu_2 \end{cases}$

where H_1 represents the geological hypothesis and H_0 just states that the former is false, or not true. In either formulation, it is always H_0 that is actually tested, and if H_0 is rejected – the alternative, H_1, is accepted.

(2) Every test function has a specified number of the degrees of freedom (DF). DF is the number of data used (n) minus the number of parameters specific to particular probability function (TF).

(3) The so-called significance level, α, is the risk of an error of first kind (i.e., rejection of H_0 when it is true), or the test's risk of error, in short. Important: the TF_{crit} is read off for α in a one-tail test, but is read off for $\frac{1}{2}\alpha$ in a two-tail test; the risk of error in either case is α, and the confidence level is $(1-\alpha)100\%$.

• The Snedecor F-test (for difference between variances)

Two-tail hypothesis:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

One-tail hypothesis:

$$\begin{cases} H_0: \sigma_1^2 \geq \sigma_2^2 \\ H_1: \sigma_1^2 < \sigma_2^2 \end{cases} \quad \text{or} \quad \begin{cases} H_0: \sigma_1^2 \leq \sigma_2^2 \\ H_1: \sigma_1^2 > \sigma_2^2 \end{cases}$$

The test function is:

$$F = \frac{S_1^2}{S_2^2}$$

← Larger variance in numerator
← Smaller variance in denominator

and it has twofold degrees of freedom:

$$\begin{cases} DF_1 = n_1 - 1 & \leftarrow \text{corresponding to } S_1^2 \\ DF_2 = n_2 - 1 & \leftarrow \text{corresponding to } S_2^2 \end{cases}$$

The critical value, F_α for one-tail test or $F_{\frac{1}{2}\alpha}$ for two-tail test, is taken from TABLE 3 for (DF_1, DF_2) and acceptable α. If $F \geq F_\alpha$ (or $F_{\frac{1}{2}\alpha}$), H_0 is rejected and H_1 is accepted. Otherwise, H_0 cannot be rejected.

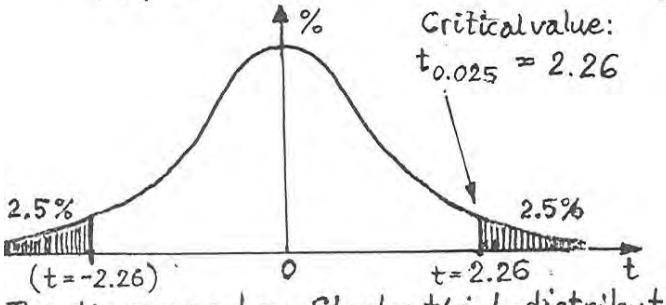
• The Student* t-test (for difference between means)

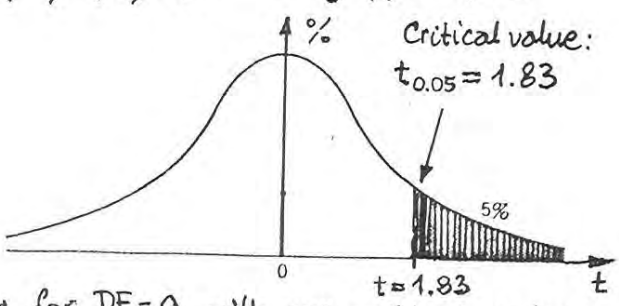
Two-tail hypothesis:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

One-tail hypothesis:

$$\begin{cases} H_0: \mu_1 \geq \mu_2 \\ H_1: \mu_1 < \mu_2 \end{cases} \quad \text{or} \quad \begin{cases} H_0: \mu_1 \leq \mu_2 \\ H_1: \mu_1 > \mu_2 \end{cases}$$





The diagrams show Student's t-distribution for DF=9, with one critical region (shaded) containing $\alpha = 5\%$ of the area under the curve (one-tail test); and with two critical regions, each containing $\frac{1}{2}\alpha = 2.5\%$ of the area (two-tail test).

---

* Mathematician W.S. Gosset published his early papers under pseudonym "Student"; he established the t-function of probability distribution.

If $\sigma_1^2 = \sigma_2^2$ (H_0 accepted in F-test), the t-test function is:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{with } DF = n_1 + n_2 - 2$$

where: s^2 = combined ("pooled") variance

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

n_1, n_2 and s_1^2, s_2^2 are the sizes and variances of the two sample populations compared.

If $\sigma_1^2 \neq \sigma_2^2$ (H_0 rejected in F-test), the t-test function is:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{with } DF = n_1 + n_2 - 2$$

The critical value (t_α for one-tail test or $t_{1/2\alpha}$ for two-tail test) is taken from TABLE 2 for DF and an acceptable α-level. If $t \geq t_\alpha$ (or $t_{1/2\alpha}$), H_0 is rejected and H_1 is accepted. If $t < t_\alpha$ (or $t_{1/2\alpha}$), H_0 cannot be rejected and is accepted. The confidence level of the test is $(1-\alpha)100\%$.

- **The Student t-test for the conformity between population mean and a theoretical (assumed) mean**

For two-tail test:

$$\begin{cases} H_0: \mu = \mu^* \\ H_1: \mu \neq \mu^* \end{cases}$$

For one-tail test:

$$\begin{cases} H_0: \mu \leq \mu^* \\ H_1: \mu > \mu^* \end{cases} \quad \text{or} \quad \begin{cases} H_0: \mu \geq \mu^* \\ H_1: \mu < \mu^* \end{cases}$$

where: μ^* = a theoretical (assumed) mean value.

The test function is:

$$t = \frac{|\bar{x} - \mu^*|}{s_{\bar{x}}} \quad \text{with } DF = n - 1$$

where: $s_{\bar{x}} = \sqrt{\frac{s_x^2}{n}}$ is the standard deviation of the mean (see p. 19).

The critical value (t_α or $t_{1/2\alpha}$) is taken from TABLE 2 for DF and an acceptable significance level (α). If $t \geq t_\alpha$ (or $t_{1/2\alpha}$), H_0 is rejected and H_1 is accepted. If $t < t_\alpha$ (or $t_{1/2\alpha}$), H_0 cannot be rejected and is accepted. The test's confidence level is $(1-\alpha)100\%$.

• Chi-square test for the conformity between population variance and a theoretical (assumed) variance

For two-tail test:
$$\begin{cases} H_0: \sigma^2 = \sigma_*^2 \\ H_1: \sigma^2 \neq \sigma_*^2 \end{cases}$$

For one-tail test:
$$\begin{cases} H_0: \sigma^2 \geq \sigma_*^2 \\ H_1: \sigma^2 < \sigma_*^2 \end{cases} \quad \text{or} \quad \begin{cases} H_0: \sigma^2 \leq \sigma_*^2 \\ H_1: \sigma^2 > \sigma_*^2 \end{cases}$$

where: σ_*^2 = theoretical (or assumed) general variance value.

The test function is:

(Greek letter chi)
$$\boxed{\chi^2 = \frac{(n-1)S^2}{\sigma_*^2}} \quad \text{with } DF = n-1$$

The critical value, χ^2_α (for one-tail test) or $\chi^2_{1/2\alpha}$ (two-tail test) is from TABLE 4 for DF and acceptable significance level α.
If $\chi^2 \geq \chi^2_\alpha$ (or $\chi^2_{1/2\alpha}$), H_0 is rejected and H_1 is accepted.

When $n > 30$ (which means a "large" sample population), the following test function can be used:
$$\boxed{Z = \sqrt{2\chi^2} - \sqrt{2n-3}}$$

which is based on standardized normal distribution.
The critical value is taken from TABLE 1B for acceptable significance level α.

If $Z \geq Z_\alpha$ (or $Z_{1/2\alpha}$), H_0 is rejected and H_1 is accepted. Otherwise, H_0 cannot be rejected. The test's confidence level is $(1-\alpha)100\%$

• Confidence interval for the difference between two means

When H_0 has been rejected in the Student t-test, the difference can be estimated as:
$$\boxed{[(\bar{x}_1 - \bar{x}_2) - L] < (\mu_1 - \mu_2) < [(\bar{x}_1 - \bar{x}_2) + L]}$$
with $(1-\alpha)100\%$ confidence coefficient

where: $\bar{x}_1$ = larger mean;
$\bar{x}_2$ = smaller mean;
L = the half-interval value, calculated as shown below.

If $\sigma_1^2 = \sigma_2^2$ (H_0 accepted in F-test):
$$L = t_{1/2\alpha}\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{where } s^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$
($t_{1/2\alpha}$ from TABLE 2 for $DF = n_1 + n_2 - 2$)

If $\sigma_1^2 \neq \sigma_2^2$ (H_0 rejected in F-test):
$$L = Z_{1/2\alpha}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$
($Z_{1/2\alpha}$ from TABLE 1B)

The Tukey test for the difference between two means

This is a simple alternative to the Student t-test, based on confidence interval. Hypothesis formulation is as before:

$$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$$

for "two-tail" test (with $t_{\frac{1}{2}\alpha}$ used as critical value)

$$\begin{cases} H_0: \mu_1 - \mu_2 \leq 0 \\ H_1: \mu_1 - \mu_2 > 0 \end{cases} \quad \text{or} \quad \begin{cases} \mu_1 - \mu_2 \geq 0 \\ \mu_1 - \mu_2 < 0 \end{cases}$$

for "one-tail" test (with t_α used as critical value)

Procedure:
Calculate the confidence interval for $(\mu_1 - \mu_2)$ as shown on the previous page and consider the signs (plus vs. minus) of its limits.
If both limits of the interval are positive, H_0 is rejected.
If one limit is positive and the other is negative (i.e., the interval includes zero), H_0 cannot be rejected and is accepted.
The confidence level of this easy test is the same as that of the confidence interval used, $(1-\alpha)100\%$.

Confidence interval for the difference between two proportions

When the sample populations are sufficiently large (n_1 and $n_2 > 100$):

$$\left[(P_1 - P_2) - 1.96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}\right] < (\pi_1 - \pi_2) < \left[(P_1 - P_2) + 1.96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}\right]$$

with confidence level of 95%.

When one or both of the sample populations are of medium size (roughly, $25 < n < 100$), a cruder estimate is made by replacing 1.96 in the formula above with $t_{\frac{1}{2}\alpha}$ for $\alpha = 5\%$ or 10% (which will mean confidence level of 95% or 90%, respectively).

When one or both of the sample populations are small ($n \leq 25$), and also in any other case, the following estimation can be used:

$$\left[(P_1 - P_2) - \sqrt{L^2_{(1)\alpha} + L^2_{(2)\alpha}}\right] < (\pi_1 - \pi_2) < \left[(P_1 - P_2) + \sqrt{L^2_{(1)\alpha} + L^2_{(2)\alpha}}\right]$$

with confidence level of $(1-\alpha)100\%$
and the values of $L_{(1)}$ and $L_{(2)}$ for the two populations read off from TABLE 5 (diagram) for $\alpha = 0.05$.

NORMALITY TESTS

• The Cramer-von Mises test for distribution normality

This test is for <u>non-grouped</u> data, and the hypotheses are:

H_0: The sampled population conforms to normal distribution
H_1: The sampled population does not conform to normal distrib.

This is a two-tail test. The values of $\bar{x}$ and s_x are calculated in advance.

Index ↓ — ↓ Data ordered from the smallest to largest value

Calculatory table

i	x_i	$x_i - \bar{x}$	$z_i = \frac{x_i - \bar{x}}{s_x}$	$P_i = P(z_i)$	$\frac{2i-1}{2n}$	$P_i - \frac{2i-1}{2n}$	$\left(P_i - \frac{2i-1}{2n}\right)^2$
1							
2							
3							
⋮							
n							
	$\sum^n x_i$						$\sum^n \left(P_i - \frac{2i-1}{2n}\right)^2$

z_i = standardized variable

The corresponding cumulative frequency of standard normal distrib.:
$P_i = P\left(\frac{x_i - \bar{x}}{s_x}\right) = P(z_i)$
which are the P-values derived for particular Z-values from TABLE 1A

directly into the formula

The test function is:

omega-square symbol →

$$\omega^2 = \left[\sum^n \left(P_i - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}\right]\left(1 + \frac{1}{2n}\right)$$

and its critical value is taken, for $\frac{1}{2}\alpha$, from TABLE 6.

If $\omega^2 \geq \omega^2_{\frac{1}{2}\alpha}$, H_0 is rejected and H_1 is accepted (the population's distribution is not normal).

If $\omega^2 < \omega^2_{\frac{1}{2}\alpha}$, H_0 is accepted (the population's distribution is normal).

The test's confidence level is: $(1-\alpha)100\%$.

● The Pearson chi-square test for distribution normality

H_0: The sampled population conforms to normal distribution
H_1: The sampled population does not conform to normal distrib.

This test is an equivalent of the previous one, but for grouped data. The number frequency of each class should be greater than 5 ($n_i > 5$). This is a two-tail test. (For calculation of $\bar{x}$ and s_x from grouped data, see p. 17, bottom).

The number of classes is k

Class number frequencies

Class upper limits

Calculatory table

Classes	n_i	x_i	$z_i = \frac{x_i - \bar{x}}{s_x}$	P_i	p_i	np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
...	...							
	$\Sigma n_i = n$							$\sum_{}^{k} \frac{(n_i - np_i)^2}{np_i}$

From TABLE 1A for each standardized value z_i (using extrapolation, if needed). These are cumulative frequencies of standard normal distribution.

Non-cumulative frequencies ("decumulated" from P_i column). Converted into number frequencies

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \qquad \text{with } DF = k - 3$$

where: k = number of classes.

The critical value is taken from TABLE 4 (for DF and and an acceptable significance level α).

If $\chi^2 \geq \chi^2_{\frac{1}{2}\alpha}$, H_0 is rejected and H_1 accepted.

If $\chi^2 < \chi^2_{\frac{1}{2}\alpha}$, H_0 cannot be rejected (the population's distribution is normal).

The test's confidence level is $(1-\alpha)100\%$

NON-PARAMETRIC COMPARISON OF TWO POPULATIONS

The following tests are independent of the variable's type of distribution and can be used for populations with normal, non-normal or unknown distributions.

• Runs test for the difference between two populations

H_0: The two populations are the same (have the same distribution)
H_1: The two populations are different (have different distributions).
This statistical hypothesis will be reformulated below to render the test "one-tail". The test procedure is as follows:
Take the data from $\{X_1\}$ and $\{X_2\}$ jointly and order them from the smallest to the largest value, but mark also to which population (1 or 2) does particular value belong. Count the number of runs (a "run" is an uninterrupted occurrence of data from the same population). For example, $n_1 = 8$ and $n_2 = 9$ with:

smallest value ↓ … largest value ↓

x_1	x_1	x_1	x_2	x_2	x_1	x_1	x_2	x_2	x_2	x_2	x_1	x_2	x_1	x_1	x_2	x_2
1			2		3		4				5	6	7		8	

Number of runs (u_o) = 8

Calculate:

u_o = the observed number of runs;

$\bar{u}_e$ = the expected mean number of runs in a random series of data from two populations:

$$\bar{u}_e = 1 + \frac{2n_1 n_2}{n_1 + n_2}$$

$s_{\bar{u}}$ = the expected standard deviation of the mean:

$$s_{\bar{u}} = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

The hypothesis can be formulated for one-tail test as follows:

$$\begin{cases} H_0: & u_o \leq \bar{u}_e \\ H_1: & u_o > \bar{u}_e \end{cases} \quad \text{or} \quad \begin{matrix} u_o \geq \bar{u}_e \\ u_o < \bar{u}_e \end{matrix}$$

and the test function is:

$$Z = \frac{|u_o - \bar{u}_e|}{s_{\bar{u}}}$$

← Absolute value of the difference

with the critical value Z_α to be read off from TABLE 1B.
If $Z \geq Z_\alpha$, H_0 is rejected with confidence $(1-\alpha)100\%$, which means significant difference between the two populations. If $Z < Z_\alpha$, H_0 cannot be rejected, which means lack of significant difference

Example: Turbidite sandstone thicknesses (cm) in two outcrops.

$\{X_1\}$, $n_1 = 10$: 28, 30, 24, 42, 33, 24, 31, 29, 35, 45

$\{X_2\}$, $n_2 = 9$: 27, 26, 34, 39, 25, 26, 27, 38, 43

The Mann-Whitney test (rank sum U-test)

An alternative test similar to the previous one, with the same hypothesis:

H_0: The two populations are the same
H_1: The two populations are different

For this test, the sample populations $\{X_1\}$ and $\{X_2\}$ must be larger than 13 (i.e., n_1 and $n_2 > 13$)*. If either of them is ≤ 20, $\alpha = 0.01$ has to be used, but when both are >20, $\alpha = 0.05$ is sufficient.

The data from $\{X_1\}$ and $\{X_2\}$ are taken jointly and ordered from the smallest to the largest value. Numbers, or <u>ranks</u> $R(x_i)$, are given to the successive data x_i (that is, $R = 1, 2, 3, \ldots$ etc.).

The ranks corresponding to data from $\{X_1\}$ are totalled as $T_1 = \sum R(x_{1i})$ and the ranks corresponding to $\{X_2\}$ are totalled as $T_2 = \sum R(x_{2i})$.

The following two values are then calculated:

$$u_1 = n_1 n_2 + n_1(n_1+1)/2 - T_1$$
$$u_2 = n_1 n_2 + n_2(n_2+1)/2 - T_2$$

To avoid calculatory errors, check if $u_1 + u_2 = n_1 n_2$

and the <u>smaller</u> u is taken for further consideration.

The mean $\bar{u}_e$ expected for a random data series and the corresponding standard deviation of the mean $s_{\bar{u}}$ are calculated:

$$\bar{u}_e = \frac{n_1 n_2}{2} \qquad s_{\bar{u}} = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$$

The hypothesis can then be formulated in one-tail version:

H_0: $u \leq u_e$ or $u \geq u_e$
H_1: $u > u_e$ or $u < u_e$

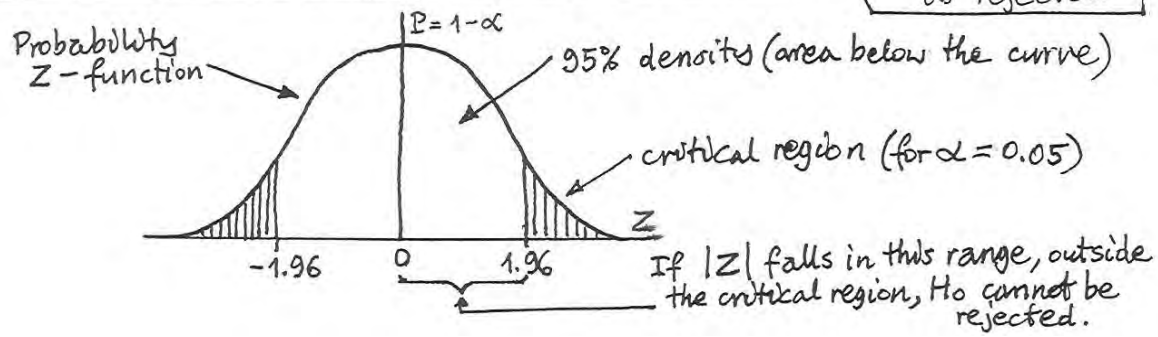
and the test function is calculated:

$$Z = \frac{u - \bar{u}_e}{s_{\bar{u}}}$$

u = the smaller value of u_1 and u_2

with the critical value taken from TABLE 1B for suitable α. If $Z \geq Z_\alpha$, H_0 is rejected with confidence $(1-\alpha)100\%$. If $Z < Z_\alpha$, H_0 cannot be rejected.

* If n_1 and/or $n_2 \leq 13$, the u-value (smaller one) is used directly as the test function's value and the critical value u_α in such case is taken from TABLE 7 for $\alpha = 0.05$ or 0.01. The confidence level of the test is $(1-\alpha)100\%$. When $u < u_\alpha$, the H_0 is rejected.



The Kolmogorov-Smirnov test (D-test)

$\begin{cases} H_0: \text{The two populations are the same (have similar distribution).} \\ H_1: \text{The two populations are different.} \end{cases}$

This is a non-parametric test for grouped (frequency) data. The hypothesis (H_1) can be formulated as for one-tail test if all D-values are positive, or all are negative.

Calculatory table:

Classes (intervals)	Population 1			Population 2			D
	Number freq.	Cumul. freq.	Cumul. freq. %	Number freq.	Cumul. freq.	Cumul. freq. %	
... – ...			*			*	
... – ...			*			*	
... – ...			*			*	
etc.			100			100	
Sums:	n_1	—	—	n_2	—	—	—

Difference between the cumulative freq. % values (those marked with the asterisks in the table).

The critical value of D_α is calculated as below:

$$D_\alpha = d_\alpha \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

where d_α is a constant

	$\alpha = 0.05$	$\alpha = 0.01$	
d_α	136	163	For two-tail test
	122	151	For one-tail test

If at least one $|D| > D_\alpha$, then H_0 is rejected (the populations are different); otherwise, H_0 cannot be rejected.

Contingency tables with chi-square test

This is a test for qualitative variable. The hypothesis is the same:

$\begin{cases} H_0: \text{The two populations are the same.} \\ H_1: \text{The two populations are different.} \end{cases}$

- When the variable studied is a binary (two-state) variable, the data table is 2 × 2:

	States of the variable A	States of the variable B	
Population 1	a	c	(a+c)
Population 2	b	d	(b+d)
	(a+b)	(c+d)	$n = a+b+c+d$

Number frequency of data

The test function is:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(b+d)(a+c)}$$

with $DF = (k-1)(w-1) = 1$, where k = number of table columns and w = number of table rows.
This is a two-tail test.

If $\chi^2 \geq \chi^2_{\frac{1}{2}\alpha}$ (value from TABLE 4), H_0 is rejected with confidence $(1-\alpha)100\%$.

Note: When a, b, c or d ≤ 5, the following formula should be used:

$$\chi^2_Y = \frac{(|ad - bc| - n/2)^2 n}{(a+b)(c+d)(b+d)(a+c)}$$

Chi-square test with Yates's correction

■ When the qualitative variable studied is a multistate variable, the data table is $2 \times k$:

	(1)	(2)	(3)	...	(k)	Row sums:
	states of the variable					
Population 1	a_1	a_2	a_3	...	a_k	$= n_1$
Population 2	b_1	b_2	b_3	...	b_k	$= n_2$
Column sums:	$n_{(1)}$	$n_{(2)}$	$n_{(3)}$	...	$n_{(k)}$	N

The test function is:

$$\chi^2 = \frac{N}{n_1} \sum_{i=1}^{k} \frac{a_i^2}{n_{(i)}} + \frac{N}{n_2} \sum_{i=1}^{k} \frac{b_i^2}{n_{(i)}} - N$$

with $DF = k - 1$

The critical value is taken from TABLE 4. This is a two-tail test. When $\chi^2 \geq \chi^2_{1/2\alpha}$, H_0 is rejected with confidence $(1-\alpha)100\%$. Otherwise, H_0 cannot be rejected.

■ SIMULTANEOUS COMPARISON OF SEVERAL UNIVARIATE POPULATIONS

● The Bartlett chi-square test (parametric test for homogeneity of variances)

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$

H_1: At least one of the variances is significantly different from the others, $\sigma_i^2 \neq \sigma_j^2$

The following calculations are done (table below). The test is two-tail.

Sample popul.	$(\sum x_i)^2$	$\sum x_i^2$	DF	$1/DF$	s_i^2	$\log s_i^2$	$(n_i-1)\log s_i^2$	$(n_i-1)s_i^2$
1	$\left(\sum_{j=1}^{n_1} x_{1j}\right)^2$	$\sum_{j=1}^{n_1} x_{1j}^2$	$n_1 - 1$	$1/(n_1-1)$	s_1^2	$\log s_1^2$	$(n_1-1)\log s_1^2$	$(n_1-1)s_1^2$
2	$\left(\sum_{j=1}^{n_2} x_{2j}\right)^2$	$\sum_{j=1}^{n_2} x_{2j}^2$	$n_2 - 1$	$1/(n_2-1)$	s_2^2	$\log s_2^2$	$(n_2-1)\log s_2^2$	$(n_2-1)s_2^2$
3	$\left(\sum_{j=1}^{n_3} x_{3j}\right)^2$	$\sum_{j=1}^{n_3} x_{3j}^2$	$n_3 - 1$	$1/(n_3-1)$	s_3^2	$\log s_3^2$	$(n_3-1)\log s_3^2$	$(n_3-1)s_3^2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$\left(\sum_{j=1}^{n_k} x_{kj}\right)^2$	$\sum_{j=1}^{n_k} x_{kj}^2$	$n_k - 1$	$1/(n_k-1)$	s_k^2	$\log s_k^2$	$(n_k-1)\log s_k^2$	$(n_k-1)s_k^2$
Sums:	$\sum(\sum x_i)^2$	$\sum\sum x_i^2$	$\sum(n_i-1)$	$\sum \frac{1}{n_i-1}$			$\sum_{i=1}^{k}(n_i-1)\log s_i^2$	$\sum_{i=1}^{k}(n_i-1)s_i^2$

Common (pooled) variance: $s^2 = \dfrac{\sum (n_i-1)s_i^2}{\sum (n_i-1)}$

$s_i^2 = \dfrac{n_i \sum x_i^2 - (\sum x_i)^2}{n_i(n_i-1)}$ — For pooled variance

$$B = (\log s^2)\sum(n_i-1) \longrightarrow \chi^2 = \frac{2.3026}{\ln 10}\left[B - \sum_{i=1}^{k}(n_i-1)\log s_i^2\right]$$

Correction factor: $C = 1 + \left[\frac{1}{3}(k-1)\right]\left[\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{\sum_{i=1}^{k}(n_i-1)}\right]$

Corrected test function: $\chi_c^2 = \frac{1}{C}\chi^2$

The critical value for $DF = k-1$ and $\frac{1}{2}\alpha$ is taken from TABLE 4. If $\chi_c^2 \geq \chi^2_{1/2\alpha}$, H_0 is rejected with confidence $(1-\alpha)100\%$.

• The Tukey-Cramer test (parametric test for homogeneity of means)

This test is applied when variances appear to be homogeneous (i.e., when H_0 is accepted in Bartlett's test). It's a two-tail test.

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_m$

H_1: At least one of the means is significantly different from the others ($\mu_i \neq \mu_k$).

Calculatory table:

Sample popul. (i)	n_i	$\sum^{n_i} x_{ij}$	$(\sum^{n_i} x_{ij})^2/n_i$	Test function
1	n_1	$\sum^{n_1} x_{1j}$	$(\sum^{n_1} x_{1j})^2/n_1$	$F = \dfrac{S_B^2}{S_W^2}$ with $\begin{cases} DF_1 = m-1 \\ DF_2 = N-m \end{cases}$
2	n_2	$\sum^{n_2} x_{2j}$	$(\sum^{n_2} x_{2j})^2/n_2$	
3	n_3	$\sum^{n_3} x_{3j}$	$(\sum^{n_3} x_{3j})^2/n_3$	
⋮	⋮	⋮	⋮	
m	n_m	$\sum^{n_m} x_{mj}$	$(\sum^{n_m} x_{mj})^2/n$	
Sums:	N	$\sum^{m}\sum^{n_i} x_{ij}$	$\sum^{m}[(\sum^{n_i} x_{ij})^2/n_i]$	Critical value $F_{1/2\alpha}$ from TABLE 3

$$(\sum\sum^{n_i} x_{ij})^2/N$$

Variation between populations:

Sum of squares: $SS_B = \sum^{m}[(\sum^{n_i} x_{ij})^2/n_i] - (\sum^{m}\sum^{n_i} x_{ij})^2/N$

Mean square (variance): $\boxed{S_B^2 = SS_B/(m-1)}$

Variation within populations:

Sum of squares: $SS_W = (\sum^{m}\sum^{n_i} x_{ij})^2 - \sum^{m}[(\sum^{n_i} x_{ij})^2/n_i]$

Mean square (variance): $\boxed{S_W^2 = SS_W/(N-m)}$

If $F \geq F_{1/2\alpha}$, H_0 is rejected and H_1 is accepted.
If $F < F_{1/2\alpha}$, H_0 cannot be rejected. The test's confidence level is $(1-\alpha)100\%$.

When H_0 has been rejected, the Tukey test (see p. 29) can be used further, for each pair of populations, to identify those whose means differ from the others.

One-way ANOVA

ANOVA (ANalysis Of VAriance) is a parametric method. In this simpliest ("one-way") version, ANOVA resembles the Tukey-Cramer test and may serve as its alternative for testing the homogeneity of population means – with two-tail hypothesis:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_m \\ H_1: \text{At least one mean is significantly different from the others } (\mu_i \neq \mu_k). \end{cases}$$

However, ANOVA focuses on the sources of variance and their relative contribution to the overall variation of variable x.

One-way ANOVA table:

Source of variance	Sum of squares	DF	F-test
Variation between populations	SS_B	$m-1$	$F = \frac{S_B^2}{S_W^2}$ with $\begin{cases} DF_1 = m-1 \\ DF_2 = N-m \end{cases}$
Variation within populations	SS_W	$N-m$	
Total variation	SS_T	$N-1$	

with: $n_1 = n_2 = n_3 = \ldots = n_m = n$; N = total number of data.

The sums of squares are calculated as follows:

$$SS_B = \sum^{m}\left[\frac{(\sum^{n} x)^2}{n}\right] - \frac{(\sum^{m}\sum^{n} x)^2}{N} \longrightarrow \text{variance } S_B^2 = \frac{SS_B}{m-1}$$

$$SS_W = \sum^{m}\sum^{n} x^2 - \sum^{m}\left[\frac{(\sum^{n} x)^2}{n}\right] \longrightarrow \text{variance } S_W^2 = \frac{SS_W}{N-m}$$

$$SS_T = \sum^{m}\sum^{n} x^2 - \frac{(\sum^{m}\sum^{n} x)^2}{N} \longrightarrow \text{variance } S_T^2 = \frac{SS_T}{N-1}$$

Note: $SS_T = SS_B + SS_W$

The critical value $F_{1/2\alpha}$ is taken from TABLE 3. If $F \geq F_{1/2\alpha}$, H_0 is rejected with $(1-\alpha)100\%$ confidence; otherwise H_0 can't be rejected.

Note: ANOVA requires that the sample populations are of equal size ($n_1 = n_2 = n_3 = \ldots = n$). If the population sizes are unequal, the problem can be solved in one of the following ways:

(1) The "outsized" n_i values can be appropriately reduced by replacing some pairs of data (closest) with their means.

(2) An "average" size ($\bar{n}$) can be used, calculated by the following recommended formula:

$$\bar{n} = \frac{1}{m-1}\left(N - \frac{\sum^{m} n_i^2}{N}\right),$$

or simply as an arithmetic mean: $\bar{n} = \frac{\sum^{m} n_i}{m}$.

EXAMPLE

Factor →

$\{x_1\}$ $n_1=6$	$\{x_2\}$ $n_2=6$	$\{x_3\}$ $n_3=6$	$\{x_4\}$ $n_4=6$	$\{x_5\}$ $n_5=6$
19.2	18.7	12.5	20.3	19.9
18.7	14.3	14.3	22.5	24.3
21.3	20.2	8.7	17.6	17.6
16.5	17.6	11.4	18.4	20.2
17.3	19.3	9.5	15.9	18.4
22.4	16.1	16.5	19.0	19.1

$m=5$
$N=30$

$\sum^n X =$					$\sum^m \sum^n X =$	
$\left(\sum^n X\right)^2 =$					$\sum^m \left(\sum^n X\right)^2 =$	
$\sum^n X^2 =$					$\sum^m \sum^n X^2 =$	
$\frac{\left(\sum^n X\right)^2}{n} =$					$\sum^m \frac{\left(\sum^n X\right)^2}{n} =$	

$SS_B =$
$SS_W =$
$SS_T =$

$S_B^2 =$
$S_W^2 =$
$S_T^2 =$

$F = \frac{S_B^2}{S_W^2} =$

● Two-way ANOVA

In this approach, the relative influence of <u>two</u> factors on the variation of x is considered simultaneously. The data table is:

Factor 2 →

Factor 1 ↓	1	2	3	...	k	
1	x_{11}	x_{12}	x_{13}	...	x_{1k}	$\bar{x}_{1*}$
2	x_{21}	x_{22}	x_{23}	...	x_{2k}	$\bar{x}_{2*}$
3	x_{31}	x_{32}	x_{33}	...	x_{3k}	$\bar{x}_{3*}$
⋮	⋮	⋮	⋮		⋮	⋮
m	x_{m1}	x_{m2}	x_{m3}	...	x_{mk}	$\bar{x}_{m*}$
	$\bar{x}_{*1}$	$\bar{x}_{*2}$	$\bar{x}_{*3}$	...	$\bar{x}_{*k}$	

$N = k \cdot m$
$\bar{x}$ = grand mean

row means

column means

and the two-tail hypothesis is as follows:

H_0: $\mu_{1*} = \mu_{2*} = \mu_{3*} = \ldots = \mu_{m*} = \mu_{*1} = \mu_{*2} = \mu_{*3} = \ldots = \mu_{*k}$
(i.e., the means of all row populations and column populations are equal)

H_1: Not all of the general means are equal.

Two-way ANOVA table:

Source of variance	Sum of squares	DF	F-test
Variation between rows (factor 1)	SS_1	$m-1$	$F_1 = \frac{S_1^2}{S_R^2}$, with $\begin{cases} DF_1 = m-1 \\ DF_2 = (m-1)(k-1) \end{cases}$
Variation between columns (factor 2)	SS_2	$k-1$	$F_2 = \frac{S_2^2}{S_R^2}$, with $\begin{cases} DF_1 = k-1 \\ DF_2 = (m-1)(k-1) \end{cases}$
Residual variation within columns	SS_R	$(m-1)(k-1)$	Critical value $F_{1/2\alpha}$ from TABLE 3
Total variation	SS_T	$N-1$	

$$SS_1 = k\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x})^2 \qquad S_1^2 = \frac{SS_1}{m-1}$$

$$SS_2 = m\sum_{j=1}^{k}(\bar{x}_{*j} - \bar{x})^2 \qquad S_2^2 = \frac{SS_2}{k-1}$$

$$SS_T = \sum x_{ij}^2 - \frac{1}{N}\left(\sum x_{ij}\right)^2 \qquad S_T^2 = \frac{SS_T}{N-1}$$

$$SS_R = SS_T - SS_1 - SS_2 \qquad S_R^2 = \frac{SS_R}{(m-1)(k-1)}$$

Values to use in F-test

EXAMPLE

Factor 1 \ Factor 2	1	2	3	
1	25	30	23	$\bar{x}_{1*} = 26$
2	20	40	18	$\bar{x}_{2*} = 26$
3	30	40	20	$\bar{x}_{3*} = 30$
4	25	50	27	$\bar{x}_{4*} = 34$
	$\bar{x}_{*1} = 25$	$\bar{x}_{*2} = 40$	$\bar{x}_{*3} = 22$	

$N = 12$
$\bar{x} = 29$

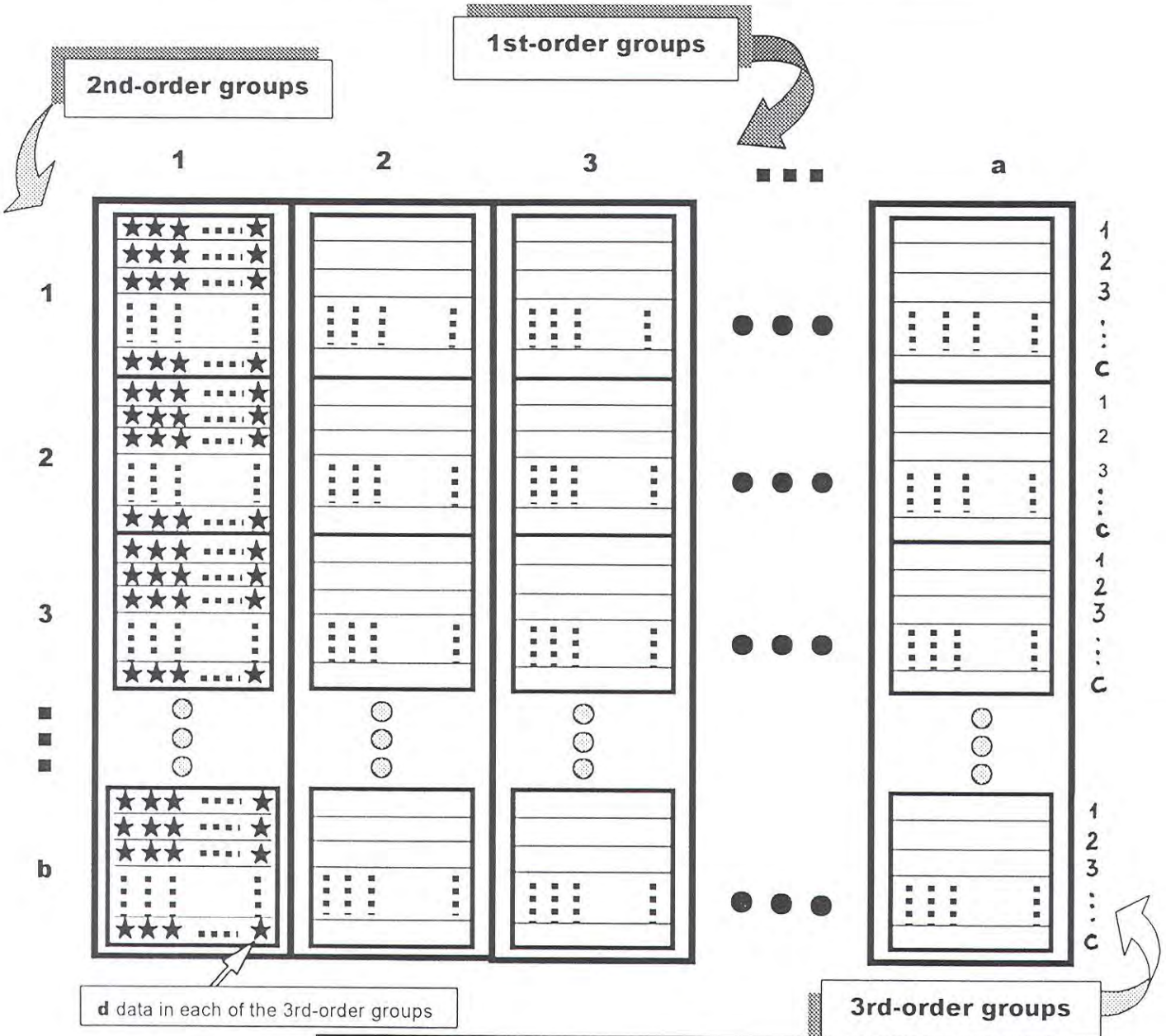
$SS_1 =$ $S_1^2 =$
$SS_2 =$ $S_2^2 =$ $F_1 =$
$SS_T =$ $S_T^2 =$ $F_2 =$
$SS_R =$ $S_R^2 =$

● Three-way ANOVA

In this approach, the relative influence of three factors on the variation of x is considered simultaneously. The data are organized into 1st-, 2nd- and 3rd-order groups, each order corresponding to one factor.

The structure of data table is shown, in a generalized form, on the next page. Note that the table consists of rows and columns, as before; only the data are considered to represent 3 levels of hierarchical groups.

The sample population is organized into hierarchical groups.
Three levels of data grouping are used in 3-way ANOVA.



N – total number of data
a – number of 1st-order groups
b – number of 2nd-order groups
c – number of 3rd-order groups
d – number of data in each 3rd-order group
N = a b c d

The questions addressed by ANOVA are:

- Do the 1st-order groups differ more among themselves than do the 2nd-order groups?
- Do the 2nd-order groups differ more among themselves than do the 3rd-order groups?
- Is the population sampled "homogeneous" at any of the levels?
- Is the variation between the 1st-order groups greater than the variation between the 2nd-order groups?
- Is the variation between the 2nd-order groups greater than the variation between the 3rd-order groups (i.e., within the 2nd-order groups)?

The F-tests used are comparing the groups at any particular hierarchical level without regard to their relative variability in respect to the other levels.

Three-way ANOVA table:

Source of variance	Sum of squares	DF	Mean square (variance)	F-test
Variation between 1st-order groups	SS_1	$DF_1 = a-1$	$S_1^2 = \frac{SS_1}{DF_1}$	$F_1 = \frac{S_1^2}{S_2^2}$ with $\begin{cases} DF_1 \\ DF_2 \end{cases}$
Variation between 2nd-order groups	SS_2	$DF_2 = a(b-1)$	$S_2^2 = \frac{SS_2}{DF_2}$	$F_2 = \frac{S_2^2}{S_3^2}$ with $\begin{cases} DF_2 \\ DF_3 \end{cases}$
Variation between 3rd-order groups	SS_3	$DF_3 = ab(c-1)$	$S_3^2 = \frac{SS_3}{DF_3}$	$F_3 = \frac{S_3^2}{S_4^2}$ with $\begin{cases} DF_3 \\ DF_4 \end{cases}$
Variation within 3rd-order groups	SS_4	$DF_4 = abc(d-1)$	$S_4^2 = \frac{SS_4}{DF_4}$	Critical value $F_{1/2\alpha}$ from TABLE 3
Total variation	SS_T	$DF_T = N-1$	$S_T^2 = \frac{SS_T}{DF_T}$	

The sums of squares (SS) are calculated as follows:

$$SS_1 = \frac{\sum^{a}\left(\sum^{b} x_i\right)^2}{b \cdot c \cdot d} - \frac{\left(\sum^{N} x_i\right)^2}{N} \qquad SS_2 = \frac{\sum^{a}\sum^{b}\left(\sum^{c} x_i\right)^2}{c \cdot d} - \frac{\sum^{a}\left(\sum^{b} x_i\right)^2}{b \cdot c \cdot d}$$

$$SS_3 = \frac{\sum^{a}\sum^{b}\sum^{c}\left(\sum^{d} x_i\right)^2}{d} - \frac{\sum^{a}\sum^{b}\left(\sum^{c} x_i\right)^2}{c \cdot d} \qquad SS_4 = \sum^{a}\sum^{b}\sum^{c}\sum^{d}(x_i)^2 - \frac{\sum^{a}\sum^{b}\sum^{c}\left(\sum^{d} x_i\right)^2}{d}$$

$$SS_T = \sum^{a}\sum^{b}\sum^{c}\sum^{d}(x_i)^2 - \frac{\left(\sum^{N} x_i\right)^2}{N}$$

where:
- a = number of 1st-order groups
- b = number of 2nd-order groups
- c = number of 3rd-order groups
- d = number of data in each 3rd-order group (all 3rd-order groups should be of equal size, d = const.)
- N = total number of data ($a \cdot b \cdot c \cdot d$).
- $SS_T = SS_1 + SS_2 + SS_3 + SS_4$

● The Kruskal–Wallis test (rank sum H-test)

This is a non-parametric and two-tail test for hypotheses:

H_0: The populations have the same distribution (are non-different).

H_1: At least one population has a distribution different from others.

The test procedure is the same as in the Mann-Whitney test (p. 33), but the following test function is used (m = number of populations):

$$\boxed{H = \frac{12}{N(N+1)} \sum^{m} \frac{T_i^2}{n_i} - 3(N+1)} \quad \text{with } DF = m-1$$

Its critical value is taken from TABLE 2 as $\chi^2_{1/2\alpha}$.

If $H \geq \chi^2_{1/2\alpha}$, H_0 is rejected with confidence $(1-\alpha)100\%$.

Test requirement: $N \geq 50$, or at least each $n_i \geq 5$.

THE CONSTRUCTION OF ROSE DIAGRAM (CIRCULAR HISTOGRAM)

The circular histogram, more often referred to as *rose diagram* by geologists (Potter and Pettijohn, 1963), is the most commonly used graphical means of presenting the frequency distribution of two-dimensional orientation (vectorial) data. A large number of sedimentary rock attributes, particularly the directional sedimentary features used as palaeocurrent indicators, are measured in terms of their azimuths and subsequently analysed graphically with the use of a rose diagram (Potter and Pettijohn, 1963). Such a plot is relatively easy to construct and gives a good visual impression of the data density variations.

It is useful to plot the two-dimensional vectorial data in a rose-diagram form before an estimation of statistical parameters (such as a vector mean, its magnitude, circular variance, etc.) is attempted. Moreover, rose diagrams are often used by researchers as the only means of summarising their data sets, especially when the frequency distribution appears to be non-unimodal and statistics cannot readily be calculated. In other words,

It is, therefore, rather surprising and quite unfortunate that the vast majority of the rose diagrams that appear in the sedimentological literature, and also in many textbooks, are *incorrect*. Simply, the way in which such diagrams are constructed by most geologists is, from a methodological point of view, wrong. This paper is intended to make the geologists aware of the error and to prevent its continuation in the literature.

It is the principle of histogram construction that the *area* of each column, or circular sector, of the diagram should be made proportional to the class frequency (density). Instead, the circular histogram is often thoughtlessly constructed with the *radius* of the sector proportional to the class frequency. This common mistake apparently derives from a misunderstanding of the nature of a histogram, with a falsely drawn analogy between a simple, columnar histogram and a circular histogram (see Fig. 1 and enclosed comment). The columnar histograms usually have linear horizontal scales (constant class widths), so that the heights of the columns are made linearly proportional to the class frequencies; for example, a column of 20% frequency is made twice as high as the column of 10% (Fig. 1A). This makes the "areal" aspect of the histogram less apparent, and leads to a serious error once the same linear approach is mechanically used to construct a circular histogram.

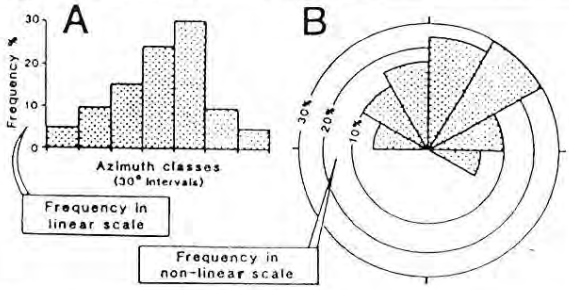
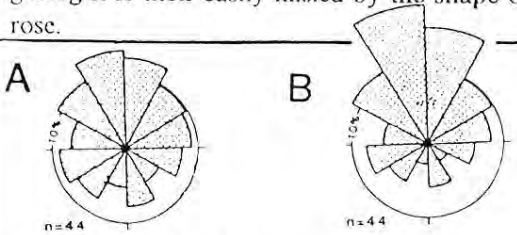


Fig. 1. Simple columnar histogram (A) and its circular equivalent (B). The diagrams, and their equivalent columns (A) and sectors (B), have the same areas. Note that the frequency scale becomes non-linear when histogram A is squeezed into a circular form B.

Comment: In the columnar histogram (A), the horizontal scale is divided into classes and the class widths remain constant for the entire height of the columns. The area of the column is proportional to the number of data in that class and represents the class frequency (density); note carefully that it is the *area*, not the height, of the column. The vertical (frequency) scale here is linear because the horizontal scale is linear and invariable. In the circular histogram (B), the classes are the same as in (A), but their linear scale (actual width) varies along the diagram radius; in such case, the frequency scale can no longer be linear.

Few geologists seem to be aware that the use of a linear (arithmetic) frequency scale in circular histograms leads to gross distortions of the diagram's shape (false relative densities are produced). This is illustrated by the example shown in Fig. 2. When the diagram in Fig. 2B is considered, the eye is likely to be unduly impressed by the dominating size of the modal sector, completely out of proportion to its true significance. The degree of preferred orientation is greatly exaggerated by this kind of plot (Fig. 2B), and the geologist is then easily misled by the shape of the rose.



SAMPLE STATISTICS
Mean vector: $\Theta = 8°$
Vector magnitude (length): $\bar{R} = 0.27$ (27%)
Vector strength: $\bar{a} = 0.006$
Batschelet's circular standard deviation: $s = 81°$

RAYLEIGH'S TEST
Critical values: $\bar{R}_{0.01} = 0.32$, $\bar{R}_{0.05} = 0.26$.
Result: no preferred orientation at 1% significance (risk-of-error) level; a northerly trend can be inferred at 10% or higher level, but is still doubtful at 5% level (since $\bar{R} < \bar{R}_{0.05}$)

Fig. 2. Azimuths of ripple foresets from an early Cretaceous bay-fill sequence (Helvetiafjellet Formation, Svalbard) plotted in a rose diagram; the number of data is $n = 44$. In diagram A, a correct frequency scale is used and the rose shows little preferred orientation; no preferred orientation at $\alpha = 0.01$ significance level is revealed by statistical test. In diagram B, a linear (incorrect) frequency scale is used and the rose shows an apparent "strongly preferred" orientation (false shape).

The correct radius of each sector (or frequency circle) of a rose diagram can be calculated in at least two ways:

(a) The radius, $r_{1\%}$, of a unit frequency ($f = 1\%$) circle is assumed, and the radius of any other frequency circle (or rose's sector), $r_{f\%}$, is then given by:

$$r_{f\%} = r_{1\%}\sqrt{f}\,. \tag{1}$$

For example: if $r_{1\%} = 5$ mm is assumed, then the radius of $f = 25\%$ sector is $r_{25\%} = 5\sqrt{25} = 25$ mm.

The same formula applies when pure number frequency (number of data in classes), rather than number frequency per cent, is used in the diagram; only the % symbol has to be omitted in such instance.

(b) The total area, A, of a rose diagram is assumed, and the radius of a particular frequency circle (or rose's sector), $r_{f\%}$, is then given by:

$$r_{f\%} = \sqrt{(2 \cdot A \cdot f)/(n \cdot w)}\,, \tag{2}$$

where the other two symbols are: n = total number of data (if number frequency used) or 100% (if number frequency per cent used); and w = sector width in radians (1 rad = $180/\pi$ degree).

As in formula (1), the % symbol has to be omitted if pure number frequency, instead of percentages, is used. For example: if $n = 60$ (measurements) and if one assumes $A = 3000$ mm^2 (30 cm^2) and $w = 0.523$ rad (30°), then the radius of $f = 25$ frequency sector is:

$$r_{25} = \sqrt{(2 \cdot 3000 \cdot 25)/(60 \cdot 0.523)} = 69 \text{ mm}.$$

Some computational effort is thus necessary to construct the circular histogram. Although the computation itself is simple and relatively rapid, many geologists may find it somewhat discouraging or inconvenient to use a numerical procedure for purely graphical purposes, especially when a large number of diagrams are to be constructed. It is, therefore, more practical to produce a standard frequency net (Fig. 3) for plotting the data. With the use of such a net, the construction of a rose diagram becomes as rapid as the construction of a simple, columnar histogram.

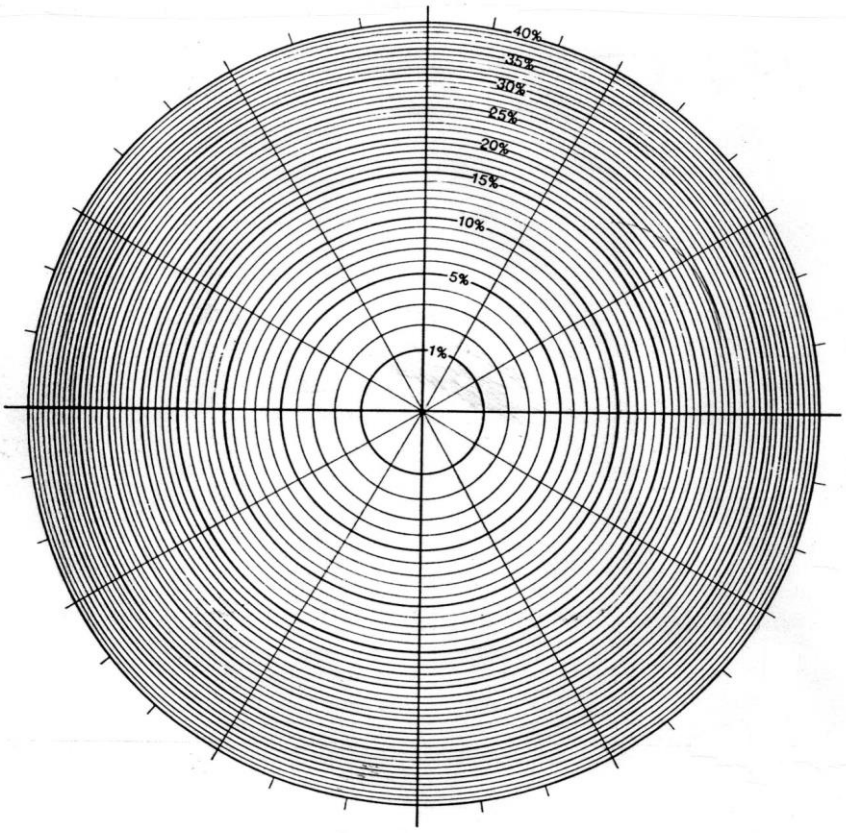


Fig. 3. A frequency net produced to facilitate rapid construction of circular histograms. The net (frequency circles) can be extended to higher values, if needed, with the use of formula (1) given in text; $r_{1\%} = 8$ mm in the present case. The % symbols have to be omitted if pure number frequencies (not percentages) are plotted.

From Nemec (1988) in the journal *Sedimentary Geology*, vol. 59, pp. 149-152.

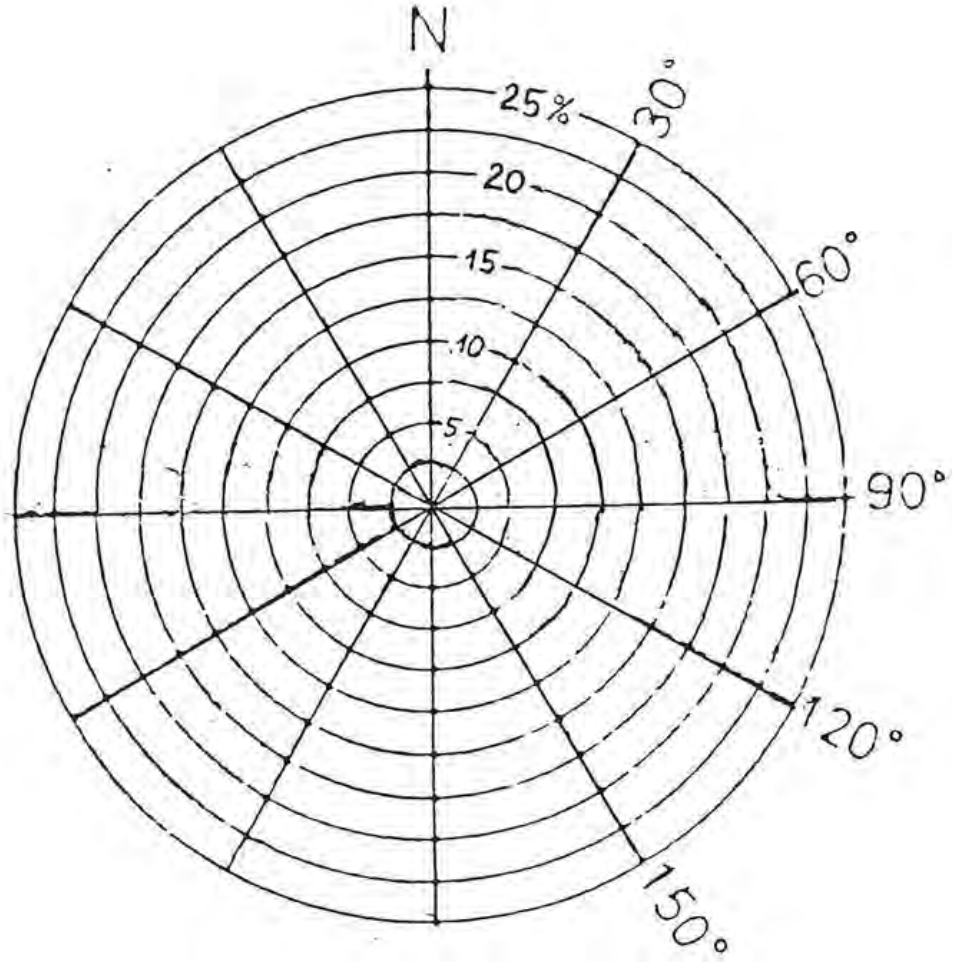
A computer software EZ-ROSE (an application of Excel and CorelDraw plotting correct rose diagrams and calculating all the related statistics and tests) was published by Jaco Baas (2000) in the journal *Computers & Geosciences*, vol. 26, pp. 153-166. The software can be downloaded from the following website
http://www.iamg.org/index.php/publisher/articleview/frmArticleID/103
or can be received upon request by e-mail from <j.baas@bangor.ac.uk>

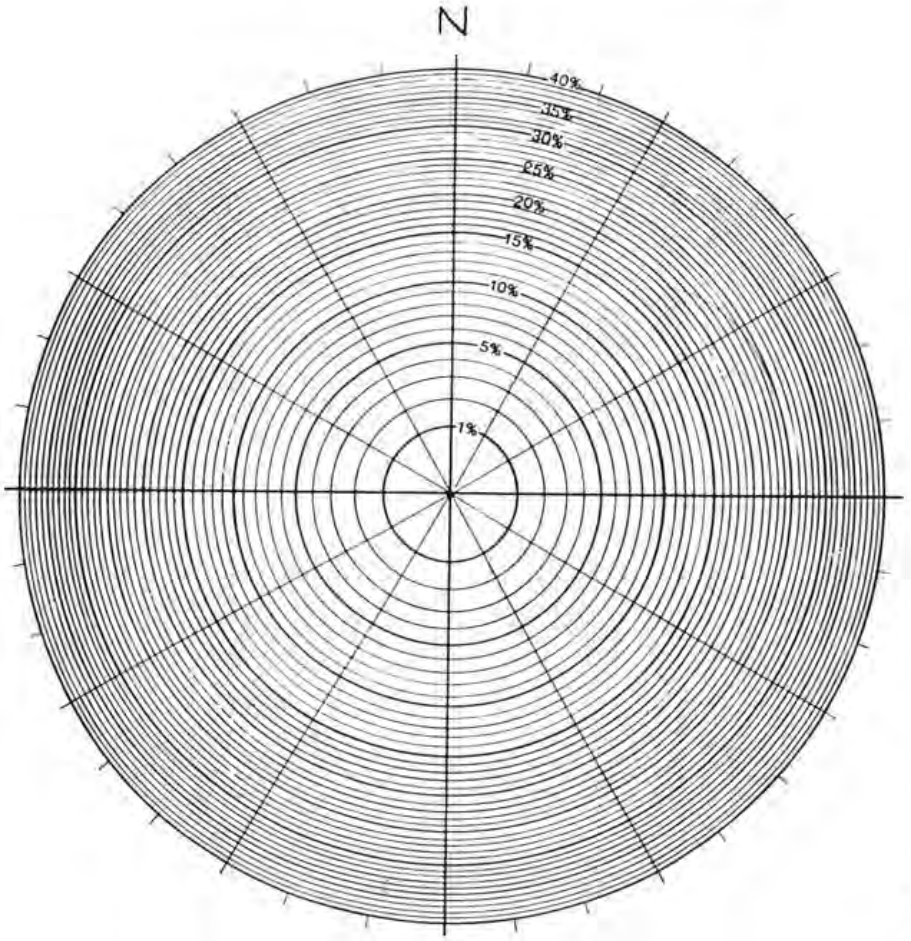
EXAMPLE

Measurements of the azimuth of ripple cross-laminae dip direction in a bay-fill sedimentary succession (Early Cretaceous Helvetiafjellet Fm., Spitsbergen). The data (n=44 measurements) have been grouped in 30°-classes.

Azimuth class (degrees)	Frequency %
0-30	9.5
30-60	8.0
60-90	6.0
90-120	4.0
120-150	6.0
150-180	4.0
180-210	6.0
210-240	8.0
240-270	5.5
270-300	11.0
300-330	18.0
330-360	14.0

n = 44





ANALYSIS OF TWO-DIMENSIONAL (AZIMUTHAL) POPULATION

When dealing with directional data, the first issue is to assess as to whether the general population sampled has a preferential trend or is chaotic (circularly uniform).

The Kuiper test of uniformity (non-parametric)

H_0: The sample population is derived from a general population with uniform (non-preferential) azimuthal distribution.

H_1: The sample population is from a general population with non-uniform (preferential) azimuthal distribution.

The test function is:

$$V_n = \max\left(u_i - \frac{i}{n}\right) - \min\left(u_i - \frac{i}{n}\right) + \frac{1}{n}$$

largest positive deviation

largest negative deviation (which comes here with its minus sign, and thus is added to the positive dev.)

The two deviations can be either calculated (see table below) or found graphically (see graph below). The data in either case must be "modulo 360°" (i.e., lineation measurements need to be all multiplied by 2).

Calculatory table:

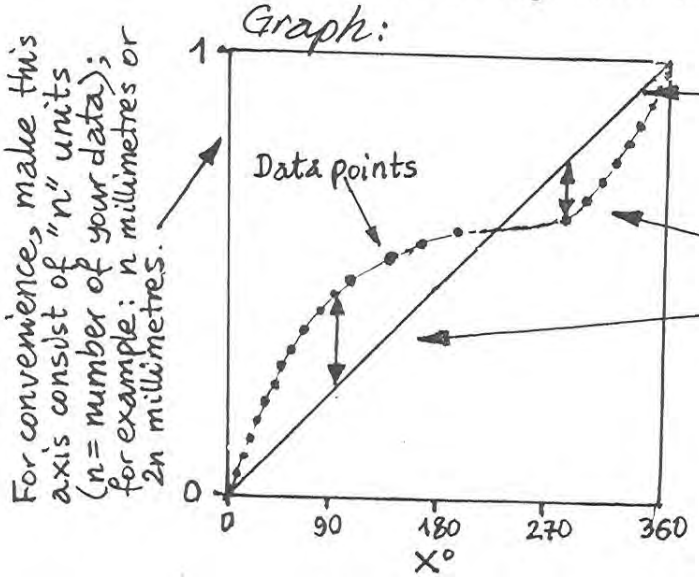
Index → i	x_i°	u_i	i/n	$u_i - i/n$
1	x_1	u_1	$1/n$	$u_1 - 1/n$
2	x_2	u_2	$2/n$	$u_2 - 2/n$
3	x_3	u_3	$3/n$	$u_3 - 3/n$
⋮	⋮	⋮	⋮	⋮
Total number of data → n	x_n	u_n	1	$u_n - 1$

Find the largest positive and the largest negative value in this last column.

$u_i = x_i^\circ / 360^\circ$

Data arranged from the smallest to the largest value (degrees)

Graph:



The straight line represents a perfectly uniform distribution.

From the graph, read off:
the largest negative deviation and
the largest positive deviation

EXAMPLE: $n = 36$
largest positive deviation = + 0.067
largest negative deviation = − 0.164
$V_n = [+0.067 - (-0.164) + \frac{1}{36}] = 0.259$

The test's critical value is calculated as follows:

$$V_\alpha = V^* / (\sqrt{n} + 0.155 + 0.24/\sqrt{n})$$

where: $V^* = 1.75$ for $\alpha = 0.05$
$V^* = 2.00$ for $\alpha = 0.01$

If $V_n \geq V_\alpha$, H_0 can be rejected with $(1-\alpha)100\%$ confidence, which will mean that the population has a non-uniform (preferential) distribution and is suitable for further statistical analysis.

■ The Watson test of uniformity (non-parametric)

This test is somewhat 'stronger', but more labourious, and can serve as an alternative to the Kuiper test.
The data have to be arranged "modulo 360°", as before, and the hypotheses H_0 and H_1 are the same as in the Kuiper test.
The test function (U^2) is calculated as follows.

Calculatory table

i	x_i°	$u_i = \frac{x_i^\circ}{360^\circ}$	u_i^2	iu_i
1				
2				
3	(data)			
⋮				
n		$\sum_1^n u_i$	$\sum_1^n u_i^2$	$\sum_1^n (iu_i)$

Mean: $\bar{u} = \frac{\sum_1^n u_i}{n}$

Partial calculation

$$u_*^2 = \sum_1^n u_i^2 - \frac{2}{n}\sum_1^n (iu_i) + \bar{u} + n\left[\bar{u} - (\bar{u})^2 + \frac{1}{12}\right]$$

where: x_i°, u_i and n are as in the Kuiper test (see table on p.45).

The test function is:

$$U^2 = \left[u_*^2 + \left(\frac{1}{n^2} - \frac{1}{n}\right)/10\right]\left[1 + \frac{0.8}{n}\right]$$

and its critical value U_α^2 is: $\begin{cases} U_{0.05}^2 = 0.187 & \text{for } \alpha = 0.05 \\ U_{0.01}^2 = 0.267 & \text{for } \alpha = 0.01 \end{cases}$

If $U^2 \geq U_\alpha^2$, H_0 is rejected with $(1-\alpha)100\%$ confidence, which means a non-uniform (preferential) distribution, suitable for further analysis.

■ Chi-square test of uniformity (non-parametric, for grouped data)

The H_0 and H_1 are the same as in the two previous tests.
The data set contains N measurements grouped in k azimuthal classes (groups). Requirements: $N/k \geq 5$ and $N/15 \leq k \leq N/5$.

The test function is:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i \quad \text{with } DF = k-1$$

where: O_i = the observed group frequencies (number frequencies);
E_i = expected group frequency for uniform distribution ($E_i = N/k$).

The critical value is taken from TABLE 4.
If $\chi^2 \geq \chi^2_\alpha$, H_o is rejected with $(1-\alpha)100\%$ confidence and the population distribution can be regarded as non-uniform.

EXAMPLE: N = 66 measurements grouped in k = 6 classes.

Azimuth class	Number freq. (O_i)	Uniform freq. (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0-60°	8			
60-120°	9			
120-180°	15			
180-240°	20			
240-300°	7			
300-360°	7			

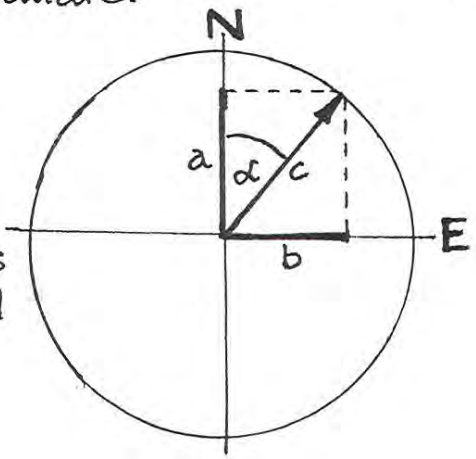
General comment

If the uniformity test (whether Kuiper's or Watson's for non-grouped data, or the chi-square for grouped data) indicates a non-uniform distribution (H_o rejected) – we can tentatively assume that the general population sampled has a circular-normal (von Mises) distribution and we can analyze the data tentatively in a parametric manner, by calculating the directional statistics (next page). These statistics will subsequently be used in the Rayleigh test to check if the assumption of circular-normality is justified.

Circular-normal distribution (see p. 9-10) is characterized by a mean vector (which has an azimuthal direction and a length) and a concentration of vectors. There are thus three parameters, and three statistics (estimators) to calculate.

For the purpose of numerical calculations, the position of a data vector (c) is defined by its "north" co-ordinate (a) and "east" co-ordinate (b). In terms of the vector's azimuth (α), these co-ordinates become the vector's directional cosine and sine:

$$\begin{cases} \cos\alpha = a/c \\ \sin\alpha = b/c \end{cases}$$



EXAMPLE: Measurements of cross-strata dip azimuth in a nearshore sandstone succession ($n = 36$). The data have been arranged from the smallest to the largest value.

Data		Watson's test			Directional sine & cosine	
i	X_i°	u_i	u_i^2	iu_i	$\sin X_i^\circ$	$\cos X_i^\circ$
1	1					
2	8					
3	12					
4	14					
5	17					
6	21					
7	23					
8	32					
9	36					
10	47					
11	58					
12	65					
13	72					
14	81					
15	98					
16	119					
17	146					
18	164					
19	176					
20	188					
21	198					
22	214					
23	236					
24	245					
25	259					
26	274					
27	294					
28	304					
29	313					
30	320					
31	330					
32	332					
33	339					
34	345					
35	349					
36	354					

Parameters and statistics

The parameters and their estimators (statistics) for two-dimensional vectorial population are:

- The mean vector direction (azimuth): γ, estimated as $\hat{\gamma}$.
- The mean vector magnitude (length): R, estimated as $\hat{R}$.
- The vectorial concentration of the population: K, estimated as $\hat{K}$.

The concentration (K) is a parameter basically analogous to variance (σ), but reverse; variance is a measure of dispersion. For the purpose of more conventional statistical analysis (e.g., the use of F-test and/or ANOVA), the statistic $\hat{R}$ can be converted into the Batschelet circular variance (s_B^2) and circular standard deviation (s_B):

$$s_B^2 = 2(1-\hat{R}) \quad \text{[in radians]} \qquad s_B = \frac{180^\circ}{\pi}\sqrt{2(1-b\cdot\hat{R})} \quad \text{[degrees]}$$

where $\hat{R}$ is given as a fraction of one (not as percentage) and $b \simeq 1$ (e.g., this correction coefficient is $b = 1.0013$ for 36 azimuthal classes of 10° width and $b = 1.0015$ for 12 classes of 30° width).

The first three statistics are calculated with the formulae given below (table).

STATISTIC	NON-GROUPED DATA	GROUPED DATA
$\hat{\gamma}$	For circular data (modulo 360°): $\hat{\gamma} = \arctan \frac{\sum_{i}^{n} \sin X_i^\circ}{\sum_{i}^{n} \cos X_i^\circ}$	For circular data (modulo 360°): $\hat{\gamma} = \arctan \frac{\sum_{i}^{k} f_i \cdot \sin X_i^\circ}{\sum_{i}^{k} f_i \cdot \cos X_i^\circ}$
	For semicircular data (lineations): $\hat{\gamma} = \frac{1}{2}\arctan \frac{\sum_{i}^{n} \sin 2X_i^\circ}{\sum_{i}^{n} \cos 2X_i^\circ}$	For semicircular data: $\hat{\gamma} = \frac{1}{2}\arctan \frac{\sum_{i}^{k} f_i \cdot \sin 2X_i^\circ}{\sum_{i}^{k} f_i \cdot \cos 2X_i^\circ}$
	where: X_i° is the measurement (azimuth value) in degrees.	where: k = number of classes; X_i° = class midpoint value; f_i = class frequency; $n = \sum_{i}^{k} f_i$
	When $\hat{\gamma}$ comes out to be negative (e.g., -72°), plot it counter-clockwise (i.e., as $360^\circ - 72^\circ = 288^\circ$ azimuth).	
$\hat{R}$	$\hat{R} = \sqrt{(\bar{C})^2 + (\bar{S})^2}$ The value for plotting in rose is in per cent (multiplied by 100%)	
	where: $\bar{C} = \frac{1}{n}\sum_{i}^{n} \cos X_i^\circ$, $\bar{S} = \frac{1}{n}\sum_{i}^{n} \sin X_i^\circ$	where: $\bar{C} = \frac{1}{n}\sum_{i}^{k} f_i \cos X_i^\circ$, $\bar{S} = \frac{1}{n}\sum_{i}^{k} f_i \sin X_i^\circ$
	For semicircular (lineation) data, $2X_i^\circ$ should be used, as above.	
$\hat{K}$	When $\hat{R} \leq 0.65$, $\hat{K} = \hat{R}(12 + 6\hat{R}^2 + 5\hat{R}^4)/6$ When $\hat{R} > 0.65$, $\hat{K} = [2(1-\hat{R}) - (1-\hat{R})^2 - (1-\hat{R})^3]^{-1}$	

Based on the statistics, we can now check if the population distribution is circular-normal, or of von Mises type (see test on next page).

The Rayleigh test of circular normality

This is a parametric test for the significance of the population's mean resultant vector, but when combined with a uniformity test (pp. 33-34) – it effectively verifies the distribution's circular normality (see further explanation below). The hypothesis pertains to concentration:

H_0: $\kappa = 0$ (the vectors have no concentration and their mean has no significance)
H_1: $\kappa > 0$ (the mean vector is significant)

The test uses the value of statistic $\hat{R}$ as function, and its critical value can be calculated when $n \geq 15$:

$$R_{0.10} = \sqrt{2.35/n} \quad \text{for } \alpha = 0.10$$
$$R_{0.05} = \sqrt{3.00/n} \quad \text{for } \alpha = 0.05$$
$$R_{0.01} = \sqrt{4.61/n} \quad \text{for } \alpha = 0.01$$

When $n < 15$, the critical values R_α are as listed below (table).

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	0.86	0.96	1.13
4	0.75	0.84	1.03
5	0.67	0.75	0.94
6	0.62	0.69	0.87
7	0.57	0.64	0.82
8	0.53	0.60	0.77
9	0.50	0.57	0.73
10	0.48	0.54	0.70
11	0.46	0.52	0.67
12	0.44	0.49	0.64
13	0.42	0.48	0.62
14	0.40	0.46	0.60

n = number of data

If $\hat{R} \geq R_\alpha$, H_0 is rejected and the mean vector is considered to be significant, with confidence $(1-\alpha)100\%$. Otherwise, H_0 cannot be rejected.

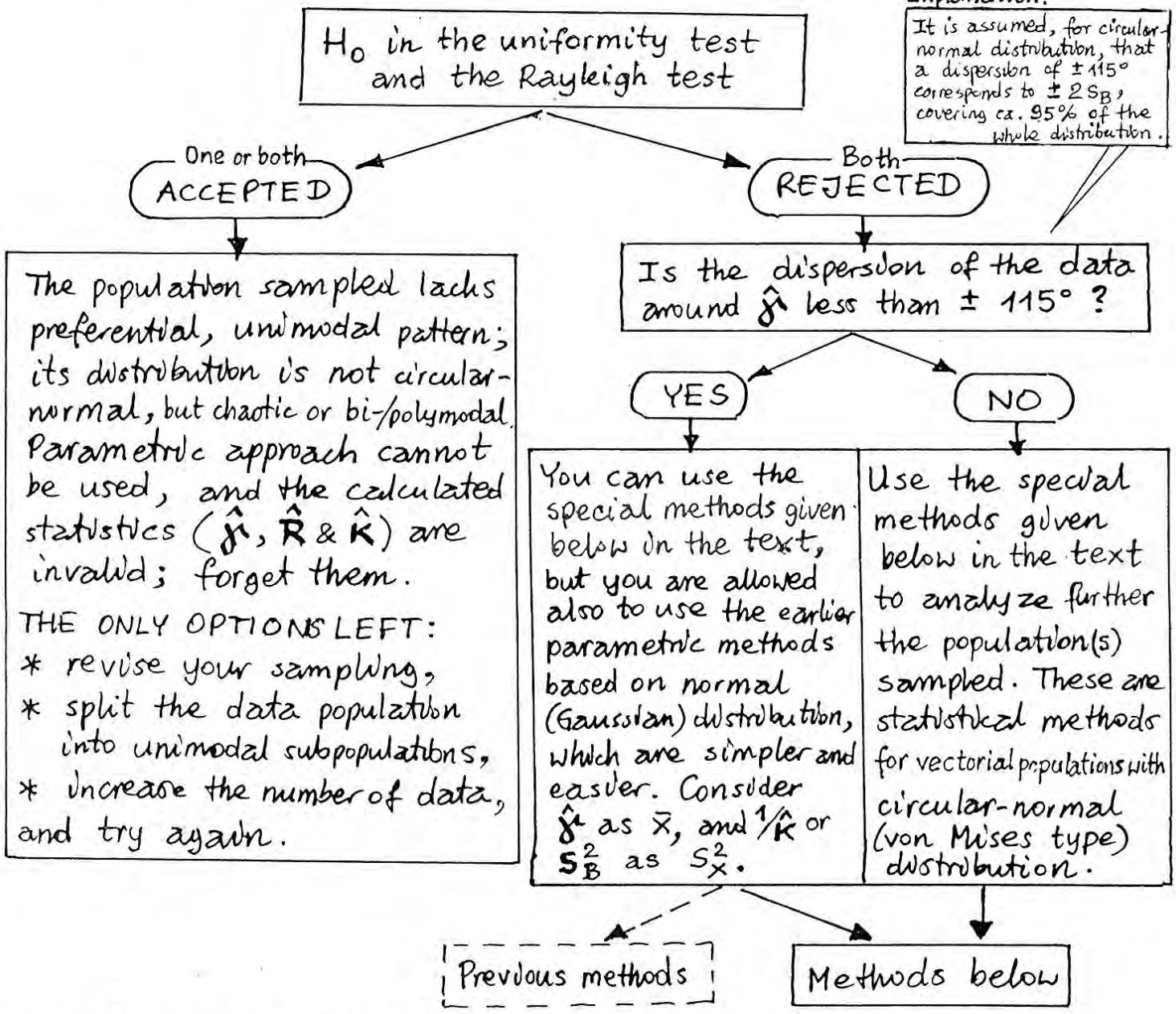
General evaluation of sample population

Result of the uniformity test	Result of the Rayleigh test	Conclusion about the popul. distribution
H_0 rejected	H_0 rejected	circular-normal
H_0 rejected	H_0 accepted	bimodal or polymodal *
H_0 accepted	H_0 accepted	non-preferential (chaotic)

* The sample population should be carefully split into two or more component subpopulations and each of them re-analyzed separately; a rose diagram or other histogram will help to make the splitting.

■ Analytical strategy

Depending upon the result of the circular-normality test, further analysis of the data will be as shown below.



■ Confidence interval (sector) for the mean vector direction

The calculation requires $(n \cdot \hat{R} \cdot \hat{\kappa}) \geq 6$ and is as follows:

$$(\hat{\gamma} - d^\circ) < \gamma < (\hat{\gamma} + d^\circ)$$

where: $\hat{\gamma}$ = the calculated statistic (mean azimuth);

$$d^\circ = \frac{57.2956^\circ \cdot Z_{1/2\alpha}}{\sqrt{n \cdot \hat{R} \cdot \hat{\kappa}}}$$

← where: 57.2956 = conversion factor to change radians into degrees;
$Z_{1/2\alpha}$ from TABLE 1B.

The confidence coefficient of the interval is $(1-\alpha)100\%$.

Comparison of two vectorial populations

The following tests can be used for a comparison of circular-normal vectorial populations in terms of their mean vectors, vector concentrations or – in a non-parametric approach – the whole frequency distributions.

• Test t for the difference between two mean vectors

$$\begin{cases} H_0: \gamma_1 \geq \gamma_2 \\ H_1: \gamma_1 < \gamma_2 \end{cases} \quad \text{or} \quad \begin{matrix} \gamma_1 \leq \gamma_2 \\ \gamma_1 > \gamma_2 \end{matrix} \qquad \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{one-tail formulation, depending on } \hat{\gamma}_1 \text{ and } \hat{\gamma}_2$$

The test function is:

$$t = \left\{ \frac{[1 + 3/(8\hat{\kappa}_*)](n-2)(n_1\hat{R}_1 + n_2\hat{R}_2 - n\hat{R}_*)}{n - n_1\hat{R}_1 - n_2\hat{R}_2} \right\}^{1/2} \qquad DF = n-2$$

where: $\hat{R}_1, \hat{R}_2$ = the mean vector lengths of the two sample populations;
n_1, n_2 = the sample population sizes;
$n = n_1 + n_2$
$\hat{R}_*, \hat{\kappa}_*$ = the mean vector length and vector concentration of the two sample populations combined ("pooled").

The critical value, t_α, is taken from TABLE 2. If $t \geq t_\alpha$, H_0 is rejected and H_1 is accepted with confidence of $(1-\alpha)100\%$; otherwise, H_0 cannot be rejected.

Comment: $\hat{R}_*$ and $\hat{\kappa}_*$ can be calculated by taking both populations as one. Alternatively $\hat{R}_*$ can be calculated by the formula given in the third test below, and $\hat{\kappa}_*$ can be derived from this value (see formula on p. 49, bottom).

• Test F for the difference between two mean vectors

An alternative test, with the same hypothesis, based on Snedecor's F-distribution. The test function has two forms:

if $\hat{\kappa}_* \geq 10$	if $\hat{\kappa}_* < 10$
$F = \dfrac{(n-2)(\hat{R}_1 + \hat{R}_2 - \hat{R}_*)}{n - \hat{R}_1 - \hat{R}_2}$	$F = \left(1 + \dfrac{3}{8\hat{\kappa}_*}\right) \dfrac{(n-2)(\hat{R}_1 + \hat{R}_2 - \hat{R}_*)}{n - \hat{R}_1 - \hat{R}_2}$

with $\begin{cases} DF_1 = 1 \\ DF_2 = n-2 \end{cases}$ and symbols as above.

The critical value is taken from TABLE 3; but if $\hat{\kappa}_* < 2$, a special table is necessary (see Mardia, 1972). If $F \geq F_\alpha$, H_0 is rejected.

• Test for the difference between two vectorial concentrations

$$\begin{matrix} H_0: \kappa_1 \geq \kappa_2 \\ H_1: \kappa_1 < \kappa_2 \end{matrix} \quad \text{or} \quad \begin{matrix} \kappa_1 \leq \kappa_2 \\ \kappa_1 > \kappa_2 \end{matrix} \qquad \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{one-tail formulation, depending on } \hat{\kappa}_1 \text{ and } \hat{\kappa}_2$$

The following quantities are calculated for combined ("pooled") sample populations:

$$\left.\begin{aligned} \bar{C}_* &= \frac{1}{n}(n_1\bar{C}_1 + n_2\bar{C}_2) \\ \bar{S}_* &= \frac{1}{n}(n_1\bar{S}_1 + n_2\bar{S}_2) \end{aligned}\right\} \text{where } \bar{C}_1, \bar{C}_2, \bar{S}_1 \text{ \& } \bar{S}_2 \text{ are calculated separately for each population (p. 49).}$$

$\hat{R}_*$ is calculated: $\hat{R}_* = (\bar{C}_*)^2 + (\bar{S}_*)^2$

and test function is selected

→ When $\hat{R}_* \leq 0.45$, the test function is:

$$Z = 1.155\,(w_1 - w_2)/\sqrt{(n_1-4)^{-1} + (n_2-4)^{-1}}$$

where: $w_1 = \arcsin(1.225\hat{R}_1)$, $w_2 = \arcsin(1.225\hat{R}_2)$ } with $\hat{R}_1$ and $\hat{R}_2$ calculated for the two sample populations individually (see p. 49).

The critical value is taken from TABLE 1B.
If $Z \geq Z_\alpha$, H_0 is rejected with confidence $(1-\alpha)100\%$.

→ When $0.45 < \hat{R}_* < 0.70$, the test function is:

$$Z = (v_1 - v_2)/\left[0.893\sqrt{(n_1-3)^{-1} + (n_2-3)^{-1}}\right]$$

where: $v_1 = \ln[f_1 + \sqrt{(1+f_1^2)}]$ with $f_1 = (\hat{R}_1 - 1.089)/0.258$

$v_2 = \ln[f_2 + \sqrt{(1+f_2^2)}]$ with $f_2 = (\hat{R}_2 - 1.089)/0.258$

The critical value is taken from TABLE 1B, as above.

→ When $\hat{R}_* \geq 0.70$, the test function is:

$$F = \frac{n_1(1-\hat{R}_1)(n_2-1)}{n_2(1-\hat{R}_2)(n_1-1)}$$

← Larger value in numerator
← smaller value in denominator

with DF: $DF_1 = n_1 - 1$ ← for $\hat{R}$ in numerator
$DF_2 = n_2 - 1$ ← for $\hat{R}$ in denominator

The critical value is taken from TABLE 3. If $F \geq F_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence.

• Non-parametric test for the difference between two vectorial populations

The runs test (p. 32) can be used for a bulk comparison.
The two populations are mixed and the data (azimuth measurements) are ordered from the smallest to the largest, being marked also as to which population does particular value belong.
The following statistics are then calculated:

u_o = the observed number of runs;
$\bar{u}_e$ = the expected mean number of runs for a random mixture of data from two populations;
$s_{\bar{u}}$ = the expected standard deviation of the mean.

(For formulae, see p. 32.)
The test function is:

$$Z = \frac{|u_o - \bar{u}_e|}{s_{\bar{u}}}$$

← Absolute difference

The critical value is taken from TABLE 1B.
If $Z \geq Z_\alpha$, H_0 (postulating lack of difference) is rejected with confidence $(1-\alpha)100\%$. Otherwise, H_0 cannot be rejected.

Simultaneous comparison of several vectorial populations

One-way ANOVA can be used as a parametric test for the homogeneity of mean vectors, with the following hypothesis:

$$\begin{cases} H_0: \gamma_1 = \gamma_2 = \gamma_3 = \dots = \gamma_m \\ H_1: \text{At least one population has a mean vector significantly different from the others } (\gamma_i \neq \gamma_k) \end{cases}$$

One-way ANOVA table:

Source of variance	Sum of squares	DF	F-test
Variation between populations	$\sum^m \hat{R}_i - \hat{R}_*$	$m-1$	$F = \frac{S_B^2}{S_W^2} = \frac{(N-m)(\sum \hat{R}_i - \hat{R}_*)}{(m-1)(N - \sum \hat{R}_i)}$ with $\begin{cases} DF_1 = m-1 \\ DF_2 = N-m \end{cases}$
Variation within populations	$N - \sum^m \hat{R}_i$	$N-m$	
Total variation	$N - \hat{R}_*$	$N-1$	Critical value $F_{1/2\alpha}$ from TABLE 3

where: $\hat{R}_i$ = the magnitude of the mean vector of particular population;
$\sum^m R_i$ = sum of the magnitudes of all mean vectors;
$\hat{R}_*$ = the magnitude of the mean vector of all populations combined ("pooled"); see formula on p. 53 (top);
N = total number of data;
m = number of populations compared.

If $F \geq F_{1/2\alpha}$, H_0 is rejected with $(1-\alpha)100\%$ confidence; if $F < F_{1/2\alpha}$, H_0 cannot be rejected.

Optimization of sampling of a vectorial population

$\Delta\gamma$ = the required accuracy of the estimation of mean vector direction, γ; this may be considered as a confidence semi-interval (analogous to $d°$ on p. 51), and taken to be $\pm 5°$, $\pm 10°$ or $\pm 20°$.
m = number of outcrops sampled;
n = number of measurements collected per outcrop (mean $\bar{n}$ can be used).

The data ($N = m \cdot n$) have been subjected to ANOVA test (as above) and H_0 has been accepted: what we have been sampling is a single population. How to optimize the sampling, while meeting the accuracy requirement?

Find the values of W and V from the following set of equations (note that their right-hand parts are mean squares from the ANOVA table):

$$\begin{cases} \dfrac{1}{2W} = \dfrac{N - \sum_{i}^{m} \hat{R}_i}{N - m} \\ \dfrac{1}{2}\left(\dfrac{1}{W} + \dfrac{\bar{n}}{V}\right) = \dfrac{\sum_{i}^{m} \hat{R}_i - \hat{R}_*}{m - 1} \end{cases}$$

calculate W from the first equation, then insert the W-value into the second equation to obtain V

Assume an acceptable risk-of-error level (α) and find the corresponding value of K_o from TABLE 8, for $\Delta\bar{\gamma}$ and α. The K_o-value can also be calculated from the Rao formula:

$$K_o = Z_\alpha / \Delta\bar{\gamma}$$

where Z_α is taken from TABLE 1B for the selected α.

Based on the values of W, V and K_o, we can estimate the values of n or m that will satisfy the requirement for the accuracy of mean vector estimation: $\bar{\gamma} \pm \Delta\bar{\gamma}$.

If m_{max} is the number of available outcrops, then the necessary number of measurements per outcrop should be:

$$n_{opt} = \frac{1}{W\left(m_{max} \cdot K_o - \frac{1}{V}\right)}$$

If there are numerous outcrops and we wish to take only a certain minimal number of measurements (n_{min}) per outcrop, then the number of outcrops to be visited is:

$$m_{opt} = \frac{1}{K_o}\left(\frac{1}{W \cdot n_{min}} + \frac{1}{V}\right)$$

The same approach can be used to optimize the sampling procedure for three-dimensional vectorial data (next section).
[From Watson (1966) in Journal of Geology, v. 74, p. 786-797.]

ANALYSIS OF THREE-DIMENSIONAL VECTORIAL DATA

Three-dimensional (spherical) vectorial measurements are data pairs:

$$\begin{cases} A^\circ = \text{dip direction (azimuth)}^* \\ D^\circ = \text{dip angle} \end{cases}$$

* If strike has been measured, this can be recalculated as follows:

$$A^\circ = \text{Strike} + 90^\circ\ (\pm 180^\circ)$$

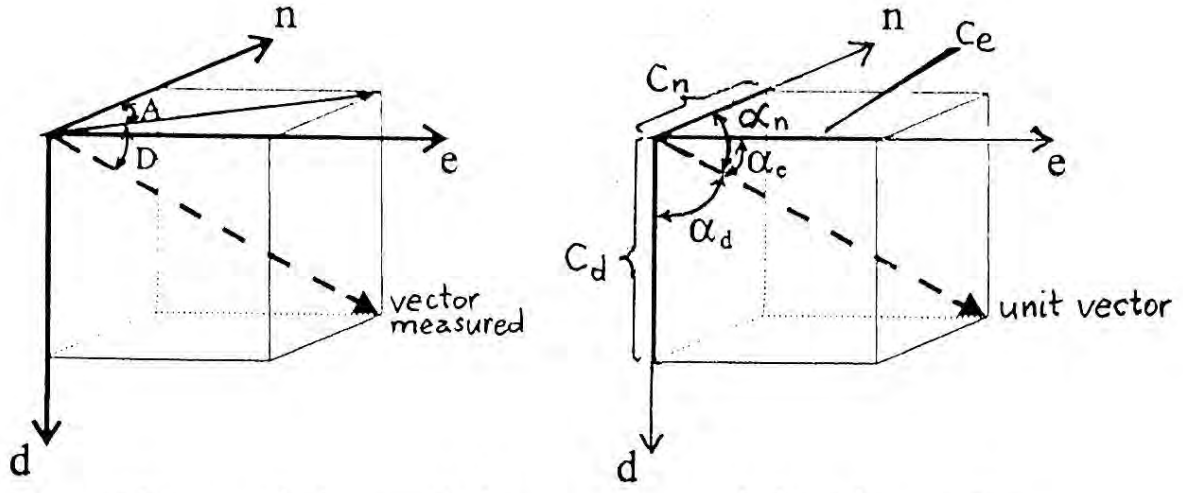
(← Correction for the right-hand rule)

When using Excel or some other standard software, it may be convenient to recalculate the data in degrees into radians:

$$\begin{cases} A_{rad} = A^\circ \cdot \frac{\pi}{180^\circ} \\ D_{rad} = D^\circ \cdot \frac{\pi}{180^\circ} \end{cases}$$

For the purpose of numerical operations, each data pair (A°, D°) is represented by a unit vector, c, which is further decomposed into three component orthogonal vectors corresponding to the "north", "east" and "down" axes (see diagram below):

$$\left.\begin{cases} c_n = \cos A^\circ \cdot \cos D^\circ \\ c_e = \sin A^\circ \cdot \cos D^\circ \\ c_d = \sin A^\circ \end{cases}\right\} \leftarrow (A_{rad} \text{ and } D_{rad} \text{ can be used, if more convenient})$$



The difference between field measurement (A, D) and its representation for numerical purpose (c_n, c_e, c_d).

In order to check the correctness of recalculation, the following formula can be used:

$$\left[(c_n)^2 + (c_e)^2 + (c_d)^2\right]^{1/2} = 1$$

For the whole data set (sample population), the mean values are calculated:

$$\bar{c}_n,\ \bar{c}_e \text{ and } \bar{c}_d.$$

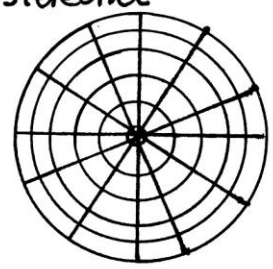
Chi-square test of spherical uniformity

The Fisher distribution is a spherical equivalent of the von Mises circular-normal distribution and the Gaussian linear (one-dimensional) normal distribution. A parametric approach assumes that the three-dimensional vectorial population sampled has a spherical-normal distribution. To meet this requirement, the population's distribution should at least be non-uniform and unimodal, and its vectorial concentration should be $\kappa > 0$ (a uniform distribution has $\kappa = 0$).

It has been demonstrated that a non-uniform unimodal spherical population, in terms of its stereo-net plot, is approximated by the Poisson distribution (see p. 9-10). This property can be used to test the distribution uniformity, which means the following hypothesis:

H_0: The population's distribution conforms to Poissonian distribution (i.e., is non-uniform)
H_1: The population's distribution does not conform to Poissonian distribution (i.e., is uniform or polymodal)

Stereonet



The stereoplot is divided into equal-area blocks, or couting circles, and the number of data (projection points) in each block is counted. The results are summarized in the form of a frequency table.

EXAMPLE

Class, x_i	0	1	2	3	4	5	≥6
Frequency, f_i	28	16	8	4	2	2	0

N = 60

N = total number of blocks

↑ These two values mean that there are 8 blocks containing 2 data points each.

The density function of Poissonian distribution is:

$$P_i(x_i, \lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

where: P_i = the density (frequency) of class value x_i;
λ = the distribution's mean;
x_i = particular class value;
e = algebraic constant (2.71828);
$x_i!$ = x_i factorial (e.g., $4! = 1 \cdot 2 \cdot 3 \cdot 4$).

In this distribution, $\frac{\sigma_x}{\lambda} = 1$ and the two parameters are estimated as:

$$\hat{\lambda} = \bar{x} = \frac{\sum f_i x_i}{N}, \quad \text{where } N = \sum^N f_i$$

$$\hat{\sigma}_x = s_x = \sqrt{\frac{\sum^N f_i x_i^2 - N\bar{x}^2}{N-1}}$$

The chi-square test function is:

$$\chi^2 = \frac{(Ofi - Efi)^2}{Efi}$$

with DF = k − 2 where: k = number of classes

where: Ofi = observed frequency (fi);
Efi = expected frequency for ideal Poissonian distribution.

Calculatory table ($Efi = P_i \cdot N$)

x_i	Ofi	P_i	Efi	$Ofi - Efi$	$(Ofi - Efi)^2$	$\frac{(Ofi-Efi)^2}{Efi}$
0	28	0.356	21.36	6.64	44.09	2.06
1	16	0.368	22.08	−6.08	36.96	1.68
2	8	0.190	11.40	−3.40	11.56	1.02
3	4	0.065	3.90	0.10	0.01	0.002
4	2	0.017	1.02	0.98	0.96	0.94
5	2	0.004	0.24	1.76	3.09	12.84
≥6	0	0.000	0.00	0.00	0.00	0.00

Sum: 18.54

$N = 60$
$\hat{\lambda} = 1.033$
$\hat{\sigma}_x = 1.325$

$\frac{\hat{\sigma}_x}{\hat{\lambda}} = 1.282$, which is greater than 1. But is it greater enough to render the distribution non-Poissonian?

calculation

$P(x=0) = e^{-\hat{\lambda}} = 2.71828^{-1.033} = 0.356$
$P(x=1) = \hat{\lambda} e^{-\hat{\lambda}} = 1.033 \cdot 0.356 = 0.368$
$P(x=2) = \frac{1}{2}(\hat{\lambda})^2 e^{-\hat{\lambda}} = 0.533 \cdot 0.356 = 0.190$
$P(x=3) = \frac{1}{6}(\hat{\lambda})^3 e^{-\hat{\lambda}} = 0.184 \cdot 0.356 = 0.065$
$P(x=4) = \frac{1}{24}(\hat{\lambda})^4 e^{-\hat{\lambda}} = 0.047 \cdot 0.356 = 0.017$
$P(x=5) = \frac{1}{120}(\hat{\lambda})^5 e^{-\hat{\lambda}} = 0.010 \cdot 0.356 = 0.004$
$P(x \geq 6) = 0.000$

Calculatory result:
$\chi^2 = 18.54$
$DF = 7 - 2 = 5$

← Calculated by formula on previous page (bottom)

And what's your conclusion about this example data set? (Can it be regarded as Poissonian?)

Parameters and statistics

The mean component vectors are calculated for the sample population:

$$\begin{cases} \bar{C}_n = \left(\sum^n C_n\right)/n \\ \bar{C}_e = \left(\sum^n C_e\right)/n \\ \bar{C}_d = \left(\sum^n C_d\right)/n \end{cases}$$

where: n = number of data

- Mean vector length (R) is estimated as:

$$\hat{R} = \sqrt{(\bar{C}_n)^2 + (\bar{C}_e)^2 + (\bar{C}_d)^2}$$

- Mean vector direction ($A°, D°$) is estimated as follows:

The directional cosines of the mean vector are calculated:

$$\begin{cases} \hat{\gamma}_n = \bar{C}_n/\hat{R} \\ \hat{\gamma}_e = \bar{C}_e/\hat{R} \\ \hat{\gamma}_d = \bar{C}_d/\hat{R} \end{cases}$$

from which the mean vector direction is derived:

$$\hat{\gamma}_{3D} = \begin{cases} \hat{A}° = \arctan(\hat{\gamma}_e/\hat{\gamma}_n) + Q \\ \hat{D}° = \arctan\left(\hat{\gamma}_d/\sqrt{\hat{\gamma}_n^2 + \hat{\gamma}_e^2}\right) \end{cases}$$

where Q is the quadrant positioning quotient:

$\hat{\gamma}_n$	$\hat{\gamma}_e$	Q
≥ 0	≥ 0	0
< 0	≥ 0	180°
< 0	< 0	180°
≥ 0	< 0	360°

If the whole calculation was for data in radians, the values obtained should now be converted into degrees (with formulae on p. 56). To obtain the angle of dip: $90° - \hat{D}°$; to obtain the strike direction: $\hat{A}° - 90°$.

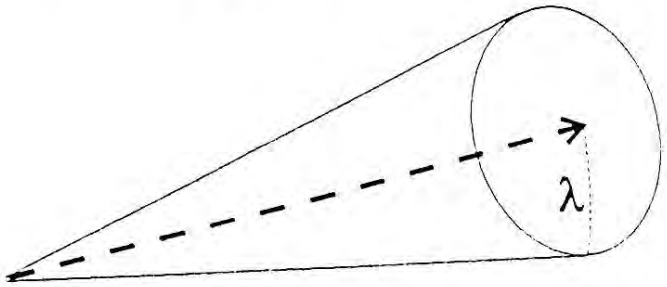
- Vectorial concentration (K) of the population is estimated as:

$$\hat{K} = \frac{1}{1 - \hat{R}}$$

This formula is valid only if $\hat{R} \geq 0.65$. Otherwise, there are problems with the calculation of $\hat{K}$.

■ Confidence interval (cone) for the mean vector

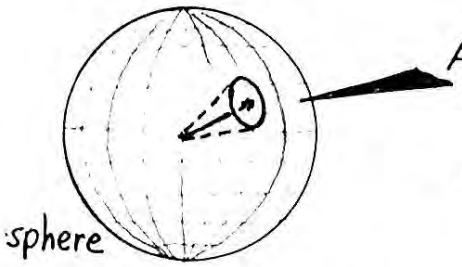
In the case of three-dimensional vectorial data, the confidence interval for the general mean (γ_{3D}) is a cone with semi-apical angle λ (see diagram below).



The angle (λ) is calculated, for an acceptable risk-of-error (α), by the following formula:

$$\lambda = \arccos\left[1 + \ln\alpha/(n \cdot \hat{R} \cdot \hat{\kappa})\right]$$

and its confidence level is $(1-\alpha)100\%$



A small circle of radius λ is plotted on the stereonet around the mean vector projection point.

■ Comparison of two vectorial populations

• Test for the difference between two mean vectors

The one-tail hypothesis tested

$$\begin{cases} H_0: \gamma_1 \leq \gamma_2 \\ H_1: \gamma_1 > \gamma_2 \end{cases} \quad \text{or} \quad \begin{cases} H_0: \gamma_1 \geq \gamma_2 \\ H_1: \gamma_1 < \gamma_2 \end{cases}$$

← Mean vectors are the same

← Mean vectors are different

The test function is:

$$F = \frac{(n-1)(n_1\hat{R}_1 + n_2\hat{R}_2 - n\hat{R}_*}{n - n\hat{R}_*}$$

with $\begin{cases} DF_1 = 2 \\ DF_2 = 2n-2 \end{cases}$

where: $n = n_1 + n_2$

$\hat{R}_1, \hat{R}_2$ = the mean vector lengths of the two sample populations;

$\hat{R}_*$ = the mean vector length of the two populations combined ("pooled"), calculated as follows:

$$\hat{R}_* = \sqrt{\left(\frac{\sum\sum c_n}{n}\right)^2 + \left(\frac{\sum\sum c_e}{n}\right)^2 + \left(\frac{\sum\sum c_d}{n}\right)^2}$$

and the sums of directional cosines ($\sum\sum c$) are calculated as totals:

	c_n	c_e	c_d
Popul. 1 (n_1)	$\sum c_{n(1)}$	$\sum c_{e(1)}$	$\sum c_{d(1)}$
Popul. 2 (n_2)	$\sum c_{n(2)}$	$\sum c_{e(2)}$	$\sum c_{d(2)}$
Totals (n_1+n_2)	$\sum\sum c_n$	$\sum\sum c_e$	$\sum\sum c_d$

The critical value, F_α, for the test is taken from TABLE 3. If the table is insufficient, the critical value can be calculated as: $F = \frac{DF_2}{2}\left(\alpha^{-2/DF_2} - 1\right)$, where $DF_2 = 2n-2$ and α = selected significance level.

If $F \geq F_\alpha$, H_0 is rejected with confidence $(1-\alpha)100\%$; otherwise, H_0 cannot be rejected.

- Test for the difference between two vectorial concentrations

The one-tail hypothesis tested is:

$$\begin{cases} H_0: K_1 \leq K_2 \\ H_1: K_1 > K_2 \end{cases} \quad \text{or} \quad \begin{cases} H_0: K_1 \geq K_2 \\ H_1: K_1 < K_2 \end{cases}$$

H_0: $K_1 \geq K_2$ ← Populations have similar dispersion

H_1: $K_1 < K_2$ ← Populations have different dispersions

The test requires that both $\hat{R}_1$ and $\hat{R}_2$ are ≥ 0.65, and the test function is:

$$F = \frac{\dfrac{n_1(1-\hat{R}_1)}{n_1-1}}{\dfrac{n_2(1-\hat{R}_2)}{n_2-1}}$$

larger value in numerator

smaller value in denominator

which can be rearranged as (see p. 53):

$$\boxed{F = \frac{n_1(1-\hat{R}_1)(n_2-1)}{n_2(1-\hat{R}_2)(n_1-1)}}$$

with $\begin{cases} DF_1 = 2(n_1-1) \leftarrow \text{for population in the numerator} \\ DF_2 = 2(n_2-1) \leftarrow \text{for population in the denominator} \end{cases}$

The critical value F_α is taken from TABLE 3. If $F \geq F_\alpha$, the H_0 is rejected with $(1-\alpha)100\%$ confidence; otherwise, H_0 cannot be rejected.

Analysis of bivariate populations

The data population to be analyzed is $\{x, y\}$, with two variables (x and y) whose mutual relationship is the focus of our interest. The data are n pairs of x- and y-values.

• Relationship between two quantitative variables: linear correlation and regression (parametric method)

The basic parameters to consider are:

σ_{xy} = covariance of the two variables, x and y; values $(+\infty, -\infty)$
- $\sigma_{xy} = 0$ means "no relationship" between x and y;
- $\sigma_{xy} > 0$ means "positive relationship" between x and y (i.e., when one variable increases, the other increases too);
- $\sigma_{xy} < 0$ means "negative relationship" between x and y (i.e., when one variable increases, the other decreases).

ρ_{xy} = coefficient of linear correlation between x and y; its values are between $+1$ and -1, with $\rho_{xy} = 0$ meaning no correlation, $\rho_{xy} > 0$ meaning positive correlation and $\rho_{xy} < 0$ negative correlation.

β_{xy} = coefficient of linear regression (dependence of y on x); its value is a tangent, or the ratio of the change in y to a unit change in x; again, $\beta_{xy} = 0$ means lack of regression, $\beta_{xy} > 0$ means "positive" regression and $\beta_{xy} < 0$ means "negative" regression

Their estimators, or statistics, are

COVARIANCE (S_{xy})

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Definition formula, which gives a "biased" estimator; $n-1$ is used in denominator to obtain an "unbiased" one.

$$S_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n(n-1)}$$

Practical formulae for "unbiased" estimator

$$S_{xy} = \frac{\sum x_i y_i}{n} - \bar{x}\cdot\bar{y} = \overline{xy} - \bar{x}\cdot\bar{y}$$

PEARSON'S CORRELATION COEFFICIENT (r_{xy})

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$$

with values in the range: $-1 \le r_{xy} \le 1$

■ The Fisher test for the significance of linear correlation

One-tail hypothesis: $\begin{cases} H_0: \rho_{xy} \le 0 \\ H_1: \rho_{xy} > 0 \end{cases}$ or $\begin{cases} \rho_{xy} \ge 0 \\ \rho_{xy} < 0 \end{cases}$

The test function:

$$t = |r_{xy}| \sqrt{\frac{n-2}{1-r_{xy}^2}}$$

with $DF = n-2$

The critical value, t_α, is taken from TABLE 2. If $t \ge t_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence; otherwise, H_0 cannot be rejected.

The problem of data closure

"Closed" data, such as percentages in the case of a rock composition (mineral or chemical), induce the so-called spurious correlation between the variables – which means false values of r_{xy} caused by the effect of closure:

$$\sum_{i=1}^{p} x_i = 100 \quad \text{for each sample } (p = \text{number of variables, or components, studied})$$

Data table type:

	Variables x_1	x_2	x_3	...	x_p	Row sums
Samples 1	x_{11}	x_{12}	x_{13}	...	x_{1p}	= 100%
2	x_{21}	x_{22}	x_{23}	...	x_{2p}	= 100%
3	x_{31}	x_{32}	x_{33}	...	x_{3p}	= 100%
⋮	⋮	⋮	⋮		⋮	⋮
n	x_{n1}	x_{n2}	x_{n3}	...	x_{np}	= 100%

The row sum is 100% when the x_{ij} values are in percent, but may as well be 1 when the values are given as fractions of 1.

No matter if only two of the variables (e.g., x_1 and x_3) are correlated with each other, or if one of the variables is correlated with all the others, as in multiple correlation (p. 74-75), the effect of spurious correlation persists. The variables show apperent correlation simply because, in a closed system of values, if one value increases, the others must decrease.

The following log-ratio solution can be used to "break" the closure of data (table above):

(1) calculate a geometric mean for each row, $\overline{x}_{(G)i\bullet}$;
(2) divide the whole row by its geometric mean, $x_{i\bullet}/\overline{x}_{(G)i\bullet}$;
(3) replace these ratio values with their logarithms:

$$\log\left(x_{i\bullet}/\overline{x}_{(G)i\bullet}\right);$$

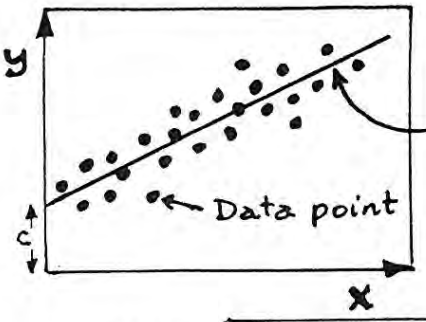
(4) perform the correlation and regression analysis on these transformed (log-ratio) data.

The calculation of geometric mean is explained earlier in the text (see p. 19).

The log-ratio transformation has been proposed and discussed in detail, along with the problem of data closure, by:

* Aitchison, J. (1986). The Statistal Analysis of Compositional Data. Chapman & Hall, London, 416 p.
* Reyment, R.A. (1988) in journal Terra Nova, v. 1, no. 1, p. 29-34.

Straight-line regression



The aim is to find a "best-fit" line (linear function) to describe the relationship between **y** and **x** ("least-squares" criterion is used)

Model: $\boxed{\hat{y} = b \cdot x + c}$

where b = the line gradient; estimator of β (see previous page);
and c = the line's intersection point with the y-axis (see diagram).

$$\boxed{b = \frac{S_{xy}}{S_x^2} = r_{xy} \cdot \frac{S_y}{S_x}}$$

b = regression coefficient (regression line's gradient)

$$\boxed{c = \bar{y} - b \cdot \bar{x}}$$

More universal way of finding **b** and **c** is by solving the following system of linear equations:

$$\begin{cases} \sum_i^n y_i = \mathbf{b} \sum_i^n x_i + \mathbf{c} \cdot n \\ \sum_i^n x_i y_i = \mathbf{b} \sum_i^n x_i^2 + \mathbf{c} \cdot \sum_i^n x_i \end{cases}$$

or the equivalent matrix equation (linear):

$$\begin{bmatrix} \sum_i^n y_i \\ \sum_i^n x_i y_i \end{bmatrix} = \begin{bmatrix} \sum_i^n x_i & n \\ \sum_i^n x_i^2 & \sum_i^n x_i \end{bmatrix} \cdot \begin{bmatrix} b \\ c \end{bmatrix}$$

The Fisher test for the significance of regression

$H_0: \beta \leq 0$ or $\beta \geq 0$
$H_1: \beta > 0$ $\beta < 0$
} one-tail formulation, depending on the sign of **b**-value

The test function: (absolute value)

$$\boxed{t = \frac{|b| \sqrt{\sum_i^n x_i^2 - n(\bar{x})^2}}{\sqrt{S_y^2 (1 - r_{xy}^2)}}} \quad \text{with } DF = n-2$$

Critical value is taken from TABLE 2.
When $t \geq t_\alpha$ (one-tail test), H_0 is rejected with confidence $(1-\alpha)100\%$.

Note: The denominator in the formula above is called the "standard error of the estimation of y on x":

$$\boxed{S_{y \cdot x} = S_y^2 (1 - r_{xy}^2)}$$

and it is simply the standard deviation of the estimation ($\hat{y}$) for the regression line. This statistic will be used later in the text.

The Fisher test for the difference between two regression coefficients (line parallelism test)

$H_0: \beta_1 \leq \beta_2$ or $\beta_1 \geq \beta_2$
$H_1: \beta_1 > \beta_2$ $\beta_1 < \beta_2$
} one-tail formulation, depending on the values of b_1 and b_2

Test function:

$$t = \frac{|b_1 - b_2|}{\sqrt{\frac{\sum z_1^2 + \sum z_2^2}{n_1 + n_2 - 4}\left(\frac{1}{(n_1-1)S_1^2} + \frac{1}{(n_2-1)S_2^2}\right)}}$$

with $DF = n_1 + n_2 - 4$

where: z = residual component (see table) and other symbols as before.

$\frac{1}{\sum(x_1-\bar{x}_1)^2} + \frac{1}{\sum(x_2-\bar{x}_2)^2}$ — Alternative expression for the term in parentheses above

Calculatory table:

i	x_i	y_i	x_i^2	$x_i \cdot y_i$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1							
2							
3							
⋮							
n							

Residuals z — $z_i = y_i - \hat{y}_i$

$\sum z^2$ is the sum of this column

Data — The same table is calculated separately for each of the two populations.

Calculated from the regression equation $\hat{y} = bx + c$ for all individual x-values.

The critical value is taken from TABLE 2.
When $t \geq t_\alpha$ (one-tail test), H_0 is rejected with confidence $(1-\alpha)100\%$.

Note: From the table above, the variance and standard deviation of the residual component are also derived:

$$S_z^2 = \frac{\sum z^2}{n-1} \quad \text{and} \quad S_z = \sqrt{S_z^2}$$

and the latter will be used further below.

Regression line's goodness-of-fit

Coefficient:

$$R^2 = r_{xy}^2 \,(\cdot 100\,\%)$$

$0 \leq R^2 \leq 100\%$

is commonly used and often called the coefficient of determination
Alternatively, one of the following coefficients can be used:

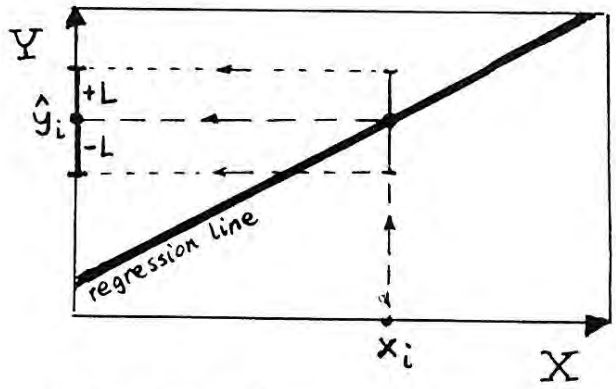
$$R_*^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}(100\%) = \frac{\sum(\hat{y} - \bar{y})^2}{(n-1)S_y^2}(100\%)$$

$$d^2 = \left(1 - \frac{\sum z^2}{(n-1)S_y^2}\right)(100\%) = \left(1 - \frac{S_z^2}{S_y^2}\right)(100\%)$$

All these coefficients indicate as to how deterministic (%) the regression equation is.

Confidence interval and confidence belt for the regression line

(1) Confidence interval (L) for single prediction



When predicting the value of y for a given x_i:

$$x_i \longrightarrow y_i = \hat{y}_i (\pm L)$$

where:

L = confidence semi-interval.

Calculate L:

$$\boxed{L = t_{1/2\alpha} \cdot S_z \cdot \sqrt{R_i}}$$

with confidence coefficient $(1-\alpha)100\%$

From TABLE 2 for $1/2\alpha$ and DF = n-2

See previous page

$$R_i = 1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)\,S_x^2}$$

and the prediction will be:

$$\boxed{(\hat{y}_i - L) < y_i < (\hat{y}_i + L)}$$

with confidence coefficient as above

(2) Confidence belt for regression line

Select a series of x-values, or at least the following five values:

$(\bar{x} - 2S_x)$, $(\bar{x} - S_x)$, $\bar{x}$, $(\bar{x} + S_x)$ and $(\bar{x} + 2S_x)$,

and for each of them (x_k) calculate L using one of the following formulae:

$$\boxed{L_k = t_{1/2\alpha} \cdot S_z \cdot \sqrt{R_k}}$$

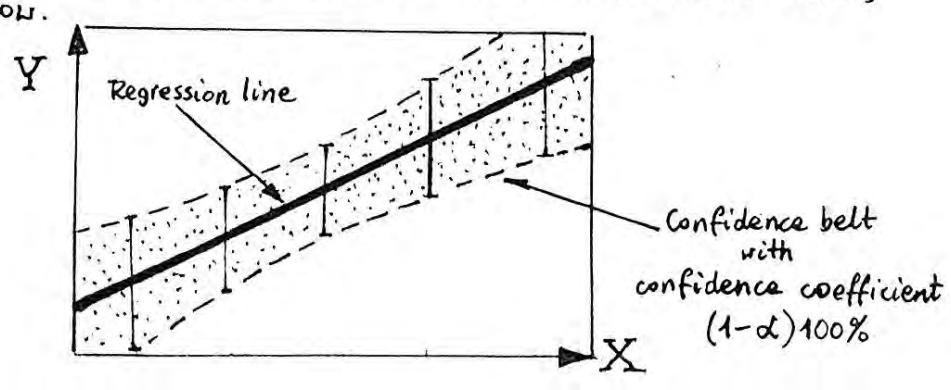
$$R_k = \frac{1}{n} + \frac{(x_k - \bar{x})^2}{(n-1)\,S_x^2}$$

or

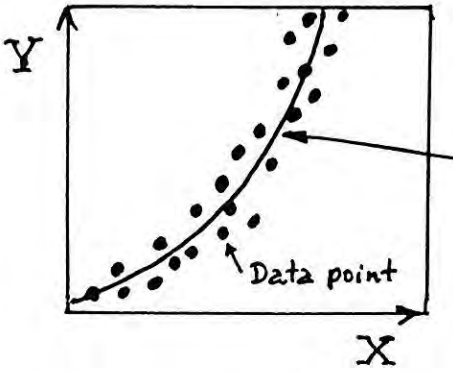
$$\boxed{L_k = \sqrt{2F_{1/2\alpha}} \cdot S_z \cdot \sqrt{R_k}}$$

From TABLE 3 for $\begin{cases} DF_1 = 2 \\ DF_2 = n-2 \end{cases}$ and $1/2\alpha$

Plot the calculated intervals and link them to make a "belt", as shown below.



Curvilinear regression



The aim is to find a "best-fit" curve to describe the relationship between **y** and **x** (the so-called least-squares criterion is applied).

Polynomial models:

$\hat{y} = bx + c$ ← straight-line model

(2nd degree)	$\hat{y} = b_1 x + b_2 x^2 + c$
(3rd degree)	$\hat{y} = b_1 x + b_2 x^2 + b_3 x^3 + c$
⋮	⋮ ⋮ ⋮ ⋮
(k th degree)	$\hat{y} = b_1 x + b_2 x^2 + b_3 x^3 + \cdots + b_k x^k + c$

For example: In order find the coefficients of equation $\hat{y} = b_1 x + b_2 x^2 + c$, the following system of linear equations has to be solved:

$$\begin{cases} \sum y_i = \boxed{b_1} \sum x_i + \boxed{b_2} \sum x_i^2 + \boxed{c} \cdot n \\ \sum x_i y_i = \boxed{b_1} \sum x_i^2 + \boxed{b_2} \sum x_i^3 + \boxed{c} \sum x_i \\ \sum x_i^2 y_i = \boxed{b_1} \sum x_i^3 + \boxed{b_2} \sum x_i^4 + \boxed{c} \sum x_i^2 \end{cases}$$

(3 equations with 3 unknowns)

or in matrix form:

$$\begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix} = \begin{bmatrix} \sum x_i & \sum x_i^2 & n \\ \sum x_i^2 & \sum x_i^3 & \sum x_i \\ \sum x_i^3 & \sum x_i^4 & \sum x_i^2 \end{bmatrix} \cdot \begin{bmatrix} \boxed{b_1} \\ \boxed{b_2} \\ \boxed{c} \end{bmatrix}$$

The goodness-of-fit is measured by the determination coefficient:

$$\boxed{d^2 = \left(1 - \frac{S_z^2}{S_y^2}\right) \cdot 100\%} \qquad 0 \leq d^2 \leq 100\%$$

where:

$$S_z^2 = \frac{\sum z_i^2}{n - k - 1}$$

$\sum z_i^2$ ← Sum of squared residuals (see table on p. 64)

k ← The polynomial's degree

The statistic S_z^2 is the variance of residuals (z).

Exponential model: $\boxed{\hat{y} = b^x \cdot c}$

This has to be transformed into a linear form:

$$\ln y = x \cdot \ln b + \ln c$$

and the coefficients (b and c) are found by solving the following system of linear equations:

$$\begin{cases} \sum^n \ln y_i = \boxed{\ln b} \cdot \sum^n x_i + \boxed{\ln c} \cdot n \\ \sum^n x_i \cdot \ln y_i = \boxed{\ln b} \cdot \sum^n x_i^2 + \boxed{\ln c} \cdot \sum^n x_i \end{cases}$$

or in matrix form:

$$\begin{bmatrix} \sum^n \ln y_i \\ \sum^n x_i \cdot \ln y_i \end{bmatrix} = \begin{bmatrix} \sum^n x_i & n \\ \sum^n x_i^2 & \sum^n x_i \end{bmatrix} \cdot \begin{bmatrix} \boxed{\ln b} \\ \boxed{\ln c} \end{bmatrix}$$

The goodness-of-fit is measured by:

$$\boxed{d^2 = \left[1 - \frac{\sum^n (y_i - \hat{y})^2}{\sum^n (\ln y_i - \overline{\ln y})^2}\right] \cdot 100\%}$$

Numerator can as well be: $\sum^n (\ln y_i - \ln \hat{y})^2$

$0 \leq d^2 \leq 100\%$

$\overline{\ln y}$ = mean logarithm $= \frac{\sum^n \ln y_i}{n}$

Logarithmic model: $\boxed{\hat{y} = b \cdot \ln x + c}$

Here, the following system of linear equations has to be solved to find "b" and "c":

$$\begin{cases} \sum^n y_i = \boxed{b} \sum^n \ln x_i + \boxed{c} \cdot n \\ \sum^n x_i y_i = \boxed{b} \sum^n x_i \ln x_i + \boxed{c} \cdot \sum^n x_i \end{cases}$$

or in equivalent matrix form (constructed as before).

The goodness-of-fit is measured by:

$$\boxed{d^2 = \frac{\left[\sum^n y_i \ln x_i - \frac{1}{n} \sum^n \ln x_i \cdot \sum^n y_i\right]^2}{\left[\sum^n (\ln x_i)^2 - \frac{1}{n}\left(\sum^n \ln x_i\right)^2\right] \cdot \left[\sum^n y_i^2 - \frac{1}{n}\left(\sum^n y_i\right)^2\right]} \cdot 100\%}$$

$0 \leq d^2 \leq 100\%$

Power model: $\boxed{\hat{y} = c \cdot x^b}$

This is transformed (linearised) as: $\ln y = \ln c + b \cdot \ln x$

and the following system of linear equations has to be solved:

$$\begin{cases} \sum^n \ln y_i = \boxed{b} \cdot \sum^n \ln x_i + \boxed{\ln c} \cdot n \\ \sum^n x_i \cdot \ln y_i = \boxed{b} \cdot \sum^n x \ln x_i + \boxed{\ln c} \cdot \sum^n x_i \end{cases}$$

or in equivalent matrix form (constructed as before).

The goodness-of-fit is measured by: $0 \leq d^2 \leq 100\%$

$$\boxed{d^2 = \frac{\left[\sum^n (\ln x \cdot \ln y) - \frac{\sum^n \ln x \cdot \sum^n \ln y}{n}\right]^2}{\left[\sum^n (\ln x)^2 - \frac{(\sum^n \ln x)^2}{n}\right] \cdot \left[\sum^n (\ln y)^2 - \frac{(\sum^n \ln y)^2}{n}\right]} \cdot 100\%}$$

Non-parametric alternative for linear correlation and regression

Spearman's rank-correlation coefficient (r_s)

The samples $\{x, y\}$ are ranked according to the increasing values of variable x and ranked separately according to the increasing values of variable y (using ranks: 1, 2, 3... in either case). The differences in sample ranks (d) are then used to calculate:

$$r_s = 1 - \frac{6\sum^{n} d^2}{n(n^2-1)}$$ n = number of samples

whose values ($-1 \leq r_s \leq +1$) have the same meaning as those of the Pearson correlation coefficient, r_{xy} (see p. 62).

Note: If there are ties in the sample ranks (i.e., when two or more samples have the same rank according to x-value and/or y-value), a mean rank is used for the tied samples and the formula is modified as:

$$r_s = \frac{E_x + E_y - \sum^{n} d^2}{2\sqrt{E_x \cdot E_y}}$$

where: $E_x = \frac{n^3-n}{12} - T_x$ with $T_x = \frac{m_x^3 - m_x}{12}$;

$E_y = \frac{n^3-n}{12} - T_y$ with $T_y = \frac{m_y^3 - m_y}{12}$;

m_x = number of samples tied according to x-value;

m_y = number of samples tied according to y-value.

Test for the significance of Spearman rank correlation

The one-tail hypothesis tested:

$$\begin{cases} H_0: \rho_s \leq 0 \\ H_1: \rho_s > 0 \end{cases} \quad \text{or} \quad \begin{matrix} \rho_s \geq 0 \\ \rho_s < 0 \end{matrix}$$

The test function:

$$t = |r_s|\sqrt{\frac{n-2}{1-r_s^2}} \quad \text{with } DF = n-2$$

The critical value, t_α, is taken from TABLE 2 if $n > 8$, but from TABLE 9 if $n \leq 8$.

If $t \geq t_\alpha$, H_0 is rejected with confidence $(1-\alpha)100\%$; otherwise, H_0 cannot be rejected.

Non-parametric linear regression

The calculatory procedure is as follows:

(1) Arrange the samples $\{x, y\}$ according to increasing x-values (n = number of samples).

(2) Calculate the "local gradients" between all possible pairs of samples, (x_j, y_j) and (x_i, y_i), for $1 \leq i < j \leq n$:

$$\boxed{b_{ij} = (y_j - y_i)/(x_j - x_i)}$$

There will be $N = n(n-1)/2$ such local gradients to obtain.

(3) Make a cumulative frequency plot of the b_{ij}-values (i.e., a cumulative curve, as in Fig. 5A on p. 7) and use the median, b_m, as an estimator of the regression line's gradient (β). (Median is the 50th percentile read from the curve.)

(4) Identify "outlier" samples and possibly delete these data from the sample population, using the criterion of median absolute deviation (MAD)

The MAD criterion is used as follows:

(a) Calculate the residuals (z)

$$\boxed{z_i = y_i - b_m x_i} \text{ for each sample } (1 \leq i \leq n).$$

(b) Make a cumulative plot of z_i-values and use the median, z_m, to calculate the absolute deviation from the median $\boxed{D_i = |z_i - z_m|}$ for each sample $(1 \leq i \leq n)$.

(c) Make a cumulative plot of D_i-values and read off the median, D_m, which is called the median absolute deviation (MAD) of the population.

(d) "Outliers" are samples with residuals (z_i) falling outside the limits $\boxed{z_m \pm k D_m}$, where k is taken to be 3 or 4, or an intermediate value, so that the number of outliers is kept small and minimum data are deleted.

(e) After elimination of outliers, the initial procedure (1)-(3) is repeated and a new representative b_m is found as median. (If no outliers have been identified, use the original b_m.)

(4) Estimate the regression line's intercept with the y-axis (see p. 64, top): $\boxed{c = z_m (\pm D_m)}$

The interval $\pm D_m$ is an approximate confidence belt, representing the "bulk" of the population (α-level uncertain).

(5) A <u>confidence belt</u> for the regression line can be estimated from the cumulative frequency plot of b_{ij}-values, by reading off the following two percentiles (gradients):

$$b_{(N+N^*)/2} \quad \text{and} \quad b_{(N-N^*)/2}$$

where: N = as defined above; $N = n(n-1)/2$

$$N^* = Z_{1/2\alpha}\left[n(n-1)(2n+5)/18\right]^{1/2}$$

↑ value from TABLE 1B ($Z_{0.05}=1.64$, $Z_{0.025}=1.96$)

The confidence coefficient of the belt will be $(1-\alpha)100\%$.

- <u>Relationship between a quantitative and a qualitative variable</u>

The values of the quantitative variable (x) are split into groups corresponding to the "states" (categories) of the qualitative variable. The groups of x-values are then compared.

Qualitative variable y (binary)	Quantitative variable (x)	
(1)	$x_{(1)}$ – values, $\{x_{(1)}\}$	$\bar{x}_{(1)}$
(2)	$x_{(2)}$ – values, $\{x_{(2)}\}$	$\bar{x}_{(2)}$

One-tail hypothesis:

$$\begin{cases} H_0: \mu_{(1)} \leq \mu_{(2)} \\ H_1: \mu_{(1)} > \mu_{(2)} \end{cases} \quad \text{or} \quad \begin{matrix} \mu_{(1)} \geq \mu_{(2)} \\ \mu_{(1)} < \mu_{(2)} \end{matrix}$$

↑ $x_{(1)}$ = values of x corresponding to state (1)
$x_{(2)}$ = values of x corresponding to state (2)
$n_{(1)} + n_{(2)} = n$

The <u>Snedecor F-test</u> and <u>Student t-test</u> can be used to verify the hypothesis (see p. 26-27), which will be a parametric approach, assuming normality of $\{x_{(1)}\}$ and $\{x_{(2)}\}$ distributions.

<u>Biserial correlation coefficient</u> (r_b) can alternatively be used as a parametric method for correlation between quantitative variable x (assumed here to be normally distributed) and qualitative variable y (binary or binarized). The variable y has two categories: 1 (smaller/minor/weaker/poorer) and 2 (larger/major/stronger/richer). The coefficient is calculated as follows:

$$r_b = \frac{(\bar{x}_{(2)} - \bar{x}_{(1)})\, p \cdot q}{Z \cdot s_x}$$

where: $\bar{x}_{(1)}$, $\bar{x}_{(2)}$ – the x-variable means corresponding to $y=1$ and $y=2$, respectively;
p = fraction of x-data corresponding to $y=2$ ($p = n_{(2)}/n$);
q = fraction of x-data corresponding to $y=1$ ($q = n_{(1)}/n$);
Z = ordinate of standard normal distribution corresponding to q (from TABLE 1A);
s_x = standard deviation of all the x-data.

Coefficient r_b assumes values between -1.0 and +1.0, which have the same meaning as those of the Pearson correlation coefficient (p. 62).

● Relationship between two qualitative variables (association analysis)

Association analysis is a non-parametric equivalent of correlation analysis. The so-called contingency tables (number frequency data) are used, with hierarchical or nonhierarchical rows and columns, depending on the type (**ordinal** or **nominal**) of the variables (p. 1).

■ Ordinal-type variables

A hierarchical contingency table and parameter gamma (γ), estimated by statistic G, can be used in such a case.

variable 1

variable 2	(1)	(2)	(3)	...	(p)	Row sums
(1)	C_{11}	C_{12}	C_{13}	...	C_{1p}	$\Sigma C_{1.}$
(2)	C_{21}	C_{22}	C_{23}	...	C_{2p}	$\Sigma C_{2.}$
(3)	C_{31}	C_{32}	C_{33}	...	C_{3p}	$\Sigma C_{3.}$
⋮	⋮	⋮	⋮	...	⋮	⋮
(k)	C_{k1}	C_{k2}	C_{k3}	...	C_{kp}	$\Sigma C_{k.}$
Column sums	$\Sigma C_{.1}$	$\Sigma C_{.2}$	$\Sigma C_{.3}$	...	$\Sigma C_{.p}$	

Contingency table k × p

k = number of hierarchical states of variable 2 (rows)

p = number of hierarchical states of variable 1 (columns)

C_{ij} = number frequency of samples showing state i of variable 2 and state j of variable 1.

The following two values are calculated:

$$f_R = \Sigma\left(C_{ij} \cdot \Sigma C_{mn}\right) \quad \text{for} \begin{cases} m<i \\ n>j \end{cases}$$

(C_{ij}: non-zero entry; ΣC_{mn}: the remaining entries to the "lower-right" in table)

$$f_L = \Sigma\left(C_{ij} \cdot \Sigma C_{mn}\right) \quad \text{for} \begin{cases} m<i \\ n<j \end{cases}$$

(C_{ij}: non-zero entry; ΣC_{mn}: the remaining entries to the "lower-left" in table)

The statistic G is calculated as follows:

$$\boxed{G = \frac{f_R - f_L}{f_R + f_L}}$$

The meaning of its values, $-1 \leq G \leq +1$, is similar to the meaning of correlation coefficient (p. 62).

Test for the significance of association

One-tail hypothesis $\begin{cases} H_0: \gamma \leq 0 \\ H_1: \gamma > 0 \end{cases}$ or $\begin{matrix} \gamma \geq 0 \\ \gamma < 0 \end{matrix}$ can be test by the function:

$$\boxed{Z = \frac{\Delta f}{M}}$$

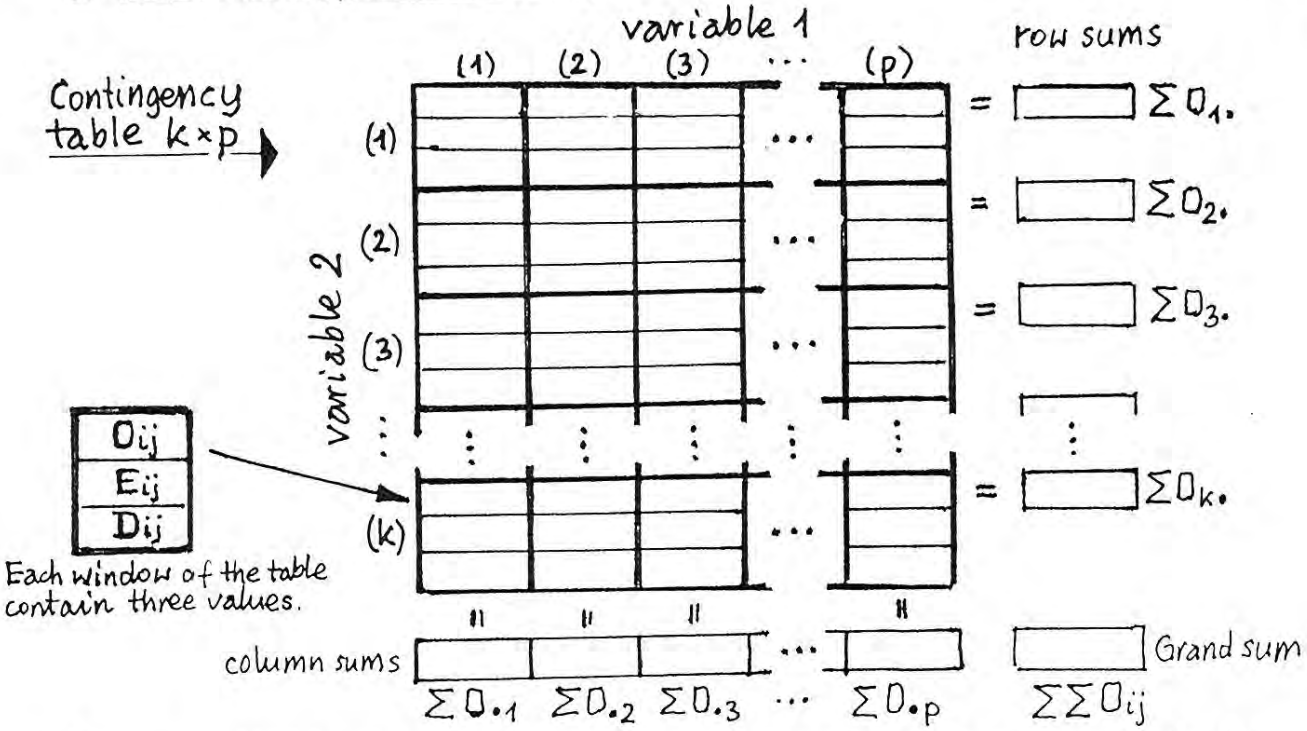
where: $\Delta f = f_R - f_L - \frac{n}{2(k-1)(p-1)}$

$$M = \sqrt{\frac{K_2 P_2}{n-1} - \frac{K_2 P_3 + P_2 K_3}{n(n-1)} + \frac{K_3 P_3}{n(n-1)(n-2)}}$$

K_2, P_2 are the sums of products of all pairs of row sums and column sums;
K_3, P_3 are similar sums of products for all triads of sums.

The critical value, Z_α, is taken from TABLE 1B.

■ Nominal-type variables



O_{ij} = observed number frequency of samples showing state i of variable 2 and state j of variable 1;

E_{ij} = expected number of such samples if there was no association between the two variables (i.e., if the frequency was due to random chance), calculated as:

$$E_{ij} = \frac{\sum O_{i\cdot} \sum O_{\cdot j}}{n} = \frac{(\text{sum of row } i)(\text{sum of column } j)}{n}$$

D_{ij} = the difference between the observed and the expected (random) frequency: $D_{ij} = O_{ij} - E_{ij}$

The significance of the association can be verified by two-tail chi-square test:

$$\boxed{\chi^2 = \sum D_{ij}^2 / E_{ij}} \quad \text{with} \quad DF = (k-1)(p-1)$$

The critical value, $\chi^2_{½\alpha}$, is taken from TABLE 4. If $\chi^2 \geq \chi^2_{½\alpha}$, the null hypothesis of non-association (H_0: $\Delta = 0$) can be rejected with $(1-\alpha)100\%$ confidence; otherwise, H_0 cannot be rejected.

Note: When $n \leq 40$, or when one or more $O_{ij} < 10$, the Yates correction is used in the test formula:

$$\boxed{\chi^2_Y = \sum (D_{ij} - 0.5)^2 / E_{ij}}$$

and the critical value, $\chi^2_{½\alpha}$, is taken from the same TABLE 4.

Pearson's contingency coefficient $C = \sqrt{\chi^2/(n+\chi^2)}$, with values between 0 and 0.8, allows the "strength" of association to be assessed and compared with that of other datasets (possibly involving other variables).

Analysis of multivariate populations

Relationship between a variable and several other variables: multicorrelation and multiregression

This is a parametric analysis of a population involving several quantitative variables, where the focus is on a linear dependence of one of them (y) on the others ($x_1, x_2, x_3, \dots x_p$):

$$y = f(x_1, x_2, x_3 \dots x_p)$$

"Dependent" variable ↑ — "Independent" variables; Functional linear relationship

The analytical issue is similar to straight-line regression (p. 6), but involving more than one independent variable. Closed data should be transformed (see p. 63).

Multiregression

The linear multiregression model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p + \lambda$$

is estimated as:

$$\hat{y} = b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots + b_p x_p + c$$

The sample population is in the form of the following data table:

Samples	y	x_1	x_2	x_3	…	x_p
1	y_1	x_{11}	x_{12}	x_{13}	…	x_{1p}
2	y_2	x_{21}	x_{22}	x_{23}	…	x_{2p}
3	y_3	x_{31}	x_{32}	x_{33}	…	x_{3p}
⋮	⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	x_{n3}	…	x_{np}
	Σy	Σx_1	Σx_2	Σx_3	…	Σx_p

Variables

Matrix $n \times p$ with $(p+1)$ variables measured in n samples.

Requirement: $n > p$

where: y_j = the value of variable y measured in sample j;
x_{ij} = the value of variable x_i measured in sample j.

The matrix is transformed into a matrix equation:

$$\underset{n \times 1}{[Y]} = \underset{n \times (p+1)}{[X]} \cdot \underset{(p+1) \times 1}{[B]}$$

$[Y]$: Columnar vector of the y-values

$[B]$: Columnar vector of the unknowns $(p+1)$

$[X]$: Extended matrix $[X]$; that is, all the x-value columns with a column of values 1 added as the last.

from which: $[B] = ([X]' \cdot [X])^{-1} ([X]' \cdot [Y])$

is found, and its values are the regression line's coefficients:

$$[B] = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_p \\ c \end{bmatrix}$$

This calculatory method follows the least-squares criterion for "best-fit" linear regression.

Multicorrelation

The coefficient of multicorrelation (r_m), as an estimate of the linear relationship between variable y and the other variables ($x_1, x_2, x_3, \dots x_p$), is calculated by one of the following formulae:

$$r_m = \sqrt{\frac{\sum^n (\hat{y}_i - \bar{y})^2}{\sum^n (y_i - \bar{y})^2}} = \sqrt{1 - \frac{\sum^n (y_i - \hat{y}_i)^2}{\sum^n (y_i - \bar{y})^2}} = \sqrt{1 - \frac{\sum^n z_i^2}{(n-1)\, s_y^2}}$$

where: $\hat{y}_i$ = the values of y calculated from the regression formula;
$y_i - \hat{y}_i$ = are residuals (z), similar as used on p. 64;
s_y^2 = the variance of variable y;
other symbols as used throughout the text.

The meaning of the coefficient's values, $-1 \leq r_m \leq +1$, is the same as that of the Pearsonian correlation coefficient (p. 62).

The value of r_m is used to estimate the regression line's goodness-of-fit:

$$\boxed{R^2 = r_m^2 \cdot 100\%}$$ coefficient of determination

Sequential multiregression and multcorrelation

The multiregression and multicorrelation are often performed in a stepwise manner, by linking the dependent variable, y, with successively more of the independent variables:

Step 1	$y = f(x_1)$	$\rightarrow R_1^2$
Step 2	$y = f(x_1, x_2)$	$\rightarrow R_2^2$
Step 3	$y = f(x_1, x_2, x_3)$	$\rightarrow R_3^2$
⋮	⋮	⋮
Step p	$y = f(x_1, x_2, x_3, \dots x_p)$	$\rightarrow R_p^2$

By comparing the successive values of R^2, the relative influence of the individual independent variables on y is estimated and an optimal multiregression model can be selected.

Cluster analysis and dendrograms

This is a method for numerical classification of samples on the basis of variables (Q-mode), or of variables on the basis of samples (R-mode).

From the data matrix $[X]_{n \times p}$, a correlation matrix $[C]_{n \times n}$ for Q-mode or $[C]_{p \times p}$ for R-mode is derived, and the correlation coefficient used, c_{ij}, is either a measure of similarity (s_{ij}) or a measure of taxonomic distance (d_{ij}). The method is non-parametric, unless the Pearson correlation coefficient is used as s_{ij}, particulary in R-mode analysis.

Samples \ Variables	X_1	X_2	X_3	...	X_p
1	X_{11}	X_{12}	X_{13}	...	X_{1p}
2	X_{21}	X_{22}	X_{23}	...	X_{2p}
3	X_{31}	X_{32}	X_{33}	...	X_{3p}
⋮	⋮	⋮	⋮		⋮
n	X_{n1}	X_{n2}	X_{n3}	...	X_{np}

← Data table (matrix)

$= [X]_{n \times p}$

The variables may be of diverse types, measured in very different units, hence the data table is usually standardized at the outset of the analysis. Each column is standardized as follows:

$$(X_{\cdot i})_{std} = \frac{X_{\cdot i} - \bar{X}_{\cdot i}}{S_{\cdot i}}$$

where: $X_{\cdot i}$ = particular value in column i; $\bar{X}_{\cdot i}$ = the mean value of column i; $S_{\cdot i}$ = the standard deviation of column i.

The data table is recalculated, using a formula for s_{ij} or d_{ij} (see next page), to obtain one of the following matrices:

FOR **Q**-MODE ANALYSIS

Samples \ Samples	1	2	3	...	n
1	C_{11}	C_{12}	C_{13}	...	C_{1n}
2	C_{21}	C_{22}	C_{23}	...	C_{2n}
3	C_{31}	C_{32}	C_{33}	...	C_{3n}
⋮	⋮	⋮	⋮		⋮
n	C_{n1}	C_{n2}	C_{n3}	...	C_{nn}

$= [Q]_{n \times n}$

FOR **R**-MODE ANALYSIS

Variables \ Variables	X_1	X_2	X_3	...	p
X_1	C_{11}	C_{12}	C_{13}	...	C_{1p}
X_2	C_{21}	C_{22}	C_{23}	...	C_{2p}
X_3	C_{31}	C_{32}	C_{33}	...	C_{3p}
⋮	⋮	⋮	⋮		⋮
p	C_{p1}	C_{p2}	C_{p3}	...	C_{pp}

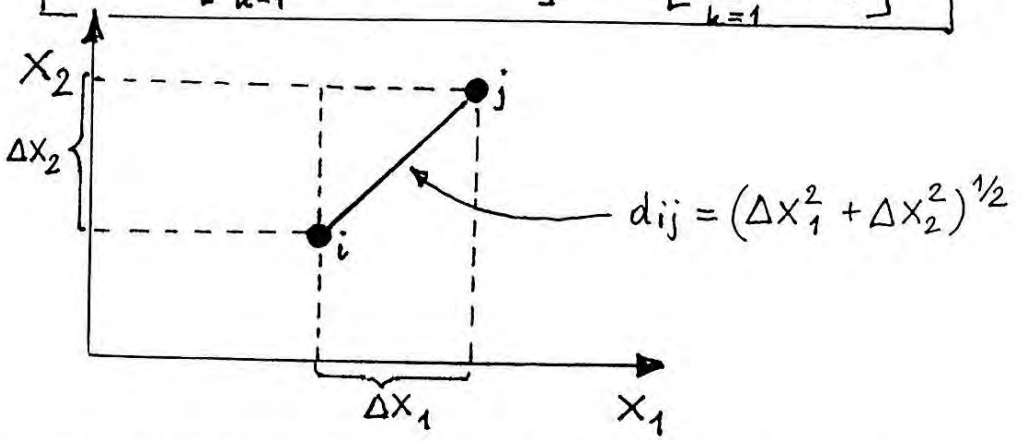
$= [R]_{p \times p}$

A wide range of possible distance (d_{ij}) and similarity (s_{ij}) coefficients can be used, and there is also a range of methods for the grouping (clustering) of items — whether samples (Q-mode) or variables (R-mode) — on the basis of these coefficients. Therefore, there are many "methods" of cluster analysis, and the results for the same data set may differ markedly from one to another.

Distance coefficients:

- Coefficient of Pitagorean distance

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right]^{1/2} = \left[\sum_{k=1}^{p} \Delta x_k^2\right]^{1/2}$$



- Coefficient of mean taxonomic distance (or the Cain-Harrison mean character distance)

$$d_{ij} = \frac{1}{p}\sum_{k=1}^{p} |x_{ik} - x_{jk}| = \frac{1}{p}\sum_{k=1}^{p} |\Delta x_k|$$

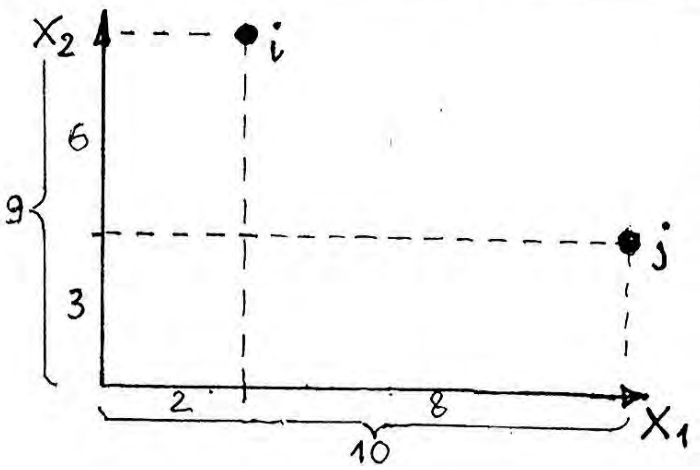
- Coefficient of Canberra taxonomic distance (or the Lance-Williams Canberra metric)

$$d_{ij} = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}} = \sum_{k=1}^{p} \frac{|\Delta x_k|}{x_{ik} + x_{jk}}$$

Similarity coefficients:

- The Sørensen similarity coefficient

$$s_{ij} = \frac{2\sum_{k=1}^{p} \{\min(x_{ik}, x_{jk})\}}{\sum_{k=1}^{p} x_{ik} + \sum_{k=1}^{p} x_{jk}}$$



The smallest coordinate on X_1-axis

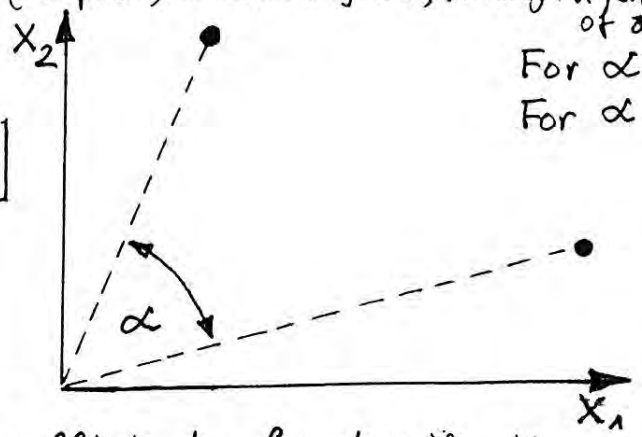
The smallest coordinate on X_2-axis

$$s_{ij} = \frac{2(2+3)}{\underbrace{(2+9)}_{\text{coordinates of } i} + \underbrace{(10+3)}_{\text{coordinates of } j}} = \frac{10}{24} = 0.416$$

• The Pearson coefficient of linear correlation

This coefficient (see p. 62) is commonly used, although it relies on distribution normality of all the variables involved.*

$$r_{ij} = \cos\alpha$$



For $\alpha = 0$, $r_{ij} = 1$

For $\alpha = 90°$, $r_{ij} = 0$

*Note: Other correlation coefficients, such as Spearman's r_s or Kendall's r_k, can also be used as s_{ij}.

• Similarity coefficients for classification of samples (Q-mode)

A wide range of such is used in palaeontology.

Example data:

Fauna species	samples 1	samples 2
(a)	10	20
(b)	4	12
(c)	–	7
(d)	–	15
(e)	32	15
(f)	15	–
(g)	2	–
(h)	1	1
SUM:	64	70

$NS_1 = 6$
$NS_2 = 6$
$NSC = 4$
$C_1 = 64$
$C_2 = 70$
$C = 10 + 4 + 15 + 1 = 30$

Symbol definitions:

NS_1 = total number of species in sample 1;
NS_2 = total number of species in sample 2;
NSC = total number of species common to both samples;
C_1 = coverage in sample 1 (total number of individuals counted);
C_2 = coverage in sample 2 (total number of individuals counted);
C = total coverage common to both samples, using the lowest value.

Jaccard's coefficient of communality:

$$s_{ij} = \frac{NSC}{NS_1 + NS_2 - NSC}$$

$$s_{12} = \frac{4}{6+6-4} = 0.50$$

Jaccard's coefficient weighted for coverage:

$$s_{ij} = \frac{C}{C_1 + C_2}$$

$$s_{12} = \frac{30}{64+70} = 0.22$$

Sørensen's coefficient:

$$s_{ij} = \frac{2\,NSC}{NS_1 + NS_2}$$

$$s_{12} = \frac{2 \cdot 4}{6+6} = 0.75$$

Sørensen's coefficient weighted for coverage:

$$s_{ij} = \frac{2C}{C_1 + C_2}$$

$$s_{12} = \frac{2 \cdot 30}{64+70} = 0.44$$

Simpson's similarity index:

$$s_{ij} = \frac{NSC}{NSC + NS_1} \qquad s_{12} = \frac{4}{4+6} = 0.40$$

Dice's coincidence index:

$$s_{ij} = \frac{2NSC}{2NSC + NS_1 + NS_2} \qquad s_{12} = \frac{4}{4+6+6} = 0.25$$

Sokal-Michener's matching coefficient:

$$s_{ij} = \frac{NSC + NSA}{TNS}$$

where:
NSA = number of species absent in both samples, but present in others;
TNS = total number of species present in the two samples.

Buser's coefficient:

$$s_{ij} = \frac{\sqrt{NSC \cdot NSA} + NSC}{\sqrt{NSC \cdot NSA} + NSC + NS_1 + NS_2}$$

NSA – as above

Legendres's similarity coefficient:

$$s_{ij} = \frac{3NSC}{3NSC + NS_1 + NS_2} \qquad s_{12} = \frac{3 \cdot 4}{3 \cdot 4 + 6 + 6} = 0.50$$

Ochiai's coefficient of closeness:

$$s_{ij} = \frac{NSC}{\sqrt{(NSC+NS_1)(NSC+NS_2)}} \qquad s_{12} = \frac{4}{\sqrt{(4+6)(4+6)}} = 0.40$$

The values of d_{ij}- and s_{ij}-type coefficients are all in the range of -1 to $+1$. However, they are sometimes multiplied by 100% – chiefly for one reason: to make small values better "visible" in digital display (e.g., $s_{ij} = 0.0012$ may not show up if only two digits after the decimal point are displayed/printed, and may thus be taken mistakenly for $s_{ij} = 0.00$; if multiplied by 100, it will be $s_{ij} = 0.12$).

Clustering methods

Some of the commonly used clustering method are demonstrated below, with small-size data matrices $[d_{ij}]$ and $[s_{ij}]$ as calculatory examples.

- The nearest-neighbour (single-link) method

  This method is suitable for both $[s_{ij}]$ and $[d_{ij}]$ matrices. The items to be clustered, whether samples (Q-mode analysis) or variables (R-mode analysis), are grouped progressively according to the similarity/distance to their closest neighbours. Each step of the merger process reduces by one the number of items. As groups are formed, the similarity/distance between them is defined as the similarity/distance between their nearest members.

For example, suppose the following matrix $[d_{ij}]$, involving five items, is to be clustered.

	1	2	3	4	5
1	0.00	0.02	0.06	0.10	0.09
2	0.02	0.00	0.05	0.09	0.08
3	0.06	0.05	0.00	0.04	0.05
4	0.10	0.09	0.04	0.00	0.03
5	0.09	0.08	0.05	0.03	0.00

Items 1 and 2 are merged ($d_{12} = d_{21} = 0.02$, which is the smallest distance in the matrix) and the distances between this group and the three remaining items are recalculated as follows:

$$d_{(12)3} = \min\{d_{13}, d_{23}\} = d_{23} = 0.05$$
$$d_{(12)4} = \min\{d_{14}, d_{24}\} = d_{24} = 0.09$$
$$d_{(12)5} = \min\{d_{15}, d_{25}\} = d_{25} = 0.08$$

which gives a new matrix:

	(12)	3	4	5
(12)	0.00	0.05	0.09	0.08
3	0.05	0.00	0.04	0.05
4	0.09	0.04	0.00	0.03
5	0.08	0.05	0.03	0.00

The smallest d_{ij} in this matrix is $d_{45} = 0.03$, and so items 4 and 5 are merged into a second group. The distances now become:

$$d_{(12)3} = 0.05 \text{ (as before)}$$
$$d_{(12)(45)} = \min\{d_{14}, d_{15}, d_{24}, d_{25}\} = d_{25} = 0.08$$
$$d_{(45)3} = \min\{d_{34}, d_{35}\} = d_{34} = 0.04$$

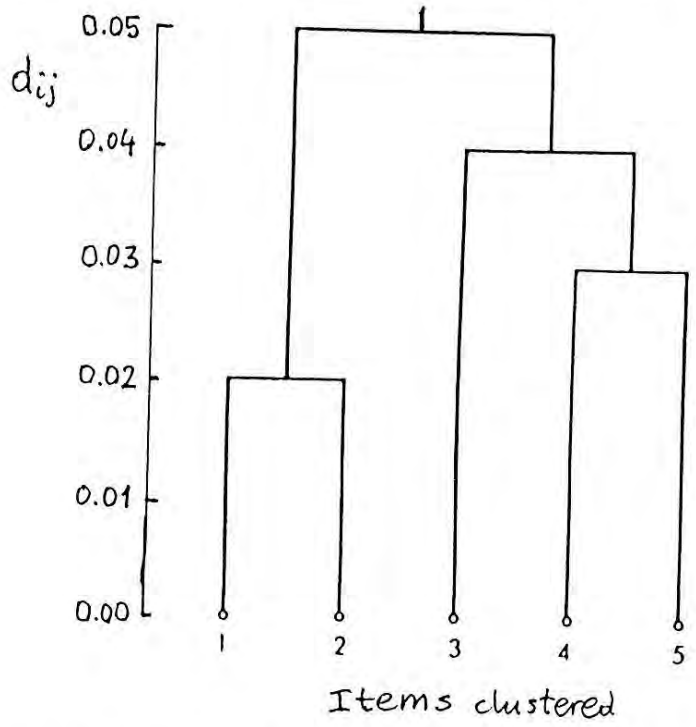
and the new matrix is:

	(12)	3	(45)
(12)	0.00	0.05	0.08
3	0.05	0.00	0.04
(45)	0.08	0.04	0.00

The smallest d_{ij} now is $d_{3(45)} = 0.04$, and so item 3 is merged with group (45). This merger completes the clustering, because the matrix becomes:

	(12)	(345)
(12)	0.00	0.05
(345)	0.05	0.00

which means two groups linked into one at $d_{(12)(345)} = 0.05$ level. The resulting dendrogram, or diagram displaying the result of the clustering, is shown below.



- The farthest-neighbour (complete linkage) method

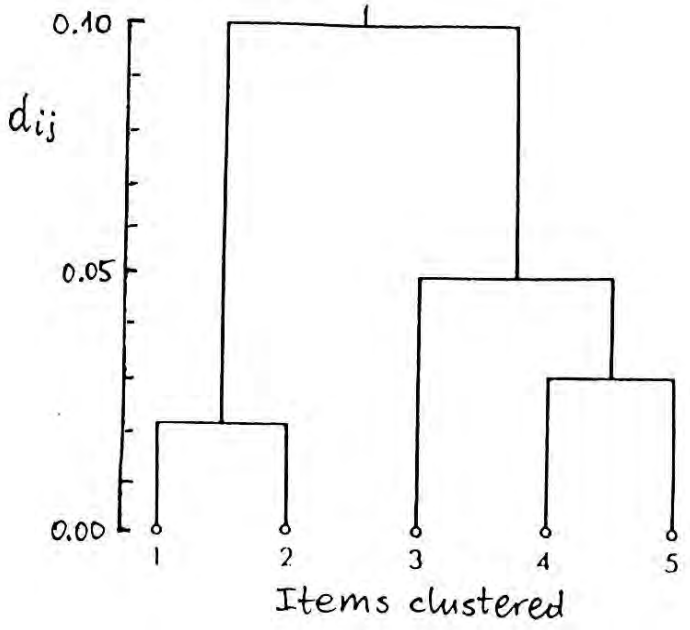
This method is suitable for both $[d_{ij}]$ and $[s_{ij}]$ matrices, and exactly opposite of the single-link method, in that a distance between groups is now defined as the distance between their most remote pair of members.

Suppose we apply this method to the matrix $[d_{ij}]$ from previous example. We begin, as in the previous method, by merging the two closest items, 1 and 2. The distances between this group and the three remaining items are obtained as follows:

$$d_{(12)3} = \max\{d_{13}, d_{23}\} = d_{13} = 0.06$$
$$d_{(12)4} = \max\{d_{14}, d_{24}\} = d_{14} = 0.10$$

$$d_{(12)5} = \max\{d_{15}, d_{25}\} = d_{15} = 0.09$$

The final result is the dendrogram shown below. Although the result is similar to that obtained by the single-link method (cf. previous page), such is not always the case.



- Centroid method

This method is suitable for both $[d_{ij}]$ and $[s_{ij}]$ matrices. The groups are considered to be in Euclidean space, and are replaced on formation by the co-ordinates of their centroid. The distance/similarity between groups is defined as the distance/similarity between the group centroids. Groups with the smallest distance, or highest similarity, are merged first.

For example, suppose the following matrix $[d_{ij}]$, pertaining to five items, is to be clustered:

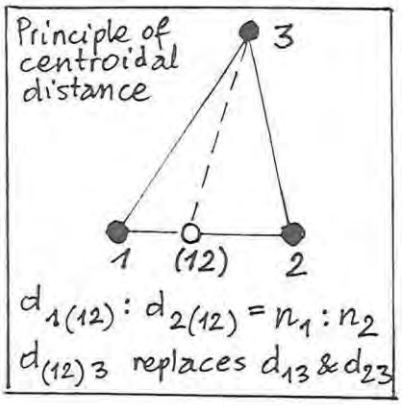
	1	2	3	4	5
1	0.00	0.01	0.29	0.50	0.50
2	0.01	0.00	0.26	0.49	0.53
3	0.29	0.26	0.00	0.05	0.13
4	0.50	0.49	0.05	0.00	0.04
5	0.50	0.53	0.13	0.04	0.00

The first to be merged are items that are closest, which means $d_{12} = 0.01$ in this case. Items 1 and 2 are thus merged and

the centroid of this group is computed. The calculation of of the distance between group (12) and item 3 is:

$$d_{(12)3} = \left[\frac{n_1}{n_1+n_2} d_{13}^2 + \frac{n_2}{n_1+n_2} d_{23}^2 - \frac{n_1 n_2}{(n_1+n_2)^2} d_{12}^2\right]^{1/2}$$

where: n_1, n_2 – numbers of items in "groups" 1 and 2 (which are $n_1 = 1$ and $n_2 = 1$ at this early stage);



and the new matrix is:

	(12)	3	4	5
(12)	0.00	0.2725	0.4925	0.5125
3	0.2725	0.00	0.05	0.13
4	0.4925	0.05	0.00	0.04
5	0.5125	0.13	0.04	0.00

The smallest d_{ij} now is $d_{45} = 0.04$, and items 4 and 5 are merged to form a second group, with its members replaced by the co-ordinates of the group centroid.

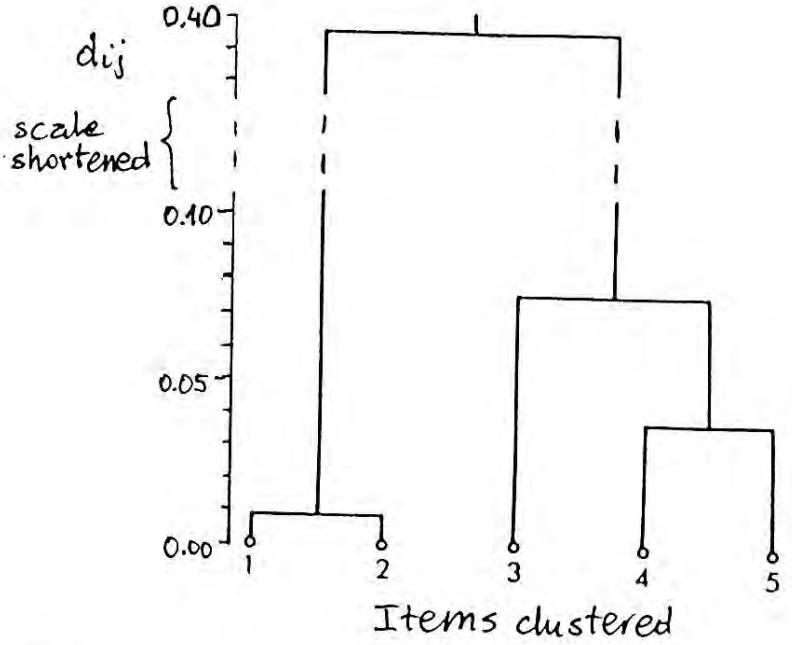
	(12)	3	(45)
(12)	0.00	0.2725	0.4925
3	0.2725	0.00	0.08
(45)	0.4925	0.08	0.00

The smallest d_{ij} now is $d_{3(45)} = 0.08$, and so item 3 is merged with group (45). The last step is to merge the two groups, (12) and (345) into a single group.

	(12)	(345)
(12)	0.00	0.3625
(345)	0.3625	0.00

The resulting dendrogram is shown below (next page).

A disadvantage of the centroid method is that if the sizes of any two groups to be merged are very different, the centroid of the new group will be very close to that of the larger of these groups and may remain within that group. The properties of the smaller group are then virtually lost (obliterated).
Furthermore, the centroid method, even if modified to render it independent of group size, is incompatible with such measures of similarity as correlation coefficients (particularly Pearson's r_{ij}).



- Pair-group method

This method is suitable for both [dij] and [sij] matrices. It is somewhat similar to the previous one, but is much less dependent on group size. The items are grouped in pairs. Suppose the following matrix [sij], pertaining to six items, is to be clustered. In this example, Spearman's r_s has been used as coefficient s_{ij}, hence the positive and negative values in the matrix, and the 1.00 values on its main diagonal.

	1	2	3	4	5	6
1	1.00	0.57	0.12	-0.65	-0.62	-0.39
2	0.57	1.00	0.46	-0.79	-0.72	-0.72
3	0.12	0.46	1.00	-0.58	-0.61	-0.52
4	-0.65	-0.79	-0.58	1.00	0.66	0.41
5	-0.62	-0.72	-0.61	0.66	1.00	0.40
6	-0.39	-0.72	-0.52	0.41	0.40	1.00

The highest s_{ij} is $s_{45} = 0.66$ and the next highest is $s_{12} = 0.57$. Items 4 and 5 and items 1 and 2 are the first pairs to be merged into groups, and the matrix is recalculated as follows:

$$s_{(12)3} = \frac{0.12 + 0.46}{2} = 0.29$$

$$s_{(12)(45)} = \frac{(-0.65) + (-0.62) + (-0.79) + (-0.72)}{4} = -0.70$$

$$s_{(12)6} = \frac{(-0.39) + (-0.72)}{2} = -0.55$$

$$s_{3(45)} = \frac{(-0.58)+(-0.61)}{2} \simeq 0.59$$

$$s_{(45)6} = \frac{0.41+0.40}{2} \simeq 0.41$$

and the new matrix is:

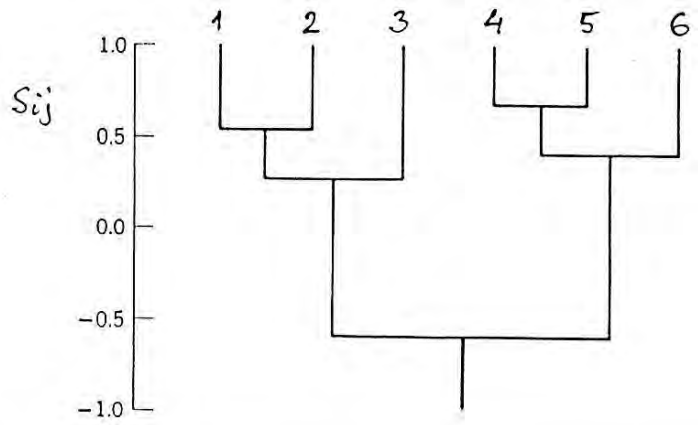
	(12)	3	(45)	6
(12)	1.00	0.29	-0.70	-0.55
3	0.29	1.00	-0.59	-0.52
(45)	-0.70	-0.59	1.00	0.41
6	-0.56	-0.52	0.41	1.00

The next pairs with highest s_{ij} are $s_{(45)6} = 0.41$ and $s_{(12)3} = 0.29$. Item 6 is thus merged with group (45) and item 3 with group (12), and matrix is recalculated as follows:

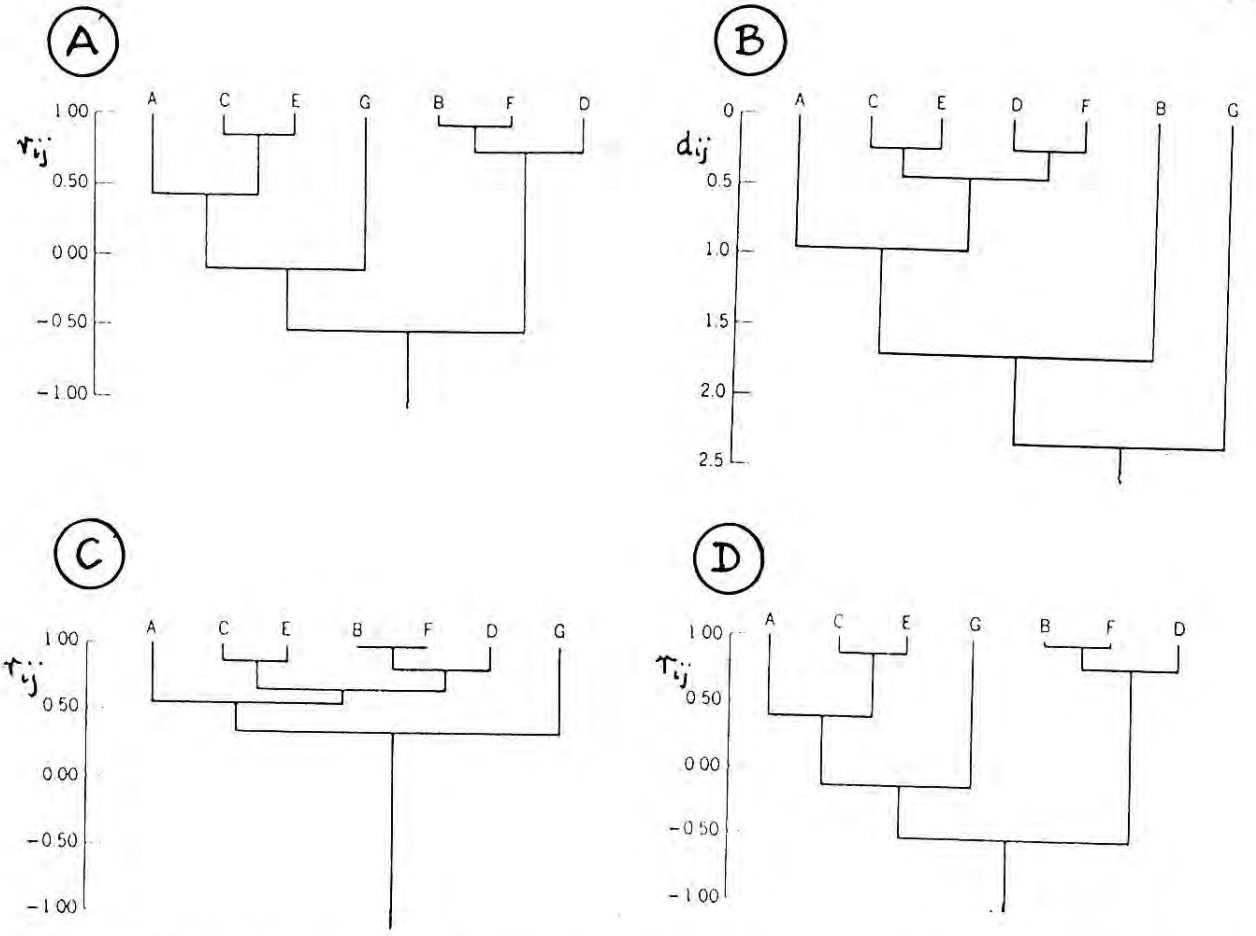
$$s_{(123)(456)} = \frac{(-0.70)+(-0.55)+(-0.59)+(-0.52)}{4} = -0.59$$

	(123)	(456)
(123)	1.00	-0.59
(456)	-0.59	1.00

The resulting dendrogram is shown below.



The combination of different coefficients d_{ij} or s_{ij} with different clustering techniques provides a wide range of methods for cluster analysis. However, different methods generally give different results, as is illustrated by the dendrograms on the next page (same example as above, analyzed by different methods; the six items have been relabelled as A, B, C, ... G); compare those results with the dendrogram above.

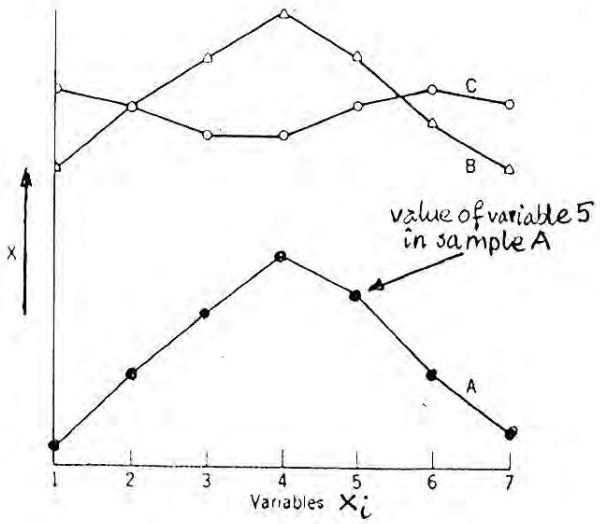


(A) Dendrogram constructed with Pearson's correlation coefficient as s_{ij} and a weighted pair-group method.

(B) Dendrogram constructed by the same clustering method, but using a d_{ij} coefficient.

(C) Dendrogram constructed with Pearson's correlation coefficient as S_{ij} and the single-link clustering method.

(D) Dendrogram constructed with Pearson's correlation coefficient as S_{ij} and an unweighted averaging method for clustering.



The differences are due to different definitions of the "closest" groups in different clustering techniques (see earlier in the text), but also due to definitions of "similarity/distance". The diagram to the left shows three samples (A, B & C) compared on the basis of seven variables measured in each of them. Samples A and B are strongly correlated (high r_{AB}), but are separated by a large distance in the variable space (large d_{AB}). Samples B and C are negatively correlated (negative r_{BC} means high dissimilarity), but are close in distance (low d_{BC}).

From:
Swan A.R.H. & Sandilands M. (1995)
Introduction to Geological Data Analysis.
Blackwell Science, Oxford.
Section 5 (pp. 360-370)

Principal components analysis

8.5.1.1 **Introduction**

Suppose that we have a bivariate data scatter ordinated against two perpendicular axes representing variables with a fairly strong linear correlation (Fig. 8.5). We might have similar variances for each variable. If we were to find a vector which passed along the 'long axis' of the data scatter, plus a second vector perpendicular to the first, we would have a new pair of perpendicular axes that could provide an alternative reference frame for the data scatter. The new axes are different in that the amount of variance in the direction of the first axis is maximised and variance along the second axis is minimised. The useful properties of the result are that we can reduce dimensionality of the data scatter in an optimal way by using the first axis only, and we may be able to make special interpretations about the reason for the elongation of the data scatter in this direction. This may seem mundane with a two-dimensional example, but the method is extremely useful for viewing and understanding high-dimensionality data scatters.

PCA is a technique for finding such linear compounds of correlated variables. These linear compounds are called principal components (PCs) and they have two useful properties.

1 In general, most of the total variance of the p variables in a data set can be accounted for by a comparatively small number k of the new variables. We say that the dimensionality of the data is reduced from p to k.

2 They are uncorrelated, which facilitates examination and analysis of the data.

There are two important implications of these remarks.

1 If the original variables are uncorrelated there is no point in applying the technique: always look at the correlation matrix and scatter diagrams before performing PCA.

2 When the number of variables is small – three or four, perhaps – the dependence can usually be resolved by visual inspection of the correlation matrix and again PCA is not necessary.

Consider the artificial correlation matrix based on 30 observations on five variables which have been labelled 1 to 5:

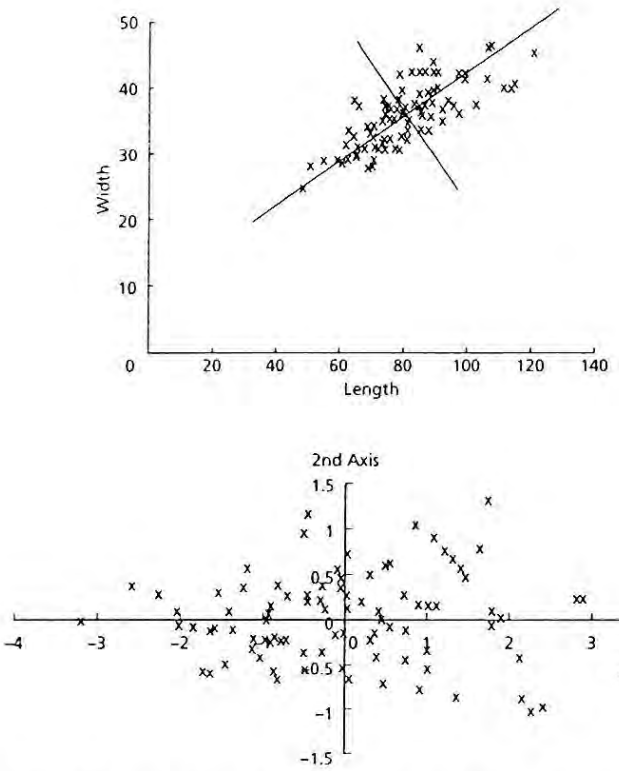


Fig. 8.5 The bivariate scatter of dimensions of echinoids (data as for Fig. 3.1) shown in (a) can be replotted against the new axes (b). In this case, axis 1 corresponds to size and axis 2 to shape.

$$\begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.85 & & & \\ 0.14 & 0.21 & & \\ 0.23 & 0.19 & 0.90 & \\ 0.78 & 0.95 & 0.25 & 0.32 \end{pmatrix} \end{array}$$

It is easy to see that the variables are in two groups, particularly if the order of variables is changed:

$$\begin{array}{c} \\ 2 \\ 5 \\ 4 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 5 & 4 \end{array} \\ \begin{pmatrix} 0.85 & & & \\ 0.78 & 0.95 & & \\ 0.23 & 0.19 & 0.32 & \\ 0.14 & 0.21 & 0.25 & 0.90 \end{pmatrix} \end{array}$$

The variables 1, 2 and 5 form one group; the other group consists of variables 3 and 4. The example is artificial but the procedure used is worth carrying out in practice.

8.5.1.2 Geometrical interpretation

We may regard n observations on p variables as n points in p-dimensional space. We try to define a space in fewer dimensions to explain nearly all the variation.

The process is equivalent to rotating the axes in p-dimensional space in directions such that projections of the points on them are spread out to the maximum extent. This idea is illustrated for two variables in Fig. 8.5(a), where the first axis has been rotated to pass through the data points. The origin of the axes has been moved to the point $(\bar{x}_1, \bar{x}_2)$ but this does not affect the covariance or the PCs. Notice that the second axis has been rotated through the same angle in order to preserve the right angle between axes. Rotations which maintain right angles between all the axes are said to be orthogonal. Notice how in Fig. 8.5(b) the variables show little or no correlation when referred to the new axes.

We shall now assume that the origin of axes has been moved to the point in p-dimensional space which corresponds to the sample means. Further, we shall refer only to the sample covariance and correlation matrices, since the 'true' matrices are never known in practice.

8.5.1.3 Definition of the components

The first PC, which we shall denote by Y_1, is that compound

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

or

$$Y_1 = a_1^T X$$

which has the greatest possible variance.

The variance could be made arbitrarily large by choosing large values of the weights, so a restriction is imposed: the sum of squares of weights is constrained to be 1:

$$a_{11}^2 + a_{12}^2 + \ldots + a_{1p}^2 = 1$$

Subject to this condition, then, there is no other linear compound of the original variables which has a higher variance than Y_1. As we saw in Section 8.2.1.4, this variance can be written

$$\mathrm{Var}(Y_1) = a_1^T S_X a_1$$

where S_X is the sample covariance of the original variables.

The second PC is the linear compound

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

which has the next highest variance, $a_2^T S_X a_2$. It can be shown that the components Y_1, Y_2 are uncorrelated.

The number of components which can be obtained is equal to the number of correlated variables in the data set. In practice, all the sample covariances (and correlations) are non-zero, so computation will produce p components from p variables.

Now the total variance of the components is the same as that of the original variables so that all components are needed to explain their total variance. It is usual, however, for most of the variance to be concentrated in the first few components and to this extent we can say that most of the variation in the data is explained by them.

8.5.1.4 Computation of the weights (loadings) a_i

The weights, commonly known as loadings, are derived from a covariance or a correlation matrix. They are affected by the scales of the variables, so if different types of variable are involved (for example, ppm silver, percentage organic material, pH) the correlation matrix would be appropriate. When the variables are measured on the same scale, the covariance matrix is often used because the results are easier to interpret and statistical tests are easier if they are at all appropriate. The weights obtained from the covariance and correlation matrices differ and it is not possible to compute weights from one type of matrix by transforming those from the other.

The set of p transformations may be combined in a single statement:

$$Y = AX$$

where A is a matrix whose rows are the vectors of weights

$$\begin{array}{l} a_1^T = (a_{11} \ldots a_{1p}) \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ a_p^T = (a_{p1} \ldots a_{pp}) \end{array}$$

We have remarked that the new variables Y_i are uncorrelated and that their variances are given by $a_i^T S a_i$. The covariance matrix of the PCs is, then,

$$AS_XA^T = \begin{pmatrix} \mathrm{Var}(Y_1) & 0 & \ldots 0 \\ 0 & \mathrm{Var}(Y_2) & \ldots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \mathrm{Var}(Y_p) \end{pmatrix}$$

or

$$AS_XA^T = S_Y$$

The transformation $Y = AX$ of the variables X has caused a transformation

of their covariance matrix S_X to a matrix S_Y in diagonal form.

The computation of the weights and variances requires the use of a computer and it is important to choose reliable software. Some packages use mathematical rather than statistical terms for the vectors and matrices we have described here, making the output rather abstract; for that reason we explain them now.

In mathematical language, the rows of A (or equivalently the columns of A^T) are called eigenvectors of the matrix S_X and the elements in the leading diagonal of S_Y are called its eigenvalues. In statistical terms the elements of the eigenvectors are weights a_{ij} applied to the original variables and the eigenvalues are variances of the new, uncorrelated variables. PCA is an example of what are sometimes called eigenvector techniques.

Worked example: see Box 8.8.

8.5.1.5 Principal component scores

The n values of the new variables are computed by using the weights a_{ij} to calculate linear combinations of the values of the original variables in each observation: for the rth object, for example, the value of the kth PC Y_k will be

$$Y_{kr} = a_{k1}x_{k1} + a_{k2}x_{k2} + \ldots + a_{kp}x_{kp}$$

Box 8.8 Worked example: principal components analysis

Using the data on skull measurements from fossil oreodont mammals (Appendix 3.11), we obtain the following from a standard statistics package.

Eigenanalysis of the correlation matrix:

	PC1	PC2	PC3	PC4
Eigenvalue	3.4428	0.3898	0.1116	0.0557
Proportion	0.861	0.097	0.028	0.014
Cumulative	0.861	0.958	0.986	1.000

Loadings:

Variable	PC1	PC2	PC3	PC4
Brainw.	−0.497	−0.488	0.705	−0.135
Toothl.	−0.501	−0.468	−0.598	0.414
Bullal.	−0.519	0.290	−0.307	−0.743
Bullad.	−0.482	0.677	0.226	0.508

Box 8.8 *Continued*

PC1 is extremely strong (86.1% of total variance) and is decribed by loadings that are similar for each variable. It is clearly interpretable as due to intercorrelations resulting from size variation: larger skulls tend to have larger component parts. PC2 accounts for about three-quarters of the remaining variation and corresponds to a type of shape variation, in which the bulla has varying relative size. PC3 and PC4 are minor and may be picking up just random effects.

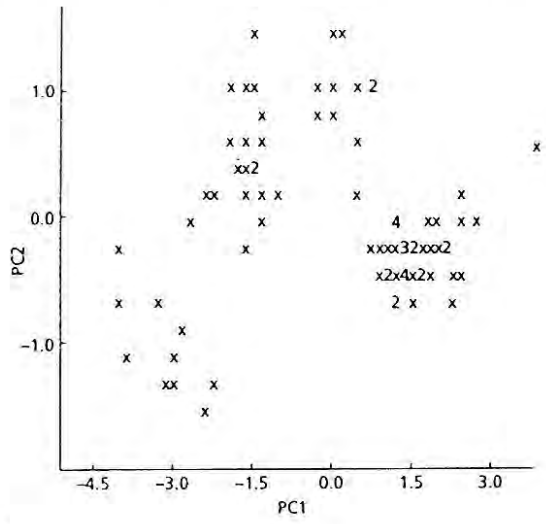


Fig. B8.8.1

The PC scores plot for PC1 and PC2 is given in Fig. B8.8.1. The PC1 axis is now readable as size and PC2 as shape, so the scatter is a useful picture of the relationship between the two. Also, the scatter suggests the existence of a number of clusters, perhaps different species.

The values are called the kth PC scores and they constitute the new data. These would correspond to the coordinates in Fig. 8.5(b), and indeed it is routine to produce scatter plots of PC scores, particularly PC1 vs. PC2. For example, see Box 8.8.

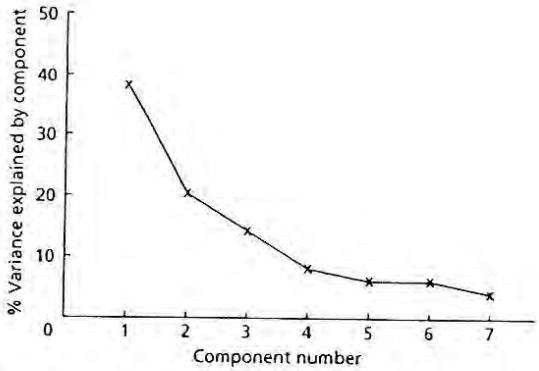


Fig. 8.6 A 'scree' diagram, showing the decline in variance explained with increasing principal component number.

8.5.1.6 Reduction of dimensionality of the space

Normally, some components will be found to contribute little to the total variance. The implication is that the values of the component scores are roughly constant and that their dimensions may be ignored – in effect, their values are replaced by their means, which we arranged to be zero. In this sense the dimensionality of the space is reduced, simplifying the analysis and interpretation of the data. When k components are omitted, only the first $(p - k)$ columns of the transforming matrix A are used to compute component scores. The components are still expressed as compounds of the original p variables at this stage but methods of reducing the number of variables are also available. Indications of these are given in Section 8.5.1.8.

Among the common criteria used for omitting components are the following.

1 Increase in the percentage of the total variance explained when the next most important component is included. If k components have been included and the addition of the next does not affect the proportion of total variance significantly, that component and the remainder are omitted. Judgement of the significance of the increase is often subjective and may be aided by a plot of the cumulative proportion against the number of components, illustrated in Fig. 8.6. It is thought that there is a tendency to retain too many components when this criterion is used.

2 The proportion of total variance explained. It has been suggested that enough components should be included to explain a high proportion (90%) of the total variance. This criterion may lead to the omission of too many components.

3 A component is omitted if its variance is less than the average of all p components (less than 1 when the correlation matrix is used).

4 Approximate equality of the last k eigenvalues (variances). If the variances of the last k components are equal the variation in them is described as spherical or isotropic because none contains more information than any other. As the important aspects of the data are summarised in the first $(p - k)$ components the last k are omitted.

8.5.1.7 Correlations between variables and components

The square of the sample correlation coefficient between variable X_i and PC Y_j can be shown to equal

$$r_{ij}^2 = a_{ij}^2 \cdot \mathrm{Var}(Y_j)/s_{ii}$$

where s_{ii} is the sample variance of variable X_i (which will be, when the correlation matrix is used to produce PCs, equal to 1). This shows that the squared correlation is a measure of the proportion of the variance of X_i which is explained by the jth PC. We shall return to this point in the next section.

8.5.1.8 Retention and discarding of variables

When several variables are investigated together, a large quantity of data is required. If some of the variables studied can be omitted the amount of computation and material will be reduced, and PCs have been used for this purpose.

Frequently, a component corresponding to a small eigenvalue – that is, a component with small variance – is considered to be of little importance. A variable with a high weighting on this component may therefore be declared redundant. A procedure for reducing the number of components used from p to k which is based on such reasoning is as follows: discard the variable with the highest absolute value of the coefficients in the component (assuming that it has not been removed at an earlier stage), and then repeat for each of the $(p - k)$ smallest eigenvalues.

The components to be considered for this treatment can be chosen according to their contributions to the total variance. When the components are derived from a correlation matrix, the total variance is equal to the number of variables, since each variable contributes 1 to the sum. It has been suggested that PCs whose contributions fall below 0.7 should be omitted (subject to leaving at least four components).

8.5.1.9 Some common geological interpretations of principal components

(refer to Box 8.9)

Box 8.9 Examples of geological interpetations of PC eigenvalues and loadings

1 Principal components analysis of soil geochemistry data from Masson Hill, Derbyshire (Appendix 3.4). The soil overlies dolomitised limestone and basalt with mineralised veins (first six PCs only shown).

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	4.9035	1.4766	1.1953	0.6156	0.3732	0.1734
Proportion	0.545	0.164	0.133	0.068	0.041	0.019
Cumulative %	54.5	70.9	84.2	91.0	95.2	97.1

Loadings:

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Zn	−0.325	0.429	0.221	−0.003	0.461	0.585
Pb	−0.419	−0.151	0.067	−0.109	−0.230	0.261
Cd	−0.309	0.317	0.476	0.083	0.279	−0.619
Mg	0.064	0.650	−0.296	−0.596	−0.262	−0.006
Ca	−0.365	0.026	−0.389	0.437	−0.028	0.135
Cu	−0.360	−0.327	−0.033	−0.505	0.096	−0.255
Ag	−0.385	−0.331	−0.084	−0.316	0.110	0.105
Sr	−0.286	0.178	−0.628	0.213	0.141	−0.332
Ba	−0.357	0.145	0.280	0.187	−0.739	−0.033

PC1 is very strong and has similar loadings on all but one of the variables: it may be interpretable as a dilution effect. High positive PC1 scores indicate low amounts of all of the variables (except Mg, which is neutral); this could be due to high quantities of a further major component −perhaps SiO_2 or organic material. Negative PC scores will indicate less of the diluting component, and hence higher percentages of the listed elements. Alternatively, Mg may be neutral because it occurs equally in veins and in the host dolomite and basalt. Samples with negative scores, then, would have high amounts of all the other analysed components: all of these would be expected to be higher in veins.

PC2 points to an unexpected association of Ag and Cu, perhaps in higher-temperature veins, and of Zn, Cd and Mg. The interpretation is uncertain.

PC3 includes a correlation of Mg, Ca and Sr: this is likely to be the signature of the host carbonates.

2 Principal components analysis of measurements of forams (Appendix 3.12) (first six PCs only shown).

Box 8.9 *Continued*

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	6.1684	1.3435	0.5182	0.4109	0.2771	0.1536
Proportion	0.685	0.149	0.058	0.046	0.031	0.017
Cumulative	0.685	0.835	0.892	0.938	0.969	0.986

Loadings:

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Length	−0.327	−0.170	0.562	−0.221	−0.587	−0.360
Width	−0.384	−0.020	−0.048	−0.167	−0.101	0.634
Wdthlc	−0.394	−0.017	0.061	0.040	−0.129	0.315
Hghtlc	−0.372	0.021	0.083	0.468	0.249	−0.316
Hght2lc	−0.379	0.076	−0.012	0.402	0.203	−0.208
Diampro	−0.301	0.247	−0.729	−0.320	−0.267	−0.372
Breadth	−0.385	0.075	−0.075	0.136	0.051	0.255
Wdthfor	0.088	−0.765	−0.366	0.392	−0.341	0.039
Locfor	−0.247	−0.558	−0.001	−0.516	0.579	−0.146

PC1 is readily interpretable as size: higher positive PC1 scores will correspond with smaller specimens. Interestingly, 'Wdthfor', the width of the foramen, is size-independent.

PC2 is a shape factor: it seems that the two measurements of the foramen (width and location) vary to an extent independently of the other variables.

PC1 is only useful as a size index, but PC2 may contain taxonomically or stratigraphically useful information.

Geochemistry and petrology

1 PCs are the result of strong correlations between variables. If we are using compositional data expressed as percentage or ppm, then data are closed (Section 3.2.2) and it is inevitable that many induced correlations will result. PC1, then, is very often strong and entirely interpretable as the result of data closure. The PC loadings of trace and minor elements, oxides or phases will tend to be similar, and these are likely to be opposite in sign to the dominant components, e.g. silica. The positive correlation of minor elements is the result of their being equally diluted by the major component: the interpretation of PC1 is this dilution effect. Unfortunately, the closure effect is not linear, so it doesn't all come out in PC1: we can't interpret the other PCs as entirely due to real geological effects!

2 PCs may result from correlations of suites of variables repesenting a particular geological origin. Variables with similar loadings in the same PC may originate from mineralisation at the same temperature; from the same parent magma; from the same sedimentary environment; from the same provenance area, etc. Interpretation can only be done with knowledge of the geological system.

Palaeontology

1 Where data are measurements of various anatomical components on fossils, PC1 is almost inevitably interpretable as size: all variables tend to be positively correlated because a bigger organism is likely to have bigger parts! The PC1 score can be used as an index of size.
2 Once size is removed in PC1, PC2 and other PCs must represent aspects of shape, perhaps related to mode of life.

Wireline logs and remote sensing

These and other methods of automated data retrieval have the common attribute of recovering large amounts of information on various variables which have a subtle relationship with the geological properties. Information on rock type does occur in gamma, sonic, etc. electric logs, and in pixel values in certain bandwidths on satellite images; but the data need processing to enhance such information. It is often found that a geological attribute such as clay content will affect multiple variables similarly, and hence cause correlations. Consequently, it is routine procedure to investigate the PCs to find those related to the required geological property.

8.5.2 Factor analysis

We have seen in the previous section how PCs may have geological interpretations: each of such components can be taken to represent a geological factor. In a sense, PCs can be described as a method for factor analysis, but the term factor analysis should be reserved for methods involving rotation of eigenvectors. The principle is that we can represent the total variation in the data set by a small number (q, say) of geological factors, each manifested in a degree of correlation (elongation) in the data scatter. In factor analysis, we only find q vectors (whereas in PCA the number of PCs equals the number of variables, p), and it is found that the variance in those q directions can be maximised by rotating the axes away from the basic PC eigenvector solution. In the VARIMAX method, orthogonality of the axes is preserved, and rotation of the axis system is attempted to account for as much of the variance as possible. The OBLIQUE method allows the axes to rotate out of perpendicular: consequently they become, to an extent, correlated. It is argued that this is reasonable because real geological factors are likely to be correlated: grain size may be correlated with quartz percentage, water depth with current velocity, etc.

The idea of factor analysis is a reasonable one, but the method has been severely criticised on the following grounds.
1 There is no clear criterion for the number of factors that exist in the data: the number is chosen by the user. If we choose different numbers of factors, the factors found are different. So the result is dependent on the choice of the user, who may be influenced by preconceived ideas.
2 It is easy to systematically rotate one axis in two dimensions through 360° in order to find the direction of maximum variance, but it is difficult to systematically search all directions in multidimensional space! In practice, the user has some control over the rotation, and subjectivity can result.
3 As a result of points 1 and 2, factor analysis can be a complicated procedure and it is more of an interactive modelling exercise. Modelling often suffers from uniqueness problems: it is difficult to know whether or not a result is the best or the only solution.

Discrimination analysis of two populations

The calculatory procedure is shown here in the matrix form.
We have two data sets, [A] and [B], representing two general populations, [α] and [β], with p variables measured in each sample:

$$\underset{p\times n}{[A]} = \underset{p\times n}{[a_{ij}]} \quad \begin{cases} n = \text{number of samples} \\ p = \text{number of variables measured} \end{cases}$$

$$\underset{p\times m}{[B]} = \underset{p\times m}{[b_{ij}]} \quad \begin{cases} m = \text{number of samples} \\ p = \text{number of variables measured} \end{cases}$$

The corresponding mean vectors and mean-difference vector are calculated:

$$\underset{p\times 1}{[\bar{a}_i]} = \begin{bmatrix} \bar{a}_1 \\ \bar{a}_2 \\ \bar{a}_3 \\ \vdots \\ \bar{a}_p \end{bmatrix} \quad \underset{p\times 1}{[\bar{b}_i]} = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \bar{b}_3 \\ \vdots \\ \bar{b}_p \end{bmatrix} \quad \underset{p\times 1}{[\bar{a}_i - \bar{b}_i]} = \begin{bmatrix} \bar{a}_1 - \bar{b}_1 \\ \bar{a}_2 - \bar{b}_2 \\ \bar{a}_3 - \bar{b}_3 \\ \vdots \\ \bar{a}_p - \bar{b}_p \end{bmatrix}$$

and the following hypothesis is tested:

$$\begin{cases} H_0: [\bar{\alpha}_i] = [\bar{\beta}_i] \\ H_1: [\bar{\alpha}_i] \neq [\bar{\beta}_i] \end{cases}$$

Two-tail hypothesis about the general populations' mean vectors, estimated by $[\bar{a}_i]$ and $[\bar{b}_i]$, respectively.

with test function:

$$\boxed{F = \frac{n\cdot m(n+m-p-1)}{p(n+m)(n+m-2)} \sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}(\bar{a}_i-\bar{b}_i)(\bar{a}_j-\bar{b}_j)} \quad \text{with } DF = n+m-p-1$$

The critical value, $F_{\frac{1}{2}\alpha}$, is taken from TABLE 3.

Note: Only if H_0 has been rejected, can the two populations be considered separable and further analysis can be performed.

The combined ("pooled") variance-covariance matrix is calculated for [A] and [B]:

$$\underset{p\times p}{[C]} = \underset{p\times p}{[c_{ij}]}$$

← NB: This is a symmetrical matrix, with variances at the main diagonal and covariances outside the diagonal.

which is further inverted:

$$\underset{p\times p}{[c_{ij}]^{-1}} = \underset{p\times p}{[d_{ij}]}$$

and used to derive the following p coefficients:

$$W_k = \sum_{i=1}^{p} d_{ki}(\bar{a}_i - \bar{b}_i)$$

which make matrix $\underset{p\times 1}{[W_k]}$

Elements of the mean-difference vector (see above)

The discriminant function for separating [α] and [β] is:

$$\boxed{D = w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_px_p}$$

and its discriminating value is: $D_0 = \frac{1}{2}\sum_{k=1}^{p} W_k(\bar{a}_k + \bar{b}_k) = \frac{1}{2}\underbrace{\left[\underset{p\times 1}{[W_k]}\left(\underset{p\times 1}{[\bar{a}_i]} + \underset{p\times 1}{[\bar{b}_i]}\right)\right]}$

calculate and add up the resulting vector (column)

The D-value is calculated for each sample of unknown affiliation. If $D > D_0$, the sample is classified to population [α]; if $D \leq D_0$, it is classified to population [β]

Analysis of sequential data populations

Tests for randomness of data series

Runs test

This is a non-parametric test for binomial data series, and the procedure is similar as described earlier in the text (p. 32). The runs are counted (example below):

series 1 1 1 0 0 1 1 0 0 0 1 0 0 1 1 1 0 0 1 1 0 0 0 1 0 1 1

runs 1 2 3 4 5 6 7 8 9 10 11 12 13

A "run" is an uninterrupted occurrence of one of the two states of the binary variable. The observed number of runs in this example is: $u_o = 13$. The expected mean number of runs in a random series, $\bar{u}_e$, and its expected standard deviation, $s_{\bar{u}}$, are then calculated (see formulae on p. 32). The hypothesis tested is:

$$\begin{cases} H_0: \text{The data series is random} \\ H_1: \text{The data series is non-random} \end{cases}$$

but can be formulated in one-tail version:

$$\begin{cases} H_0: u_o \leq \bar{u}_e \\ H_1: u_o > \bar{u}_e \end{cases} \quad \text{or} \quad \begin{cases} u_o \geq \bar{u}_e \\ u_o < \bar{u}_e \end{cases}$$

The test function is:

$$\boxed{Z = \frac{|u_o - \bar{u}_e|}{s_{\bar{u}}}}$$

where:

$$\bar{u}_e = 1 + \frac{2n_1 n_2}{N} \quad \text{and} \quad s_{\bar{u}} = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - N)}{N^2 (N-1)}}, \quad N = n_1 + n_2$$

The critical value, Z_α, is taken from TABLE 1B. If $Z \geq Z_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence; otherwise, H_0 cannot be rejected.

Runs up-and-down (RUD) test

Another version of the same test, for a data series of non-binomial variable, but whose quantitative or ordinal qualitative values can be converted into a binary series. The data series is turned into a series of "plus" and "minus" signs, according to the "up" (increase) or "down" (decrease) changes in the successive values of the variable. The number of runs is counted and the test and formulae are the same as above.

Some authors use the following formulae instead:

$$\bar{u}_e = \frac{2N-1}{3} \quad \text{and} \quad s_{\bar{u}} = \sqrt{\frac{16N-29}{90}}$$

The RUD test focuses on the occurrence of local rising or falling trends within the data series.

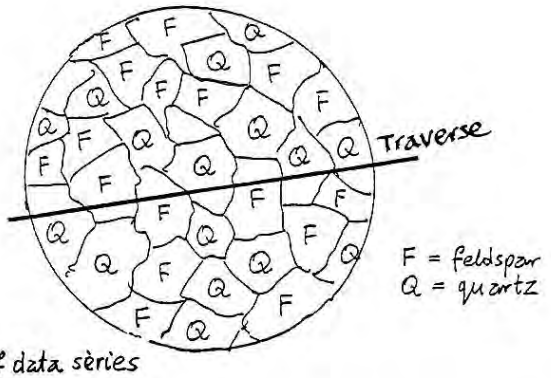
Runs about median (RAM) test

Another version of runs test, with similar procedure, but with the successive data values being compared with the median value of the whole data series (for definition of median, see p. 18). The resulting binary series will then consist of n "plus" and "minus" signs (depending on whether a given data value is larger or smaller than the median), and the observed number of runs is evaluated in an analogous way (see p. 94, bottom).

This test focuses on the occurrence of clusters of data that are larger or smaller than the median value.

(Some authors use mean, instead of median, but this renders the test semiparametric.)

Exercise (simple runs test, p. 94)



Sequence of 100 grains in a pegmatite

feldspar: $n_1 = 64$ grains
quartz: $n_2 = 36$ grains

FQQFQQFFQFQFFFFFFFQQFQFFFQFFFFQFFFQQFQFQQQFFFFFQFFFFFQQQQ
FFQQQFFFFFFQFQFFFFFQFQFQFFQFFFFFQFFFQQFQFFQ

End of data series

Calculate: $u =$

$\bar{u} =$

$s^2_{\bar{u}} =$

$s_{\bar{u}} =$

$$z = \frac{|u - \bar{u}|}{s_{\bar{u}}} =$$

Conclusion:

Kendall's turning-points test

This non-parametric test is conceptually identical to the up-and-down runs test. The data in the series $\{x_i\}$ are replaced with 1 if $\begin{cases} x_i > x_{i-1} \\ x_i > x_{i+1} \end{cases}$ or $\begin{cases} x_i < x_{i-1} \\ x_i < x_{i+1} \end{cases}$; otherwise, x_i is replaced with 0. The data series thus becomes binary. An isolated 1 represents a "turning point". The total number of turning points is counted (T_o) and the mean number of turning points expected for a random series ($\bar{T}_e$) is calculated, together with the standard deviation ($S_{\bar{T}}$) for this mean:

$$\bar{T}_e = \frac{2(n-2)}{3} \qquad S_{\bar{T}} = \sqrt{\frac{16n-29}{90}}$$

where: n = number of elements in the binarized series.

The hypothesis tested in one-tail version is:

$$\begin{cases} H_0: T_o \leq \bar{T}_e \\ H_1: T_o > \bar{T}_e \end{cases} \quad \text{or} \quad \begin{matrix} T_o \geq \bar{T}_e \\ T_o < \bar{T}_e \end{matrix}$$

and the test function is:

$$Z = \frac{|T_o - \bar{T}_e|}{S_{\bar{T}}} \longleftarrow \text{absolute difference}$$

The critical value, Z_α, is taken from TABLE 1B. If $Z \geq Z_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence.

Fisz's median-crossing test

An alternative non-parametric test for a binary or binarized data series. The data series is binarized with the use of median value (see p. 18), $x_m = Md$. If $x_i \leq x_m$, x_i is replaced with 0; if $x_i > x_m$, x_i is replaced with 1; and if $x_i = x_m$, no value is given and this x_i is disregared.

M_o = the observed total number of cases where either 0 is followed by 1 or 1 is followed by 0.

$\bar{M}_e = \frac{n-1}{2}$ Mean number expected for random series

$S_{\bar{M}} = \sqrt{\frac{n-1}{4}}$ Stand. deviation of the mean

n = total number of elements in the binary series.

The one-tail hypothesis is:

$$\begin{cases} H_0: M_o \leq \bar{M}_e \\ H_1: M_o > \bar{M}_e \end{cases} \quad \text{or} \quad \begin{matrix} M_o \geq \bar{M}_e \\ M_o < \bar{M}_e \end{matrix}$$

The test function is:

$$Z = \frac{|M_O - \bar{M}e|}{S_{\bar{M}}}$$ ← absolute difference

and the critical value, Z_α, is taken from TABLE 1B. If $Z \geq Z_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence.

■ Gold's length-of-runs test

Also this non-parametric test uses median value as a reference. A "run length", L, is defined as the number of consecutive data in the series that are either all above or all below the median. The total number of runs of every particular length, N_L, is counted, and the corresponding expected mean number of runs for a random series is calculated:

$$N_{L(e)} = \frac{n + 3 - L}{2^{L+1}}$$

The test function is:

$$\chi^2 = \sum_{L=1}^{L_{max}} \frac{(N_L - N_{L(e)})^2}{N_{L(e)}}$$ with $DF = L_{max} - 1$

L_{max} = the maximum run length in the series

This is a one-tail test, and the critical value, χ^2_α, is taken from TABLE 4. If $\chi^2 \geq \chi^2_\alpha$, the null hypothesis of series randomness is rejected with $(1-\alpha)100\%$ confidence.

■ Meacham's rank-difference test

The values, X_i, in the data series are given ranks, VR, with the smallest value being assigned rank VR = 1, the next being assigned rank VR = 2, etc. The following statistic is calculated:

$$VR = \sum_{i=2}^{n} |VR_i - VR_{i-1}|$$

where: i = the successive number of value X_i in the series;
n = the total number of x-values in the series;
VR = sum of the absolute values of rank differences between the successive x-values.

The expected mean $\bar{VR}_e$ and its standard deviation $S_{\bar{VR}}$ for a random series are calculated:

$$\bar{VR}_e = \frac{(n+1)(n-1)}{3}$$

$$S_{\bar{VR}} = \sqrt{\frac{(n-2)(n+1)(4n-7)}{90}}$$

The test function is:

$$Z = \frac{|VR - \overline{VR}e|}{S_{\overline{VR}}}$$

← absolute difference

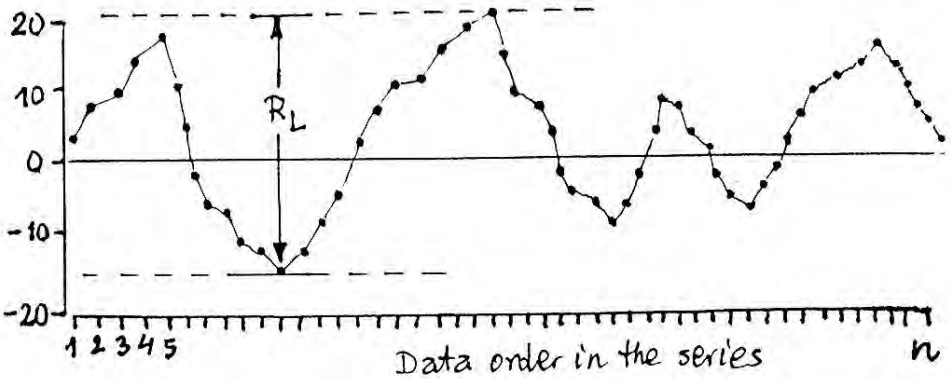
and its critical value is taken from TABLE 1B. If $Z \geq Z_\alpha$, the null hypothesis of series randomness is rejected with $(1-\alpha)100\%$ confidence.

Note: If a tie occurs in the ranking ($X_i = X_k$), a mean rank is given to each of these x-values: $\frac{1}{2}(VR_i + VR_k)$.

Hurst's statistic (rescaled range analysis)

The Hurst statistic, K, is used to recognize the degree of clustering, or grouping, of low and high x-values in the data series. The series should be long, $n > 100$.

The data series $\{X_i\}$ is first converted into series $\{\log X_i\}$ by replacing each value with its logarithm (base 10). The mean ($\bar{X}_L$) and standard deviation (S_L) of this transformed series are calculated, and the series is plotted by using the cumulative unit deviations from the mean, $(\bar{X}_L - X_i)$.



R_L = the maximum range of cumulative departures from the mean for logarithmized data series.

The following relationship holds: $R_L/S_L \sim n^h$, where the coefficient h is a useful parameter estimated by the Hurst statistic K. If the data series is long ($n > 100$) and spans several clusters, then $R_L \sim n^h$ and h is expected to be between 0.5 and 0.8. The value $D = 2 - h$ is the series fractal dimension (expected to be between 1.5 and 1.2). The Hurst statistic is calculated as:

$$K = \frac{\log(R_L/S_L)}{\log(n/2)}$$

A long random (chaotic) data series is expected to have $K = 0.5$.

Values $K > 0.5$ indicate a serial dependence, which means a tendency for high and/or low values to occur in irregular clusters, or groupings. The K-value can be used to quantify and compare the degree of long-term clustering in different series. $K < 0.5$ indicates anticlustering (regularly-spaced data)

● Tests for asymmetric trends in data series

■ Spearman's rank-correlation test

The elements of the data series $\{x_i\}$ are given position ranks (PR) and value ranks (VR), as shown in the example below.

$\{x_i\}$	9.61	5.25	6.35	5.80	5.54	1.71	0.25	1.21	0.43	1.11	0.21
PR:	1	2	3	4	5	6	7	8	9	10	11
VR:	11	9	10	8	7	6	2	5	3	4	1
D:	-10	-7	-7	-4	-2	0	5	3	6	6	10

where D is the rank difference: $D_i = PR_i - VR_i$.
The sum of squared differences is calculated $\sum_{i=1}^{n} D_i^2 = 424$, and the correlation between VR and PR is estimated by:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

Spearman rank-correlation coefficient (see p. 69)

$$-1 \leq r_S \leq +1$$

where: n = number of data in the series.

In the example above, $r_S = 1 - \frac{6 \cdot 424}{1320} = -0.927$.

Since the PRs were set in an increasing order (1, 2, 3, ...), a negative correlation indicates a decreasing trend of x-values. A positive correlation in such a case would indicate an increasing trend of x-values. (PRs can as well be set in decreasing order.)

The significance of correlation can be evaluated by testing a one-tail hypothesis:

$$\begin{cases} H_0: \rho_S \geq 0 \\ H_1: \rho_S < 0 \end{cases} \quad \text{or} \quad \begin{matrix} \rho_S \leq 0 \\ \rho_S > 0 \end{matrix}$$

The test function t is given earlier in the text (p. 69). If $t \geq t_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence; otherwise, H_0 cannot be rejected.

Note: There can be no ties in the PRs. But if ties occur in the VRs, the formula for r_S is slightly more

complicated:

$$r_S = \frac{E_{PR} + E_{VR} - \sum_i^n D_i^2}{2\sqrt{E_{PR} \cdot E_{VR}}}$$

where: $E_{PR} = \frac{n^3 - n}{12}$, $E_{VR} = \frac{n^3 - n}{12} - \frac{m^3 - m}{12}$,

n = total number of elements in the data series,
m = total number of elements tied according to their VR.

■ Kendall's rank-correlation test

The elements of the data series $\{X_i\}$ are given value ranks (VR), and the rank of each element is compared with the ranks of all subsequent elements in the series. For each comparison, a negative score ($VR_k - VR_i < 0$ for $k > i$) or a positive score ($VR_k - Vi > 0$) is obtained. The Kendall correlation coefficient is calculated as follows:

$$r_K = \frac{S_{(+)} - S_{(-)}}{n(n-1)/2} \qquad -1 \le r_K \le +1$$

where: $S_{(+)}$ = total number of positive scores;
$S_{(-)}$ = total number of negative scores;
n = total number of data in the series.

If ties occur, the formula is slightly more complicated:

$$r_K = \frac{S_{(+)} - S_{(-)}}{\sqrt{S_{(+)} + S_{(-)}}\sqrt{S_{(+)} + S_{(-)} + T}}$$

where: T = total number of the "additional" elements tied, for all VRs.

The significance of correlation can be evaluated by testing a one-tail hypothesis:

$$\begin{cases} H_0: \varrho_K \le 0 \\ H_1: \varrho_K > 0 \end{cases} \quad \text{or} \quad \begin{matrix} \varrho_K \ge 0 \\ \varrho_K < 0 \end{matrix}$$

with the test formula:

$$Z = |r_K| \Big/ \sqrt{\frac{2(2n+5)}{9n(n-1)}}$$

and critical value Z_α from TABLE 1B.
If $Z \ge Z_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence and H_1 is accepted.

The Moore-Wallis signs test

The data series is converted into a series of "plus" and "minus" signs (as in the up-and-down runs test). This non-parametric test focuses on the number of "minus" signs. Quantities used:

M_0 = the observed total number of "minus" signs;

$\bar{M}_e$ = the expected mean number of "minus" signs if the series was random (chaotic);

$S_{\bar{M}}$ = standard deviation of the latter mean.

$$\boxed{\bar{M}_e = (n-1)/2} \qquad \boxed{S_{\bar{M}} = \sqrt{(n+1)/12}}$$

where: n = total number of data in the series (minus all ties).

Requirement: $n \geq 12$

The following one-tail hypothesis:

$$\begin{cases} H_0: & M_0 \leq \bar{M}_e \\ H_1: & M_0 > \bar{M}_e \end{cases} \quad \text{or} \quad \begin{matrix} M_0 \geq \bar{M}_e \\ M_0 < \bar{M}_e \end{matrix}$$

is verified with the following test function:

$$\boxed{Z = \frac{|M^* - \bar{M}_e|}{S_{\bar{M}}}}$$

where: $M^* = M_0 - 0.5$ if $M_0 \geq \bar{M}_e$

$M^* = M_0 + 0.5$ if $M_0 < \bar{M}_e$

The critical value, Z_α, is taken from TABLE 1B.

If $Z \geq Z_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence and H_1 is accepted.

Note: The sign of $(M^* - \bar{M}_e)$ has a meaning if H_0 has been rejected. Positive sign means a decreasing trend, whereas negative sign means an increasing trend in the series.

Before applying this test, it may be helpful to "filter" the original data series with a k-term moving average (p. 103) in order to remove "noise", or random variation component. The formulae in such a case are as follows:

$$\boxed{\bar{M}_e = (n-k)/2} \qquad \boxed{S_{\bar{M}} = \sqrt{(n+k)12}} \qquad \boxed{Z = \frac{|M_0 - \bar{M}_e|}{S_{\bar{M}}}}$$

where: n = as above

k = the order of the moving average used, required to be $k \leq n/12$.

Mean-square successive difference test

The data series $\{X\}$ for this test should involve a quantitative variable and contain $n \geq 20$ data. The test is non-parametric and evaluates the following two-tail hypothesis:

$\begin{cases} H_0: \text{The data series is random.} \\ H_1: \text{The data series is non-random and shows some trend.} \end{cases}$

The test function is:

$$z = \frac{\left|1 - \frac{1}{2}k\right|}{\sqrt{\frac{n-2}{n^2-1}}}$$

Instead of the mean value, median (Md, p. 18) may be used to render the test fully non-parametric.

where:

$$k = \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The critical value is taken from TABLE 1B. If $|z| \geq z_{1/2\alpha}$, H_0 is rejected with $(1-\alpha)100\%$ confidence.

If H_0 has been rejected, a positive value of $(1-\frac{1}{2}k)$ will indicate a long-scale trend or oscillation in the series, whereas a negative value will indicate short-scale oscillations.

The Wald-Bertram test

This non-parametric test is for the comparison of two equal-length data series involving a binary or binarized variable (series of "zero"/"one" values or "plus"/"minus" signs).

The hypothesis tested is:

$\begin{cases} H_0: \text{The two data series have the same trend.} \\ H_1: \text{The two series have significantly different trends.} \end{cases}$

The test function is the ratio of inter-series sign mismatches (RSM):

$$RSM = \frac{PMM}{MPM}$$

where: PMM = number of "plus-minus" mismatches
MPM = number of "minus-plus" mismatches

If $RSM \neq 1$, H_0 is rejected with approximately 90% confidence.

EXAMPLE

series 1: + − − − + − + − − + + + + − − + − −
series 2: + − − + + − + + − + + − + + − + − +

PMMs: 3 MPMs: 5

$RSM = \frac{3}{5} = 0.60$

Statistical filtering techniques

If the data series $\{x\}$ is assumed to contain a trend and a random "noise" component:

$$\boxed{\text{Data series} = \text{Trend} + \text{Noise}}$$

one can attempt to remove at least some of the noise by "filtering", or smoothing, the data series with a statistical formula.
The data series is $\{x_1\ x_2\ x_3 \cdots x_n\}$, where the x-values represent a quantitative variable, and a wide range of formulae can be used.

Moving-average technique

$$\bar{x}_i = \frac{x_{i-1} + x_i + x_{i+1}}{3} \qquad \text{3-term formula}$$

$$\bar{x}_i = \frac{x_{i-2} + x_{i-1} + x_i + x_{i+1} + x_{i+2}}{5} \qquad \text{5-term formula}$$

$$\bar{x}_i = \frac{x_{i-3} + x_{i-2} + x_{i-1} + x_i + x_{i+1} + x_{i+2} + x_{i+3}}{7} \qquad \text{7-term formula}$$

Higher-order (9-term, 11-term, etc.) formulae can also be used if the data series is sufficiently long. Note that the higher the order (k) of the formula, the shorter the "filtered" data series. If k is the number of terms in the formula used, the series will lose $(k-1)/2$ elements on its either end, or $(k-1)$ elements in total.

Special empirical formulae

Sheppard's 5-term formula ($k=5$):

$$\bar{x}_i = \frac{1}{35}\left[17x_i + 12(x_{i-1} + x_{i+1}) - 3(x_{i-2} + x_{i+2})\right]$$

Davis's 7-term formula ($k=7$):

$$\bar{x}_i = \frac{1}{21}\left[7x_i + 6(x_{i-1} + x_{i+1}) + 3(x_{i-2} + x_{i+2}) - 2(x_{i-3} + x_{i+3})\right]$$

Agterberg's 9-term formula ($k=9$):

$$\bar{x}_i = 0.31x_i + 0.16(x_{i-1} + x_{i+1}) + 0.08(x_{i-2} + x_{i+2}) + 0.04(x_{i-3} + x_{i+3}) + 0.02(x_{i-4} + x_{i+4})$$

Spencer's 21-term formula ($k=21$):

$$\bar{x}_i = \frac{1}{350}\left[60x_i + 57(x_{i-1} + x_{i+1}) + 47(x_{i-2} + x_{i+2}) + 33(x_{i-3} + x_{i+3}) + 18(x_{i-4} + x_{i+4}) + 6(x_{i-5} + x_{i+5}) - 2(x_{i-6} + x_{i+6}) - 5(x_{i-7} + x_{i+7}) - 5(x_{i-8} + x_{i+8}) - 3(x_{i-9} + x_{i+9}) - (x_{i-10} + x_{i+10})\right]$$

The amount of original variance left after smoothing (trend) and the amount removed (noise) can be estimated:

$$\text{Trend}(\%) = \frac{SS_2 \cdot 100\%}{SS_1}$$

$$\text{Noise}(\%) = \frac{(SS_1 - SS_2)100\%}{SS_1}$$

where:

$$SS_1 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \quad \text{(original series)}$$

$$SS_2 = \sum \bar{x}_i^2 - \frac{(\sum \bar{x}_i)^2}{n-(k-1)} \quad \text{(filtered series)}$$

● Autocorrelation method (for quantitative data)

Data sequence: $\{x_1, x_2, x_3, \cdots, x_n\}$

This method is for the recognition of repetitiveness, or periodicity, in the data sequence. Autocorrelation coefficient (r_L) is calculated and plotted as a correlogram for successive "lag" (L) values:

$$r_L = \frac{\left[(n-L)\sum^{n} x_i x_{i+L} - \sum^{n} x_i \sum^{n} x_{i+L}\right] n(n-1)}{\left[n\sum^{n} x_i^2 - \left(\sum^{n} x_i\right)^2\right](n-L)(n-L-1)} \qquad -1 \le r_L \le 1$$

In other form:

$$r_L = \frac{S_{xx}}{S_x^2} = \frac{COV(x_i, x_{i+L})}{S_x^2}$$

where:

$$COV(x_i, x_{i+L})\Big|_{\text{for given } L} = \frac{(n-L)(\Sigma x_i x_{i+L}) - \Sigma x_i \Sigma x_{i+L}}{(n-L)(n-L-1)}$$

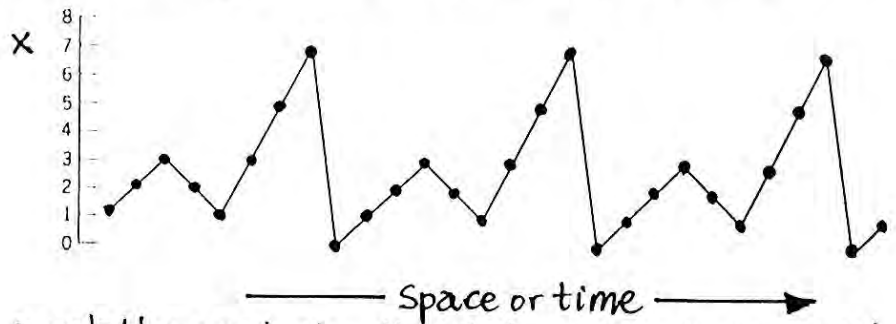
with the summation changing progressively

$\sum_{i=1}^{n}$ for $L=0$, $\sum_{i=2}^{n-1}$ for $L=1$, $\sum_{i=3}^{n-2}$ for $L=2$, etc.

and

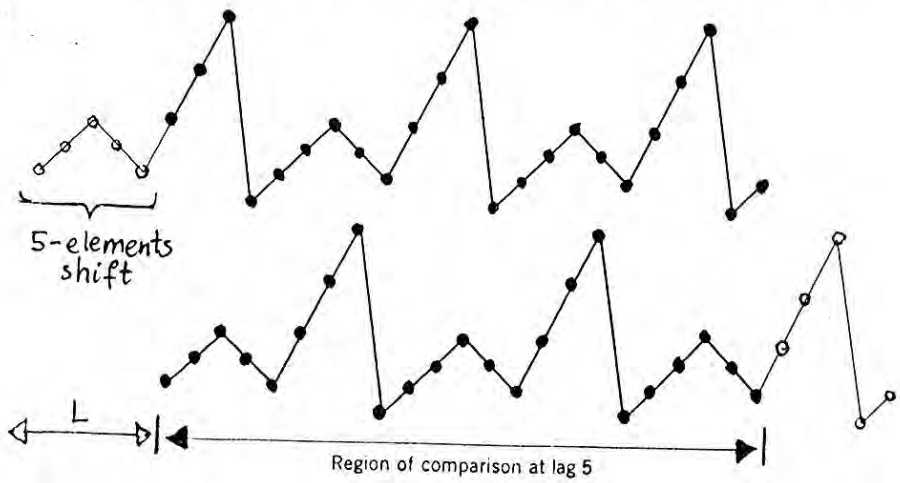
$$S_x^2 = \frac{n\sum^{n} x_i^2 - \left(\sum^{n} x_i\right)^2}{n(n-1)}$$, which is the ordinary variance of all the data sequence.

Example data sequence: a series of measurement of variable x along a distance in space or time



$n = 28$
$L = 0$

In autocorrelation analysis, the data series is compared with itself (or its own "copy") — as is shown below at lag $L=5$.



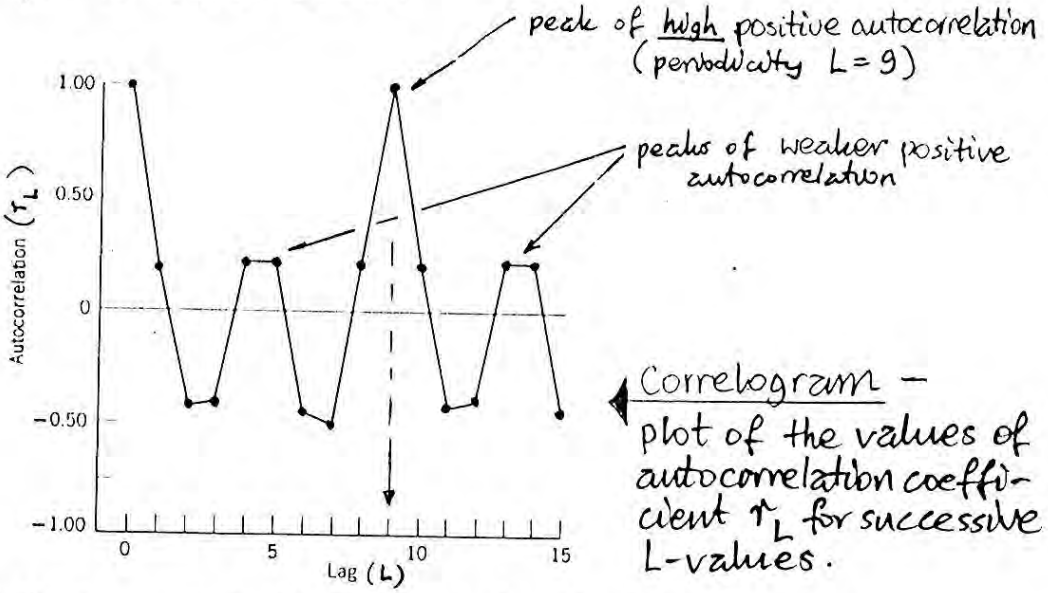
$L = 5$

The procedure is as follows: The data series is shifted relative to its "copy" step by step, with L=0, L=1, L=3, etc.

At lag L=0, the sequence is identical with itself ($r_0 = 1.00$).

At lag L=1, we compute the correlation between elements 1, 2, 3, ... (n-1) of the sequence and elements 2, 3, 4, ..., n of its "copy".
At lag L=2, we compute the correlation between elements 1, 2, 3, ..., (n-2) of the sequence and elements 3, 4, 5, ..., n of its "copy".
And so on, and so on.

The correlogram (for L=0 to L=15) for the data sequence in this example is shown below.



Correlogram – plot of the values of autocorrelation coefficient r_L for successive L-values.

Correlograms help to reveal the periodicity of data sequences.

The significance of any r_L (e.g., correlogram peaks) can be tested by verifying the hypothesis:

$$\begin{cases} H_0: & \rho_L \leq 0 \\ H_1: & \rho_L > 0 \end{cases}$$ "one-tail" hypothesis for positive autocorrelation

or

$$\begin{cases} H_0: & \rho_L \geq 0 \\ H_1: & \rho_L < 0 \end{cases}$$ "one-tail" hypothesis for negative autocorrelation (if needed to be assessed)

Test function:

$$\boxed{Z_L = |r_L| \sqrt{n-L}}$$

with critical value, Z_α, taken from TABLE 1B.

If $Z_L \geq Z_\alpha$, H_0 is rejected with confidence $(1-\alpha)100\%$ and the correlation can be regarded as significant.

• Cross-correlation method (for quantitative data)

The same method as described in the previous section, but now used for the comparison of two different data sequences, in order to find their "best-match" relative position (which we would call "correlation" in lithostratigraphy, for example).

The cross-correlation coefficient (r_{CL}):

$$r_{CL} = \frac{cov(X_{1,i}, X_{2,i+L})}{S_1 \cdot S_2}$$

($X_{1,i}$, $X_{2,i+L}$: elements of respectively; S_1: stand. deviation for sequence 1; S_2: stand. deviation for sequence 2)

is calculated with the following formula:

$$r_{CL} = \frac{n^* \sum x_{1i}x_{2i} - \sum x_{1i} \sum x_{2i}}{\left[n^* \sum x_1^2 - (\sum x_1)^2\right]\left[n^* \sum x_2^2 - (\sum x_2)^2\right]}$$

where: x_{1i} and x_{2i} are the "overlapped" elements of the two sequences (labelled 1 and 2);

n^* = number of the pairs of "overlapped" elements.

As before, the values of r_{CL} for the successive lags (L) are plotted in the form of a correlogram, and can be tested for significance using the following test (same hypotheses as before, previous page). The test function here is:

$$t = |r_{CL}| \sqrt{\frac{n^* - 2}{1 - r_{CL}^2}} \quad \text{with } DF = n^* - 2$$

The critical value t_α taken from TABLE 2. If $t \geq t_\alpha$, H_0 is rejected with $(1-\alpha)100\%$ confidence and the correlation is deemed significant.

• Autoassociation and cross-association methods (for qualitative variable)

Methods analogous to autocorrelation and cross-correlation, but for series of qualitative data. In autoassociation method:

N = total number of elements in the series
m = number of states (element types) involved
n_i = number of occurrences of state i

→ Probability of randomness: $P = \frac{\sum n_i^2 - N}{N^2 - N}$

For each lag L, the following quantities are calculated:

C = number of comparisons (pairs compared)
M = number of matches and MM = number of mismatches
R = ratio M/C (→ plotted for each L as "associogram")
Expected number of matches due to random chance: $\hat{M} = C \cdot P$
Expected number of mismatches due to random chance: $\hat{MM} = C(1-P)$

Test: $\chi^2 = \frac{(M-\hat{M})^2}{\hat{M}} + \frac{(MM-\hat{MM})^2}{\hat{MM}}$ with $DF = 1$

In cross-association method, one data series is moved relative to the other, step by step, and for each comparison the following values are calculated (the example below is for position 3):

```
          E A B A B E B C B D B E
C B A E B E A B A B C B
```

Number of comparisons = 3 (also called "position")
Number of matches = 1
Ratio: 1/3 = 0.33

In this example, sequence 1 (upper) involves 5 "states", or categories, of the qualitative variable (A, B, C, D & E) and consists of $n_1 = 12$ elements. Sequence 2 (lower) involves 4 "states" (A, B, C & E) and has $n_2 = 12$ elements.

First, the probability of a match at any position is calculated for random sequences with similar characteristics. The formula used is:

$$P = \frac{\sum_{i=1}^{k} n_{1i}\, n_{2i}}{n_1\, n_2}$$

where n_{1i} and n_{2i} are the numbers of occurrences of state i in sequences 1 and 2, respectively.

State (i)	n_{1i}	n_{2i}	$(n_{1i}\, n_{2i})$
A	2	3	6
B	5	5	25
C	1	2	2
D	1	0	0
E	3	2	6
Sums:	$n_1 = 12$	$n_2 = 12$	$\sum n_{1i} n_{2i} = 39$

In our example:

$$P = \frac{39}{12 \cdot 12} = 0.27$$

The probability of a mismatch is $(1.00 - 0.27) = 0.73$

The expected number of matches ($\hat{M}$) for an overlapped interval can be calculated as P times the number of comparisons.

In our example, $\hat{M} = (0.27)(3) = 0.81$ for position 3.

The corresponding number of expected mismatches is:

$\hat{MM} = (0.73)(3) = 2.19$.

For each comparison (position), the value of chi-square test function is calculated:

$$\chi^2 = \frac{(M - \hat{M})^2}{\hat{M}} + \frac{(MM - \hat{MM})^2}{\hat{MM}} \quad \text{with DF} = 1$$

or

$$\chi^2_Y = \frac{(|M - \hat{M}| - \tfrac{1}{2})^2}{\hat{M}} + \frac{(|MM - \hat{MM}| - \tfrac{1}{2})^2}{\hat{MM}}$$

(with Yate's correction)

When $\hat{M}$ is small, as for the comparison positions near ends of sequences.

where: M and MM are the observed numbers of matches and mismatches at a given position of comparison.

The critical value, χ^2_α, is taken from TABLE 4. If $\chi^2 \geq \chi^2_\alpha$, the null hypothesis of non-association is rejected with $(1-\alpha)100\%$ confidence.

In the example above, for position 3:

$O = 1 \quad O' = 2$
$E = 0.81 \quad E' = 2.19$

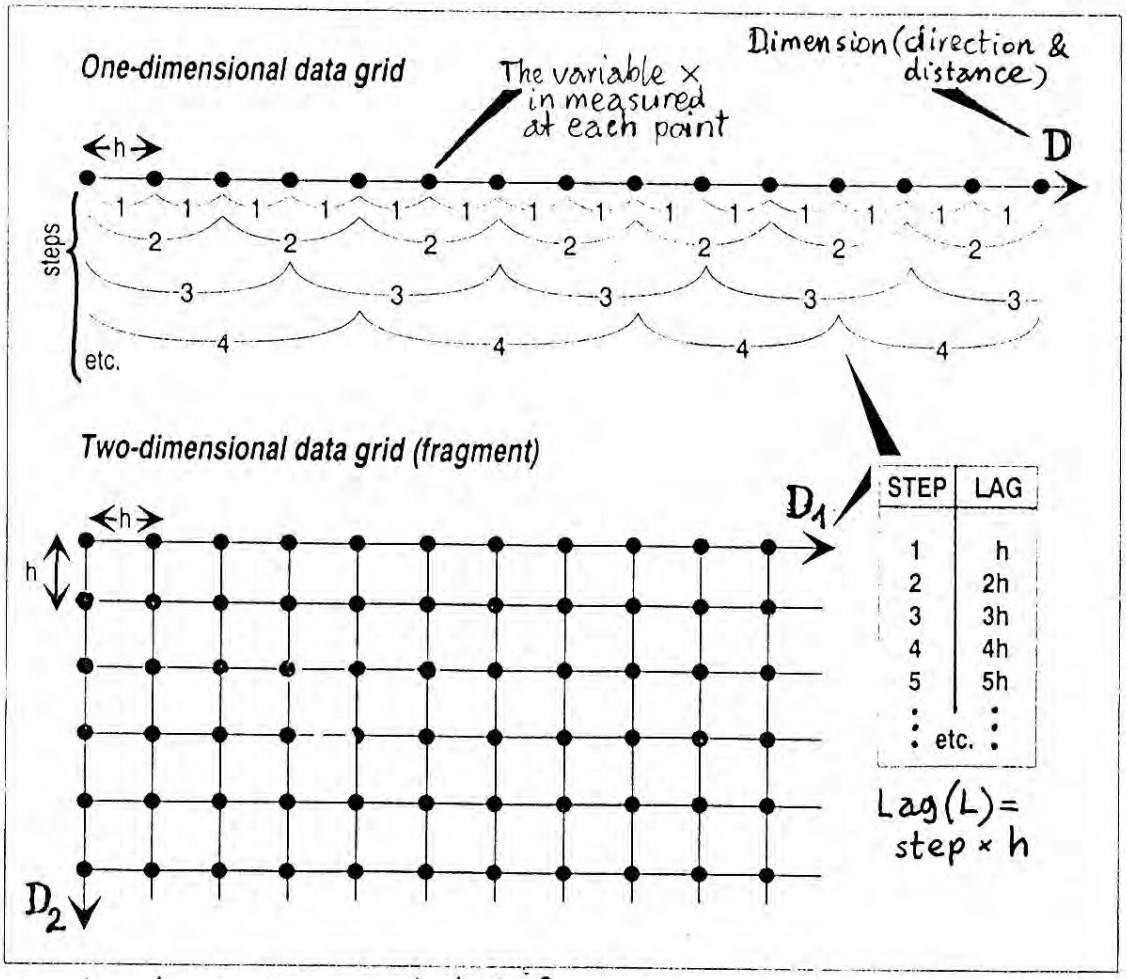
$$\chi^2 = \frac{(1-0.81-0.5)^2}{0.81} + \frac{(2-2.19-0.5)^2}{2.19} = 0.34$$

The critical value $\chi^2_{0.10} = 2.71$, which means that H_0 cannot be rejected with 90% confidence, and not even with 80% confidence ($\chi^2_{0.20} = 1.64$); there is no significant association in position 3.

The successive χ^2-values can be plotted in a form analogous to correlogram (see p.105), to obtain an associogram, where positions with various degree of series association can be depicted and compared.

● Semivariogram

This method is for quantitative data, to recognize the spatial pattern of changes in the variance of variable X. The data are collected at equally-spaced points along a one-dimensional profile or two-dimensional grid (see diagram below). The spacing between the data points is **h**, but the data population is subsequently analyzed by the calcution of $\gamma^*(L)$ for $L = h, 2h, 3h, \ldots$ etc.

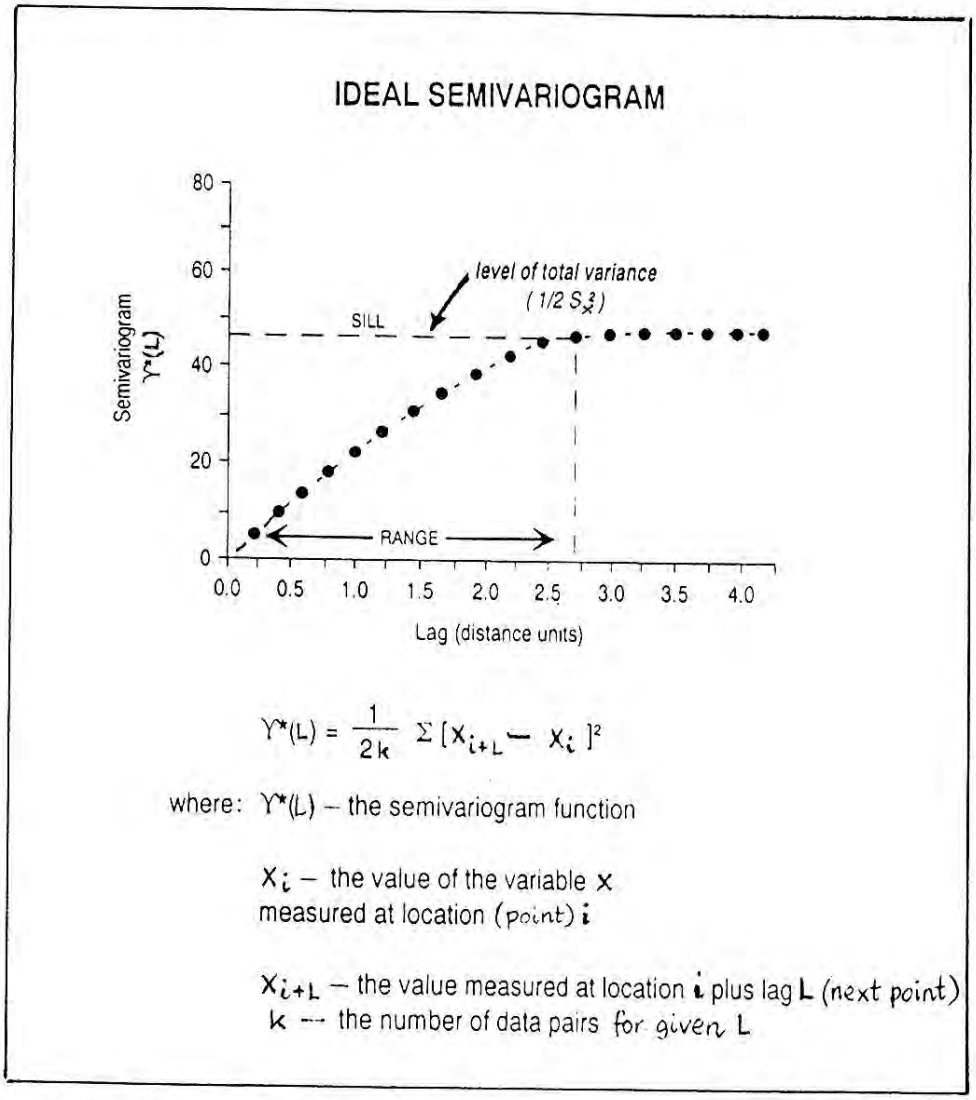


The semivariance is calculated for every

$$\gamma^*(L) = \text{semivar } X_{(L)} = \frac{1}{2k}\sum_{i=1}^{n}(X_{i+L} - X_i)^2$$

where: k = number of data pairs (X_{i+L}, X_i) compared.

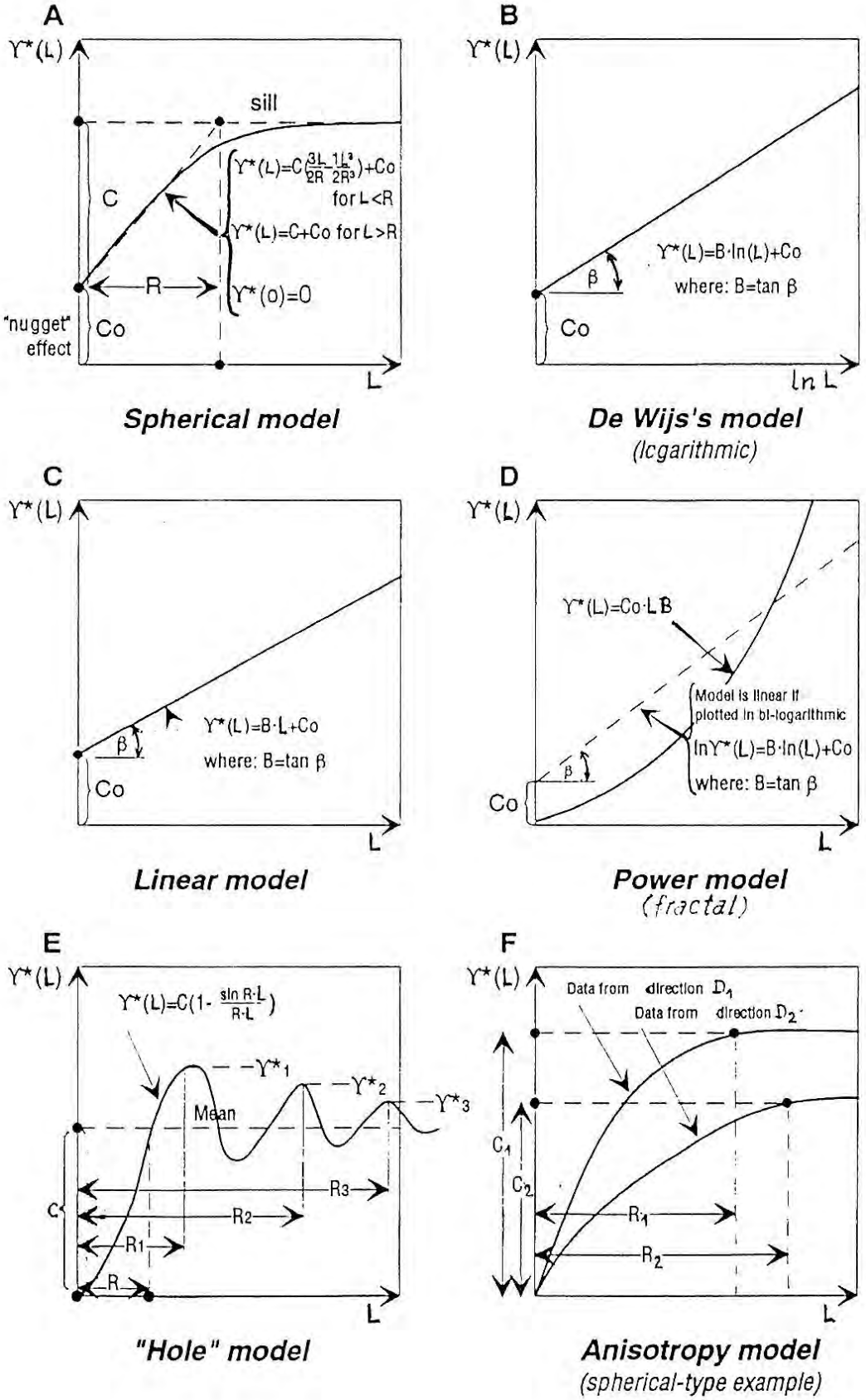
The calculated values of $\gamma^*(L)$ are plotted as a two-axial diagram, called <u>semivariogram</u>.



The <u>range (R)</u> is the distance (L value) at which the function $\gamma^*(L)$ reaches the level of the bulk half-variance ($\frac{1}{2} S_x^2$), called sill. The range, referred to also as the "radius of influence", is the distance – around any given point i – within which the x_i-value is influenced by the x-values in the adjacent points; beyond that distance, the influence of the other, farther-lying points on a given point's value is negligible.

Semivariogram can be calculated for more than one dimension (see diagram F next page) to study <u>anisotropy</u> in the variation of **x**.

Semivariograms may have different forms (see review on the next page), revealing different types of spatial structure of the variation in **x**-values.



. Review of different semivariogram models (based on Agterberg, 1974).

● Markov-chain analysis

This method is for the recognition of "modal cycles" in a data series of qualitative variable. The data series consists of a large number of elements (N) representing a smaller finit number of "states" (s) of the qualitative variable (e.g., lithofacies in a stratigraphic log), such that $N \gg s$. The transitions from one state to another in the series are counted, with two possible variants of taking observations (see next page) – resulting in either an embedded chain or a regular chain model; the former takes into account only changes of state, whereas the latter allows for transition from one state into the same state.
The following four matrices are calculated: (**A**) transition count matrix; (**B**) observed transition probability matrix; (**C**) expected transition probability matrix; and (**D**) difference matrix.

(A)

states	1	2	3	…	s	$\sum_s n_i$ ← Row sum
1	n_{11}	n_{12}	n_{13}	…	n_{1s}	$\sum n_1$
2	n_{21}	n_{22}	n_{23}	…	n_{2s}	$\sum n_2$
3	n_{31}	n_{32}	n_{33}	…	n_{3s}	$\sum n_3$
⋮	⋮	⋮	⋮		⋮	⋮
s	n_{s1}	n_{s2}	n_{s3}	…	n_{ss}	$\sum n_s$
Column sums →	$\sum n_1$	$\sum n_2$	$\sum n_3$	…	$\sum n_s$	$\sum\sum n_i = N$

n_{ij} = number of the observed transitions from state i to state j;
$\sum n_i$ = number of occurrences of state i;
n_{ii} = diagonal elements (= 0 in embeded chain or ≥ 0 in regular chain).

(B)

	1	2	3	…	s	Row sums ↓
1	p_{11}	p_{12}	p_{13}	…	p_{1s}	= 1
2	p_{21}	p_{22}	p_{23}	…	p_{2s}	= 1
3	p_{31}	p_{32}	p_{33}	…	p_{3s}	= 1
⋮	⋮	⋮	⋮		⋮	⋮
s	p_{s1}	p_{s2}	p_{s3}	…	p_{ss}	= 1

$p_{ii} = p_{jj} = 0$ in embedded chain;
p_{ij} = observed probability (frequency, as fraction of 1) of the transition of state i to state j;

$$p_{ij} = n_{ij} / \sum n_i$$

(C)

	1	2	3	…	s
1	e_{11}	e_{12}	e_{13}	…	e_{1s}
2	e_{21}	e_{22}	e_{23}	…	e_{2s}
3	e_{31}	e_{32}	e_{33}	…	e_{3s}
⋮	⋮	⋮	⋮		⋮
s	e_{s1}	e_{s2}	e_{s3}	…	e_{ss}

e_{ij} = expected probability of the transition of state i to state j if the series was random;

$$e_{ij} = \frac{\sum n_j}{N - \sum n_i}$$

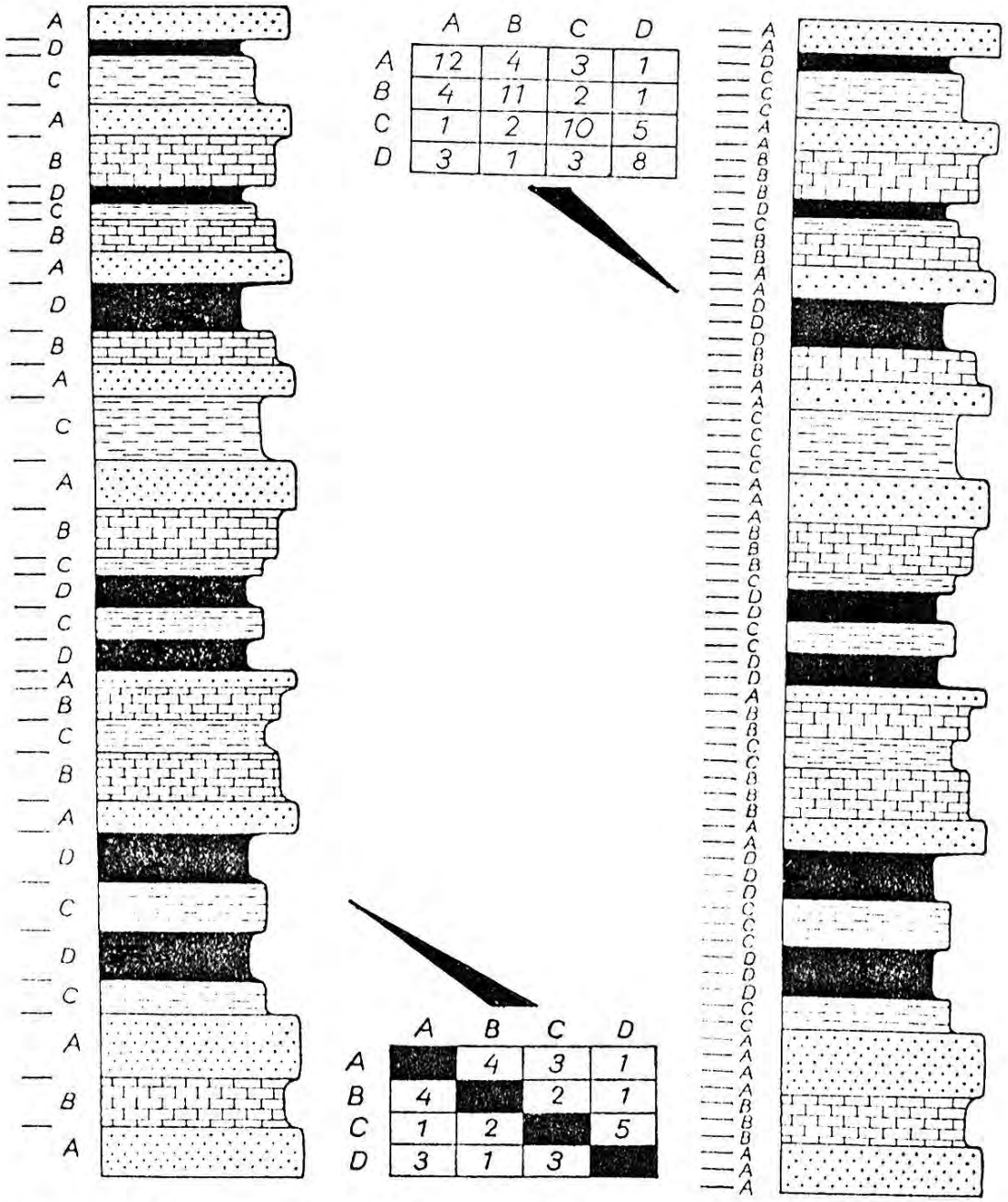
$e_{ii} = e_{jj} = 0$ in embedded chain.

(D)

	1	2	3	…	s
1	d_{11}	d_{12}	d_{13}	…	d_{1s}
2	d_{21}	d_{22}	d_{23}	…	d_{2s}
3	d_{31}	d_{32}	d_{33}	…	d_{3s}
⋮	⋮	⋮	⋮		⋮
s	d_{s1}	d_{s2}	d_{s3}	…	d_{ss}

$$d_{ij} = p_{ij} - e_{ij}$$

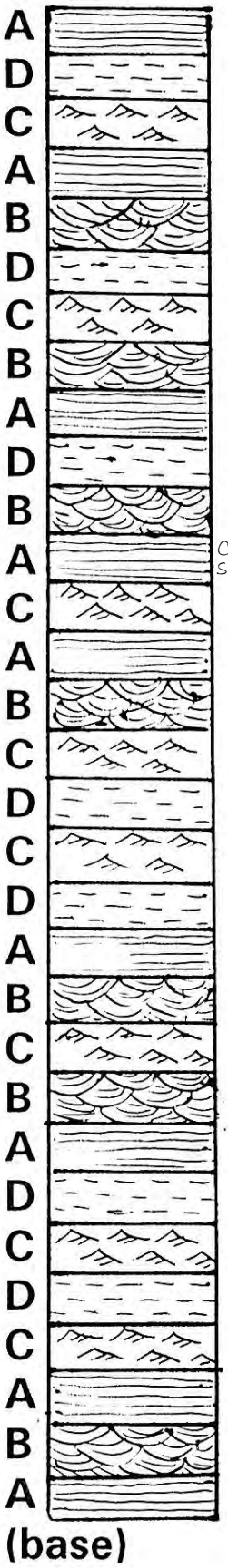
$d_{ii} = d_{jj} = 0$ in embedded chain.
Values $d_{ij} > 0$ are of main interest, and the matrix normally needs to be tested (see further text).



The pattern of data collection for an embedded Markov chain (left) and a regular Markov chain (right); the short horizontal lines on the log's left-hand margin are points of observation.

In an embedded chain, only the changes from one state to another are recognized and counted. The transition count matrix thus has zeros on the main diagonal. This model is most commonly used.
In a regular chain, the observations are taken at regular intervals and the transitions from one state to the same state are included. The transition count matrix thus has values ≥ 0 on the main diagonal.
The advantage of regular chain is that it takes into account the "duration" (thickness) of states. However, the transition counts depend strongly on the "step" (interval) used. In principle, an optimal step is such which gives a balance between the count values on and outside the matrix's diagonal.

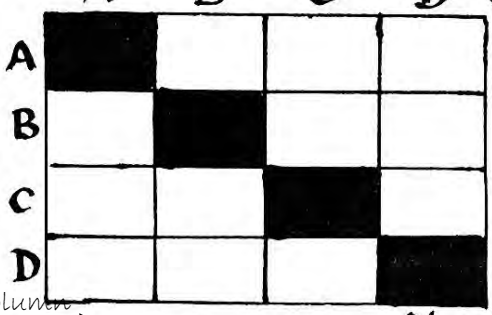
FACIES SUCCESSION IN A FLUVIAL SANDSTONE BODY



EXERCISE

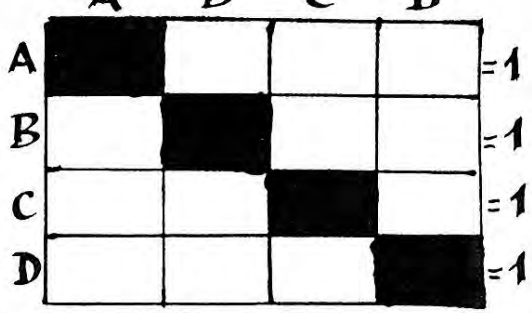
- A = sandstone with plane-parallel stratification
- B = sandstone with large-scale trough cross-stratification
- C = sandstone with current-ripple cross-lamination
- D = mudstone

$[n_{ij}]$



Row sums ↓

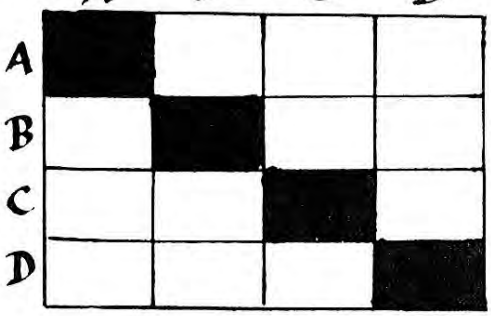
$[p_{ij}]$



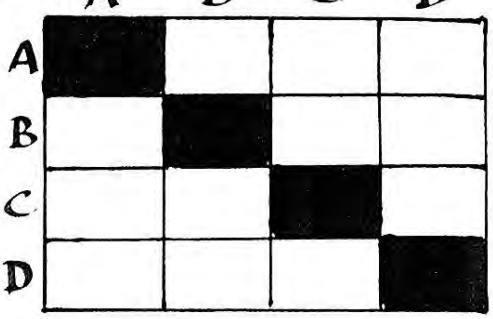
Column sums →

N =

$[e_{ij}]$

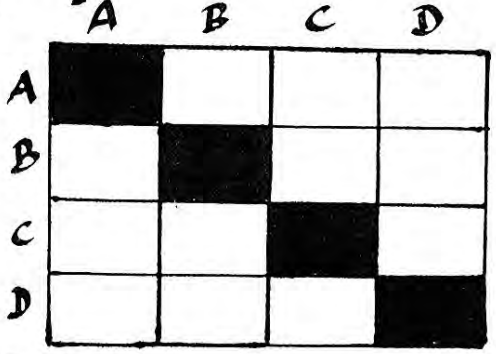


$[d_{ij}]$



BULK TEST:

$\left[\frac{(n_{ij} - n_i e_{ij})^2}{n_i e_{ij}}\right]$



row sums

=
=
=
=

DF = ☐

Critical value (for α = ☐): $\chi^2_{\frac{1}{2}\alpha}$ = ☐ χ^2 = ☐

■ Test for the significance of Markovian difference matrix

This test is for the bulk difference matrix, and the hypothesis tested is:

$\begin{cases} H_0: \text{The series of states is random} \\ H_1: \text{The series of states is "cyclic" (formed by Markov process, with some degree of "memory")} \end{cases}$

The test function is:

($\sum_{j}^{s} n_{ij}$: Row sum from the count matrix $[n_{ij}]$)

$$\chi^2 = \sum_{ij}^{s} \left[(n_{ij} - e_{ij} \textstyle\sum_{j}^{s} n_{ij})^2 / e_{ij} \sum_{j}^{s} n_{ij} \right] \quad \text{with } DF = (S-1)^2$$

This is a two-tail test, and the critical value $\chi^2_{½\alpha}$ is taken from TABLE 4. If $\chi^2 \geq \chi^2_{½\alpha}$, H_0 is rejected with confidence of $(1-\alpha)100\%$ and the process that formed the series is deemed to have had Markovian memory; the Markov-chain model derived is valid.

■ Test for the significance of the individual elements ("windows") of the difference matrix

The difference matrix may fail the previous "bulk" test, but this does not necessarily mean that all the transitions n_{ij} are non-preferential, due just to random chance. The following hypothesis may further be tested:

$\begin{cases} H_0: \text{The frequency of particular state transition is no higher than expected for a random series } (\delta_{ij} \leq 0). \\ H_1: \text{The frequency of this transition is significantly higher than expected for a random series } (\delta_{ij} > 0). \end{cases}$

This is a one-tail test, and the calculatory procedure is as follows. Select a significance level ($\alpha = 0.05$, but $\alpha = 0.10$ may be quite enough). For each element $d_{ij} > 0$ of the difference matrix calculate the probability of finding at least the observed number of transitions, $n = n_{ij}$, in $N = \sum_{j}^{s} n_{ij}$ occurrences of state i in a random series. (n_{ij} and $\sum n_{ij}$ are values from the transition count matrix, and symbols n and N are as used in the formula below):

$$P_{ij}(N,n) = \sum_{n=n_{ij}}^{n=N} C(N,n)\, e_{ij}^{\,n}\, q_{ij}^{\,N-n}$$

where: $C(N,n) = \dfrac{N!}{(N-n)!\,n!}$

e_{ij} = the corresponding element of the expected transition probability matrix;

$q_{ij} = 1 - e_{ij}$

EXAMPLE: Suppose we wish to test the transition from state A to state B, having $n_{AB}=4$, $p_{AB}=0.57$, $e_{AB}=0.22$, $d_{AB}=0.35$ and $\sum n_A = 7$ (data from Markovian matrices). For the test, the following values are needed:

$$n_{AB}=4 \qquad e_{AB}=0.22$$
$$\sum n_A = 7 \qquad q_{AB}=0.78$$

to calculate $P_{AB}(7,4)$. The calculation will be from $n=4$ to $n=7$.

n	$C(7,n)$	e_{ij}^{n}	q_{ij}^{7-n}	$P_{ij}(7,n)$
4	35	0.0023	0.475	0.038
5	21	0.0005	0.608	0.006
6	7	0.0001	0.780	0.0005
7	1	0.00002	1.000	0.00002
			Sum:	0.045

$$P_{AB}(7,4)=0.045 < \alpha=0.10$$
$$\text{and} < \alpha=0.05$$

The null hypothesis (H_0) is rejected with 95% confidence; the element d_{AB} in the difference matrix indicates a preferential transition from state A to state B.

Test for stationarity of Markov chain

H_0: The Markovian process that formed the series was stationary*.
H_1: The process was non-stationary.

* This means steady (not changing with time) and/or uniform (not changing laterally in space). To test steadiness, the series is split into $m \geq 2$ time segments. To test uniformity, $m \geq 2$ coeval data series are compared; e.g., from different localities. This is a two-tail test.

The test function is given below (critical value $\chi^2_{1/2\alpha}$ from TABLE 4):

$$\chi^2 = 2\sum_{k=1}^{m}\sum_{ij}^{s}\left[n_{ij(k)}\ln(p_{ij(k)}/p_{ij})\right] \quad \text{with } DF=(m-1)(s^2-s)$$

where: $n_{ij(k)}$ and $p_{ij(k)}$ are the transition counts and observed probabilities from the series interval k;
p_{ij} = the observed transition probabilities from the whole series or from all the coeval series jointly;
m = number of series intervals or number of coeval series;
s = number of states involved (total).

Note: The test requires that Markov matrices are calculated for the individual segments or coeval series, and also for all of them jointly.

Spectral Analysis

Spectral analysis is the partitioning of the variation in a time series into components according to the duration or length of the intervals within which the variation occurs. This is done by considering the time series to be the sum of many simpler time series that have the form of regular sinusoids of differing amplitudes, wavelengths, and starting points. These sinusoids are calculated so they are orthogonal, or statistically independent of one another. Since the sum of all of the sinusoids is equal to the original time series, the sum of the variation in all of the sinusoids must also be equal to the total variation in the series.

Spectral analysis is known, in its various guises, as harmonic analysis, Fourier analysis, and frequency analysis. The roots of the method are interwoven with those of musical theory, reflecting a common concern with vibratory motion. Kepler, in the seventeenth century, applied the harmonic relationships he found in arithmetic, geometry, and music to the positions of the planets and discovered the laws of planetary motion. However, it was Jean Baptiste Fourier (1768–1830) who provided proof that any continuous, single-valued function could be represented by a series of sinusoids, the relationship that now bears Fourier's name. Before examining spectral analysis in detail, a few terms must first be defined. Most of these terms were coined by electrical engineers and are used in the analysis of electrical signals. Although an electrical signal is an energy wave that changes with time, engineers are accustomed to seeing the signal "frozen" on an oscilloscope. Engineering discussions of Fourier analysis always assume that the signal is time-variant; however, the fact that engineers examine the signal as a spatial phenomenon on an oscilloscope indicates that time and space can be considered equivalent. This is, of course, mathematically correct, and the unconscious ease with which electrical engineers make the transition should assure us that the two are conceptually interchangeable as well.

Figure 4.48 shows a regularly repeating signal that can be described as a pure sinusoidal wave. The distance from a point on one wave form to the equivalent point on the next wave form is called the *wavelength*, λ. *Frequency*, f, is the reciprocal of wavelength, or $f = 1/\lambda$. It is the number of wave forms in a unit of length or time. In most engineering problems the signal is expressed in terms of frequency. In geologic problems, we usually are interested in wavelengths. The time required for a regularly repeating signal to exactly repeat itself is called its *period*. The term period is equivalent to wavelength, but the period is measured in units of time such as milliseconds rather than units of distance such as centimeters. With some phenomena, as for example sea waves, it is appropriate to use both wavelength and period. Wavelength in this instance refers to the spatial distance from crest to crest between adjacent wave forms. Period refers to the length of time required for two successive waves to pass a fixed reference point, such as the end of a pier. It is the duration in time between the appearance of one wave and the appearance of the next. (Use of the word period has given rise to the expression *periodic*, which means a signal that repeats at regular intervals.) Half of the height from trough to crest of a wave form is called its *amplitude*, A.

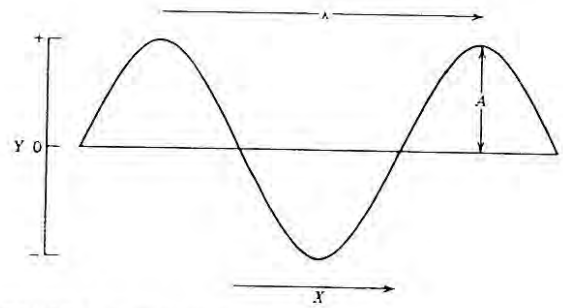


FIGURE 4.48 A regularly repeating sinusoidal wave; λ = wavelength, A = amplitude.

Figure 4.49 shows two identical sine waves that have been displaced with respect to one another. The amplitudes and wavelengths of the two signals are the same. The difference in Y_1 and Y_2 between the two signals, at a specified value of x_i, is attributable entirely to the offset between the two wave forms. The offset is described as a difference in the *phase angle*, ϕ, and its derivation can be shown with the help of Figure 4.50. A simple mechanical device that draws a sine wave consists of a disk of radius r revolving at constant speed around its center. A pen fastened to a push rod connected to the edge of the disk will trace a line on a sheet of paper being moved at constant speed under the end of the rod. The line is a sine wave having an amplitude $A = r$ and a wavelength that is a function of the rate of rotation of the disk and the velocity of the paper strip. The value of Y at any position is equal to $Y = A \sin \phi$.

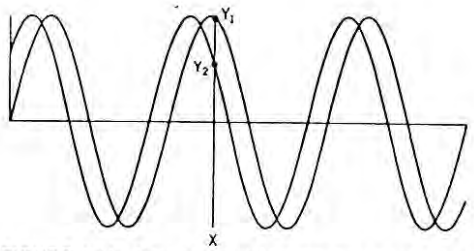


FIGURE 4.49 Two sinusoidal waves that are identical in form. Difference between Y_1 and Y_2 for a specified value X is attributable to difference in phase between the two wave forms.

The following short description of this method is from Davis (1986) "Statistics and Data Analysis in Geology", Wiley & Sons, New York, pp. 248-69.

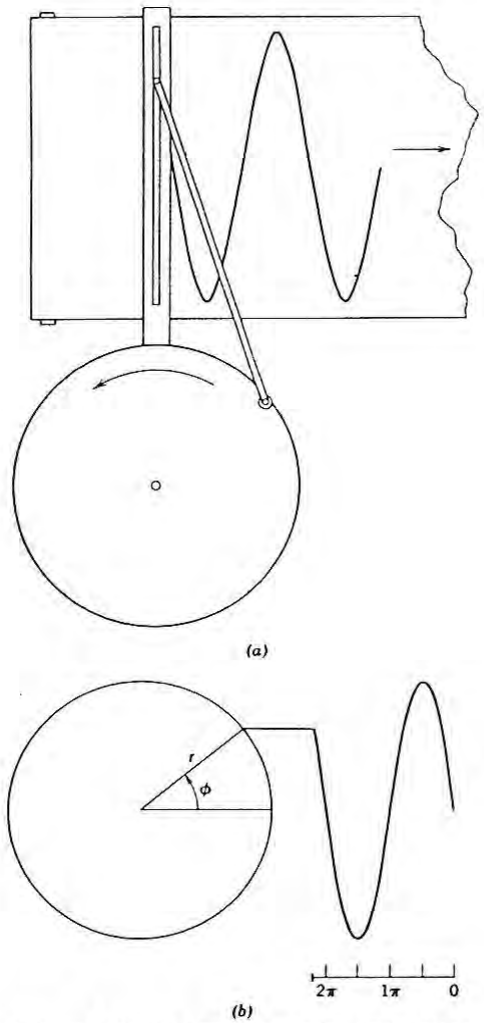


FIGURE 4.50 (*a*) Simple mechanical device to produce a sine wave by transforming rotary motion into linear motion. (*b*) Relation between angle and radius of disk to amplitude and phase angle of sine wave. Record shown from $\phi = 0$ to slightly more than 2π radians, or over one complete revolution of the disk.

To draw one complete wave form, the pen must revolve completely around the disk, moving through 360° or 2π radians. Suppose we start the device operating with the pen initially resting at an arbitrary location on the paper that we will call 0. The angle α between the pen, the center of the disk, and the center line on the paper is some value ϕ. These are shown in Figure 4.51. If we allow the device to operate for a distance x, down the record and then stop, the pen will be resting

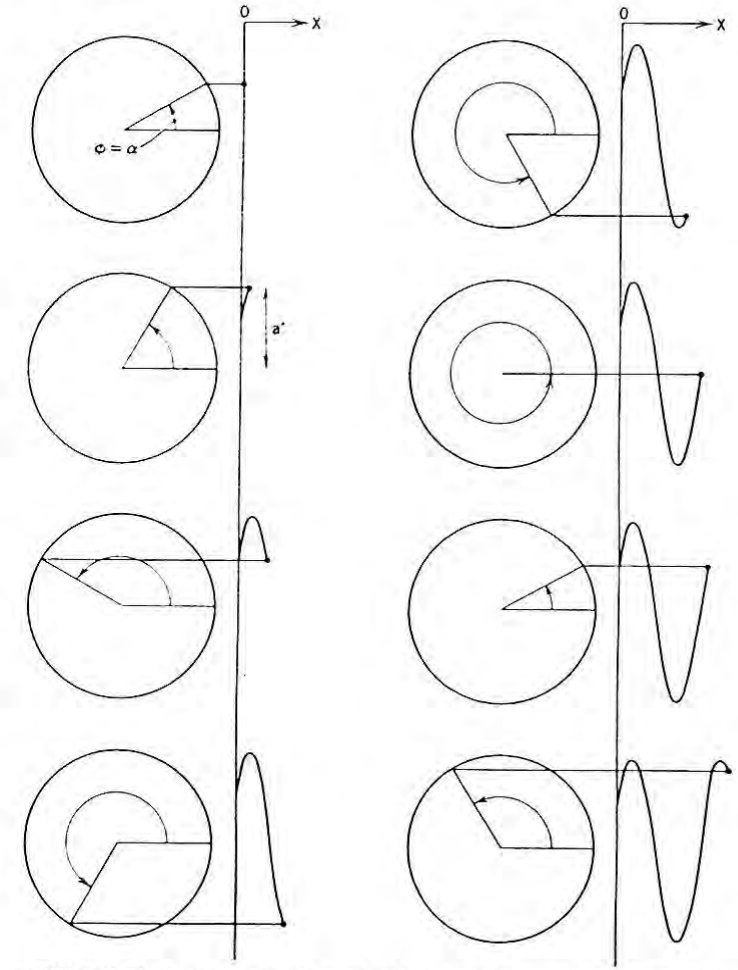


FIGURE 4.51 Progressive changes in phase angle from an initial value ϕ. At successive values of x, $Y_i = A \sin(2\pi x_i/X + \phi)$ and $\alpha_i = (2\pi x_i/X + \phi)$.

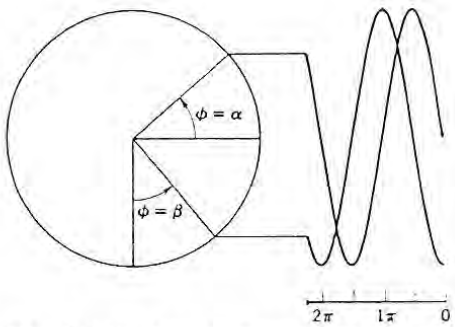


FIGURE 4.52 Sine and cosine waves related to rotation of disk. Phase angles for sine and cosine waves are α and β, respectively. Record is shown starting at $\phi = 0$ and continuing for slightly more than 2π radians, or one revolution of the disk.

at a point whose Y coordinate is equal to $A \sin(2\pi x,/X + \phi)$. The angle α, will be $(2\pi x,/X + \phi)$. In these expressions, X is the total length of the record. The important thing to note is that the initial angle $\alpha = \phi$ at the starting point appears as a constant in all subsequent determinations of the Y coordinate and the angle α. The constant ϕ is known as the *phase*, *phase angle*, or *phase constant*. The three parameters of amplitude, wavelength, and phase angle completely describe the wave form. Note that the phase angle of a sine wave is measured from the horizontal center line.

To construct a series of different wave forms, we may alter the amplitude, the wavelength, or the phase angle. We can also describe a cosine wave, identical in form to a sine wave, but having a phase angle β measured from a vertical line through the origin of the record. That is, a cosine wave is 90° or $\pi/2$ radians out of phase with a sine wave. If we add another pen to our mechanical signal generator, as in Figure 4.52, we can create both a cosine wave and a sine wave simultaneously. Obviously, the two cannot be distinguished except by the phase angle.

We can summarize the basic trigonometric relationships with the help of Figure 4.53, which shows a sinusoidal curve. This development is given in terms of the cosine, but an equivalent development, leading to the same relationship, can be based on the sine. Assume the wave makes one complete fluctuation in a time or distance X. Any point x within that interval can be expressed in radians by the conversion

$$\theta = \frac{2\pi x}{X} \tag{4.82}$$

The X coordinates now cover the interval from 0 to 2π radians. For convenience, we will assume that the amplitude of the curve is unity. Then, the equation of the

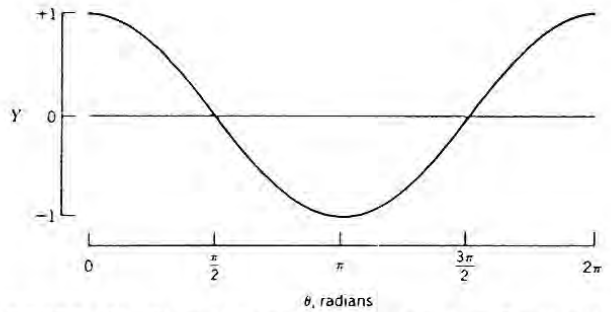


FIGURE 4.53 Sinusoidal curve expressed in radians. Total length of curve, X, is defined as equal to 2π.

curve is

$$Y = \cos\theta \tag{4.83}$$

The amplitude can be changed to any desired magnitude simply by multiplying by a coefficient A,

$$Y = A\cos\theta \tag{4.84}$$

The number of cycles that occur within the basic interval (the frequency) may be increased or decreased by multiplying θ by a coefficient k,

$$Y = \cos(k\theta) \tag{4.85}$$

As drawn, the crest of the wave form occurs at the origin, but it may be shifted to any position by subtracting a phase angle, ϕ.

$$Y = \cos(\theta - \phi) \tag{4.86}$$

Combining all of these modifications, we find that any curve of regular sinusoidal form may be written as

$$Y_k = A_k \cos(k\theta - \phi_k) \tag{4.87}$$

The subscript k represents the *harmonic number*, or the number of cycles per basic interval. Since in spectral analysis many different harmonics will be combined to reproduce the original time series, this subscript is necessary to designate a particular wave form.

Next, we may utilize the trigonometric identity for the difference between two angles,

$$\cos(R - S) = \cos S \cos R + \sin S \sin R$$

to rewrite eq. (4.87). The equation then becomes

$$Y_k = A_k \cos\phi_k \cos(k\theta) + A_k \sin\phi_k \sin(k\theta) \tag{4.88}$$

This expression can be simplified by defining two coefficients, $\alpha_k = A_k \cos \phi_k$ and $\beta_k = A_k \sin \phi_k$, giving

$$Y_k = \alpha_k \cos(k\theta) + \beta_k \sin(k\theta) \tag{4.89}$$

Any time series, regardless of how complex in form (except that it must be continuous or without breaks, and there can only be one value of Y for each value of x), can be represented as the sum of a series of cosine wave forms defined in this manner, as illustrated in Figure 4.54. This is an expression of the Fourier relationship,

$$Y = \sum_{k=0}^{\infty} \alpha_k \cos(k\theta) + \beta_k \sin(k\theta) \tag{4.90}$$

You will note that (4.90) is a linear equation (that is, all of the terms are added together) and resembles in form the polynomial regressions we considered earlier in this chapter. In the Fourier equation, the trigonometric terms $\cos(k\theta)$ and $\sin(k\theta)$ are equivalent to the power terms of a polynomial, such as X^2 and X^3. The α_k and β_k coefficients of these terms can be found by least squares. However, if we attempt to estimate the coefficients by matrix methods, we quickly face insurmountably large computational problems if many harmonics are to be determined.

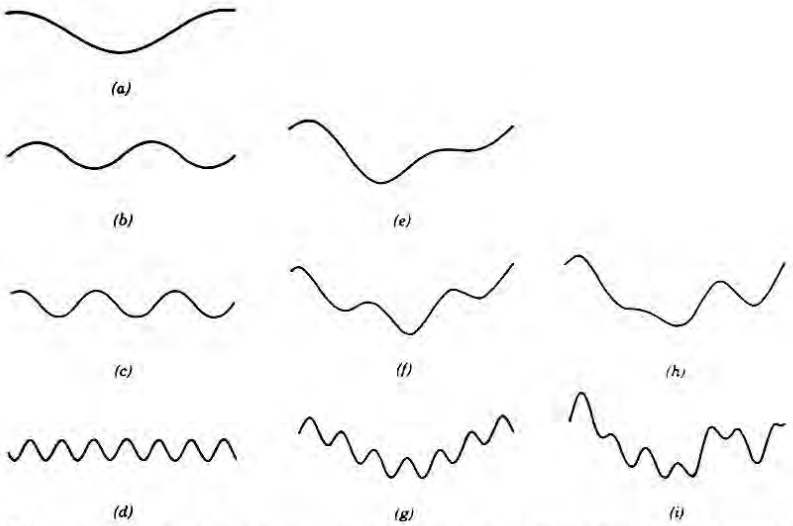


FIGURE 4.54 Effect of adding successive harmonics of a cosine wave. (*a*) Fundamental or first harmonic, $A = 0.5$, $k = 1$, $\phi = 0°$. (*b*) Second harmonic, $A = 0.3$, $k = 2$, $\phi = 0°$. (*c*) Third harmonic, $A = 0.3$, $k = 3$, $\phi = 45°$. (*d*) Seventh harmonic, $A = 0.25$, $k = 7$, $\phi = 180°$. (*e*) First plus second harmonic, (*a*) + (*b*). (*f*) First plus third harmonic, (*a*) + (*c*). (*g*) First plus seventh harmonic, (*a*) + (*d*). (*h*) First plus second and third harmonics, (*a*) + (*b*) + (*c*). (*i*) First plus second, third and seventh harmonics, (*a*) + (*b*) + (*c*) + (*d*).

Harmonic Analysis

If a time series represents a truly periodic phenomenon, we may use a technique called harmonic analysis to decompose the series into its constituent parts. Naturally occurring time series that are periodic include tidal fluctuations, seasonal variation in temperatures, and varve sequences. They also include records generated by instruments such as X-ray diffractometers, although the Fourier transformation of such data is a specialized topic. The time series must be sampled at discrete points which are equally spaced a distance Δ apart. If there are n of these points, one of which is j, the α_k and β_k coefficients of the Fourier equation can be found by the computational formulas:

$$\beta_k = \frac{2}{k} \sum_{j=0}^{n-1} Y_j \sin\left(\frac{2\pi jk}{n}\right) \tag{4.91}$$

$$\alpha_k = \frac{2}{k} \sum_{j=0}^{n-1} Y_j \cos\left(\frac{2\pi jk}{n}\right) \tag{4.92}$$

The equations are not so formidable as they may appear at first glance; the expressions in parentheses simply convert the position of the jth point into radians, defining the total length of the time series as 2π radians. Because of trigonometric relationships, the coefficient β_0 is always zero, and α_0 simplifies to

$$\alpha_0 = \frac{1}{n} \sum_{j=0}^{n-1} Y_j \tag{4.93}$$

which is just the mean of the time series. The subscripts range from 0 at the beginning of the time series to $n - 1$ at the end, rather than from 1 to n, in order to correctly convert the position of the first observation to zero radians.

Once we have determined the α_k and β_k coefficients for the kth harmonic, we can determine the amplitude of the wave form by

$$A_k = \sqrt{\alpha_k^2 + \beta_k^2} \tag{4.94}$$

The phase angle of the kth harmonic is

$$\phi_k = \tan^{-1}\left(\frac{\beta_k}{\alpha_k}\right) \tag{4.95}$$

The expression $\tan^{-1}$ means "find the angle whose tangent is given."

If a regular sinusoid is uniformly sampled at discrete points, the variance of these samples is related to the amplitude of the wave form. In the limit, the variance is simply half the square of the amplitude, or

$$s_k^2 = \frac{A_k^2}{2} = \frac{\alpha_k^2 + \beta_k^2}{2} \tag{4.96}$$

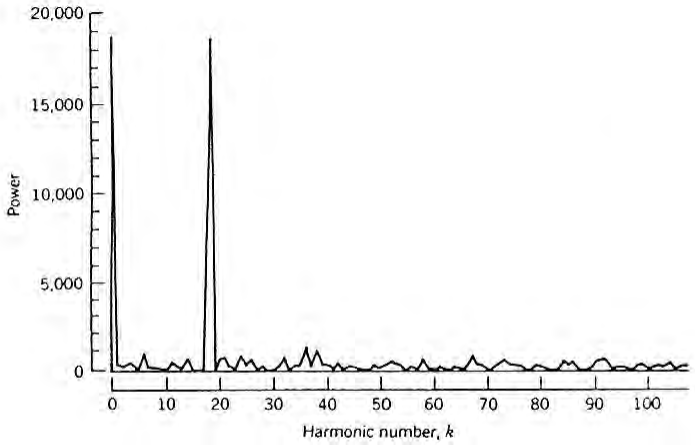


FIGURE 4.55 Periodogram of monthly flow of Cave Creek, Kentucky. Spectral values indicate that flow has a single periodic annual cycle. Power is given as the square of the measurement units, which are hundredths of inches.

Since the Fourier theorem holds that a time series can be regarded as the sum of many sinusoidal functions, or harmonics, the variance of the time series must be composed of the sum of the variances of these same harmonics. We can therefore express the variance of the kth harmonic as a proportion of the total variance of the time series from which it has been derived. If we compute a large series of harmonics, the variances of the successive harmonics can be plotted in a *periodogram*, or, as it is sometimes called, a *discrete* or *line power spectrum*. This is simply a plot of power or variance versus harmonic number, k, and traditionally is shown in the form given in Figure 4.55. (The terms *power* and *variance* are synonymous; the former arose in electrical engineering and is now widely used in all fields of engineering, signal analysis, and pattern recognition.) The original time series is shown in Figure 4.56. The time series represents the monthly flow of water in Cave Creek in Kentucky; data are given in Table 4.29, taken from Haan (1977). The periodogram indicates that the variation in monthly flow has an annual (12-month) period, plus a random component.

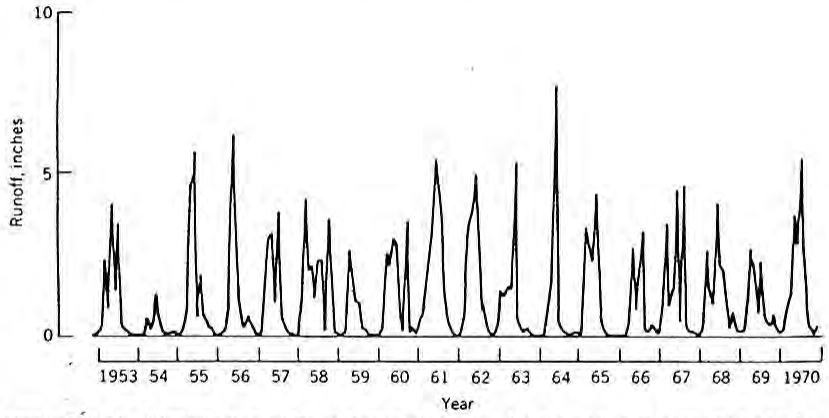


FIGURE 4.56 Monthly flow of Cave Creek, Kentucky, for the period October 1952 to September 1970.

TABLE 4.29 Monthly Runoff of Cave Creek, Kentucky, Given in Hundredths of Inches

Water Year	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.
1953	2	5	19	240	86	416	147	354	31	18	7	1
1954	0	2	4	54	22	40	139	35	8	7	6	14
1955	2	4	30	73	463	579	59	197	55	24	28	3
1956	4	6	13	59	637	469	192	28	32	64	38	8
1957	7	10	172	308	325	103	392	68	24	6	5	2
1958	3	106	432	200	221	117	235	236	19	369	170	12
1959	6	9	17	270	195	112	102	24	24	5	4	2
1960	3	36	269	219	313	291	68	19	364	138	14	30
1961	12	52	79	204	295	532	476	414	159	48	18	4
1962	2	6	76	346	401	508	330	79	96	30	8	7
1963	39	141	124	150	146	548	52	25	14	29	11	3
1964	1	4	3	87	173	788	45	21	11	8	2	16
1965	15	7	347	276	230	449	146	31	8	5	1	2
1966	4	2	2	48	281	79	202	332	25	14	41	11
1967	7	119	357	97	161	466	50	476	33	14	15	7
1968	9	38	271	135	98	425	238	199	91	29	75	16
1969	14	22	112	278	216	73	237	74	40	27	66	17
1970	7	25	91	130	389	291	568	206	38	14	6	27

Data are collected during a "water year" beginning in October of the preceding year and ending in September in order to avoid breaking the sequence in the middle of the winter. From Haan (1977).

The spectrum computed in the preceding manner is sometimes called the *raw spectrum*, and is an estimate of the true or population spectrum. These estimates have associated with them very large standard errors, which cannot be reduced by increasing n for reasons that will be explained in the next section. Some of this error may represent *aliasing*, or the incorporation of irresolvable high frequencies into lower frequencies. These high frequencies, whose wavelengths are less than twice the spacing between sample points, cannot be detected. The highest frequency that can be estimated is the *Nyquist frequency*, whose wavelength is exactly 2Δ, where Δ is the distance between successive observations. The reason why the variances of frequencies beyond the Nyquist limit are confounded with those of lower frequencies can be seen readily in Figure 4.57.

Nevertheless, it is possible to statistically test for the existence of the dominant periodic component of the periodogram. The test, devised by Fisher, calculates the probability that a spectral value s_i^2 will exceed the value σ_i^2 of a time series

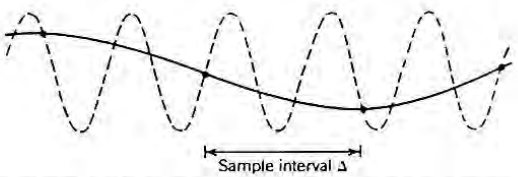


FIGURE 4.57 High-frequency sinusoidal wave (dashed line) sampled at discrete points yields apparent lower frequency wave (heavy line).

composed of independent random points. The test involves calculation of the ratio

$$\hat{g} = \frac{s^2_{max}}{2s^2} \tag{4.97}$$

where s^2_{max} is the largest peak in the periodogram and s^2 is the variance of the entire time series. The critical value of g for a specified probability p is given by

$$g \approx 1 - e^{\frac{\ln p - \ln m}{m - 1}} \tag{4.98}$$

where $m = n/2$ if the time series contains an even number of observations and $m = (n-1)/2$ if n is odd. If the test statistic $\hat{g}$ exceeds the critical value g, the periodic component may be presumed to exist. If the test value does not exceed the critical value, the observed spectral peak s^2_{max} could have arisen simply by random chance.

The Continuous Spectrum

Harmonic analysis and the construction of the periodogram or line spectrum is appropriate if the time series being investigated is truly periodic. There are few naturally occurring phenomena that possess true periodicity, except for those related to astronomical cycles such as monthly tides or seasonal changes. Most geologic time series are not periodic, but are stochastic. A sequence is *stochastic* if it can be characterized only by its statistical properties; this is in contrast to a deterministic sequence whose state can be predicted exactly from its coefficients.

Even though a time series may not contain truly periodic components, the methods of spectral analysis can provide valuable insights into the behavior of the process that has generated the sequence. Most time series can be regarded as continuous sequences, even though they ordinarily are sampled at discrete points. It is possible to calculate a *continuous spectrum* or *spectral density function* for such sequences, in which the variance of the time series is apportioned among a set of *frequency bands*. The continuous spectrum has the form of a continuous plot of variance versus frequency, and is analogous to a continuous probability distribution, with variances proportional to the areas under the spectral curve between limiting frequencies. The total area under the spectrum is equal to the total variance of the time series from which it is calculated. In contrast, the line spectrum of a periodic time series shows the variances attributable to defined individual frequencies.

The specific sequence of observations we wish to analyze can be considered a random sample from a large, perhaps infinite, set of such time series that could be produced by the process under study. The complete set of time series is called an *ensemble*, and is the population from which our particular sample is taken.

A time series is said to be *stationary* (or time-invariant or nonevolutionary) if its properties do not change with time. A spatial series with the same characteristics is said to be *homogeneous*. If a time series is divided into small segments and the means of all these segments tend to be the same (and the same as the mean of the entire time series), it is referred to as being *first-order stationary* or stationary in the mean. If in addition the autocovariance changes only with lag and not with position along the time series, the series is *second-order stationary*. This is also called *weak stationarity*, or stationarity in the wider sense. If all higher moments are dependent only on lag and not on position, the series is *strongly stationary*, or possesses stationarity in the strict sense.

If a time series is not only strongly stationary, but all statistics are invariant from time series to time series within the ensemble, the ensemble is *ergodic*. Many statistical tests of time series assume ergodicity, just as univariate statistical tests may assume uniformity of variance. We may check for ergodicity if we have several time series that are all realizations of the same stochastic process. Commonly, however, we have only one time series available, and then no check is possible. We can test for stationarity in the mean and variance of a single time series by regression, or by dividing the series into segments and testing to see if the statistics for the segments are the same. If they are, the series is *self-stationary* and ergodicity may be assumed. If the series is not self-stationary, the ensemble from which it is derived cannot be ergodic.

Sometimes a time series can be made stationary by *levelling*, or subtracting a linear trend from the observations. That is, a linear regression of Y_j on j is computed, using least squares as described earlier. Then, a new time series is defined as the deviations,

$$Y'_j = Y_j - \hat{Y}_j \tag{4.99}$$

where $\hat{Y}_j$ are the predicted values given by the regression equation. If the original series was characterized by a slow change in its average value, the new series will have a stationary mean of zero.

For a stationary, stochastic time series that is continuous and sampled at discrete, equally spaced points, the continuous variance (or power) spectrum may be calculated by either of two methods. The newer, and more widely used approach involves calculating many values of the line spectrum by the Fast Fourier Transform (FFT) computer algorithm. These spectral values are then averaged across frequencies to produce a smoothed estimate of the continuous spectrum. The Fast Fourier Transform, as its name implies, is extremely rapid and requires only $n \cdot \log_2 n$ arithmetic operations rather than n^2 operations as do alternative procedures. This approach requires that the Fourier relationship be expressed in complex form,

as

$$s_k^2 = \frac{1}{n} \sum_{j=-n/2}^{n/2} Y_j e^{i(2\pi jk/n)} \tag{4.100}$$

where i is the imaginary number $\sqrt{-1}$. Developing the FFT algorithm using complex arithmetic requires mathematical and computational skills beyond those assumed in this book. Therefore, we will not pursue the topic further, even though this approach is now one of the most widely used in spectral analysis. An excellent introduction to the Fast Fourier Transform, including a discussion of complex arithmetic, is given in Rayner (1971), who also describes applications of the FFT in geography. The first landmark article describing the Fast Fourier Transform algorithm was published by Gentleman and Sande (1966), based upon the mathematics previously introduced by Cooley and Tukey (1965). Modern texts on time series analysis, such as Bloomfield (1976) and Bendat and Piersol (1971), treat the subject extensively.

A somewhat older procedure for calculating the continuous spectrum involves finding the Fourier transform of the autocorrelation function of a time series. Developed originally by Bartlett (1948), this approach achieves the same results as the newer FFT method, and is easier to understand even though not as computationally efficient. It is still widely used, especially when the autocorrelation function is also of interest and when the time series is not extremely long. Jenkins and Watts (1968), for example, discuss this approach at length, and Yevjevich (1972) applies it to problems in hydrology.

The Fourier transform of the autocorrelation function is the power spectrum; for a discretely sampled time series, Jenkins and Watts (1968) give the transform in the following form:

$$s_f^2 = \frac{A_f^2}{2} = \sum_{l=-(L-1)}^{L-1} r_l e^{i2\pi fl} \tag{4.101}$$

Here, s_f^2 is the variance or power in the frequency band centered at frequency f, r_l is the autocorrelation at lag l, the maximum lag in the autocorrelogram is L, and i is the imaginary number $\sqrt{-1}$.

Note that the equation is very similar to the complex Fourier transform given in eq. (4.100). However, eq. (4.101) may be greatly simplified. Although the autocorrelation function can be calculated for negative as well as positive lags, it is symmetrical about lag zero; that is, $r_l = r_{-l}$. Therefore, the calculations need only to be performed over the range of l from zero to $(L - 1)$, and then doubled. Also, because our data consist of a finite set of real numbers that are evenly spaced, we may use a cosine representation rather than complex notation. The computational form of eq. (4.101) is

$$s_f^2 = \frac{A_f^2}{2} = 2\left(\sum_{l=1}^{L-1} r_l \cos 2\pi fl\right) \tag{4.102}$$

It is necessary to work in frequencies rather than harmonics, because specific lags of the autocorrelation function do not transform into simple multiples of one another in the frequency spectrum. The relationship between lag and frequency is given by

$$f_j = \frac{j}{2\Delta L} \tag{4.103}$$

where Δ is the spacing between successive observations in the time series, and j represents both the jth lag and the jth spectral band. The frequency of the jth band is f_j and is measured in cycles per unit, where the units are the same as those of Δ, the spacing between data points. If Δ is measured in millimeters, f_j is given in cycles per millimeter.

Use of eq. (4.102) yields what is called the *raw spectrum*. In general, it will not be satisfactory because of the high standard error in each of the spectral bands. This cannot be overcome in the usual manner, by making the time series longer so that n (the total number of observations) is greater. It is true that as n increases for a fixed lag l, there is more information in the form of more crossproducts $(x_j - \bar{X})(x_{j-l} - \bar{X})$, and the autocorrelation r_l will have a smaller standard error. However, this is not true for s_f^2. Jenkins and Watts (1968) show that the information contained in each estimate s_f^2 of the spectrum is spread over a band of frequencies whose width is $\pm l/n$ about f. As n increases, the total information contained in the variance spectrum is distributed over an increasing number of bands of decreasing width. As a result, increasing n makes it possible to estimate the variance in narrower frequency bands, yielding a less biased estimate of the true spectrum. However, the standard error of these narrower bands is not improved.

Bartlett (1948) first introduced a method for reducing the standard error in spectral estimates by a process variously referred to as *windowing*, *filtering*, *smoothing*, or *spatial averaging*. The autocorrelation function may be weighted by a *lag window* or filter, and the weighted autocorrelogram transformed into the spectral domain. The weights emphasize the shorter lags, which are based on more observations than are longer lags. Alternatively, the raw spectrum may be calculated and then smoothed by weighting and averaging together adjacent spectral values. The set of weights is called a *spectral window* or filter, and is the Fourier transform of the lag window. Both methods produce equivalent results, which may be thought of as a blurring or smoothing of the variances over adjacent frequencies.

Many windows or filters have been suggested by various authors; the design of windows that have superior properties is a fine art among time series analysts. One commonly used window is the Tukey-Hanning filter, which in the spectral domain is

$$s_f^2 = \frac{1}{4} s_{f-1}^2 + \frac{1}{2} s_f^2 + \frac{1}{4} s_{f+1}^2 \tag{4.104}$$

The Fourier transform of the Tukey-Hanning filter into the lag domain is

$$W_l = (1 + \cos \pi l/L)/2 \tag{4.105}$$

This defines a set of weights for the autocorrelation function which drops smoothly

from $W_0 = 1.0$ at lag zero to $W_L = 0.0$ at the maximum lag, L. Each autocorrelation coefficient is multiplied by its corresponding weight before the Fourier transform is made. Equation (4.102) becomes

$$s_f^2 = 2\left(1 + s\sum_{l=1}^{L-1} W_l r_l \cos 2\pi f l\right) \tag{4.106}$$

Other windows have been proposed, usually either in the lag domain or in the frequency domain because their application is easier in one form or the other. The widely used Parzen window has the form

$$W_l = 1 - \frac{6l^2}{L^2}\left(1 - \frac{l}{L}\right) \tag{4.107}$$

for lags between zero and $L/2$, and

$$W_l = 2\left(1 - \frac{l}{L}\right)^3 \tag{4.108}$$

for lags from $L/2$ to L.

A practical strategy for investigating the spectral characteristics of a stochastic time series includes the following steps.

1. If necessary, de-trend or level the data by computing the coefficients of the regression of Y_j on j, then finding the residuals $Y_j' = Y_j - \hat{Y}_j$.
2. Compute the autocorrelation of the original series (if stationary) or of the series of residuals Y', up to a maximum lag L, which is no more than $n/3$. Some authors recommend more severe limits on the maximum lag to be calculated, such as $n/5$, $n/6$, or even $n/10$.
3. Compute the raw spectral density, s_f^2, using the discrete Fourier transform given in eq. (4.102).
4. Smooth the raw spectrum by using the Tukey-Hanning window (eq. 4.104) as a moving average to produce the smoothed estimate of the variance spectrum, s_f^2.

Obviously, there are many decisions that must be made in spectral analysis that may have a significant effect on the final results. These include choice of n, the length of the sequence to be analyzed (if this is under the analyst's control); L, the maximum lag; and the window to be used for smoothing. The interactions between these variables, and their effect on both the resolution and significance of the smoothed spectrum, are discussed at length by Jenkins and Watts (1968).

In performing Fourier analysis, we have transformed our data from one domain to another. We began with observations in the form of values Y_j at points in space or time, x_j. The succession of points forms a signal or wave form, defined by X and Y coordinates. The data, defined in this manner, are said to be in the time or spatial domain, depending upon whether X denotes points in time or distance. By determining the component frequencies in the signal, we have transformed the data to the frequency domain. A physical analogy can be drawn with the effect of a glass prism on sunlight, as shown in Figure 4.58. A beam of white light can be regarded as a complex wave form changing with time, and composed of many colors (or wavelengths) of light. A prism acts as a frequency analyzer, and separates the beam into its components, which appear as a rainbow display. Each colored band is separated from its neighbor by an amount proportional to the difference in their wavelengths or frequencies, and the intensity of each band is proportional to the contribution of that particular wavelength to the total intensity of the original beam. Examining the spectrum of a light source may tell us a great many things: the composition of the source, its temperature, nature of the material through which the light passed, and so forth. In a similar fashion, examining the power spectrum of a data sequence may tell us a great deal about its nature and origin, information that may not be apparent in any other way.

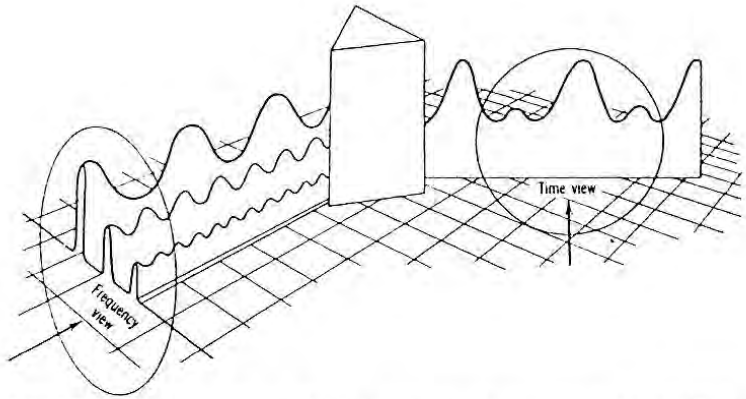


FIGURE 4.58 A prism acts as a frequency analyzer, transforming white light (time or spatial domain) into its constituent spectrum of colors (frequency domain). Courtesy Hewlett-Packard Corp.

Filters

A *filter* is a mathematical operator that changes a time series into another time series having some desired form. Such operators are called "filters" because originally they were electronic filter circuits, consisting of networks of resistors and capacitors, used to selectively suppress or enhance specific frequencies in electronic signals. Now, the same function is performed on digital signals by mathematical filters implemented in a computer. Much of the basis for filtering stems from the work of Norbert Wiener on statistical communication theory, and many practical applications occur in seismic processing and remote sensing. Because of this background, most of the terminology of filtering is derived from the fields of electrical engineering and physics, rather than classical statistics.

The most extensive applications of filtering techniques are associated with the

processing of reflection seismic data. Reflection seismology can be regarded as the creation of an input signal (the seismic energy produced by a dynamite shot or a vibrator) which is then "filtered" by absorption of energy as it travels through the Earth. Higher frequencies tend to be preferentially absorbed, so the nature of the seismic pulse changes as it travels. In addition, lower frequency waves travel through rocks more rapidly, so the initial sharp input becomes blurred and attenuated. The final output, which is detected by an array of geophones at the surface, consists of reflections returning from successively deeper horizons. Unfortunately, after their travels, the reflections are severely distorted and difficult to interpret.

The job of the geophysicist is to remove, as much as possible, the deleterious effects of the physical filtering to which the seismic signal has been subjected. This is done by processing the seismic recording with digital filters that remove unwanted frequencies, sharpen diffuse reflections, and subdue the noise in the record until the desired but attenuated signal can be seen.

Because so much of the distortion of seismic signals is frequency-related, geophysicists are especially interested in filters that preferentially pass or remove specific frequencies. Seismic filtering often is done in the frequency domain, by taking the Fourier transform of the seismic record, removing unwanted parts of the spectrum, and then retransforming the filtered spectrum back into the real domain. However, exactly the same process can be performed by transforming the Fourier domain filter into the real world and convolving its transform with the seismic record. Because these filters are designed to pass certain frequencies while suppressing others, they are referred to as high-pass, low-pass, or band-pass filters.

Although in theory it is possible to produce perfect filters that will pass only signals of specified frequencies and no others, such filters usually are not realizable. Typically, these theoretically perfect filters require an infinite number of terms. If they are truncated to manageable size, the filters themselves may introduce spurious features in the output. The design of seismic filters thus is an art of the practical, seeking filters that are sufficiently short so that their use is economically feasible, and that produce good approximations of the desired output with minimum undesirable side effects. An adequate discussion of frequency-specific filters would require much more space than is available for the topic in this text, so interested readers must be referred to the many specialized books on the subject. We will turn instead to conceptually more simple types of filters that also see wide use in many areas of time series analysis.

A time series to be filtered is referred to as the *input*, and the altered time series that emerges from the filter is called the *output*. The filtering operation convolves the input with the filter to produce the output:

$$[C] = [B] * [f] \tag{4.109}$$

If there is some form we want the output to assume, say $[D]$, a filter can be designed so that the actual output $[C]$ is as close to the desired output $[D]$ as possible. To do this, we compute elements of the filter that will minimize the sum of the squared differences between $[C]$ and $[D]$. The mathematical process is very similar to that used to determine the coefficients of a linear regression equation. Robinson and Treitel (1980) have an extensive and exceptionally lucid discussion of the process, which is called *least squares filter design*.

As a simple example, suppose the input $[B]$ is a vector having two elements, $[b_0 \quad b_1]$. The filter also contains two elements, $[f_0 \quad f_1]$. Since the output from the filter is the convolution of the input with the filter, it will consist of three elements.

$$[C] = [B] * [f]$$

$$[C] = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \begin{bmatrix} f_0 & f_1 \end{bmatrix} \quad \begin{bmatrix} f_0 b_0 & f_1 b_0 \\ f_0 b_1 & f_1 b_1 \end{bmatrix} \tag{4.110}$$

$$[C] = [f_0 b_0 \quad f_0 b_1 + f_1 b_0 \quad f_1 b_1]$$

We want the filter output to be equal to some specified output $[D]$, so the two can be equated, defining a set of simultaneous equations.

$$\begin{aligned} f_0 b_0 + 0 &= d_0 \\ f_0 b_1 + f_1 b_1 &= d_1 \\ 0 + f_1 b_1 &= d_2 \end{aligned} \tag{4.111}$$

In matrix form, this can be written as

$$\begin{bmatrix} b_0 & 0 \\ b_1 & b_0 \\ 0 & b_1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \end{bmatrix} \tag{4.112}$$

We will denote the left-hand matrix as $[\beta]$.

The matrix equation is overdetermined; that is, there are three equations but only two unknowns. However, if both sides of an equation are multiplied by the same quantity, the equation is unchanged. So, multiplying both sides by the transpose of $[\beta]$:

$$[\beta]' [\beta] [f] = [\beta]' [D]$$

or

$$\begin{bmatrix} b_0 & b_1 & 0 \\ 0 & b_0 & b_1 \end{bmatrix} \begin{bmatrix} b_0 & 0 \\ b_1 & b_0 \\ 0 & b_1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} b_0 & b_1 & 0 \\ 0 & b_0 & b_1 \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \\ d_2 \end{bmatrix}$$

which becomes

$$\begin{bmatrix} (b_0^2 + b_1^2) & b_0 b_1 \\ b_0 b_1 & (b_0^2 + b_1^2) \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} (b_0 d_0 + b_1 d_1) \\ (b_0 d_1 + b_1 d_2) \end{bmatrix} \tag{4.113}$$

Geophysicists refer to the left-hand matrix as the autocorrelation matrix of the input, and the right-hand matrix as the cross-correlation between the input and output. Actually, the elements of these matrices are not correlation coefficients, but are uncorrected sums of squares and cross-products. However, the denominators

of any correlation terms would be identical, and so would cancel out; the solution to the equation set would be the same using either correlations or sums of squares and cross-products. The solution is found by inverting the left-hand matrix and postmultiplying by the right-hand vector. If we call the left-hand matrix $[R]$ and the right-hand vector $[X]$, this is

$$[f] = [R]^{-1}[X]$$

One common use of filtering is to search through a time series for occurrences of a specific wave form, or its digital representation, which is a specific pattern of numbers. The analyst may wish to turn the occurrences of this wave form into something more conspicuous, such as a sharp peak or spike in the output record.

The input wave form $[B]$ will consist of [2 1] and the desired output will consist of $[D]$ = [3 0 0]. These are shown as wave forms in Figure 4.59. Substituting into eq. (4.113) yields

$$\begin{bmatrix} 2 & 0 \\ 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}$$

Premultiplying both sides by $[\beta]'$:

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$$

Inverting the left-hand matrix yields

$$\begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}^{-1} = \begin{bmatrix} .2381 & -.0952 \\ -.0952 & .2381 \end{bmatrix}$$

The desired filter coefficients are

$$\begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} .2381 & -.9852 \\ -.0952 & .2381 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.4286 \\ -.5712 \end{bmatrix}$$

The graphic representation of the filter is shown in Figure 4.59. We can determine its performance by using it to operate on the original input. This is done by inserting the coefficients into the definitional equation of the filtering operation given in eq. (4.109):

$$[C] = [2 \quad 1] * [1.4286 \quad -.5712]$$

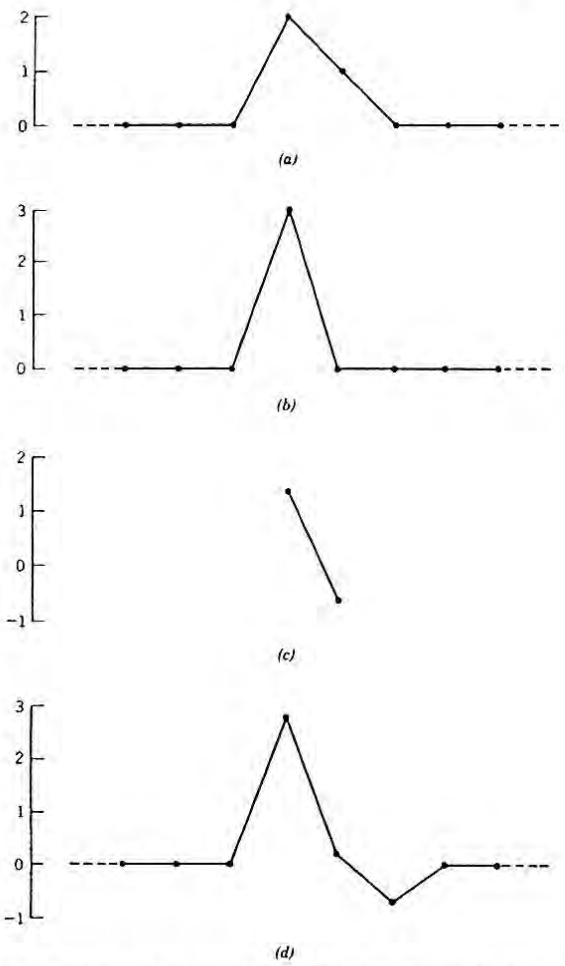


FIGURE 4.59 (*a*) Arbitrary input with values [2 1]. (*b*) Desired form of output is [3 0 0]. (*c*) Least-squares filter [1.43 −.57]. (*d*) Output from least-squares filter [2.9 .3 −.6].

The convolution is

$$[C] = \begin{bmatrix} 2 \\ 1 \end{bmatrix} * \begin{bmatrix} 1.4286 & -.5712 \\ 2.8572 & -1.1424 \\ 1.4286 & -.5712 \end{bmatrix}$$

$$= [2.8572 \quad (1.4286 - 1.1424) \quad -.5712]$$

$$= [2.8572 \quad .2862 \quad -.5712]$$

The filter output is also shown on Figure 4.59; obviously, we were not successful in creating a pure [3 0 0] spike, but the output is a fairly good approximation. In fact, it is the best approximation possible for this particular input and a filter of only two terms. We can measure just how good the filter is by comparing its output with the desired output. This is done by finding the *error*, defined as the sum of the squared differences between [*C*] and [*D*]:

$$\epsilon = (3 - 2.8572)^2 + (0 - .2862)^2 + (0 + .5712)^2$$

$$= .1428$$

Forms of output other than a spike can be specified, and in certain instances will yield a more efficient filter. Common filter designs are referred to as the boxcar (a square wave), Gaussian, and sawtooth.

Smoothing and Time-Trend Analysis

Perhaps the most familiar types of filters used in geology are those designed to reduce the variance in a time series. These are arbitrary filters whose general action is to "smooth" a data sequence; the output from the filter is a subdued approximation of the input. The justification for such a filtering process is that the time series consists of two components—a "long-term" signal or meaningful part, and superimposed random noise. By its nature, such noise must be "short-term." As the signal tends to be the same from point to point and the noise does not, an average of several adjacent points will tend to converge on the value of the signal alone.

One method of smoothing consists of approximating short segments of the original sequence with smooth lines or curves. These curves can be fitted by least squares, using regression techniques discussed earlier. Orthogonal polynomials would seem to be especially convenient for this purpose, because they require only the multiplication of a series of terms with the observations in the segment being fitted, followed by summation and a division, to yield a coefficient of the fitted curve. Once the coefficients have been determined, the equation can be evaluated and the original observations Y_i replaced with the predicted values $\hat{Y}_i$. As long as the order of the fitted curve is less than the number of points in the segment, the curve will be an approximation that is less variable than the original observations, and hence smoother.

Ordinarily, a sequence would not be smoothed by a series of nonoverlapping segments, since these would tend to have abrupt discontinuities at their joins. Instead, the smoothing operation consists of fitting an approximation to a small segment, and determining the predicted value $\hat{Y}_i$ corresponding to the middle observation in the segment. The next segment is chosen so that it overlaps all but one of the observations in the first, and the process is repeated. Eventually, the original data set is replaced by a smoothed series derived from the curves fitted to these overlapping segments.

Since a series of coefficients must be determined and evaluated for each overlapping segment, and there are almost as many segments in a time series as there are original observations, this would be a tedious process even using orthogonal polynomials. However, the coefficients themselves are not needed; all we require are the estimates $\hat{Y}_i$ at the center points of the segments. With this constraint, it is possible to rewrite the orthogonal polynomial equations to yield an alternative set of orthogonal terms that will produce the estimated values of the center points directly. Details of the derivation are given by Savitzky and Golay (1964).

Because we are interested only in estimating the central point in a short sequence of points, n must always be an odd number. Also, it happens that either a quadratic or cubic curve will have the same value at the central point, so one set of weights will do for either. The same is true for the fourth- and fifth-order curves. Of course, if we fit a linear function to the data, the value at the central point will simply be the mean of the points in the segment. This is equivalent to saying that all the terms of a linear orthogonal polynomial equation are equal to 1.

Table 4.30 reflects these considerations; it gives the orthogonal polynomial weighting functions ω for odd numbers of points in a sequence for the second- or third-order and the fourth- or fifth-order curves.

To perform smoothing using orthogonal polynomials, we simply convolve the data sequence with the set of polynomial terms, then divide by the sum of the terms. That is,

$$[C] = [B] * [\omega]/\Sigma\omega$$

Since the set of terms is symmetrical about the center term, the result is the same if we compute a moving average, which is simply a moving cross-product between the set of terms and the time series, divided by $\Sigma\omega$. This process is perhaps more familiar in geology under the name *time-trend analysis*. Among the equations used for this purpose is Shepard's five-term filter, usually written as a moving average of the form

$$\hat{Y}_i = \frac{1}{35}[17Y_i + 12(Y_{i+1} + Y_{i-1}) - 3(Y_{i+2} + Y_{i-2})]$$

If you compare the weights in Shepard's moving-average equation with the orthogonal polynomial terms given in Table 4.30, you will see that it is simply a quadratic equation fitted by least squares to five points. Other smoothing equations used in time-trend analysis are also either identical or very similar to orthogonal polynomials. These smoothing functions are described in the classic work by Whittaker and Robinson (1944, pp. 285–291). Figure 4.60 shows the results of smoothing a drilling-time log by various equations.

The methods for smoothing (filtering) of data series are given earlier in the text (see p. 103).

Fractality of Fourier series

Consider a Fourier sine series for a variable $Y_h = f(X)$ over an interval L of the values of X:

$$Y_h = \sum_{h=1}^{\infty} A_h \sin(2\pi i X/L)$$

where: A_h = amplitude associated with wavelength $\lambda_h = L/i$ (with $i = 1, 2, 3, \ldots$). The variable Y_h is a fractal if the amplitudes A_i have a power-law dependence on λ_i:

$$\boxed{A_i = c \cdot \lambda_i^{\frac{5-2D}{2}}}$$

where: c = constant; and D = fractal number (dimension) of the series.

The coefficients of the Fourier series, A_h, correspond to the wavelength $\lambda_h = 2\pi/k_h$, where k_h is the wave number and h is the harmonic number. The spectral power density of a series is given by $S_h = A_h^2 \cdot L$, where L is the series interval. A series is a fractal (self-similar) if:

$$\boxed{S_h = c_1 \cdot k_h^{-b}}$$

or

$$\boxed{A_h = c_2 \cdot \lambda_h^{b/2}}$$

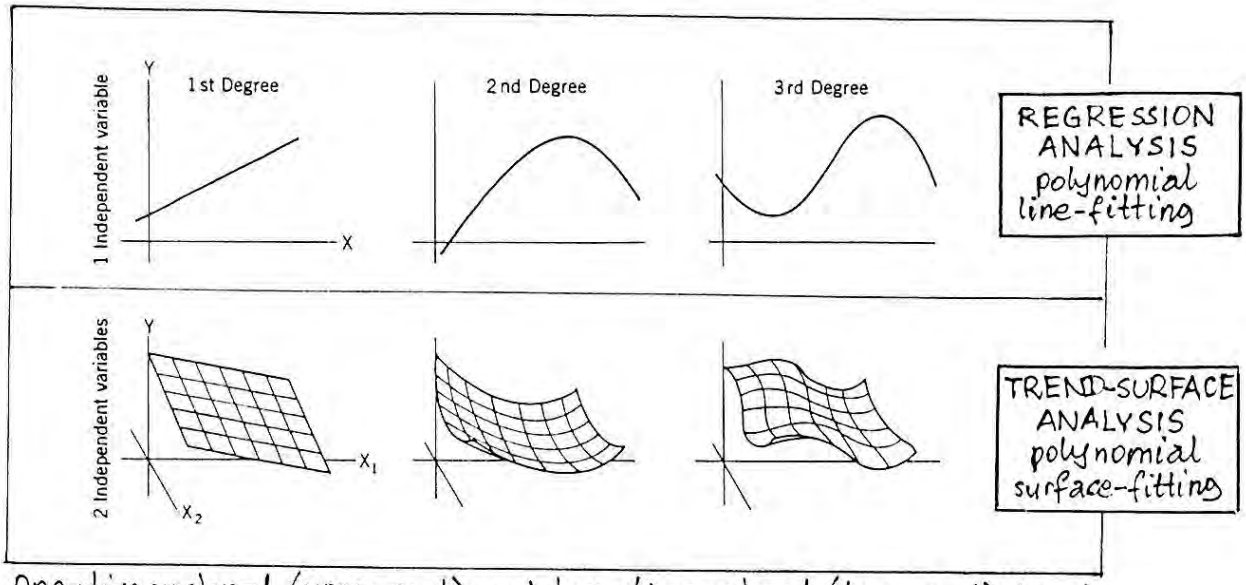
where the coefficients c and b can be derived by power-model regression (see p. 68). The corresponding fractal dimension is:

$$D = \tfrac{1}{2}(3-b)$$

From the relationship $D = 2 - H$, one obtains $H = 2 - D$, where H is the Hurst exponent. If $H > 0.5$, the series is persistent. If $H < 0.5$, the series is antipersistent (clustered). The H-value can be used to quantify and compare the persistence of different series.

Methods for quantitative data with regional/location co-ordinates.

• Trend-surface analysis



One-dimensional (upper part) and two-dimensional (lower part) trend of variable Y approximated by polynomial functions of first, second and third degree.

The data population in trend-surface analysis, $\{x_1, x_2, y\}$, is used to find a trend-surface $\hat{y} = f(x_1, x_2)$ defined by polynomial equation.

The following description of trend-surface analysis is from:

Davis J.C. (1986) Statistics and Data Analysis in Geology. 2nd edit., Wiley & Sons, New York, pp. 405-429.

Trend Surfaces

Trend analysis is the geology profession's name for a mathematical method of separating map data into two components—that of a regional nature, and local fluctuations. This has been done intuitively or graphically by geologists for years. Petroleum geologists, for example, refer to "regional dip" or "basinal configuration" as opposed to "local structures." Petrologists may speak of the "regional grain" of a metamorphic terrain. Geophysicists have long been accustomed to the concept of "regional trends" and "local anomalies." All of these expressions imply a belief that any given observation is the outcome of two interacting geologic forces or sets of forces—that which shaped the region or general geologic setting, and that which caused small areas to deviate from the regional pattern. Obvious examples may be drawn from structural geology. The Tertiary basins of Wyoming are the result of major movement along faults that extend deep into the crust. Within basins, folded structures have developed as the result of such diverse local agents as gravity sliding, minor antithetic faulting, failure of incompetent beds in areas of high dip, and so forth. If we regard the geometric shape of a basin as a regional structure, these small structures represent local deviations.

What we consider to be "regional" and "local" is largely subjective. It depends in part upon the size of the region being examined. If the buried Precambrian surface of the United States is under consideration, the basins and intervening mountain ranges of Wyoming will appear as anomalies or local deviations, as will the Black Hills, Ozark Dome, Michigan Basin, and other major features of the basement. Within an individual Wyoming basin, as we have seen, "regional" and "local" take on quite different meanings.

The availability of data exerts a very real influence on the nature of regional trends and local deviations. It is fruitless, for example, to search for meaningful local features whose suspected size approaches the spacing between sample points. Such features, whether they exist or not, simply cannot be detected. The relationship between the size of detectable features and the spacing of control points can be

calculated for regular nets of points (Singer and Wickman, 1969, pp. 2–19), but not for the less tractable situation of irregularly distributed points.

Even the purpose of the geologic examination exerts an influence on our concept of the two spatial components. In a South African gold mine, for example, the only "deviations" of interest may be those which exceed a certain arbitrary value determined by the economics of the operation. On the other hand, a petroleum exploration group reexamining an area may be searching for small structural anomalies, knowing that the larger structures in the region have been adequately tested. The larger features are then considered to be part of the "regional trend."

For purposes of illustration, we can consider the set of observations in X and Y in Figure 5.72*a*. These may be separated into "regional" and "local" components in a variety of ways. Suppose we decide that the regional trend has the form of a straight line placed through the points in some manner. The data may then be separated into a major linear trend, shown by the heavy line, and three large anomalies, with a smaller anomaly at one extreme of the sample set (Fig. 5.72*b*). However, we may decide that a parabolic function would be more representative of the regional trend than a straight line. In that case, we might arrive at the separation shown in Figure 5.72*c*. The parabolic trend is very much different from that shown in Figure 5.72*b*, and the distribution of deviations consequently is different. We could postulate still more complex forms for the regional trend, resulting in smaller and smaller deviations, as in the cubic function fit to the data in Figure 5.72*d*. Eventually, our trend and samples would coincide, at which point there would be no residuals. Then, of course, there would be no separation of the data into components, and the purpose of the exercise would be defeated.

An obvious question at this point is how can the data be objectively separated into two components if the definition of the components is entirely subjective? This may be done if we use, instead of a geologic definition of trend and deviation, an operational definition which specifies the way in which our data are to be treated. A trend may then be defined as a linear function of the geographic coordinates of a set of observations so constructed that the squared deviations from the trend are minimized. Let us examine the three parts of this definition carefully.

1. It is based on the *geographic coordinates*. This means that an observation, whether it is the elevation of a horizon or the amount of gold in a vein, is considered to be in part a function of the location of the observation.
2. The trend is a *linear function*. That is, it has the form $Y = b_1X_1 + b_2X_2 + \cdots$, where the b's are coefficients and the X's are some combination of the geographic coordinates. The equation will yield values of $\hat{Y}$ that are the trend components of an observation. Note that the terms of the equation are added together.
3. The specific linear function chosen for the trend must *minimize the squared deviations* from the trend. The appearance of squared deviations at this point may bring to mind some material from Chapter 2. The sum of the squared deviations from the mean defines the variance of the sample. If we substitute "line" or "plane" for mean in this sentence, we see that the trend can be regarded as a function having the smallest variance about it. You may also recall that we defined a line of regression in a similar fashion. Indeed, trend analysis is an adaptation of the statistical field of multiple regression, and the techniques have been borrowed directly from the discipline. In some cases, we can even use the powerful tests of hypotheses of multiple regression on geologic problems.

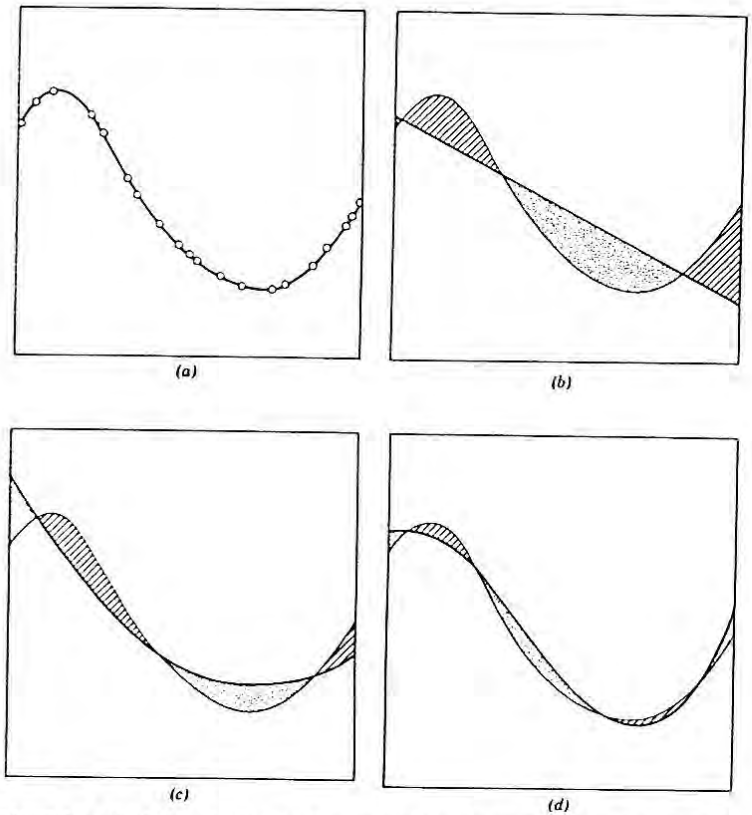


FIGURE 5.72 Concept of trend illustrated in two dimensions. (*a*) Collection of original data points and the line on which they lie. (*b*) Straight-line trend fit to the observations. (*c*) Parabolic trend. (*d*) Cubic trend. Shadings represent positive and negative residuals from the trends.

In Chapter 4, we computed a line of regression of Y on X, which was the line of best estimate of values of Y for any specified value of X. The equation of the line, $Y = b_0 + b_1X$ was found by solving the set of so-called *normal equations*

$$\begin{aligned} \Sigma Y &= b_0 n + b_1 \Sigma X \\ \Sigma XY &= b_0 \Sigma X + b_1 \Sigma X^2 \end{aligned} \tag{5.82}$$

for the unknown coefficients b_0 and b_1. The summations in this and subsequent equations extend from $i = 1$ to n. They have been omitted to simplify the written equations. We can easily expand this set of equations to the case where there are two independent variables, such as mutually perpendicular geographic coordinates. A linear trend surface is an equation of this type:

$$Y = b_0 + b_1X_1 + b_2X_2 \tag{5.83}$$

That is, a geologic observation, Y, may be regarded as a linear function of some constant value (b_0) related to the mean of the observations, plus an east-west (b_1) coordinate component and a north-south (b_2) component. Because the equation contains three unknowns, we need three normal equations to find the solution:

$$\begin{aligned} \Sigma Y &= b_0 n + b_1\Sigma X_1 + b_2\Sigma X_2 \\ \Sigma X_1Y &= b_0\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \\ \Sigma X_2Y &= b_0\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 \end{aligned} \tag{5.84}$$

Solving this series of simultaneous equations will give the coefficients of the best-fitting linear trend surface, "best-fitting" being defined by the least-squares criterion.

The equations can be rewritten into matrix form as

$$\begin{bmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1X_2 \\ \Sigma X_2 & \Sigma X_1X_2 & \Sigma X_2^2 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \Sigma Y \\ \Sigma X_1Y \\ \Sigma X_2Y \end{bmatrix} \tag{5.85}$$

The similarity between this matrix equation and (4.33) should be obvious. Both may be regarded as curve fitting using two independent variables. Here, our two variables are X_1 and X_2, mutually perpendicular geographic coordinates. In fitting a quadratic curve to data along a line, the two variables were X and X^2. However, there basically is no difference between the two procedures. As an example of the fitting of a linear trend surface, we will consider the following problem.

The Anglo-Barren Oil Company (ABOC) has been given a concession in a remote part of northeastern Africa. The area is extremely inhospitable, nearly inaccessible, and almost completely unknown geologically. Under the terms of the concession, ABOC must drill ten wells during the year or forfeit their rights. The company management has decided to drill a series of widely spaced exploratory holes to develop the geologic background necessary for continued prospecting. Well locations within the 100-km² concession are shown in Figure 5.73 and are listed in Table 5.15 (coordinates are given in kilometers from the southwest corner of the concession) along with the elevations of the bottom of the Cretaceous. We wish to determine the linear trend, and from an examination of residuals, pick areas that seem the most promising for additional exploration.

We must first accumulate the sums, sums of powers, and sums of cross-products required in (5.84). The necessary entries are

$$\begin{aligned} \Sigma X_1 &= 539 & \Sigma X_2 &= 482 & \Sigma Y &= -4579 \\ \Sigma X_1^2 &= 36{,}934 & \Sigma X_2^2 &= 31{,}692 & \Sigma X_1Y &= -211{,}098 \\ \Sigma X_1X_2 &= 27{,}030 & & & \Sigma X_2Y &= -232{,}342 \end{aligned}$$

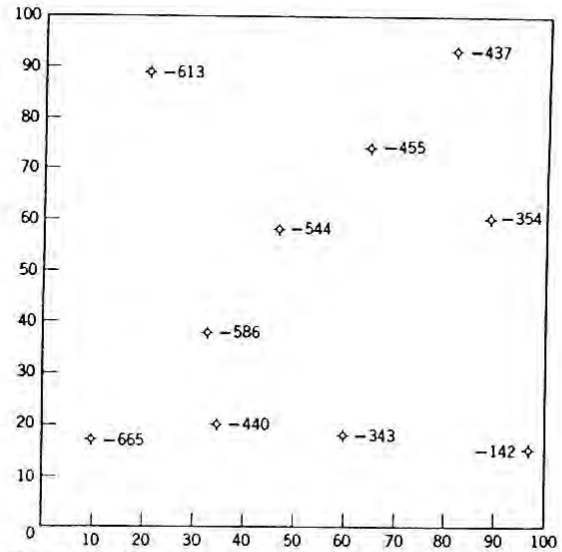


FIGURE 5.73 Map of well locations and elevations of the base of Cretaceous units in the Anglo-Barren Oil Company concession in northeastern Africa. Elevations are in meters below sea level. Coordinates of the map are in kilometers from southwest corner of concession.

Substituting these values into (5.85) gives

$$\begin{bmatrix} 10 & 539 & 482 \\ 539 & 36{,}943 & 27{,}030 \\ 482 & 27{,}030 & 31{,}692 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -4{,}579 \\ -211{,}098 \\ -232{,}342 \end{bmatrix}$$

This matrix equation can be solved by methods described in Chapter 3. The solutions

TABLE 5.15 Well Coordinates and Elevations of the Base of the Cretaceous within the Anglo-Barren Oil Company Concession

X_1 (km)	X_2 (km)	Y (m)
10.0	17.0	−665.0
21.0	89.0	−613.0
33.0	38.0	−586.0
35.0	20.0	−440.0
47.0	58.0	−544.0
60.0	18.0	−343.0
65.0	74.0	−455.0
82.0	93.0	−437.0
89.0	60.0	−354.0
97.0	15.0	−142.0

are

$$b_0 = -621.0, \qquad b_1 = 4.8, \qquad b_2 = -2.0$$

Having obtained the coefficients of the linear equation $Y = b_0 + b_1X_1 + b_2X_2$, we can calculate the expected or trend values, $\hat{Y}$, of the base of the Cretaceous at each of the ten wells. These, with the deviations $(Y_i - \hat{Y}_i)$, are listed in Table 5.16. We also can calculate measures of how well the trend surface corresponds to the observations by using (4.18)–(4.24), which we developed for the fitting of a line. In particular, we can compute the total variation as the sum of squares of the dependent variable, depth:

$$SS_T = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 215{,}324.9$$

For the estimated values of $\hat{Y}$ in Table 5.16, we can calculate the sum of squares due to the trend or regression:

$$SS_R = \Sigma \hat{Y}^2 - \frac{(\Sigma \hat{Y})^2}{n} = 193{,}861.4$$

The difference between these gives the sum of squares due to residuals or deviations from the trend and is

$$SS_D = SS_T - SS_R = 21{,}463.5$$

Now, the percentage of goodness-of-fit of the trend surface is

$$100\% \cdot R^2 = \frac{SS_R}{SS_T} = 90.0\%$$

The coefficient of multiple correlation is

$$R = \sqrt{R^2} = 0.95$$

From these extremely high values, we would conclude that the base of the Cretaceous in this area is an almost smooth, uniformly dipping plane. Deviations from the surface are relatively small, as a glance at Table 5.16 will confirm. Apparently, the basal Cretaceous contact is a gently dipping surface of low relief, as shown in Figure 5.74. Although this simple analysis seems sufficient for this example, we have already discussed the possibility that a geologic trend may not be a plane, and may be extremely complex. Furthermore, we very rarely have any prior knowledge about what the functional form of the trend should be. Physicists can state that the path of a falling projectile should be a parabola, because they know something of the controlling forces—that is, the acceleration of gravity and the conservation of momentum. Geologists can seldom speak with any authority about what form a geologic surface or distribution "should" take. Instead, they do the next best thing, and approximate the unknown function with one of arbitrary nature. In particular, they use a polynomial expansion of the linear trend surface, introducing powers and cross-products of the geographic coordinates. Polynomials are extremely flexible, and if expanded to sufficiently high orders, can conform to very complex surfaces.

A note of caution should be injected at this point. Polynomial functions are used for geologic trend analysis merely as a matter of convenience. The equations that are necessary to find the coefficients of the trend may easily be established and solved by computer. Use of polynomials in no way intimates a belief that geologic processes are polynomial functions or even that they are linear. Their nature, unknown and perhaps ultimately unknowable, can only be approximated by a polynomial expansion. Other approximations (or in rare cases, model equations) may be more appropriate in specific instances; some of these will be discussed in subsequent parts of this section.

TABLE 5.16 Geographic Coordinates, Elevation of Base of Cretaceous, Trend-Surface Estimate of Cretaceous Base, and Residual $(Y - \hat{Y})$

X_1 (km)	X_2 (km)	Y (m)	$\hat{Y}$ (m)	$(Y - \hat{Y})$ (m)
10.0	17.0	−665.0	−606.6	−58.3
21.0	89.0	−613.0	−695.7	82.7
33.0	38.0	−586.0	−537.8	−48.1
35.0	20.0	−440.0	−492.8	52.8
47.0	58.0	−544.0	−510.2	−33.7
60.0	18.0	−343.0	−369.2	26.2
65.0	74.0	−455.0	−455.5	0.5
82.0	93.0	−437.0	−411.5	−25.4
89.0	60.0	−354.0	−313.0	−40.9
97.0	15.0	−142.0	−186.1	44.1

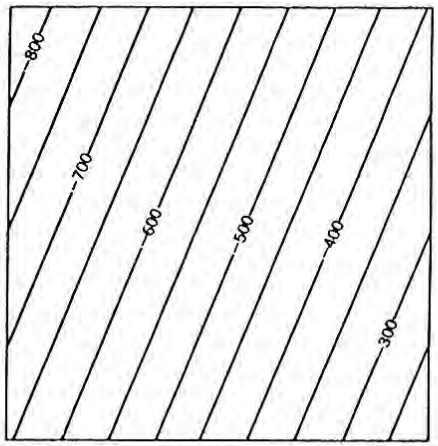


FIGURE 5.74 Trend-surface map of Anglo-Barren Oil Company concession. Contour interval is 50 m.

As we saw in Chapter 4, a least-squares line may be expanded to a second-order curve (parabola) by adding a squared term to the linear equation:

$$Y = b_0 + b_1X_1 + b_2X_1^2 \tag{5.86}$$

An equivalent expansion of (5.83) gives a second-degree trend surface:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_2^2 + b_5X_1X_2 \tag{5.87}$$

Note that the equation contains terms which are the squares of the two geographic coordinates and a cross-product term, X_1X_2. The expansion of this trend-surface equation to even higher powers should be relatively obvious. Each geographic variable is simply raised to a higher power, creating two new variables, then the appropriate cross-products of the two coordinates are calculated to give other new variables. For example,

$$Y = b_0 + \overbrace{b_1X_1 + b_2X_2}^{1} + \overbrace{b_3X_1^2 + b_4X_2^2 + b_5X_1X_2}^{2} + \overbrace{b_6X_1^3 + b_7X_2^3 + b_8X_1^2X_2 + b_9X_1X_2^2}^{3} \tag{5.88}$$

is a third-degree trend surface. The first-degree coefficients are b_1 and b_2. The coefficients b_3, b_4, and b_5 are second-degree, because the variables in these terms are of the form $X_3 = (X_1 \cdot X_1)$, $X_4 = (X_2 \cdot X_2)$, and $X_5 = (X_1 \cdot X_2)$. That is, the variables are the products of multiplying two original variables together. Similarly, b_6, b_7, b_8, and b_9 are third-degree coefficients, as the variables in these terms result from the multiplication of three original variables together; that is, $X_6 = (X_1 \cdot X_1 \cdot X_1)$, $X_7 = (X_2 \cdot X_2 \cdot X_2)$, $X_8 = (X_1 \cdot X_1 \cdot X_2)$, and $X_9 = (X_1 \cdot X_2 \cdot X_2)$.

Petroleum explorationists have made good use of trend surface analysis in the search for oil and gas in central Alberta, Canada. The Alberta Basin is extremely petroliferous, with production coming primarily from Lower Cretaceous sands and from much deeper carbonate reefs of Upper Devonian age. Many more wells penetrate the shallower Cretaceous horizons than the Devonian, especially near the Rocky Mountain front where reef prospects may be 15,000 ft or more deep. The Devonian reefs are thick accumulations of carbonates in the Leduc Formation, overlain and surrounded by shales that represent lagoonal and open marine clastic sediments. The carbonate reefs were rigid and incompressible, while the enclosing fine-grained clastic sediments of the Ireton Shale were compacted to a fraction of their original thickness as the basin subsided and deposition continued. This differential compaction created drape structures over the buried reefs; these structures persist in the overlying rocks, although their magnitudes become less on shallower horizons.

Deep-seated compaction features are not readily apparent on the Cretaceous horizons because of the overriding effect of the strong regional dip (Fig. 5.75). Closure at depth is expressed only as slight changes in local gradient that appear on structural maps of Cretaceous horizons as subtle variations in the spacing of contour lines. However, if the strong regional dip component could be removed

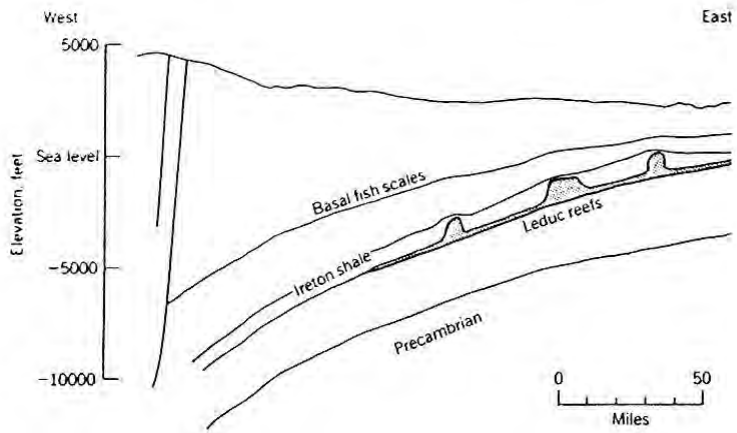


FIGURE 5.75 Diagrammatic cross-section across western margins of Alberta Basin. Leduc reefs and Ireton Shale are Upper Devonian; Basal Fish Scales is Lower Cretaceous.

from these maps, presumably the underlying drape features would appear as closed structures. Since the density of wells that bottom in the Cretaceous is relatively great, their analysis could provide valuable information about possible prospects in the Devonian even though few wells penetrate the deeper horizon.

The Basal Fish Scales is a black shale occurring near the boundary between the Lower and Upper Cretaceous. The unit derives its name from the abundant fish scales it contains. It is also characterized by numerous bentonite beds, which produce conspicuous spikes on gamma ray logs because of their high radioactivity. Bentonites are synchronous and so form excellent markers for regional correlation. The top of the Basal Fish Scales can be picked with exceptional consistency on log traces because of the pronounced response from a thick, persistent bentonite. The map shown in Figure 5.76 includes an area of about 3500 square miles in west-central Alberta, and was constructed using picks of this bentonite at the top of the Basal Fish Scales in 360 exploratory holes. The map area lies on the western margin of the Alberta Basin, immediately in front of the overthrust zone marking the Rocky Mountain front. Beds dip downward to the southwest, with increasing dip as the western edge of the basin is approached. Within the map area, depths of the Basal Fish Scales range from slightly in excess of 1000 ft below sea level in the northeast, to almost 5000 ft below sea level in the southwest.

Figures 5.77 and 5.78 are first- and second-degree polynomial trend surfaces of the Basal Fish Scales; statistics relating to the fit of the trend surfaces are given in Table 5.17. Both trend surfaces provide an extremely high fit to the observations, although the second degree is significantly better than the first degree. In this application, statistical tests of significance are not an appropriate guide for the selection of degree of trend surface, because the problem is not one of statistical

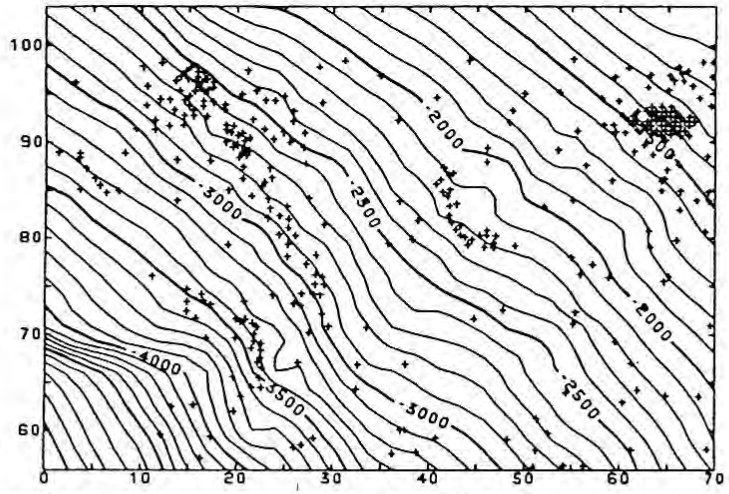


FIGURE 5.76 Structure contour map of part of Alberta Basin, drawn on top of Basal Fish Scales. Contours in feet below sea level. Well control shown by crosses.

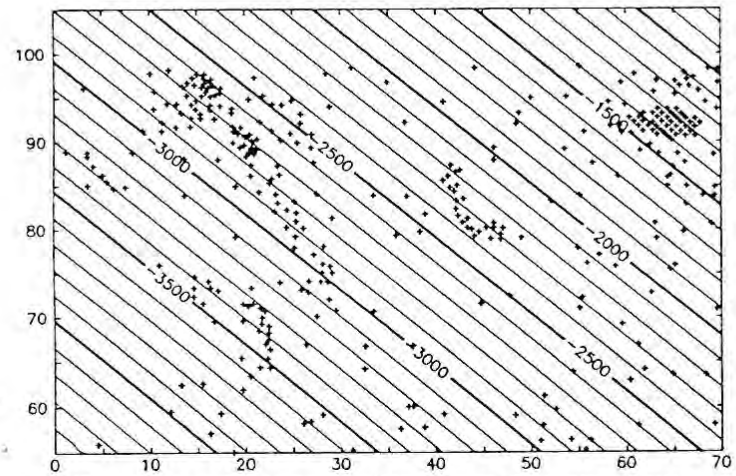


FIGURE 5.77 First-degree trend surface of Basal Fish Scales. Contour interval in feet below sea level.

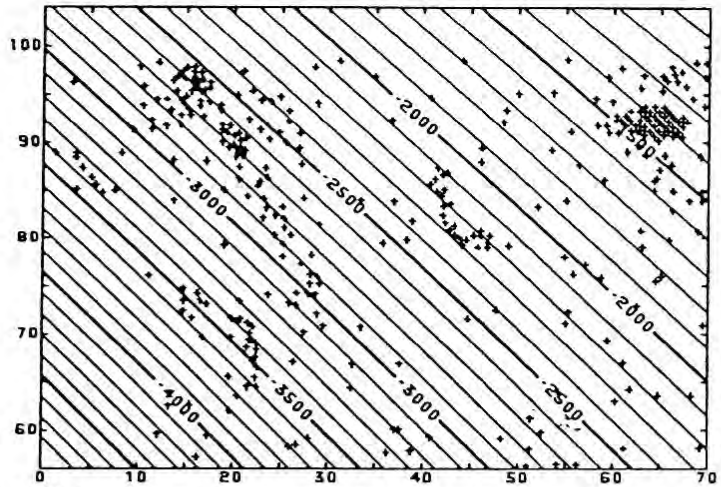


FIGURE 5.78 Second-degree trend surface of Basal Fish Scales. Contour interval in feet below sea level.

estimation. Rather, the objective is to closely simulate the regional features of the structure contour map, so subtraction of the trend will remove the regional component of structure. In effect, the trend surface is used as a high-pass filter, removing the large-scale structural variation from the map and leaving behind small-scale features.

Figure 5.79 shows residuals from the second-degree trend surface. Figure 5.80 is a paleogeographic map representing the Upper Devonian, reconstructed from well and seismic information. Note the strong coincidence between positive trend-surface residuals on the Basal Fish Scales, and the locations of major Devonian

TABLE 5.17 Statistics of First- and Second-Degree Polynomial Trend Surfaces Fitted to Elevation of Basal Fish Scales in West-Central Alberta

First-Degree Trend Surface	
Percent goodness-of-fit (R^2)	98.6%
Correlation coefficient (R)	0.993
Trend surface equation:	
$Y = -6351.4 + 29.3X_1 + 33.8X_2$	
Second-Degree Trend surface	
Percent goodness-of-fit (R^2)	99.7%
Correlation coefficient (R)	0.999
Trend surface equation:	
$Y = -7993.3 + 63.4X_1 + 59.2X_2$	
$-0.1X_1^2 - 0.3X_2^2 - 0.1X_1X_2$	

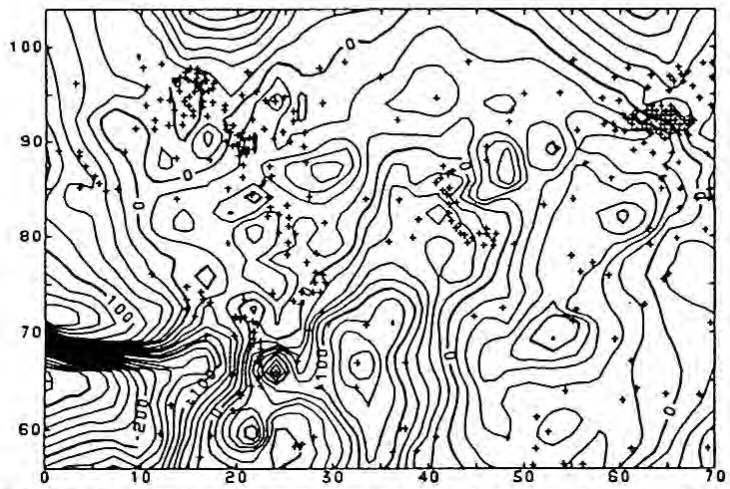


FIGURE 5.79 Residual map constructed by subtracting second-degree trend surface (Fig. 5.78) from structural contour map of Basal Fish Scales (Fig. 5.75). Positive residuals are shaded. Contour interval is 20 ft.

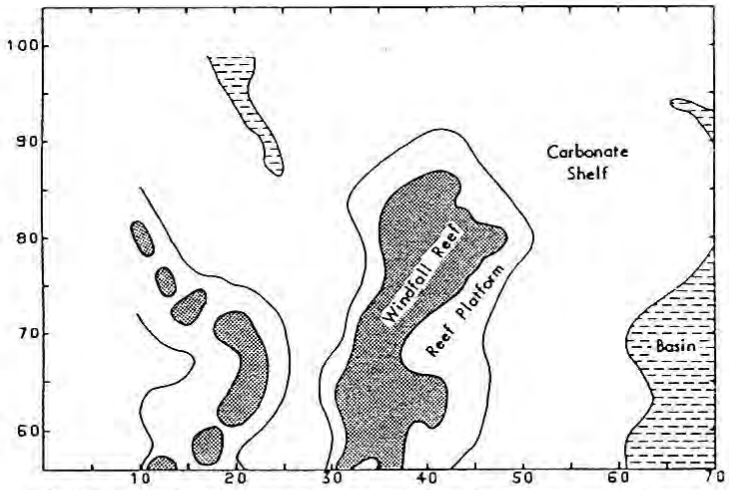


FIGURE 5.80 Paleogeographic map of Upper Devonian showing Leduc Reefs growing on carbonate platform along west margin of Devonian sea.

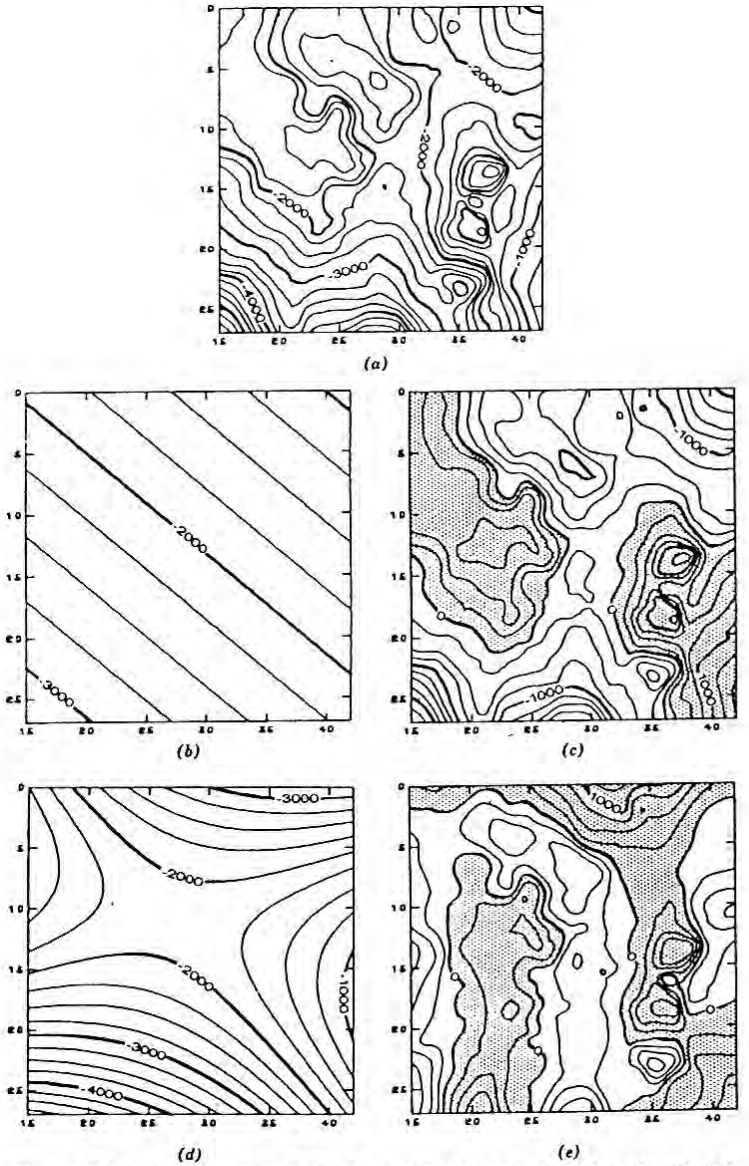


FIGURE 5.81 (*a*) Contour map of the top of Ordovician rocks (Arbuckle Gp.) in central Kansas. Contours in feet below sea level. Original well control shown in Figure 5.10. (*b*) First-degree trend surface. (*c*) Contoured residuals from first-degree trend. (*d*) Second-degree trend surface. (*e*) Contoured residuals from second-degree trend. Areas of positive residuals shown by shading.

reefs, particularly the Windfall Reef. Removal of the regional trend has successfully isolated minor components of structure representing drape over these reefs. Several major oil fields were discovered in this area using this exploration technique.

Still unanswered is the question, "what is a high (or low) goodness-of-fit or correlation in a trend surface?" In geologic studies, we often must rely on experience and intuition to provide an answer. For example, in structural analyses performed on data from Kansas, Oklahoma, Texas, Wyoming, California, England, and other localities, we have judged first- and second-degree correlations of less than 0.3 as providing a "poor fit." Correlations of 0.4–0.6 yield interpretable

TABLE 5.18 Geographic Coordinates (in Arbitrary Units) and Elevations of the Top of the Ordovician in Central Kansas

X_1 (units)	X_2 (units)	Y (ft)	X_1 (units)	X_2 (units)	Y (ft)	X_1 (units)	X_2 (units)	Y (ft)
23.36	22.10	−2961.0	35.90	9.90	−1537.0	26.86	21.17	−2951.0
29.80	10.58	−2240.0	21.54	14.90	−1667.0	29.93	16.80	−2435.0
22.18	16.24	−1872.0	33.46	12.31	−1694.0	24.62	23.12	−3177.0
22.91	3.36	−2584.0	19.40	16.57	−2300.0	26.04	16.52	−2241.0
39.33	21.14	−1119.0	35.18	19.79	−1465.0	32.86	13.52	−1775.0
15.10	18.50	−3062.0	21.20	3.50	−2349.0	15.10	25.85	−5400.0
28.96	10.30	−2540.0	23.50	13.02	−1564.0	36.99	25.28	−1852.0
18.90	16.69	−2300.0	19.09	24.40	−3657.0	22.40	13.23	−1500.0
21.11	12.26	−1505.0	34.45	19.45	−1257.0	27.91	16.48	−2481.0
29.92	4.00	−1921.0	25.60	17.18	−2337.0	19.92	11.10	−1599.0
41.99	5.31	−2056.0	28.60	26.15	−4373.0	29.62	20.27	−2875.0
21.86	4.01	−2466.0	29.32	11.48	−2109.0	36.50	10.21	−1353.0
41.30	16.38	−1077.0	32.60	24.15	−2923.0	16.58	5.12	−1608.0
22.75	21.67	−2780.0	25.90	19.05	−2607.0	34.45	6.22	−1499.0
35.93	14.53	−707.0	20.12	19.80	−2751.0	30.26	22.27	−3029.0
21.59	21.16	−2677.0	33.87	23.97	−1626.0	35.60	20.95	−1860.0
28.20	6.40	−2801.0	34.28	1.78	−2305.0	23.23	11.59	−1412.0
30.50	19.83	−2678.0	39.60	19.40	−1135.0	27.02	17.01	−2407.0
41.20	3.50	−2586.0	39.01	9.93	−1971.0	20.94	22.90	−3044.0
31.38	17.74	−2190.0	24.60	20.45	−2483.0	17.53	10.02	−1657.0
24.30	23.90	−3367.0	25.10	22.84	−3095.0	28.27	8.16	−2540.0
24.80	21.10	−2959.0	24.48	24.45	−3589.0	17.83	0.10	−1647.0
16.75	8.66	−1709.0	26.47	13.19	−1490.0	32.58	19.86	−2140.0
22.72	11.15	−1431.0	20.23	17.58	−2307.0	35.45	17.03	−883.0
20.40	21.82	−3022.0	35.27	19.23	−1037.0	21.21	14.57	−1695.0
26.49	6.43	−2431.0	25.00	9.20	−1407.0	36.20	16.25	−1746.0
33.12	17.04	−1792.0	22.95	8.42	−2133.0	28.52	17.20	−2440.0
24.67	15.61	−2146.0	35.00	23.13	−3090.0	29.80	13.96	−2346.0
30.10	12.92	−2131.0	30.07	15.18	−1890.0	30.40	10.02	−2182.0
26.12	0.50	−2295.0	40.00	16.82	−1366.0	37.45	7.18	−1934.0
28.25	14.35	−2421.0	18.45	9.01	−1651.0	25.02	17.82	−2063.0
27.21	10.58	−2204.0	38.01	26.05	−1857.0	34.20	26.20	−2834.0
19.38	17.60	−2348.0	29.05	21.08	−2998.0	16.60	1.85	−1583.0
42.00	17.90	−885.0	35.40	20.90	−1902.0	17.40	1.10	−1521.0
32.48	2.24	−1618.0	41.18	20.14	−998.0	41.12	19.86	−835.0
18.96	14.54	−1834.0	26.70	12.80	−1785.0	29.92	22.02	−3061.0
32.97	11.79	−1884.0	15.58	4.63	−1593.0	17.10	14.05	−2402.0
23.50	1.57	−2308.0	20.70	20.39	−2722.0	21.70	14.21	−1608.0
41.89	15.25	−1108.0	20.80	11.50	−1477.0	20.60	6.28	−1613.0
31.40	12.20	−2058.0	41.11	5.22	−2274.0	30.32	15.55	−2195.0
31.60	22.82	−2711.0	22.11	21.00	−2598.0	39.62	18.03	−1421.0

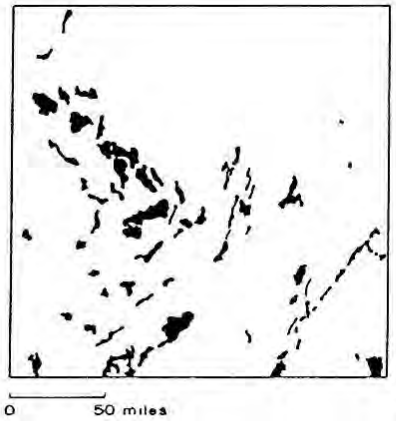


FIGURE 5.82 Locations of major oil and gas fields in central Kansas.

trend and residual maps, and those higher than 0.7 are regarded as conforming closely to the original data points. Again, the purpose of the investigation must be kept firmly in mind while interpreting goodness-of-fit. In all of these structural studies, we were examining basins having relatively simple shapes and were searching for relatively small residuals compared to the total size of the basin. Third- and fourth-degree polynomials usually provided very high fits (over 0.8). As a reference value, randomly generated data sets having the same range of values as actual data yielded correlations of up to 0.3 for fourth-degree polynomials. Realistic-appearing trends can be extracted from random data, so geologists should proceed cautiously in the interpretation of trend and residual maps having low goodness-of-fit.

Figure 5.81*a* is a contour map of the data given in Table 5.18, which are elevations of the top of the uppermost formation of Ordovician age in central Kansas. Figure 5.81*b* is a first-degree trend surface fit to this same data; a map of residuals from the surface is shown as Figure 5.81*c*. A second-degree trend surface and residuals from the same data are shown as Figures 5.81*d* and 5.81*e*.

As an exercise, using the data from Table 5.18, fit a third-degree polynomial trend surface to the Ordovician top. Compare your map to the second-degree trend. Contour the residuals from the surface and compare this map with the second-degree residuals. Does the pattern of residuals change? Figure 5.82 is a generalized map of the major oil fields discovered in central Kansas. It may be instructive to compare the field distributions with the distribution of positive residuals from the three trends.

Statistical Tests of Trends

The goodness-of-fit of a trend surface may be tested statistically, by comparing the variance due to regression or trend to the variance due to deviations from the trend. You will recall from Chapter 2 that tests of equality of variances involve

the F distribution and are valid only if the data satisfy certain conditions. If these assumptions can justifiably be made, we may regard the b_i coefficients found by least squares as estimates of the true population regression coefficients, β_i, and test hypotheses about their nature. We must assume that the population of the dependent variable is normally distributed about the regression, and that its variance does not change with changes in the independent variables. In addition, the samples must be drawn without bias from this population. The last condition often is the most difficult to satisfy, especially in a structural analysis based on oil wells, as the wells presumably were not located by random selection! Testing of statistical hypotheses about trend surfaces is perhaps most easily defended when dealing with data such as geochemical assays on samples collected according to a planned experimental design.

The significance of a trend or regression may be tested by performing an analysis of variance, which is the process of separating the total variation of a set of observations into components associated with defined sources of variation. This, or course, has been done by dividing the total variation of Y into two components, the trend (or regression) and the residuals (or deviations). The degrees of freedom associated with total variation in a trend analysis are $(n - 1)$, where n is the number of observations. The degrees of freedom associated with regression are determined by the number of terms or coefficients in the polynomial equation fit to the data. Degrees of freedom for deviation are the number of degrees of freedom associated with total variation minus those that are accounted for by regression, or $\nu_D = \nu_T - \nu_R$. A formal analysis of variance table is shown in Table 5.19.

The mean squares are found by dividing the various sums of squares by the appropriate degrees of freedom. By reducing sums of squares to mean squares, they have been converted to estimates of variance and may be compared using an F probability distribution. The MS_D is the variance about the regression line; MS_R is the variance of the regression line about its mean. If the regression is significant, the deviation about the regression will be small compared to the variance of the regression itself.

In a general test of a trend-surface equation, the ratio of interest is that between

TABLE 5.19 General ANOVA for Significance of Regression of kth-Degree Polynomial Trend Surface; Number of Coefficients in Trend-Surface Equation not Counting the b_0 Coefficient is m, Number of Data Points is n

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Square	F Test
Polynomial Regression	SS_R	m	MS_R	MS_R/MS_D
Deviation from Polynomial	SS_D	$n - m - 1$	MS_D	
Total Variation	SS_T	$n - 1$		

variance due to regression and variance due to deviation. The F test gives a probabilistic answer to the question of whether the variances being examined have been obtained by random sampling from the same population. Or, is the regression effect not significantly different from the random effect? An affirmative answer may be interpreted as meaning that (a) the distribution of Y is random and independent of values of $X_1, \ldots, X_m$, or (b) the distribution of Y may be in part a function of $X_1, \ldots, X_m$, but the wrong functional model has been fit to the data.

In more formal terms, the F test for significance of fit is a test of the hypothesis and alternative

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$$
$$H_1 : \beta_1, \beta_2, \ldots, \beta_m \neq 0 \qquad (5.89)$$

The hypothesis to be tested is that the partial regression coefficients are equal to zero, or in other words, there is no regression. If the computed value of F exceeds the table value of F, this hypothesis is rejected and the alternative, H_1, is accepted.

In polynomial trend-surface analysis, it is customary for some investigators to fit a series of equations of successively higher degrees to the data. In such an analysis, a number of regression sums of squares will be produced, each larger than the preceding sum. The analysis of variance table may be expanded to analyze the contribution of the additional partial regression coefficients and give a measure of the appropriateness of increasing the order of the equations. The test is developed by finding the difference in sums of squares due to regression of the higher polynomial equation minus the regression sums of squares due to fitting the lower order polynomial. This difference is divided by the difference in regression degrees of freedom, giving the mean square of regression due to increasing the degree of the polynomial. This mean square is then divided by the mean square due to deviation from the higher polynomial. If the resulting F value is significant, the deleted order was contributing to the regression and should be retained. If the value is not significant, nothing has been gained by fitting the higher degree polynomial. An ANOVA table for testing the significance of a higher degree polynomial trend surface is given in Table 5.20.

The F test for significance of added terms is a test of the hypothesis and alternative

$$H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_m = 0$$
$$H_1 : \beta_{k+1}, \beta_{k+2}, \ldots, \beta_m \neq 0 \qquad (5.90)$$

The null hypothesis states that partial regression coefficients after the kth term are all equal to zero, or, they do not contribute to the regression caused by the first through kth term. (Remember that the polynomial trend surface of degree p contains k coefficients, whereas the polynomial equation of the $(p + 1)$ trend contains m coefficients.) Again, if the computed value of F exceeds the table value, the hypothesis is rejected. The test procedure is given for an equivalent curvilinear model by Li (1964, pp. 176–181).

In some instances we may be interested in assessing the effect of a single partial regression coefficient. This may be done simply by omitting that term from the polynomial equation and recomputing the sums of squares due to regression and

TABLE 5.20 General ANOVA for the Significance of Increasing the Degree of a Polynomial Trend from p to $(p + 1)$ Degree; Polynomial Equation of Degree p Has k Coefficients, not Counting the b_0 Term; Equation of Degree $(p + 1)$ Has m Coefficients, not Counting the b_0 Term; Number of Observations is n

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Test
Regression of Degree $(p + 1)$	SS_{RP+1}	m	MS_{RP+1}	MS_{RP+1}/MS_{DP+1}[a]
Deviation from Degree $(p + 1)$	SS_{DP+1}	$n - m - 1$	MS_{DP+1}	
Regression of Degree p	SS_{RP}	k	MS_{RP}	MS_{RP}/MS_{DP}[b]
Deviation from Degree p	SS_{DP}	$n - k - 1$	MS_{DP}	
Regression Due to Increase from p to $(p + 1)$ Degree	$SS_{RI} = SS_{RP+1} - SS_{RP}$	$m - k$	MS_{RI}	MS_{RI}/MS_{DP+1}[c]
Total Variation	SS_T	$n - 1$		

[a]Test of significance of the $(p + 1)$-degree trend surface.
[b]Test of significance of the p-degree trend surface.
[c]Test of significance of increase in fit of the $(p + 1)$ degree over p degree.

deviation. The contribution of the deleted term is the difference in the two sums of squares due to regression. Significance of this term may be tested by computing the ratio between the mean square due to the deleted term and the mean square due to the complete regression. The F ratio has one and $(n - m - 1)$ degrees of freedom. The ANOVA for the significance of a single deleted coefficient is given in Table 5.21.

Individual terms in the regression equation also may be tested by calculating the increase in SS_R when a new variable is added. However, this is not an advisable practice, because the tendency may be to regard all further terms as nonsignificant after encountering several successive nonsignificant coefficients. This may not be the case. In trend-surface analysis, the complete set of terms for the next higher order should be added, and then the added individual terms tested one by one by elimination. Abbreviated sets of higher degree terms should not be added blindly, unless there are compelling reasons to do so. In one example, because of computer limitations, third-order "hypersurfaces" were fit to oil-gravity data with an equation that contained no cubic cross-product terms. The resulting regression was considerably different from that produced from the same data when they were rerun on a larger computer using a program that fitted a complete cubic equation. Additional terms are especially apt to be significant if the correlation coefficients of lower orders are small.

TABLE 5.21 ANOVA for Testing the Significance of a Single Deleted Coefficient; the Complete Polynomial Regression Equation Contains m Coefficients, not Counting the b_0 Term; the Regression Equation Contains Only $(m - 1)$ Coefficients after Deleting the kth Coefficient; Number of Observations is n

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Test
Regression of All Terms	SS_R	m	MS_R	MS_R/MS_D[a]
Deviation	SS_D	$n - m - 1$	MS_D	
Regression Omitting kth Term	SS_{R-1}	$m - 1$	MS_{R-1}	MS_{R-1}/MS_{D-1}[b]
Deviation	SS_{D-1}	$n - m - 2$	MS_{D-1}	
Regression of kth Term Only	$SS_{RK} = SS_R - SS_{R-1}$	1	MS_{RK}	MS_{RK}/MS_D[c]
Total	SS_T	$n - 1$		

[a]Test of significance of the p-degree trend surface.
[b]Test of significance of the p-degree trend surface fitted without the kth coefficient.
[c]Test of significance of the kth coefficient alone.

The two data sets presented previously (Tables 5.15 and 5.18) are representative of problems in structural trend-surface analysis. The purpose of both investigations was to seek areas where the structural surface departed from a polynomial model equation. In these problems, the error distribution is such that the appropriateness of tests of the significance of the regressions is suspect. However, in the following example experimental and sampling conditions seem adequate to justify application of regression tests.

Figure 5.83 shows a planar view of single crystal of sphalerite collected from a mine in northern Mexico. Investigators are interested in the distribution of iron through the crystal, so it has been carefully cleaved through its center and the surface has been polished. The iron content at 10 Å spots spaced 1 mm apart has been determined by electron microprobe. The sample points are shown as dots on Figure 5.83*a*; a contour map of the values is shown as 5.83*b*; and the data are listed in Table 5.22. Although use of iron content in sphalerite as a temperature indicator has been criticized because of the possibility of nonequilibrium conditions at the time of crystallization, the researchers postulate that the growing shell of the crystal was in equilibrium with the ore solution at all times. Therefore, the average composition of the crystal may be inadequate as a temperature indicator, but the composition of successive thin shells of the crystal may define a temperature-change curve. Excluding the possibility of zoning, the simplest model for the distribution of iron in the crystal would be a gradual increase or decrease symmetrically outward from the center. The fitting of a second-degree polynomial

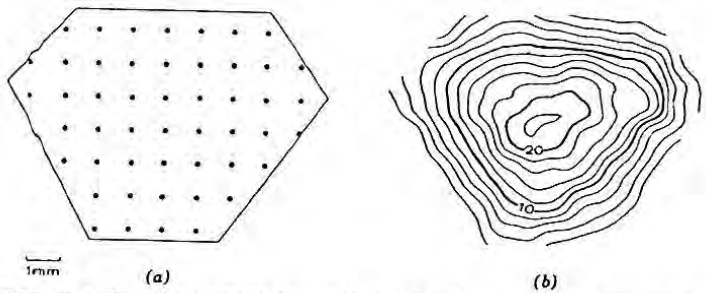


FIGURE 5.83 (*a*) Locations of sampling points in an electron microprobe study of the iron content of a sphalerite crystal. (*b*) Contours showing iron concentration, derived from data taken at points shown in part *a*.

TABLE 5.22 Iron Content of 10 Å Spots Spaced on a 1-mm Grid Across Cleavage Surface of Sphalerite Crystal

Iron Content (%)					
X_1	X_2	Y	X_1	X_2	Y
2.0	1.0	3.1	2.0	4.0	6.4
3.0	1.0	4.6	3.0	4.0	14.6
4.0	1.0	5.8	4.0	4.0	17.6
5.0	1.0	7.2	5.0	4.0	21.2
6.0	1.0	8.4	6.0	4.0	21.0
7.0	1.0	6.3	7.0	4.0	13.4
8.0	1.0	2.4	8.0	4.0	7.5
1.0	2.0	2.5	9.0	4.0	0.4
2.0	2.0	10.2	2.0	5.0	3.1
3.0	2.0	12.8	3.0	5.0	8.6
4.0	2.0	16.1	4.0	5.0	15.0
5.0	2.0	14.2	5.0	5.0	16.2
6.0	2.0	15.1	6.0	5.0	14.8
7.0	2.0	12.8	7.0	5.0	9.8
8.0	2.0	9.0	8.0	5.0	3.1
9.0	2.0	5.3	3.0	6.0	5.0
1.0	3.0	4.3	4.0	6.0	7.2
2.0	3.0	14.1	5.0	6.0	12.3
3.0	3.0	15.6	6.0	6.0	10.6
4.0	3.0	20.2	7.0	6.0	4.5
5.0	3.0	20.6	3.0	7.0	0.6
6.0	3.0	18.5	4.0	7.0	2.4
7.0	3.0	16.2	5.0	7.0	3.5
8.0	3.0	10.2	6.0	7.0	4.7
9.0	3.0	4.6			

regression of iron content on the coordinates of the sample points seems an appropriate way of testing this hypothesis. Using the data from Table 5.22, fit a series of trend surfaces to the iron content. From the various sums of squares, construct an ANOVA similar to Table 5.20 and test the significance of the polynomial regressions.

Two Trend-Surface Models

You may have noted that the preceding discussions imply that two basically different types of geologic problems are analyzed using trend-surface methods. On one hand, trend surfaces are fit to structural data in an attempt to isolate "local structures." It has been demonstrated empirically that in a sedimentary basin these residuals may be associated with structurally or hydrodynamically trapped oil. Alternatively, trend surfaces have been used to define regional trends in petrographic and geochemical data. These two applications differ in objectives and in underlying assumptions, but their common methodology has obscured these differences.

Trend surfaces fit to structural data can be represented by the model equation

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_2^p + (\gamma_i + \epsilon_i) \qquad (5.91)$$

The equation states that a given observation (elevation of a formation top) is equal to the sum of a constant term related to the means of the geographic coordinates, plus a polynomial expansion of degree p of the geographic coordinates, plus a local component, plus a randomly distributed measurement error. The latter two terms are considered to be confounded, and are the parameters of interest.

In contrast, a trend surface fit to petrographic or similar data usually is represented by an ordinary response-surface model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_2^p + \epsilon_i \qquad (5.92)$$

The response-surface model equation is similar in all respects to the trend-surface equation except that the local component γ_i is not considered to be present. Interest is centered upon the nature of the trend, as expressed in the estimates of the β's or polynomial coefficients

Petrographic and geochemical variates usually are characterized by high variance among replicates. This variance arises because of inhomogeneities within the samples analyzed, local or small-scale variation in composition (on a scale larger than the sample but smaller than the interval between samples), and analytical or instrumental errors. In a typical study, these are confounded and produce a normal distribution of error about each observation. Although each source of variance could be isolated and measured by replication at several levels within the experiment, this may not be done because of economic or other factors.

A trend surface or regression on geographic variables can be fitted appropriately to such data if certain basic assumptions seem reasonable. These require that the errors ϵ_i be randomly and normally distributed about the regression, that they have zero mean, and that they have constant variance. This in turn means that the errors are independent of one another. If these conditions are fulfilled, the regression

may be tested for significance and inferences drawn about the trend. The appropriate statistical tests are those we have presented in Tables 5.19–5.21. Many other special statistical designs have been created and are widely used in such fields as agriculture and chemical engineering; an introduction to these is given in Mendenhall (1968, Chapter 10). Koch and Link (1980, pp. 217–222) discuss a geologic application of one such design. The important aspect of such inferences is that they pertain to the trend. This is emphasized in Figure 5.84, where they observed values of Y_i can be seen to fall within the distribution of errors around the regression.

In the fitting of trend surfaces to structural data, observations (typically elevations of a geologic horizon) are not replicated. Although it is possible to relog a well, it usually is not possible to drill a second, nearby well which would constitute a replicate sample. Repeated down-hole measurements may vary by a few feet depending upon depth and amount of stretch in the cable, but this source of experimental error is always one or more orders of magnitude smaller than deviations in trend-surface analysis. Lack of replicates means that local variation cannot be assessed. However, is may be rationalized that this source of error is also inconsequential, as the drill-hole is not sampling from a population of surfaces. Only one formation top exists and the only variance is due to the negligible measurement error. Thus, all residual variance in trend-surface analysis can be ascribed to lack of fit.

In terms of the model equation, this is equivalent to saying that ϵ_i is negligibly small compared to γ_i. Although the random error ϵ_i has zero mean and is independent for all values of Y_i, it cannot be isolated because replicates have not been taken. The confounded term $(\gamma_i + \epsilon_i)$ also has a mean of zero, but it is not in general independent for values of Y_i. In fact, the purpose of the analysis is to define regions of specified size in X_1 and X_2 over which $(\gamma_i + \epsilon_i)$ are correlated. Figure 5.85 shows a theoretical distribution of ϵ_i about the structural surface. For the most part, deviations of the surface from the polynomial model do not reflect the magnitude of the error term but rather the local component γ_i.

The difference between these two model equations for trend surfaces is reflected in the way in which autocorrelation among residuals is regarded. In polynomial regression analysis, autocorrelation is considered to be a violation of model as-

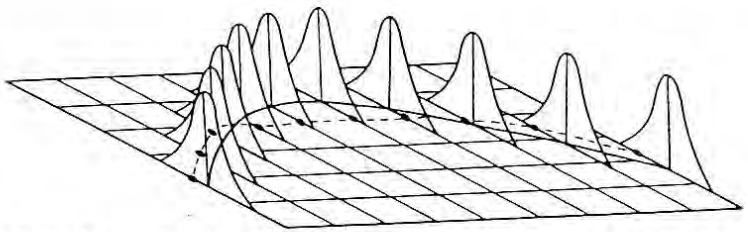


FIGURE 5.84 Distribution of error variance around the regression line according to polynomial regression model. Observations, indicated by dots, are assumed to lie within the variance distribution around the model regression line.

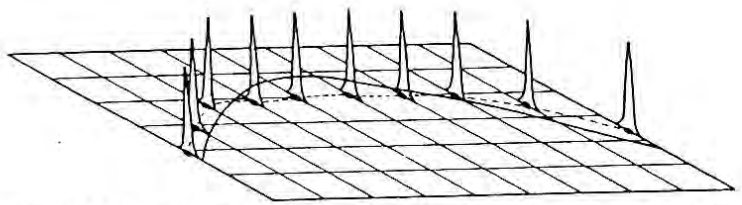


FIGURE 5.85 Distribution of error variance around observed surface according to structural model of trend surfaces. Errors in replications of points are centered about some mean value for each point, rather than around the polynomial model equation.

sumptions and is assumed to invalidate (or seriously weaken) inferences drawn from the analysis. Such an attitude may be appropriate for petrographic and geochemical data, because replication tends to reveal a normal distribution of error around relatively simple regressions. Although analyses for lack-of-fit may suggest more complex regressions, in general the error term is sufficiently large to account for all deviations.

In contrast, geologists using trend-surface analysis on structural data are seeking areas of autocorrelated residuals. As stated before, almost all structural deviations are ascribable to lack-of-fit, and the presence of autocorrelated residuals indicates a region larger than the sampling interval where the surface deviates from the polynomial model in a consistent direction. Either large areas of autocorrelated residuals or single points of large deviations are of interest in petroleum exploration, because they may indicate regions where local structures (γ_i) have a strong influence. Because the deviations are not randomly distributed, standard tests of significance of the regression cannot be used. This generally is not regarded as a drawback, as the regression itself is only of incidental interest.

Pitfalls

At this point, it may be appropriate to emphasize some factors which can adversely affect trend-surface analyses, or indeed, any type of map analysis. These cautionary statements have been repeated many times in the literature, but they seem to be ignored more times than not (Chayes, 1970). The effect of these factors can range from a slight distortion of the trend (a form of bias) to total invalidation of the results.

The first and most obvious point is that adequate control must be present. At an absolute minimum, the number of data points must exceed the number of coefficients in the polynomial equation or the results of the regression are invalid. If statistical tests are to be run, the number of control points determines the degree of freedom and these must be sufficiently large so a meaningful F test can be performed. If the degrees of freedom for deviation are small (a consequence of the number of polynomial coefficients approaching the number of data points),

only extremely high correlations will be found significant. Furthermore, the power of the test (the probability of not committing a type II error) decreases drastically with small sample sizes. Of course, the number and spacing of control points has a direct influence on the size of local deviations that can be detected in trend-surface analysis of structural data, and relates to the resolution with which these can be defined.

Ordinarily, we do not consider control points beyond the boundaries of our map. Often, the map area may even extend slightly beyond the actual lateral extent of the data points, there being few (or no) control points actually on the map boundaries. In such circumstances, there are almost no constraints on the form of the trend surface near the edges of the map. Whatever slope exists in the region of control is extrapolated without limits along the map boundaries. This creates what are called "edge effects." If a high-order trend is being fitted to data, extrapolated values near the edges of the map may reach astronomical proportions. Minor edge effects will exist even if the entire map area is uniformly covered with control points up to the boundary. Therefore, it is good practice to have data over an area in excess of the size of the area to be mapped. This forms a "buffer region" around the map in which edge effects are concentrated; the control points in this region constrain the form of the trend surface within the map area proper. The necessary width of the buffer depends primarily on the density of available control. If the map contains many control points, a narrow bordering strip will suffice. If control density is low, a much wider belt around the map will be necessary to absorb edge effects. Incidentally, it should be pointed out that edge effects are not unique to trend surfaces, but also occur in contour maps, moving-average surfaces, and other forms of fitted surfaces.

The arrangement of the data points within the map also has a pronounced effect on the form of the regression. The illustrations in Figure 5.86 are taken from an evaluation of the effect of data distributions on polynomial trend surfaces (Doveton and Parsley, 1970, pp. B198–B202). A series of points were placed at random locations on a basin-shaped surface, and a second-degree trend was calculated.

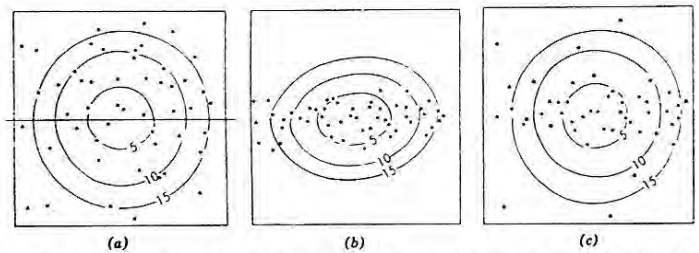


FIGURE 5.86 Influence of control-point distribution on trend surfaces. (*a*) Original test surface with randomly placed control points. (*b*) Distorted trend produced by sampling original test surface along a narrow band. (*c*) Nearly correct trend produced by adding a few outlying points to the narrow band of observations.

The regression equation was then used to calculate values of the dependent variable at points placed according to various sampling patterns. Ideally, surfaces fit to these points would be identical to the original form from which the data were derived. Figure 5.86*a* shows the surface produced by randomly distributed points; the fit is over 95% and the trend is essentially identical to the original model. In Figure 5.86*b*, however, the sample points are distributed along a narrow band. The fit of the surface is still good (93%), but the form of the regression has become badly distorted parallel to the sample pattern. Only a few control points outside the band are necessary to correct this bias, however, as can be seen in Figure 5.86*c* (also 93% fit). These tests demonstrate that the shape of the map area can seriously affect the form of a polynomial fit. If the data do not occupy an approximately equidimensional area, the trend surface will be elongated parellel to the pattern of points. Remember that these examples are for idealized models having no local or random component; distortions are more severe if small-scale "noise" is present.

Cautionary remarks about the deleterious effect of clustered data points on trend surfaces also have been published. Clustering or bunching of control points is especially apt to be troublesome in petroleum exploration by trend-surface methods, because wells are most abundant on and around known oil fields. These areas may exert an undue influence on the regional trend, although there is evidence that this effect is not as severe as sometimes feared (Doveton and Parsley, 1970, pp. B200–B201). Figure 5.87*a* shows a cubic trend-surface model used to generate data points for analysis of the clustering effect. Points were taken in various clustered patterns and were used in an attempt to recreate the original surface. Figure 5.87*b* shows a strongly clustered data-point distribution and the trend surface created.

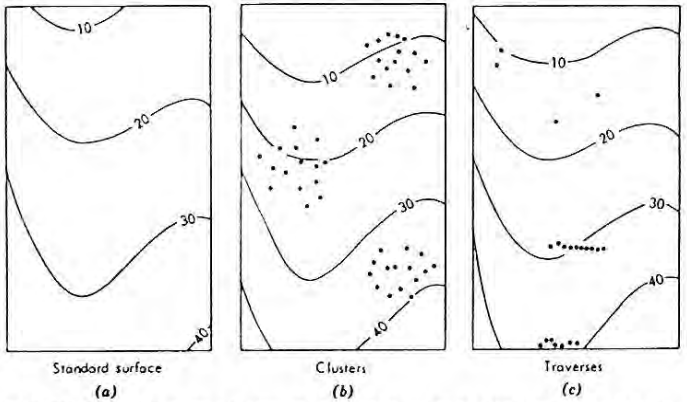


FIGURE 5.87 Influence of clustering of control points on trend surfaces. (*a*) Original third-degree test surface. (*b*) Trend surface obtained from control points placed in clusters on the test surface. (c) Trend surface obtained from control points located along traverses across parts of the test surface.

The surface accounts for 99% of the original variance. An even more radically clustered distribution is shown in Figure 5.87*c*, but the cubic surface accounts for 100% of the total variance. Both Figures 5.87*b* and 5.87*c* are essentially identical to the original surface. These experiments suggest that trend-surface methods may be more robust against effects of clustering than is commonly supposed. Again, it should be remembered that these experimental tests are essentially noise-free, and more severe distortions are to be expected in the presence of local variation.

● Comparison with other surface-fitting methods

A comparison of surface fitting algorithms for geophysical data

T. El Abbass[1], C. Jallouli[2], Y. Albouy[1] and M. Diament[2]
[1]*ORSTOM, Bondy, France;* [2]*Laboratoire de Géophysique, Université Paris Sud, Orsay, France.*

ABSTRACT

Many treatments in geophysics require regular grids of data. Since the data are generally recorded irregularly (e.g. gravity measurements along roads) or even on tracks (e.g. satellite measurements), it is necessary to grid the observed data.

We present the result of a comparison of various surface fitting algorithms carried out in order to check their reliability. Two different types of sampling have been verified: (i) clouds of points and (ii) points along tracks. Five algorithms have been extensively tested: (1) polynomial fit, (2) algorithm using a combination of spline–laplace, (3) krigeage (4) least-squares fitting method, and (5) finite element method. The suitability of each method for different sets of data and the limitations (in terms of amplitude and gradient) are discussed.

Terra Nova, 2, 467–475

INTRODUCTION

Geophysical data are generally recorded with a rather random distribution that needs interpolation in order to compute a regular grid. Regular grids are required either for representation purposes or for further treatments. The development of small computers has allowed the interpolation procedure to be easily performed and therefore a number of workers in earth sciences, not only in geophysics, now have to create regular grids from irregularly spaced data, for which, various algorithms have been proposed. Bhattacharyya (1969) applied the bicubic spline interpolation to geophysical data. Sandwell (1987) applied biharmonic spline interpolation in SEASAT Altimeter Data. A method which is based on the theory of random variables is described by Matheron (1970) and Olea (1974). Hayes and Halliday (1974) and Dierckx (1980) apply a least-squares fitting method of cubic spline surface.

A new method, described by Inoue (1986), is based on the definition for optimum function minimizing the L2 norm composed of the data residuals and the first and second derivative. The optimum function is approximated by a cubic B-spline expansion.

Many methods for solving smoothing or interpolating problems have been proposed – see reviews by Crain (1970), Brodlie (1980), Mclain (1980), and Lancaster and Salkauskas (1986).

We present here the results of a comparison of various methods which are commonly used. Our aim is to give some hints as to the most suitable method according to the shape of the data (important extrema or important gradients) and to their geographical distribution (random or along tracks or roads). For that purpose, we choose to study several algorithms for interpolating: (1) polynomial fitting (La Porte, 1962); (2) a combination of spline and laplace methods (plot 88); (3) kriging (Matheron, 1970); (4) least-squares fitting of bicubic spline surfaces (Dierckx, 1980); and (5) a finite-element approach using a cubic B-spline basis (Inoue, 1986).

We used theoretical data for testing these different algorithms. However, geophysical data were treated by the same algorithms for constructing the map of a Bouguer anomaly. All algorithms were tested on a personal computer (IBM compatible).

PRESENTATION OF THE ALGORITHMS CHECKED

Polynomial method

A well-known algorithm for fitting data on a surface using polynomials was proposed by La Porte (1962) in order to produce isovalue maps from a set of data using computers. For each point where an interpolated value must be computed (i.e. the 'nodes' of the grids), a fit of a polynomial surface to the neighbouring data is performed using a least-squares approach. Therefore the coefficients of the polynomial surface are computed and the interpolated value is deduced. The data on which the fit is performed are located inside a circle with a variable diameter. Additionally, each datum is weighted so that more weight is given to the closest datum to the interpolated point.

The weight is taken as (La Porte, 1962):

$$P_i = [(R^2 - D_i^2)/(D_i^2 + n^2)]^2$$

where R is the radius of the interpolating circle, D_i is the distance between the i data and the point to estimate, i.e. the centre of the circle, and n is a small quantity such that P_i is defined if D_i equals zero. With larger values of n, the polynomial surface will be smoother.

Once n and R are fixed, in order both to avoid an over smoothing and to take into account enough data to compute the polynomial coefficients, it appears that this interpolation method depends upon the degree of the polynomial and upon the repartition of the data. Note that if the degree of the polynomial is increased, the number of data points

needed is also increased and therefore R must be enlarged.

We used a program (Albouy and Godivier, 1981) strictly following the algorithm given by La Porte (1962).

The spline–Laplacian method

This interpolation algorithm is a combination of a Laplacian and of a spline interpolation. The equation used is:

$$[\delta^2X(z)+\delta^2Y(z)]+C[\delta^4X(z)+\delta^4Y(z)] = 0$$

The first bracket corresponds to the Laplacian interpolation and the second one to the spline interpolation. Therefore according to the value of C the computed surface is more or less controlled by one interpolation or another. For example, the computed surface is more controlled by spline interpolation for high values of C.

It must be noted that the estimated surface has to fit the observed points. An interpolation function is computed between each two data points from their values, the slope and the curvature.

We used the Plot 88 library software (Young and Van, 1987).

The kriging method

The kriging method, developed by G. Matheron and his school, is based on geostatistics. To estimate the value Z of one variable at a point M, this method uses an estimator:

$$Z^* = \sum_{i=1}^{n} P_iZ_i \text{ with } \sum_{i=1}^{n} P_i = 1$$

where n is the number of data, Z_i is the measured values, and P_i is the weight.

The weights P_i are not a function of distance from the data points to the estimated point only, but depend on the spatial continuity of the variable. The spacial continuity is defined by a function called a variogram which is used in the determination of the weighting factors P_i.

The variogram function determines the 'variance of estimate'. For a given length (h) and direction (M_1M_2), the variogram is defined by the following expression:

$$\gamma(h) = 1/2[\text{variance } (Z_2-Z_1)]$$

and experimentally,

$$\gamma(h) = 1/2N_h \Sigma (Z_2-Z_1)^2,$$

where N_h is the number of vectors $h = \overrightarrow{M_1M_2}$.

We compute the experimental variogram and we determine a theoretical model of $\gamma(h)$ which is a linear function of the distance.

Krigeage minimizes the variance of estimate σ_E^2 by differentiation with respect to weights P_i. The variance of estimate is given by the following expression:

$$\sigma_E^2 = 2\Sigma P_i\gamma(\overrightarrow{MM_i})-\Sigma P_iP_j\gamma(\overrightarrow{M_iM_j})$$

The minimum variance of krigeage gives an idea of the error made.

Finally, we note that this method is detailed by Matheron (1970); Guillaume (1970) and Haas and Viallix (1976) applied it in order to solve the problem of estimates and contouring. Further details of variograms are given in Delhomme (1976).

A least-squares fitting method using a cubic b-spline

Hayes and Halliday (1974) describe a computational method for fitting bicubic spline by least-squares for arbitrary data and Dierckx (1977) for data on a rectangular grid. However, automatic determination of knots is not considered, and because the goodness of fit with a least-squares spline function is considerably affected by the number of knots and their position, Dierckx (1980) developed this method where, the number of knots and their positions are determined automatically.

The following problem is considered. Given the function values f_r at some point $s(x_r, y_r)$, $r = 1, \ldots, m$ scattered arbitrarily in the domain D with positive weights w_r, $r = 1, \ldots, m$, we determine a spline function $s(x, y)$ while trying to find a compromise between the following objectives:

(i) The prescribed value f_r should be closely enough.
(ii) The approximating spline should be smooth enough in sense that the discontinuities in its derivatives are as small as possible.

The spline is given in the b-spline representation:

$$s(x, y) = \sum_{i=1}^{\lambda-4} \sum_{j=1}^{\mu-4} C_{ij}M_i(x)N_j(y)$$

where $M_i(x)$, $N_j(y)$ are normalized cubic b-spline (De Boor, 1972) and the coefficient C_{ij} is determined as the solution of the following constrained minimization problem:

Minimize η Subject to the constraint $\varepsilon_r^2 \leq S$,

where η is a measure of the lack of smoothness of $s(x, y)$, ε_r denotes the weighted residual $w_r(f_r-s(x, y))$, and S is a non-negative number which controls the extent of smoothing and is therefore called the 'smoothing factor'. (This parameter is discussed by Reinsch (1967) and Dierckx (1980, 1981).

We use a subroutine from NAG library (Numerical Algorithm Group, 1988).

The finite-element approach method

An algorithm using the finite element approach method was recently proposed by Inoue (1986). This method is used to compute a surface using cubic b-spline expansion and to minimize the misfit, the fluctuations and the roughness of the function. As mentioned by Inoue (1986), the solution so computed represents the deformation of a thin elastic plate under tension. The plate being pulled laterally toward datum points by springs whose stiffnesses are inversely proportional to the variances of the data.

Indeed, Inoue proposes to minimize the following function:

$$\Pi = \sum_{p=1}^{n} w_p(\Phi_p-d_p)^2 + 1/l_u^2 \iint [W_1(\Phi_x^2+\Phi_y^2) + W_2(\Phi_{xx}^2+2\Phi_{xy}^2+\Phi_{yy}^2)]dxdy$$

with $\Phi_x^2 = [\Phi_x'(x, y)]^2$

$\Phi_{xx}^2 = [\Phi_{xx}''(x, y)]^2$

$W_1 = \tau(l_u/d_u)^2/\rho$

$W_2 = (1-\tau)(l_u^2/d_u)^2/\rho$

$$\Phi(x, y) = \sum_{i=1}^{M_x-3} \sum_{j=1}^{M_y-3} C_{ij}F_i(x)G_j(y)$$

where $F_i(x)$ and $G_j(y)$ are cubic b-spline bases, w_p are the weights of the data d_p, l_u is the unit length, d_u is the unit observable, ρ is a roughness parameter. It controls the tradeoff between the fitness

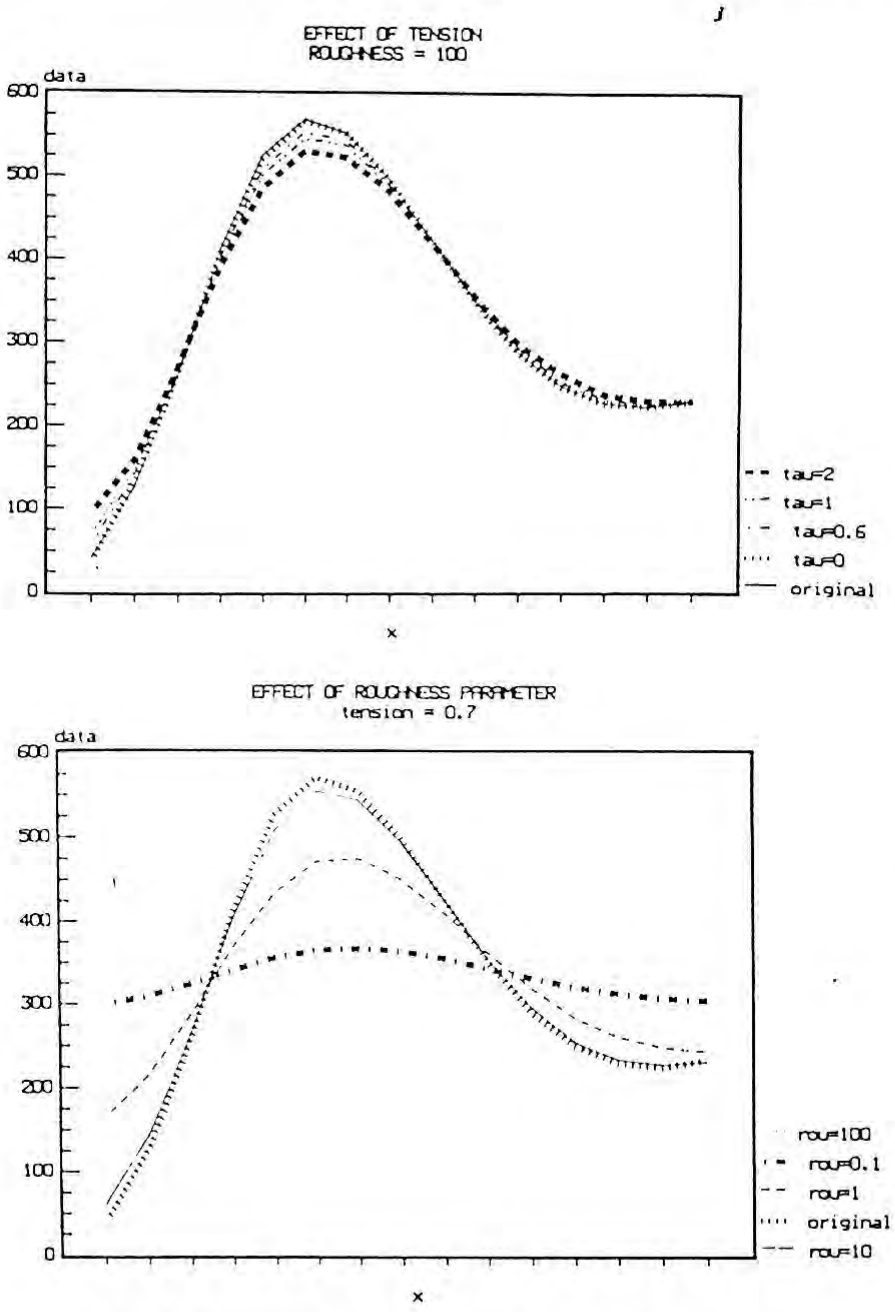


Fig. 1. *The effect of roughness and tension in the finite-element method.*

to data and the total smoothness. τ is a tension parameter. It controls the nature of the function by varying from a plate free from tension ($\tau = 0$) to a membrane under tension ($\tau = 1$).

These two parameters are discussed by Inoue (1986). In Fig. 1 we show the roll of each one when applied to geophysical data (we remark here that with a low value of roughness the calculated curve doesn't fit the observed data). The interpolation is controlled also by the number of division parameters (M_x and M_y). These parameters depend on the data distribution.

So three parameters control the interpolation: the roughness, the tension and the number of divisions of the domain.

We used the program given by Inoue (1986) in his paper.

PRINCIPLE OF OUR STUDY

Our approach to test various algorithms of surface fitting is shown in Fig. 2. We first choose a synthetic surface Fig. 3 which presents both high extrema and important slopes, these are difficult features to interpolate but ones that are often observed in the field. Indeed this surface is defined as the composition of three functions as:

$$f(x, y) = Aexp(ax^2+by^2+cx+dy+e)^n$$

The resulting function is computed over a square area.

Then a distribution of points is taken out of that synthetic surface and so defines our set of data. We interpolate this set of data in order to obtain a regular grid and we finally compare the computed grid with the one having the same spacing deduced *directly* from the synthetic surface. We can quantify the quality of the interpolation by computing the residual grid.

Several distributions have been studied: random distribution with various densities for the data and distribution along profiles varying the distances between the profiles. These two kinds of distribution were chosen in order to represent two different types of data, either data on land such as those recorded during gravity survey, or data obtained at sea or by satellites.

RESULTS AND DISCUSSION FROM A THEORETICAL EXAMPLE

Polynomial fit

As previously mentioned, the interpolation using the polynomial fit depends upon the degree of the polynomial, the interpolation radius and the distribution of experimental data.

Figure 4 shows residuals obtained for the third and four degrees. We show the results obtained with a random distribution of 400 data points and with a radius of a half of the diagonal. As it could be expected, the residual is smaller when the degree is higher since, for example, degree 3 gives a residual of about 5% (Fig. 4a), while degree 4 gives 0.5% (Fig. 4b). Similarly, it is obvious that the fit is improved with a smaller radius, displaying a linear fit with a radius of a fourth and an eighth of the diagonal (Fig. 5a and Fig. 5b, respectively).

We note that this method doesn't represent the extreme very well.

Concerning the profile distribution, it

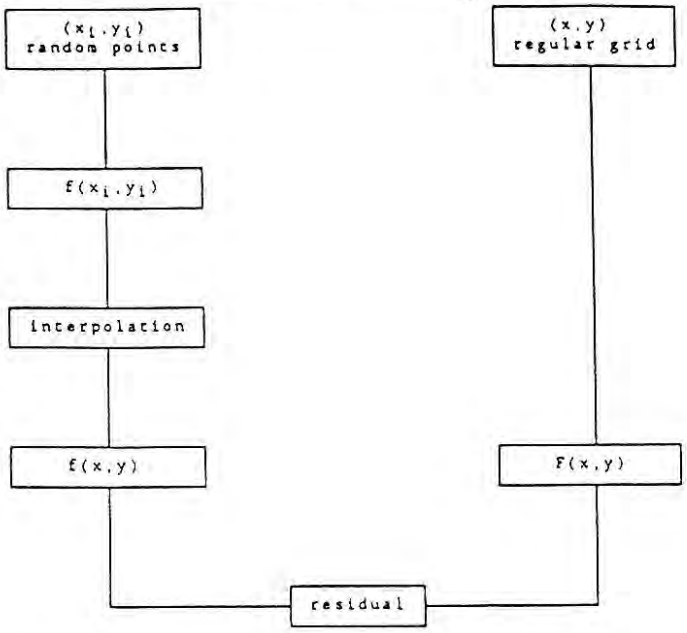


Fig. 2. *The approach used to test the various methods of interpolation.*

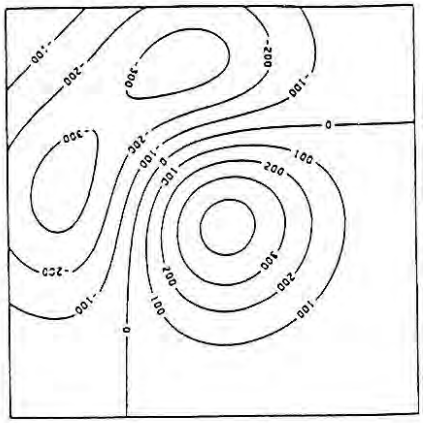


Fig. 3. *The synthetic surface computed from the regular grid. The amplitude of this surface is about 800 units.*

signal (Fig. 7), the same range as that obtained with an identical distribution of points with a linear polynomial interpolation. It must also be noted that this method always provides poor fits in the areas of pronounced slopes.

appears that the residual is controlled by the spacing between the profiles and by their orientation in respect to the features of our synthetic data as expected (Fig. 6a). But if the profiles have larger distances between them, the interpolation creates an artificial feature parallel to the profiles when the distance between them is greater than the wavelength inspected (Fig. 6b) thus there is an aliasing effect in the space domain.

It must be remarked that this much used method gives reasonably good results with a high degree of accuracy and a small interpolation radius. Nevertheless it is rather slow and respects the slopes but does not represent well the extremes.

Spline–Laplacian method

We first studied the influence of the parameter C which controls the ratio of spline and Laplacian interpolation. The main variations are in the area with large slopes but it appears that this parameter does not act in an important way. Therefore, we kept a value of C = 3, i.e. a mean value, such that both spline and Laplacian interpolations act for the following tests.

Using a random distribution with 400 points, the residual is of about 5% of the

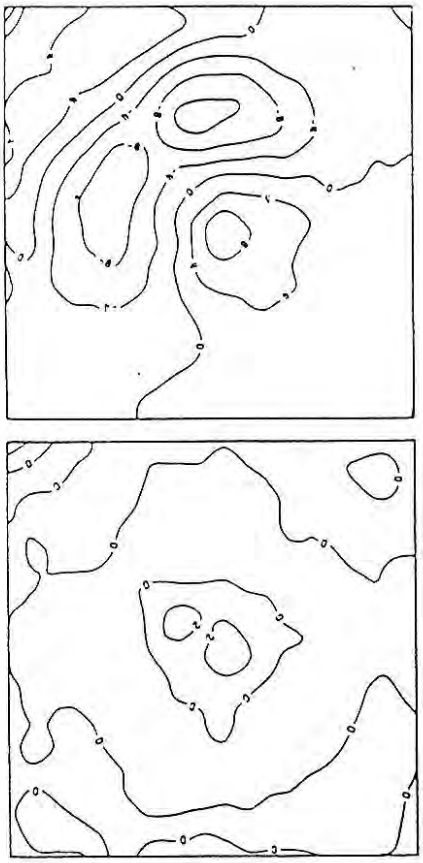


Fig. 4. *Two maps of residuals obtained by polynomial interpolation from 400 data points show the effect of degree of the polynome. (a) (top) is obtained by a third degree polynomial interpolation; (b) is obtained by fourth degree.*

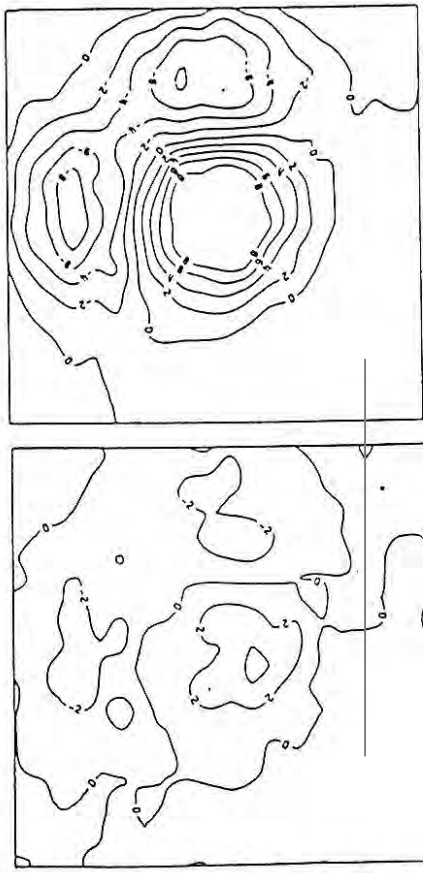


Fig. 5. *Two maps of residuals obtained by polynomial interpolation show the effect of radius parameter. (a) (top) and (b) are obtained with a radius of a fourth and an eighth of the diagonal, respectively.*

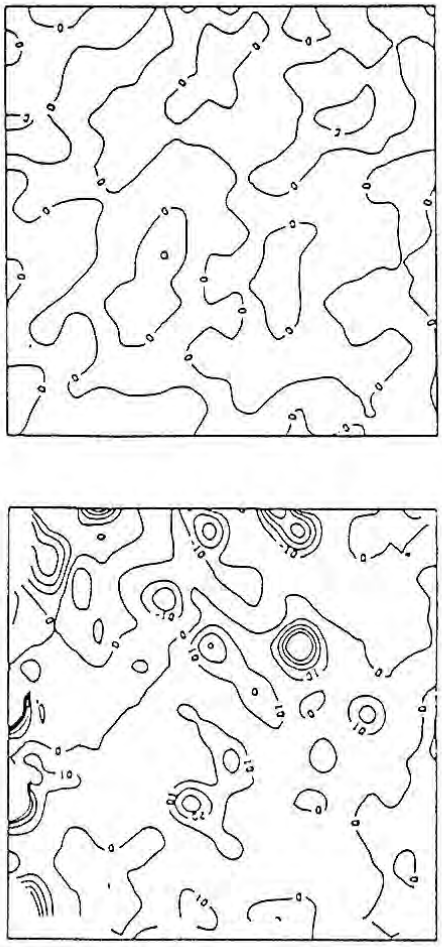


Fig. 6. *Residuals obtained by polynomial interpolation from profile distribution of points. The only difference between (a) (top) and (b) is the spacing of profiles which is wider in (b).*

With data along tracks, the results are always less accurate than with a random distribution of data.

In conclusion, although this method is quite fast, it seems to provide poor results and to be dependent on the data distribution. Nevertheless, it can be used for fast representations of data sets.

Kriging method

The first step consists in computing a variogram. Figure 8 shows the computed variogram from a random distribution of 400 data points. Interpretation of the experimental variogram is used to fix a linear model of the variogram function of distances. The first branch of the experimental variogram appears to be very linear which is logical according to the fact that data are synthetic.

Results of the interpolation are shown in Fig. 9. The high residual values are shown on the maxima and minima, so this method, like the polynomial method, doesn't represent well the extremas. Nevertheless, it requires a lot of computer time.

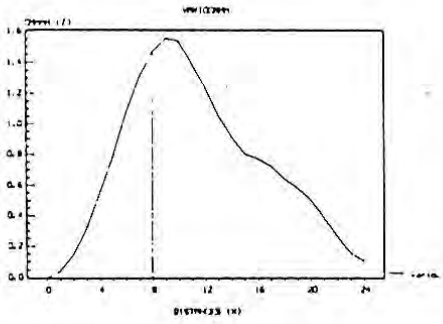


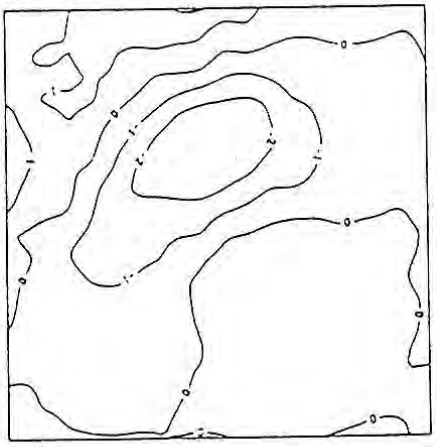
Fig. 8. *Variogram computed from random distribution of 400 data points.*

A least-squares of surfaces fitting

In this method we control the smoothness of the fit with the smoothing factor S.

We studied the influence of S in a random distribution for a set of 1600 points with S = 50; Fig. 10a shows a residual about 0.5% over the slope. In Fig. 10b a residual is obtained with S = 4000; the high value of residual is over the extremas and it is about 1% of the signal.

Figure 11a shows a residual obtained for 400 points with S = 50. We have the same results for 1600 points, but with S = 4000 the residual is about 2%



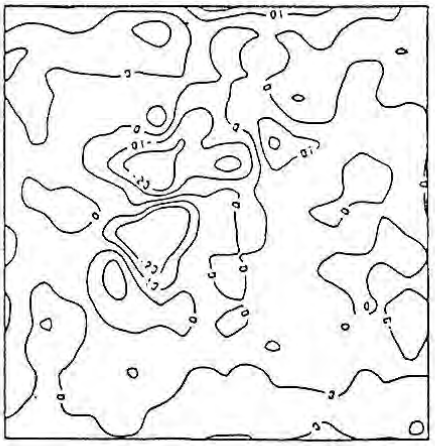


Fig. 7. *Residual maps obtained by spline–laplacian method. The residual presents a high value in area of the pronounced slope.*

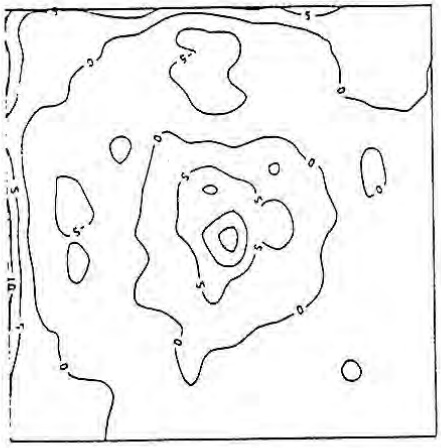


Fig. 9. *Residual obtained with Kriging method. High values in residual map are observed on the maxima and the minima of the surface.*

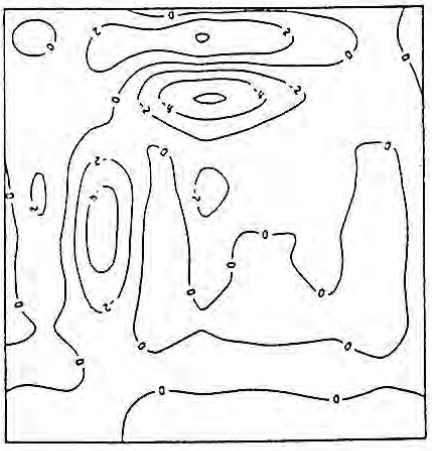
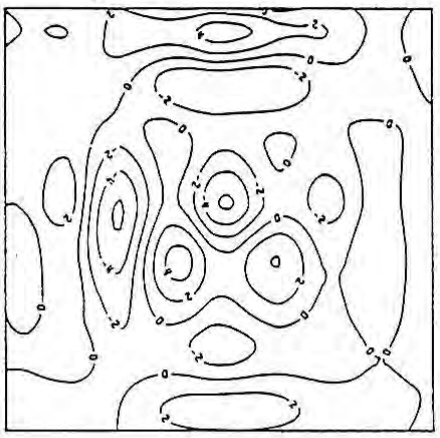
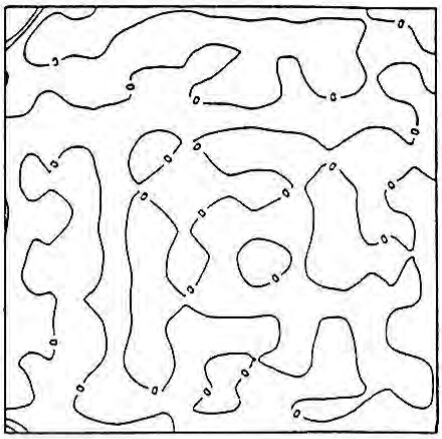
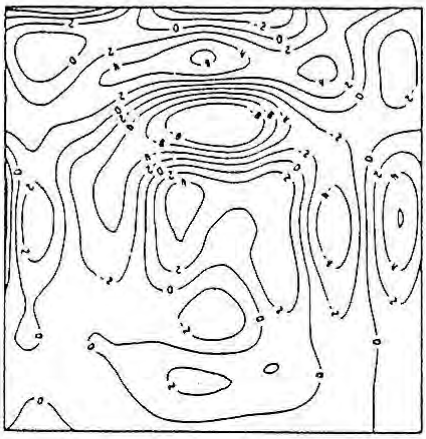


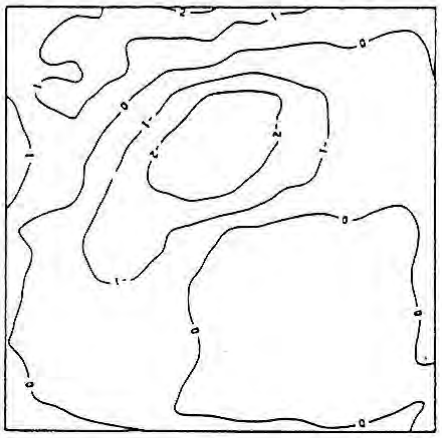
Fig. 10. *Residuals obtained by a least-squares surface fitting from 1600 points. (a) (top) the result with a smoothing factor S = 50; (b) the result with a smoothing factor S = 4000.*











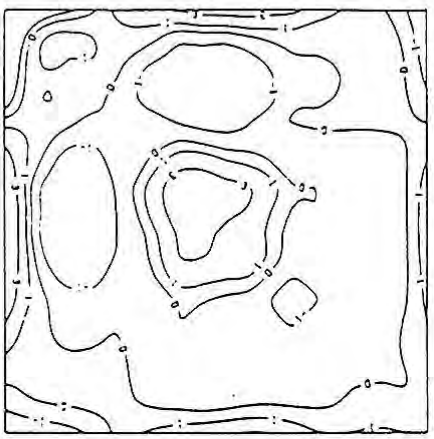


Fig. 11. *Residuals obtained by a least-squares surface fitting from 400 points. (a) the result with smoothing factor S = 50; (b) the result with smoothing factor S = 4000.*

Fig. 12. *Residuals obtained by a least-squares surface-fitting along tracks. (a) the result with a smoothing factor S = 4000; (b) the result with a smoothing factor S = 50.*

Fig. 13. *Two residuals shows the roughness effect. (a) and (b) are the results of roughness parameters of 1000 and 10, respectively. Tension parameter is fixed at 0.*

over the minimum and the maximum (Fig. 11b).

Using a distribution along tracks, the residuals show that the results are not controlled by the orientation of the tracks. With $S = 4000$, residual values are high over the extremas (Fig. 12a), and with $S = 50$ the low residual values are above the slope (Fig. 12b).

This method does not represent the extremas well when high values of S are used but gives good results – small residuals with small S parameter. However, this advantage for theoretical data is not confirmed if the method is applied to geophysical data (see later).

Finite-element approach method

As previously mentioned, three parameters control the quality of the interpolation, and their relative influence can be depicted. Figure 13 shows that the value of the roughness must be as high as possible in order to get the best fit. Figure 14 shows that the best fit is obtained with low tension values.

Concerning the number of divisions, it appears that it is closely related to the data distribution: the number of divisions must be increased when the grid interval is decreased. Also, as could be expected, the increase in the number of divisions yields better results; too high roughness and a great number of divisions have no influence on the results but do increase computer time.

This method gives good results, i.e. small residuals, even if data are distributed along tracks (Fig. 15). Moreover, it is a fast algorithm.

RESULTS AND DISCUSSION FROM A GEOPHYSICAL EXAMPLE

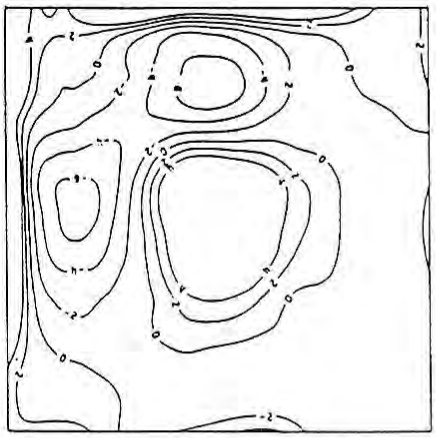
We have applied all these methods in gravity measurements to construct contour maps of the Bougeur anomaly.

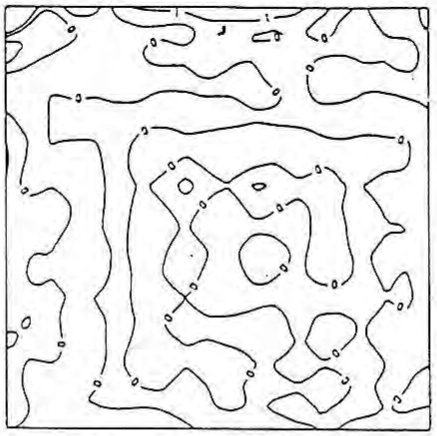
Figure 16 shows the position of the gravity measurements over a rectangle of 330×330 km containing 394 data points.

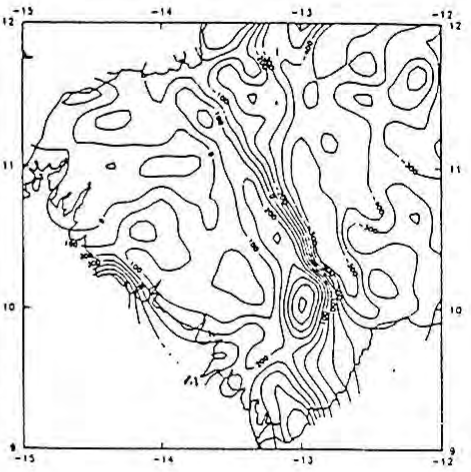
The results show that the quality of maps is not good with the spline-laplace methods (17a and b) as artificial anomalies are created.

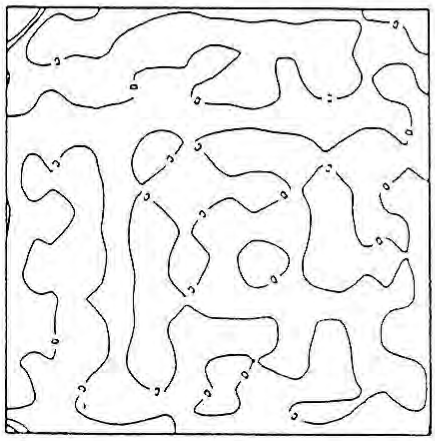
With the polynomial method the results are not satisfactory over the boundary domain where there are few data (Fig. 18) although the slope is respected.

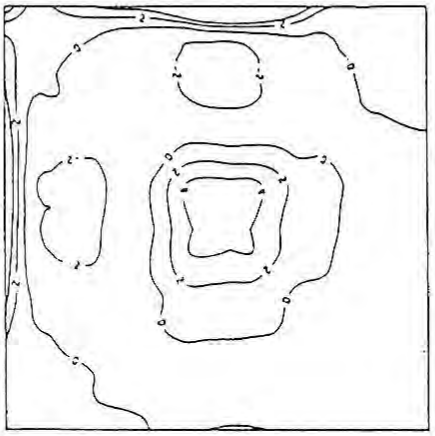
The maps obtained with the kriging method (Fig. 19) and finite element











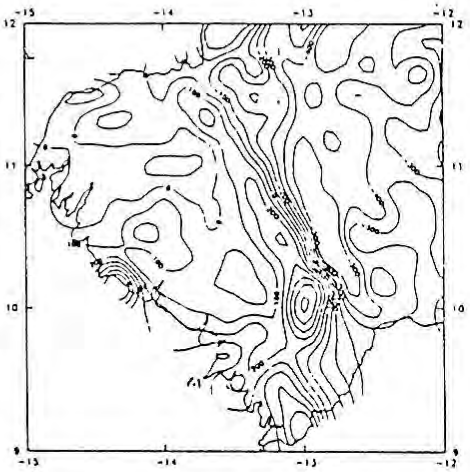


Fig. 14. *These residuals show the effect of the tension parameter. (a) (top) and (b) are the results of interpolation with tension parameters equal to 0 and 1, respectively. Roughness parameter is fixed at 1000.*

Fig. 15. *Two residuals obtained along tracks. (a) (top) is the result with roughness equal to 1000 and tension equal 0; (b) is the result with roughness equal to 1000 and tension equal to 1.*

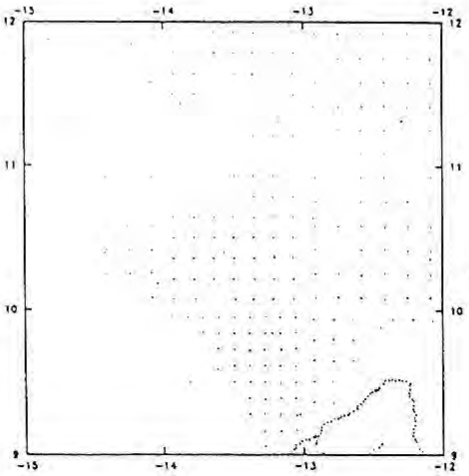
Fig. 17. *Contour maps (a) (top) with the spline method; (b) with Laplacian method. Results are not very satisfactory as they are too rough.*

approach method (Fig. 20a) are good, showing minimum curvature with respect to extremas and the slope. In addition this method gives good results in areas with limited data.

Figures 20b, c and d show the role of the three essential parameters: roughness ρ, tension τ, and the number of divisions M. They give many solutions according to the desired objective and the data application.

The results show that the roughness ρ and the number of divisions M should be determined so that the fitted function extracts the true distribution from the noisy data as efficiently as possible.

Finally, Fig. 21 shows contour maps according to different S values. Choosing S too high (S = 4000000) results in oversmoothed data (Fig. 21a). Overfitting the data results from choosing S too



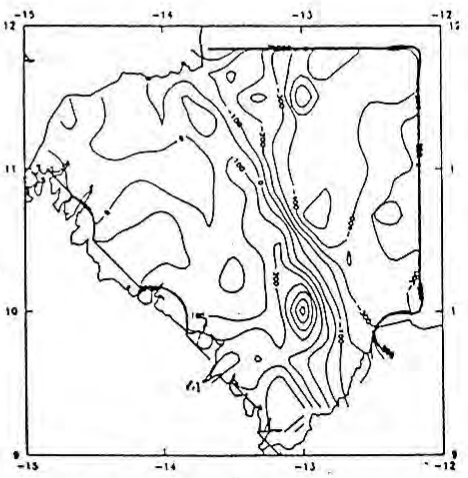


Fig. 16. *The position of the gravity measurements over the studied area.*

Fig. 18. *The contour map with a polynomial third degree: Result unsatisfactory near the boundaries.*

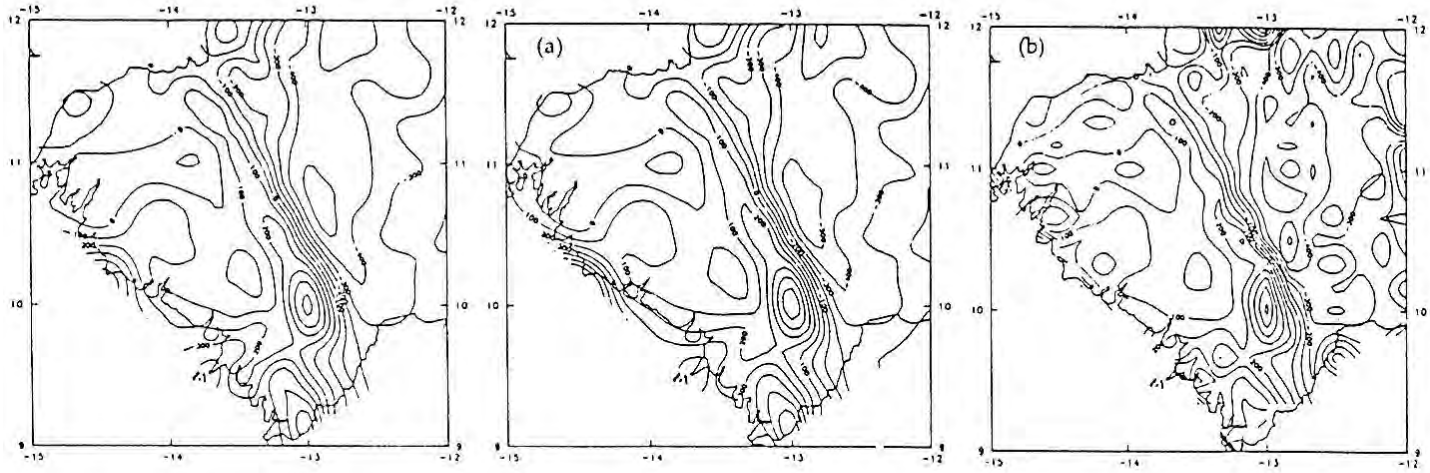


Fig. 19. *The countour map with Kriging method: A linear variogram is applied.*

small (S = 100000) (Fig. 21b). Figure 21c shows the contour map with S = 400000, which could be accepted.

The result demonstrates that the fit produced by this method is unreliable in a region with few data. The unfavourable situation is accentuated if S becomes smaller.

CONCLUSION

Our tests confirm that the interpolation technique, although widely used, is a very sensitive method and that careful attention must be paid to it. From the results presented here it is obvious that no method is universal.

Indeed the choice must be made according to the distribution of points and to the quality of interpolation needed i.e. it is more important to fit extrema or slopes, and, finally, to the computer time needed.

The polynomial method appears to be an efficient method to fit slopes if a high degree is used but can be very slow.

The spline–Laplacian method gives smoothed maps quite fast, but is far from being the best.

The kriging method is obviously a powerful one but it is heavy and expensive in calculation time.

The least-squares surface fitting method is not easy to use, because in practice this method is used with considerable interaction between the user and the computer via a graphics terminal, the reason being that the data are sensitive to the parameter S which has to be chosen carefully: too small an S value will result in an overfitting; too large an S value results in an underfitting of data. Therefore, the user should always control the fit graphically before accepting it as satisfactory. Finally it should be mentioned that the value S is not very crucial and also depends on the number of data points and the application.

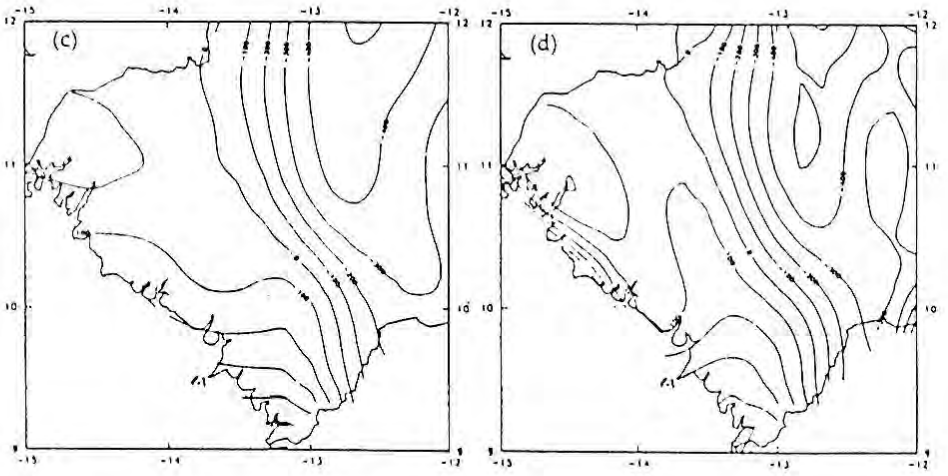


Fig. 20. *The results of contouring maps with the finite element method. (a) p = 2000, τ = 0, and the number of divisions is 19×19; (b) p = 2000, τ = 1 and the number of divisions is 19×19; (c) p = 10, τ = 0 and the number of divisions is 19×19; (d) p = 2000, τ = 0 and the number of divisions is 4×4.*

The finite-element method is reasonably fast, easy to use, even if the first choice of parameters is quite subjective, and gives a good fit, so it appears to be a good choice. In addition, including the generalized tension parameter this method has more applications in order to find a suitable solution for gridding data when these data vary rapidly with distance (Smith and Wessel, 1990), e.g. using a tension parameter is good for bathymetry data.

Both kriging and finite element methods provide an estimation of the accuracy of the interpolation, and take the standard errors into account.

As previously mentioned, it is obvious that whichever interpolation algorithm is used, data would be predicted only where no measurement has been made. Therefore, interpolated grids must always be checked with respect to the original data so that possibly artificial features introduced during the interpolation may be removed before interpretation.

ACKNOWLEDGEMENTS

We thank C.I.R.C.E. (Centre Inter-Régionale de Calcul Electronique) for

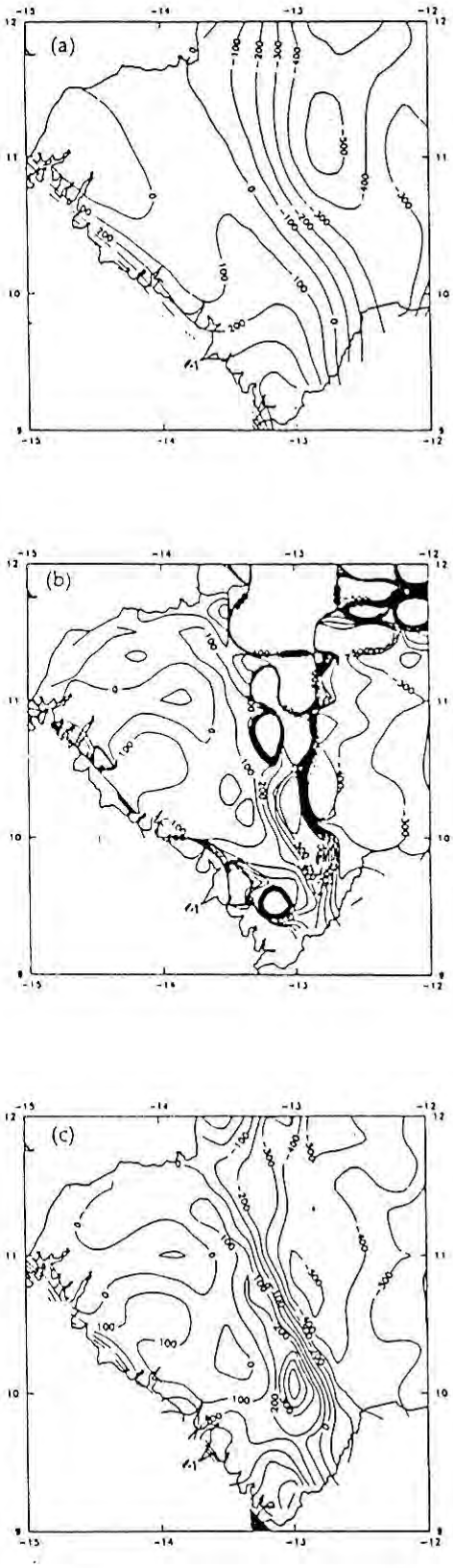


Fig. 21. *The contour maps with a least-squares surface fitting. (a) Underfitting with S = 4000000; (b) Overfitting with S = 100000; (c) Satisfactory fit with S = 400000.*

permitting us to use NAG library. Critical comments by anonymous reviewers were very helpful.

REFERENCES

Albouy Y. and Godivier R. (1981) Cartes gravimétriques de RCA, *Notice explicative*, 90, 00–00. ORSTOM, Paris.

Bhattacharyya B.K. (1969) Bicubic spline interpolation as a method for treatment of potential field data, *Geophysics*, 34, 402–423.

Boor de C. (1972) On calculating with B-splines, *J. Approx. Theory*, 6, 50–62.

Brodlie K.W. (1980) A review of methods for curve and function drawing. In: *Mathematical Methods in Computer Graphics and Design* (ed. by K.W. Brodlie), pp. 1–37. Academic Press, London.

Crain I.K. (1970) Computer Interpolation and Contouring of Two Dimensional Data: a review, *Geoexploration*, 8, 71–86.

Delhomme J.P. (1976) *Applications de la théorie des Variables Régionalisées dans les Sciences de l'eau. Thèse de Docteur Ingénieur*, Université Pierre et Marie Curie, Paris, 160 pp.

Dierckx P. (1977) An algorithm for least-squares fitting of cubic spline surfaces to functions on a rectilinear mesh over a rectangle (ALGORITHM 008), *J. Comp. Appl. Maths*, 3, 113–129.

Dierckx P. (1980) An algorithm for surface fitting with spline functions, *Report TW50, Applied Maths & Progr. Div.*, K.U. Leuven, Belgium, 36 pp.

Dierckx P. (1981) An improved algorithm for curve fitting with spline functions, *Report TW54, Dept. Computer Science*, K.U. Leuven, Belgium, 41 pp.

Guillaume A. (1970) Analyse des variables Régionalisées, Traitement du Signal en Sciences de la Terre, *Diffusion DOIN éditeur*, Paris, France, 180 pp.

Haas A.G. and Viallix J.R. (1976) Krigeage applied to Geophysics: The answer to the problem of estimates and contouring, *Geophys. Prospect.*, 24, 49–69.

Hayes J.G. and Halliday J. (1974) The Least-squares Fitting of Cubic Spline Surfaces to General Data Sets, *J. Inst. Maths Appls.*, 14, 89–103.

Inoue H. (1986) A least-squares smooth fitting for irregularly spaced data: Finite-element approach using the cubic B-spline basis, *Geophysics*, 51, 2051–2060.

Lancaster P. and Salkauskas K. (1986) *Curve and Surface Fitting: An Introduction*, Academic Press, London, 280 pp.

La Porte M. (1962) Elaboration Rapide de Cartes Gravimetriques Déduite de l'Anomalie de Bouguer à l'aide d'une Calculatrice Electronique, *Geophys. Prospect.*, X, 238–257.

Matheron G. (1970) *La Théorie des Variables Régionalisées, et ses Applications*. Les cahiers du centre de morphologie mathématique de Fontainebleau, Fascicule 5.

Mclain D.H. (1980) Interpolation methods for erroneous data. In: *Mathematical Methods in Computer Graphics and Design* (ed. by Brodlie K.W.), pp. 87–104. Academic Press, London.

Numerical Algorithms Group Limited (1988) *The NAG Fortran Library Manual—Mark 13*. NAG, Oxford.

Olea R.A. (1974) Optimal contour mapping using Universal kriging, *J. geophys. Res.*, 79, 695–702.

Plot88 (1987) *Software Library Reference Manual*. Plotworks Inc., T.L. Young and M.L. Van Woert, PhD.

Reinsch C.H. (1967) Smoothing by Spline Functions, *Num. Math.*, 10, 177–183.

Sandwell D.T. (1987) Biharmonic Spline Interpolation of GEOS-3 and SEASAT Altimeter Data, *Geophys. Res. Letts*, 14, 139–142.

Smith W.H.F. and Wessel P. (1990) Gridding with continuous curvature spline in tension, *Geophysics*, 55, 293–305.

Received 6 June 1990;
accepted 5 September 1990

Statistical tables

Sources:

Fisher, R.A. & Yates, F. (1948) *Statistical Tables for Biological, Agricultural and Medical Research* Edition, Oliver and Boyd, Edinburgh.

Hald, A. (1952). *Statistical Tables and Formulas.* John Wiley & Sons, New York.

Kellaway, F.W. (Ed.) (1968). *The Penguin-Honeywell Book of Tables.* Honeywell Controls Ltd. (E.D Division), London.

Table 1A

Table 1A. Cumulative standard normal distribution.

CUMULATIVE STANDARD NORMAL DISTRIBUTION*

z	P	z	P	z	P	z	P	z	P
−4.00	0.00003	−3.50	0.00023	−3.00	0.00135	−2.50	0.00621	−2.00	0.02275
−3.99	0.00003	−3.49	0.00024	−2.99	0.00139	−2.49	0.00639	−1.99	0.02330
−3.98	0.00003	−3.48	0.00025	−2.98	0.00144	−2.48	0.00657	−1.98	0.02385
−3.97	0.00004	−3.47	0.00026	−2.97	0.00149	−2.47	0.00676	−1.97	0.02442
−3.96	0.00004	−3.46	0.00027	−2.96	0.00154	−2.46	0.00695	−1.96	0.02500
−3.95	0.00004	−3.45	0.00028	−2.95	0.00159	−2.45	0.00714	−1.95	0.02559
−3.94	0.00004	−3.44	0.00029	−2.94	0.00164	−2.44	0.00734	−1.94	0.02619
−3.93	0.00004	−3.43	0.00030	−2.93	0.00169	−2.43	0.00755	−1.93	0.02680
−3.92	0.00004	−3.42	0.00031	−2.92	0.00175	−2.42	0.00776	−1.92	0.02743
−3.91	0.00005	−3.41	0.00032	−2.91	0.00181	−2.41	0.00798	−1.91	0.02807
−3.90	0.00005	−3.40	0.00034	−2.90	0.00187	−2.40	0.00820	−1.90	0.02872
−3.89	0.00005	−3.39	0.00035	−2.89	0.00193	−2.39	0.00842	−1.89	0.02938
−3.88	0.00005	−3.38	0.00036	−2.88	0.00199	−2.38	0.00866	−1.88	0.03005
−3.87	0.00005	−3.37	0.00038	−2.87	0.00205	−2.37	0.00889	−1.87	0.03074
−3.86	0.00006	−3.36	0.00039	−2.86	0.00212	−2.36	0.00914	−1.86	0.03144
−3.85	0.00006	−3.35	0.00040	−2.85	0.00219	−2.35	0.00939	−1.85	0.03216
−3.84	0.00006	−3.34	0.00042	−2.84	0.00226	−2.34	0.00964	−1.84	0.03288
−3.83	0.00006	−3.33	0.00043	−2.83	0.00233	−2.33	0.00990	−1.83	0.03362
−3.82	0.00007	−3.32	0.00045	−2.82	0.00240	−2.32	0.01017	−1.82	0.03438
−3.81	0.00007	−3.31	0.00047	−2.81	0.00248	−2.31	0.01044	−1.81	0.03515
−3.80	0.00007	−3.30	0.00048	−2.80	0.00256	−2.30	0.01072	−1.80	0.03593
−3.79	0.00008	−3.29	0.00050	−2.79	0.00264	−2.29	0.01101	−1.79	0.03673
−3.78	0.00008	−3.28	0.00052	−2.78	0.00272	−2.28	0.01130	−1.78	0.03754
−3.77	0.00008	−3.27	0.00054	−2.77	0.00280	−2.27	0.01160	−1.77	0.03836
−3.76	0.00008	−3.26	0.00056	−2.76	0.00289	−2.26	0.01191	−1.76	0.03920
−3.75	0.00009	−3.25	0.00058	−2.75	0.00298	−2.25	0.01222	−1.75	0.04006
−3.74	0.00009	−3.24	0.00060	−2.74	0.00307	−2.24	0.01255	−1.74	0.04093
−3.73	0.00010	−3.23	0.00062	−2.73	0.00317	−2.23	0.01287	−1.73	0.04182
−3.72	0.00010	−3.22	0.00064	−2.72	0.00326	−2.22	0.01321	−1.72	0.04272
−3.71	0.00010	−3.21	0.00066	−2.71	0.00336	−2.21	0.01355	−1.71	0.04363
−3.70	0.00011	−3.20	0.00069	−2.70	0.00347	−2.20	0.01390	−1.70	0.04457
−3.69	0.00011	−3.19	0.00071	−2.69	0.00357	−2.19	0.01426	−1.69	0.04551
−3.68	0.00012	−3.18	0.00074	−2.68	0.00368	−2.18	0.01463	−1.68	0.04648
−3.67	0.00012	−3.17	0.00076	−2.67	0.00379	−2.17	0.01500	−1.67	0.04746
−3.66	0.00013	−3.16	0.00079	−2.66	0.00391	−2.16	0.01539	−1.66	0.04846
−3.65	0.00013	−3.15	0.00082	−2.65	0.00402	−2.15	0.01578	−1.65	0.04947
−3.64	0.00014	−3.14	0.00084	−2.64	0.00415	−2.14	0.01618	−1.64	0.05050
−3.63	0.00014	−3.13	0.00087	−2.63	0.00427	−2.13	0.01659	−1.63	0.05155
−3.62	0.00015	−3.12	0.00090	−2.62	0.00440	−2.12	0.01700	−1.62	0.05262
−3.61	0.00015	−3.11	0.00094	−2.61	0.00453	−2.11	0.01743	−1.61	0.05370
−3.60	0.00016	−3.10	0.00097	−2.60	0.00466	−2.10	0.01786	−1.60	0.05480
−3.59	0.00017	−3.09	0.00100	−2.59	0.00480	−2.09	0.01831	−1.59	0.05592
−3.58	0.00017	−3.08	0.00104	−2.58	0.00494	−2.08	0.01876	−1.58	0.05705
−3.57	0.00018	−3.07	0.00107	−2.57	0.00508	−2.07	0.01923	−1.57	0.05821
−3.56	0.00019	−3.06	0.00111	−2.56	0.00523	−2.06	0.01970	−1.56	0.05938
−3.55	0.00019	−3.05	0.00114	−2.55	0.00539	−2.05	0.02018	−1.55	0.06057
−3.54	0.00020	−3.04	0.00118	−2.54	0.00554	−2.04	0.02068	−1.54	0.06178
−3.53	0.00021	−3.03	0.00122	−2.53	0.00570	−2.03	0.02118	−1.53	0.06301
−3.52	0.00022	−3.02	0.00126	−2.52	0.00587	−2.02	0.02169	−1.52	0.06426
−3.51	0.00022	−3.01	0.00131	−2.51	0.00604	−2.01	0.02222	−1.51	0.06552

* Standard normal distribution has a zero mean ($\mu = 0$) and a unit standard deviation ($\delta = 1$); that is, also a unit variance ($\delta^2 = 1$). The distribution of a particular variable can be standardized as follows: $Z_i = \frac{x_i - \bar{x}}{s_x}$

The cumulative probability of $P(\Xi_i)$ is defined:

$$P = 1 - \alpha$$

and

$$\alpha = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}Z^2\right)$$

Table 1A

Table 1A. Cumulative standard normal distribution. [Continued]

z	P	z	P	z	P	z	P	z	P
−1.50	0.06681	−0.90	0.18406	−0.30	0.38209	0.30	0.61791	0.90	0.81594
−1.49	0.06811	−0.89	0.18673	−0.29	0.38591	0.31	0.62172	0.91	0.81859
−1.48	0.06944	−0.88	0.18943	−0.28	0.38974	0.32	0.62552	0.92	0.82121
−1.47	0.07078	−0.87	0.19215	−0.27	0.39358	0.33	0.62930	0.93	0.82381
−1.46	0.07215	−0.86	0.19489	−0.26	0.39743	0.34	0.63307	0.94	0.82639
−1.45	0.07353	−0.85	0.19766	−0.25	0.40129	0.35	0.63683	0.95	0.82894
−1.44	0.07493	−0.84	0.20045	−0.24	0.40517	0.36	0.64058	0.96	0.83147
−1.43	0.07636	−0.83	0.20327	−0.23	0.40905	0.37	0.64431	0.97	0.83398
−1.42	0.07780	−0.82	0.20611	−0.22	0.41294	0.38	0.64803	0.98	0.83646
−1.41	0.07927	−0.81	0.20897	−0.21	0.41683	0.39	0.65173	0.99	0.83891
−1.40	0.08076	−0.80	0.21186	−0.20	0.42074	0.40	0.65542	1.00	0.84134
−1.39	0.08226	−0.79	0.21476	−0.19	0.42465	0.41	0.65910	1.01	0.84375
−1.38	0.08379	−0.78	0.21770	−0.18	0.42858	0.42	0.66276	1.02	0.84614
−1.37	0.08534	−0.77	0.22065	−0.17	0.43251	0.43	0.66640	1.03	0.84850
−1.36	0.08691	−0.76	0.22363	−0.16	0.43644	0.44	0.67003	1.04	0.85083
−1.35	0.08851	−0.75	0.22663	−0.15	0.44038	0.45	0.67364	1.05	0.85314
−1.34	0.09012	−0.74	0.22965	−0.14	0.44433	0.46	0.67724	1.06	0.85543
−1.33	0.09176	−0.73	0.23270	−0.13	0.44828	0.47	0.68082	1.07	0.85769
−1.32	0.09342	−0.72	0.23576	−0.12	0.45224	0.48	0.68439	1.08	0.85993
−1.31	0.09510	−0.71	0.23885	−0.11	0.45620	0.49	0.68793	1.09	0.86214
−1.30	0.09680	−0.70	0.24196	−0.10	0.46017	0.50	0.69146	1.10	0.86433
−1.29	0.09853	−0.69	0.24510	−0.09	0.46414	0.51	0.69497	1.11	0.86650
−1.28	0.10027	−0.68	0.24825	−0.08	0.46812	0.52	0.69847	1.12	0.86864
−1.27	0.10204	−0.67	0.25143	−0.07	0.47210	0.53	0.70194	1.13	0.87076
−1.26	0.10383	−0.66	0.25463	−0.06	0.47608	0.54	0.70540	1.14	0.87286
−1.25	0.10565	−0.65	0.25785	−0.05	0.48006	0.55	0.70884	1.15	0.87493
−1.24	0.10749	−0.64	0.26109	−0.04	0.48405	0.56	0.71226	1.16	0.87698
−1.23	0.10935	−0.63	0.26435	−0.03	0.48803	0.57	0.71566	1.17	0.87900
−1.22	0.11123	−0.62	0.26763	−0.02	0.49202	0.58	0.71904	1.18	0.88100
−1.21	0.11314	−0.61	0.27093	−0.01	0.49601	0.59	0.72240	1.19	0.88298
−1.20	0.11507	−0.60	0.27425	0.00	0.50000	0.60	0.72575	1.20	0.88493
−1.19	0.11702	−0.59	0.27760	0.01	0.50399	0.61	0.72907	1.21	0.88686
−1.18	0.11900	−0.58	0.28096	0.02	0.50798	0.62	0.73237	1.22	0.88877
−1.17	0.12100	−0.57	0.28434	0.03	0.51197	0.63	0.73565	1.23	0.89065
−1.16	0.12302	−0.56	0.28774	0.04	0.51595	0.64	0.73891	1.24	0.89251
−1.15	0.12507	−0.55	0.29116	0.05	0.51994	0.65	0.74215	1.25	0.89435
−1.14	0.12714	−0.54	0.29460	0.06	0.52392	0.66	0.74537	1.26	0.89617
−1.13	0.12924	−0.53	0.29806	0.07	0.52790	0.67	0.74857	1.27	0.89796
−1.12	0.13136	−0.52	0.30153	0.08	0.53188	0.68	0.75175	1.28	0.89973
−1.11	0.13350	−0.51	0.30503	0.09	0.53586	0.69	0.75490	1.29	0.90147
−1.10	0.13567	−0.50	0.30854	0.10	0.53983	0.70	0.75804	1.30	0.90320
−1.09	0.13786	−0.49	0.31207	0.11	0.54380	0.71	0.76115	1.31	0.90490
−1.08	0.14007	−0.48	0.31561	0.12	0.54776	0.72	0.76424	1.32	0.90658
−1.07	0.14231	−0.47	0.31918	0.13	0.55172	0.73	0.76730	1.33	0.90824
−1.06	0.14457	−0.46	0.32276	0.14	0.55567	0.74	0.77035	1.34	0.90988
−1.05	0.14686	−0.45	0.32636	0.15	0.55962	0.75	0.77337	1.35	0.91149
−1.04	0.14917	−0.44	0.32997	0.16	0.56356	0.76	0.77637	1.36	0.91309
−1.03	0.15150	−0.43	0.33360	0.17	0.56749	0.77	0.77935	1.37	0.91466
−1.02	0.15386	−0.42	0.33724	0.18	0.57142	0.78	0.78230	1.38	0.91621
−1.01	0.15625	−0.41	0.34090	0.19	0.57535	0.79	0.78524	1.39	0.91774
−1.00	0.15866	−0.40	0.34458	0.20	0.57926	0.80	0.78814	1.40	0.91924
−0.99	0.16109	−0.39	0.34827	0.21	0.58317	0.81	0.79103	1.41	0.92073
−0.98	0.16354	−0.38	0.35197	0.22	0.58706	0.82	0.79389	1.42	0.92220
−0.97	0.16602	−0.37	0.35569	0.23	0.59095	0.83	0.79673	1.43	0.92364
−0.96	0.16853	−0.36	0.35942	0.24	0.59483	0.84	0.79955	1.44	0.92507
−0.95	0.17106	−0.35	0.36317	0.25	0.59871	0.85	0.80234	1.45	0.92647
−0.94	0.17361	−0.34	0.36693	0.26	0.60257	0.86	0.80511	1.46	0.92785
−0.93	0.17619	−0.33	0.37070	0.27	0.60642	0.87	0.80785	1.47	0.92922
−0.92	0.17879	−0.32	0.37448	0.28	0.61026	0.88	0.81057	1.48	0.93056
−0.91	0.18141	−0.31	0.37828	0.29	0.61409	0.89	0.81327	1.49	0.93189

Table 1A

Table 1A. Cumulative standard normal distribution. [Continued]

z	P	z	P	z	P	z	P	z	P
1.50	0.93319	2.10	0.98214	2.70	0.99653	3.30	0.99952	3.90	0.99995
1.51	0.93448	2.11	0.98257	2.71	0.99664	3.31	0.99953	3.91	0.99995
1.52	0.93574	2.12	0.98300	2.72	0.99674	3.32	0.99955	3.92	0.99996
1.53	0.93699	2.13	0.98341	2.73	0.99683	3.33	0.99957	3.93	0.99996
1.54	0.93822	2.14	0.98382	2.74	0.99693	3.34	0.99958	3.94	0.99996
1.55	0.93943	2.15	0.98422	2.75	0.99702	3.35	0.99960	3.95	0.99996
1.56	0.94062	2.16	0.98461	2.76	0.99711	3.36	0.99961	3.96	0.99996
1.57	0.94179	2.17	0.98500	2.77	0.99720	3.37	0.99962	3.97	0.99996
1.58	0.94295	2.18	0.98537	2.78	0.99728	3.38	0.99964	3.98	0.99997
1.59	0.94408	2.19	0.98574	2.79	0.99736	3.39	0.99965	3.99	0.99997
1.60	0.94520	2.20	0.98610	2.80	0.99744	3.40	0.99966	4.00	0.99997
1.61	0.94630	2.21	0.98645	2.81	0.99752	3.41	0.99968		
1.62	0.94738	2.22	0.98679	2.82	0.99760	3.42	0.99969		
1.63	0.94845	2.23	0.98713	2.83	0.99767	3.43	0.99970		
1.64	0.94950	2.24	0.98745	2.84	0.99774	3.44	0.99971		
1.65	0.95053	2.25	0.98778	2.85	0.99781	3.45	0.99972		
1.66	0.95154	2.26	0.98809	2.86	0.99788	3.46	0.99973		
1.67	0.95254	2.27	0.98840	2.87	0.99795	3.47	0.99974		
1.68	0.95352	2.28	0.98870	2.88	0.99801	3.48	0.99975		
1.69	0.95449	2.29	0.98899	2.89	0.99807	3.49	0.99976		
1.70	0.95543	2.30	0.98928	2.90	0.99813	3.50	0.99977		
1.71	0.95637	2.31	0.98956	2.91	0.99819	3.51	0.99978		
1.72	0.95728	2.32	0.98983	2.92	0.99825	3.52	0.99978		
1.73	0.95818	2.33	0.99010	2.93	0.99831	3.53	0.99979		
1.74	0.95907	2.34	0.99036	2.94	0.99836	3.54	0.99980		
1.75	0.95994	2.35	0.99061	2.95	0.99841	3.55	0.99981		
1.76	0.96080	2.36	0.99086	2.96	0.99846	3.56	0.99981		
1.77	0.96164	2.37	0.99111	2.97	0.99851	3.57	0.99982		
1.78	0.96246	2.38	0.99134	2.98	0.99856	3.58	0.99983		
1.79	0.96327	2.39	0.99158	2.99	0.99861	3.59	0.99983		
1.80	0.96407	2.40	0.99180	3.00	0.99865	3.60	0.99984		
1.81	0.96485	2.41	0.99202	3.01	0.99869	3.61	0.99985		
1.82	0.96562	2.42	0.99224	3.02	0.99874	3.62	0.99985		
1.83	0.96638	2.43	0.99245	3.03	0.99878	3.63	0.99986		
1.84	0.96712	2.44	0.99266	3.04	0.99882	3.64	0.99986		
1.85	0.96784	2.45	0.99286	3.05	0.99886	3.65	0.99987		
1.86	0.96856	2.46	0.99305	3.06	0.99889	3.66	0.99987		
1.87	0.96926	2.47	0.99324	3.07	0.99893	3.67	0.99988		
1.88	0.96995	2.48	0.99343	3.08	0.99897	3.68	0.99988		
1.89	0.97062	2.49	0.99361	3.09	0.99900	3.69	0.99989		
1.90	0.97128	2.50	0.99379	3.10	0.99903	3.70	0.99989		
1.91	0.97193	2.51	0.99396	3.11	0.99906	3.71	0.99990		
1.92	0.97257	2.52	0.99413	3.12	0.99910	3.72	0.99990		
1.93	0.97320	2.53	0.99430	3.13	0.99913	3.73	0.99990		
1.94	0.97381	2.54	0.99446	3.14	0.99916	3.74	0.99991		
1.95	0.97441	2.55	0.99461	3.15	0.99918	3.75	0.99991		
1.96	0.97500	2.56	0.99477	3.16	0.99921	3.76	0.99992		
1.97	0.97558	2.57	0.99492	3.17	0.99924	3.77	0.99992		
1.98	0.97615	2.58	0.99506	3.18	0.99926	3.78	0.99992		
1.99	0.97670	2.59	0.99520	3.19	0.99929	3.79	0.99992		
2.00	0.97725	2.60	0.99534	3.20	0.99931	3.80	0.99993		
2.01	0.97778	2.61	0.99547	3.21	0.99934	3.81	0.99993		
2.02	0.97831	2.62	0.99560	3.22	0.99936	3.82	0.99993		
2.03	0.97882	2.63	0.99573	3.23	0.99938	3.83	0.99994		
2.04	0.97932	2.64	0.99585	3.24	0.99940	3.84	0.99994		
2.05	0.97982	2.65	0.99598	3.25	0.99942	3.85	0.99994		
2.06	0.98030	2.66	0.99609	3.26	0.99944	3.86	0.99994		
2.07	0.98077	2.67	0.99621	3.27	0.99946	3.87	0.99995		
2.08	0.98124	2.68	0.99632	3.28	0.99948	3.88	0.99995		
2.09	0.98169	2.69	0.99643	3.29	0.99950	3.89	0.99995		

Table 1B

Table 1B. The most commonly used critical values of Z-function.

α	0.20	0.10	0.05	0.025	0.01	0.005	0.0025	0.001
Z_α	0.84	1.28	1.64	1.96	2.33	2.57	2.81	3.09

Table 2

Table 2. The critical values of Student's t-function
(DF = degrees of freedom, α = level of significance).

DF	α							
	0.25	0.20	0.15	0.10	0.05	0.025	0.005	0.0005
1	1.0005	1.376	1.963	3.078	6.314	12.706	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	5.841	12.941
4	0.741	0.941	1.190	1.533	2.132	2.776	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	4.032	6.859
6	0.718	0.906	1.134	1.440	1.943	2.447	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	3.499	5.405
8	0.706	0.889	1.108	1.397	1.860	2.306	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.947	4.073
16	0.690	0.866	1.071	1.337	1.746	2.120	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.750	3.646
35	0.682	0.852	1.052	1.306	1.690	2.030	2.724	3.591
40	0.681	0.851	1.050	1.303	1.684	2.021	2.704	3.551
45	0.680	0.850	1.048	1.301	1.680	2.014	2.690	3.520
50	0.680	0.849	1.047	1.299	1.676	2.008	2.678	3.496
55	0.679	0.849	1.047	1.297	1.673	2.004	2.669	3.476
60	0.679	0.848	1.046	1.296	1.671	2.000	2.660	3.460
70	0.678	0.847	1.045	1.294	1.667	1.994	2.648	3.435
80	0.678	0.847	1.044	1.293	1.665	1.990	2.638	3.416
90	0.678	0.846	1.043	1.291	1.662	1.987	2.632	3.402
100	0.677	0.846	1.042	1.290	1.661	1.984	2.626	3.390
200	0.676	0.844	1.039	1.286	1.653	1.972	2.601	3.340
300	0.676	0.843	1.038	1.285	1.650	1.968	2.592	3.323
400	0.676	0.843	1.038	1.284	1.649	1.966	2.588	3.315
500	0.676	0.843	1.037	1.284	1.648	1.965	2.586	3.310
1000	0.675	0.842	1.037	1.283	1.647	1.962	2.581	3.301
∞	0.67449	0.84162	1.03643	1.28155	1.64485	1.95996	2.57582	3.29053

Table 3

Table. 3. The critical values of F-function for $\alpha = 0.10$
(DF_1 and DF_2 are degrees of freedom and α is level of significance)

Degrees of freedom for denominator, DF_2	Degrees of freedom for numerator, DF_1																			
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.864	49.500	53.593	55.833	57.241	58.204	58.906	59.439	59.858	60.195	60.705	61.220	61.740	62.002	62.265	62.529	62.794	63.061	63.328
2	8.5263	9.0000	9.1618	9.2434	9.2926	9.3255	9.3491	9.3668	9.3805	9.3916	9.4081	9.4247	9.4413	9.4496	9.4579	9.4663	9.4746	9.4829	9.4913
3	5.5383	5.4624	5.3908	5.3427	5.3092	5.2847	5.2662	5.2517	5.2400	5.2304	5.2156	5.2003	5.1845	5.1764	5.1681	5.1597	5.1512	5.1425	5.1337
4	4.5448	4.3246	4.1908	4.1073	4.0506	4.0098	3.9790	3.9549	3.9357	3.9199	3.8955	3.8689	3.8443	3.8310	3.8174	3.8036	3.7896	3.7753	3.7607
5	4.0604	3.7797	3.6195	3.5202	3.4530	3.4045	3.3679	3.3393	3.3163	3.2974	3.2682	3.2380	3.2067	3.1905	3.1741	3.1573	3.1402	3.1228	3.1050
6	3.7760	3.4633	3.2888	3.1808	3.1075	3.0546	3.0145	2.9830	2.9577	2.9369	2.9047	2.8712	2.8363	2.8183	2.8000	2.7812	2.7620	2.7423	2.7222
7	3.5894	3.2574	3.0741	2.9605	2.8833	2.8274	2.7849	2.7516	2.7247	2.7025	2.6681	2.6322	2.5947	2.5753	2.5555	2.5351	2.5142	2.4928	2.4708
8	3.4579	3.1131	2.9238	2.8064	2.7265	2.6683	2.6241	2.5893	2.5612	2.5380	2.5020	2.4642	2.4246	2.4041	2.3830	2.3614	2.3391	2.3162	2.2926
9	3.3603	3.0065	2.8129	2.6927	2.6106	2.5509	2.5053	2.4694	2.4403	2.4163	2.3789	2.3396	2.2983	2.2768	2.2547	2.2320	2.2085	2.1843	2.1592
10	3.2850	2.9245	2.7277	2.6053	2.5216	2.4606	2.4140	2.3772	2.3473	2.3226	2.2841	2.2435	2.2007	2.1784	2.1554	2.1317	2.1072	2.0818	2.0554
11	3.2252	2.8595	2.6602	2.5362	2.4512	2.3891	2.3416	2.3040	2.2735	2.2482	2.2087	2.1671	2.1230	2.1000	2.0762	2.0516	2.0261	1.9997	1.9721
12	3.1765	2.8068	2.6055	2.4801	2.3940	2.3310	2.2828	2.2446	2.2135	2.1878	2.1474	2.1049	2.0597	2.0360	2.0115	1.9861	1.9597	1.9323	1.9036
13	3.1362	2.7632	2.5603	2.4337	2.3467	2.2830	2.2341	2.1953	2.1638	2.1376	2.0966	2.0532	2.0070	1.9827	1.9576	1.9315	1.9043	1.8759	1.8462
14	3.1022	2.7265	2.5222	2.3947	2.3069	2.2426	2.1931	2.1539	2.1220	2.0954	2.0537	2.0095	1.9625	1.9377	1.9119	1.8852	1.8572	1.8280	1.7973
15	3.0732	2.6952	2.4898	2.3614	2.2730	2.2081	2.1582	2.1185	2.0862	2.0593	2.0171	1.9722	1.9243	1.8990	1.8728	1.8454	1.8168	1.7867	1.7551
16	3.0481	2.6682	2.4618	2.3327	2.2438	2.1783	2.1280	2.0880	2.0553	2.0281	1.9854	1.9399	1.8913	1.8656	1.8388	1.8108	1.7816	1.7507	1.7182
17	3.0262	2.6446	2.4374	2.3077	2.2183	2.1524	2.1017	2.0613	2.0284	2.0009	1.9577	1.9117	1.8624	1.8362	1.8090	1.7805	1.7506	1.7191	1.6856
18	3.0070	2.6239	2.4160	2.2858	2.1958	2.1296	2.0785	2.0379	2.0047	1.9770	1.9333	1.8868	1.8368	1.8103	1.7827	1.7537	1.7232	1.6910	1.6567
19	2.9899	2.6056	2.3970	2.2663	2.1760	2.1094	2.0580	2.0171	1.9836	1.9557	1.9117	1.8647	1.8142	1.7873	1.7592	1.7298	1.6988	1.6659	1.6308
20	2.9747	2.5893	2.3801	2.2489	2.1582	2.0913	2.0397	1.9985	1.9649	1.9367	1.8924	1.8449	1.7938	1.7667	1.7382	1.7083	1.6768	1.6433	1.6074
21	2.9609	2.5746	2.3649	2.2333	2.1423	2.0751	2.0232	1.9819	1.9480	1.9197	1.8750	1.8272	1.7756	1.7481	1.7193	1.6890	1.6569	1.6228	1.5862
22	2.9486	2.5613	2.3512	2.2193	2.1279	2.0605	2.0084	1.9668	1.9327	1.9043	1.8593	1.8111	1.7590	1.7312	1.7021	1.6714	1.6389	1.6042	1.5668
23	2.9374	2.5493	2.3387	2.2065	2.1149	2.0472	1.9949	1.9531	1.9189	1.8903	1.8450	1.7964	1.7439	1.7159	1.6864	1.6554	1.6224	1.5871	1.5490
24	2.9271	2.5383	2.3274	2.1949	2.1030	2.0351	1.9826	1.9407	1.9063	1.8775	1.8319	1.7831	1.7302	1.7019	1.6721	1.6407	1.6073	1.5715	1.5327
25	2.9177	2.5283	2.3170	2.1843	2.0922	2.0241	1.9714	1.9292	1.8947	1.8658	1.8200	1.7708	1.7175	1.6890	1.6589	1.6272	1.5934	1.5570	1.5176
26	2.9091	2.5191	2.3075	2.1745	2.0822	2.0139	1.9610	1.9188	1.8841	1.8550	1.8090	1.7596	1.7059	1.6771	1.6468	1.6147	1.5805	1.5437	1.5036
27	2.9012	2.5106	2.2987	2.1655	2.0730	2.0045	1.9515	1.9091	1.8743	1.8451	1.7989	1.7492	1.6951	1.6662	1.6356	1.6032	1.5686	1.5313	1.4906
28	2.8939	2.5028	2.2906	2.1571	2.0645	1.9959	1.9427	1.9001	1.8652	1.8359	1.7895	1.7395	1.6852	1.6560	1.6252	1.5925	1.5575	1.5198	1.4784
29	2.8871	2.4955	2.2831	2.1494	2.0566	1.9878	1.9345	1.8918	1.8560	1.8274	1.7808	1.7306	1.6759	1.6465	1.6155	1.5825	1.5472	1.5090	1.4670
30	2.8807	2.4887	2.2761	2.1422	2.0492	1.9803	1.9269	1.8841	1.8498	1.8195	1.7727	1.7223	1.6673	1.6377	1.6065	1.5732	1.5376	1.4989	1.4564
40	2.8354	2.4404	2.2261	2.0909	1.9968	1.9269	1.8725	1.8289	1.7929	1.7627	1.7146	1.6624	1.6052	1.5741	1.5411	1.5056	1.4672	1.4248	1.3769
60	2.7914	2.3932	2.1774	2.0410	1.9457	1.8747	1.8194	1.7748	1.7380	1.7070	1.6574	1.6034	1.5435	1.5107	1.4755	1.4373	1.3952	1.3476	1.2915
120	2.7478	2.3473	2.1300	1.9923	1.8959	1.8238	1.7675	1.7220	1.6843	1.6524	1.6012	1.5450	1.4821	1.4472	1.4094	1.3676	1.3203	1.2646	1.1926
∞	2.7055	2.3026	2.0838	1.9449	1.8473	1.7741	1.7167	1.6702	1.6315	1.5987	1.5458	1.4871	1.4206	1.3832	1.3419	1.2951	1.2400	1.1686	1.0000

Table 3

[Continued]

Table 3. The critical values of F-function for $\alpha = 0.05$
(DF_1 and DF_2 are degrees of freedom, and α is level of significance).

Degrees of freedom for numerator, DF_1

Degrees of freedom for denominator, DF_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	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	8.59	8.57	8.55	8.53
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	5.72	5.69	5.66	5.63
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	4.46	4.43	4.40	4.37
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	3.77	3.74	3.70	3.67
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	3.34	3.30	3.27	3.23
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	3.04	3.01	2.97	2.93
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	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
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	2.53	2.49	2.45	2.40
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	2.43	2.38	2.34	2.30
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	2.34	2.30	2.25	2.21
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	2.27	2.22	2.18	2.13
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	2.20	2.16	2.11	2.07
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	2.15	2.11	2.06	2.01
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	2.10	2.06	2.01	1.96
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	2.06	2.02	1.97	1.92
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	2.03	1.98	1.93	1.88
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	1.99	1.95	1.90	1.84
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	1.96	1.92	1.87	1.81
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	1.94	1.89	1.84	1.78
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	1.91	1.86	1.81	1.76
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	1.89	1.84	1.79	1.73
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	1.87	1.82	1.77	1.71
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	1.79	1.74	1.68	1.62
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	1.69	1.64	1.58	1.51
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	1.59	1.53	1.47	1.39
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	1.50	1.43	1.35	1.25
∞	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	1.39	1.32	1.22	1.00

Interpolation should be performed using reciprocals of the degrees of freedom.
Reproduced from *Biometrika*, Vol. 33 (1943), p. 73.

Table 3

[Continued]

Table 3. The critical values of F-function for $\alpha = 0.025$ (DF_1 and DF_2 are degrees of freedom, and α is level of significance).

Degrees of freedom for denominator, DF_2	Degrees of freedom for numerator, DF_1														
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	∞
1	647 79	799 50	864 16	899 58	921 85	937 11	948 22	956 66	963 28	968 63	976 71	984 87	993 10	997 25	1001 4
2	38 51	39 00	39 17	39 25	39 30	39 33	39 36	39 37	39 39	39 40	39 41	39 43	39 45	39 46	39 46
3	17 44	16 04	15 44	15 10	14 88	14 73	14 62	14 54	14 47	14 42	14 34	14 25	14 17	14 12	14 08
4	12 22	10 65	9 98	9 60	9 36	9 20	9 07	8 98	8 90	8 84	8 75	8 66	8 56	8 51	8 46
5	10 01	8 43	7 76	7 39	7 15	6 98	6 85	6 76	6 68	6 62	6 52	6 43	6 33	6 28	6 23
6	8 81	7 26	6 60	6 23	5 99	5 82	5 70	5 60	5 52	5 46	5 37	5 27	5 17	5 12	5 07
7	8 07	6 54	5 89	5 52	5 29	5 12	4 99	4 90	4 82	4 76	4 67	4 57	4 47	4 41	4 36
8	7 57	6 06	5 42	5 05	4 82	4 65	4 53	4 43	4 36	4 30	4 20	4 10	4 00	3 95	3 89
9	7 21	5 71	5 08	4 72	4 48	4 32	4 20	4 10	4 03	3 96	3 87	3 77	3 67	3 61	3 56
10	6 94	5 46	4 83	4 47	4 24	4 07	3 95	3 85	3 78	3 72	3 62	3 52	3 42	3 37	3 31
11	6 72	5 26	4 63	4 28	4 04	3 88	3 76	3 66	3 59	3 53	3 43	3 33	3 23	3 17	3 12
12	6 55	5 10	4 47	4 12	3 89	3 73	3 61	3 51	3 44	3 37	3 28	3 18	3 07	3 02	2 96
13	6 41	4 97	4 35	4 00	3 77	3 60	3 48	3 39	3 31	3 25	3 15	3 05	2 95	2 89	2 84
14	6 30	4 86	4 24	3 89	3 66	3 50	3 38	3 29	3 21	3 15	3 05	2 95	2 84	2 79	2 73
15	6 20	4 77	4 15	3 80	3 58	3 41	3 29	3 20	3 12	3 06	2 96	2 86	2 76	2 70	2 64
16	6 12	4 69	4 08	3 73	3 50	3 34	3 22	3 12	3 05	2 99	2 89	2 79	2 68	2 63	2 57
17	6 04	4 62	4 01	3 66	3 44	3 28	3 16	3 06	2 98	2 92	2 82	2 72	2 62	2 56	2 50
18	5 98	4 56	3 95	3 61	3 38	3 22	3 10	3 01	2 93	2 87	2 77	2 67	2 56	2 50	2 44
19	5 92	4 51	3 90	3 56	3 33	3 17	3 05	2 96	2 88	2 82	2 72	2 62	2 51	2 45	2 39
20	5 87	4 46	3 86	3 51	3 29	3 13	3 01	2 91	2 84	2 77	2 68	2 57	2 46	2 41	2 35
21	5 83	4 42	3 82	3 48	3 25	3 09	2 97	2 87	2 80	2 73	2 64	2 53	2 42	2 37	2 31
22	5 79	4 38	3 78	3 44	3 22	3 05	2 93	2 84	2 76	2 70	2 60	2 50	2 39	2 33	2 27
23	5 75	4 35	3 75	3 41	3 18	3 02	2 90	2 81	2 73	2 67	2 57	2 47	2 36	2 30	2 24
24	5 72	4 32	3 72	3 38	3 15	2 99	2 87	2 78	2 70	2 64	2 54	2 44	2 33	2 27	2 21
25	5 69	4 29	3 69	3 35	3 13	2 97	2 85	2 75	2 68	2 61	2 51	2 41	2 30	2 24	2 18
26	5 66	4 27	3 67	3 33	3 10	2 94	2 82	2 73	2 65	2 59	2 49	2 39	2 28	2 22	2 16
27	5 63	4 24	3 65	3 31	3 08	2 92	2 80	2 71	2 63	2 57	2 47	2 36	2 25	2 19	2 13
28	5 61	4 22	3 63	3 29	3 06	2 90	2 78	2 69	2 61	2 55	2 45	2 34	2 23	2 17	2 11
29	5 59	4 20	3 61	3 27	3 04	2 88	2 76	2 67	2 59	2 53	2 43	2 32	2 21	2 15	2 09
30	5 57	4 18	3 59	3 25	3 03	2 87	2 75	2 65	2 57	2 51	2 41	2 31	2 20	2 14	2 07
40	5 42	4 05	3 46	3 13	2 90	2 74	2 62	2 53	2 45	2 39	2 29	2 18	2 07	2 01	1 94
60	5 29	3 93	3 34	3 01	2 78	2 63	2 51	2 41	2 33	2 27	2 17	2 06	1 94	1 88	1 82
120	5 15	3 80	3 23	2 89	2 67	2 52	2 39	2 30	2 22	2 16	2 05	1 94	1 82	1 76	1 69
∞	5 02	3 69	3 12	2 79	2 57	2 41	2 29	2 19	2 11	2 05	1 94	1 83	1 71	1 64	1 57

Table 3

[Continued]

Table 3. The critical values of F-function for $\alpha = 0.01$
(DF_1 and DF_2 are degrees of freedom, and α is level of significance).

Degrees of freedom for denominator, DF_2	Degrees of freedom for numerator, DF_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4.052	5.000	5.403	5.625	5.764	5.859	5.928	5.982	6.023	6.056	6.106	6.157	6.209	6.235	6.261	6.287	6.313	6.339	6.366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Table 4

Table 4. The critical values of chi-square function
(DF is degrees of freedom, and α is level of significance).

DF	α															
	0.995	0.99	0.975	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.025	0.01	0.005
1	0.0^{3}393	0.0^{3}157	0.0^{3}982	0.0^{2}393	0.0158	0.0642	0.148	0.275	0.455	0.708	1.07	1.64	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	0.446	0.713	1.02	1.39	1.83	2.41	3.22	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	1.00	1.42	1.87	2.37	2.95	3.67	4.64	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	1.65	2.19	2.75	3.36	4.04	4.88	5.99	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	2.34	3.00	3.66	4.35	5.13	6.06	7.29	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	3.07	3.83	4.57	5.35	6.21	7.23	8.56	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	3.82	4.67	5.49	6.35	7.28	8.38	9.80	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	6.42	7.34	8.35	9.52	11.0	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	7.36	8.34	9.41	10.7	12.2	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	8.30	9.34	10.5	11.8	13.4	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	9.24	10.3	11.5	12.9	14.6	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	10.2	11.3	12.6	14.0	15.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	11.1	12.3	13.6	15.1	17.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	9.47	10.8	12.1	13.3	14.7	16.2	18.2	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	10.3	11.7	13.0	14.3	15.7	17.3	19.3	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.2	12.6	14.0	15.3	16.8	18.4	20.5	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.0	13.5	14.9	16.3	17.8	19.5	21.6	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	12.9	14.4	15.9	17.3	18.9	20.6	22.8	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	13.7	15.4	16.9	18.3	19.9	21.7	23.9	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	14.6	16.3	17.8	19.3	21.0	22.8	25.0	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	15.4	17.2	18.8	20.3	22.0	23.9	26.2	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	16.3	18.1	19.7	21.3	23.0	24.9	27.3	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	17.2	19.0	20.7	22.3	24.1	26.0	28.4	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	18.1	19.9	21.7	23.3	25.1	27.1	29.6	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	18.9	20.9	22.6	24.3	26.1	28.2	30.7	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	19.8	21.8	23.6	25.3	27.2	29.2	31.8	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	20.7	22.7	24.5	26.3	28.2	30.3	32.9	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	21.6	23.6	25.5	27.3	29.2	31.4	34.0	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	22.5	24.6	26.6	28.3	30.3	32.5	35.1	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	23.4	25.5	27.4	29.3	31.3	33.5	36.3	40.3	43.8	47.0	50.9	53.7
31	14.5	15.7	17.5	19.3	21.4	24.3	26.4	28.4	30.3	32.3	34.6	37.4	41.4	45.0	48.2	52.2	55.0
32	15.1	16.4	18.3	20.1	22.3	25.1	27.4	29.4	31.3	33.4	35.7	38.5	42.6	46.2	49.5	53.5	56.3
33	15.8	17.1	19.0	20.9	23.1	26.0	28.3	30.3	32.3	34.4	36.7	39.6	43.7	47.4	50.7	54.8	57.6
34	16.5	17.8	19.8	21.7	24.0	26.9	29.2	31.3	33.3	35.4	37.8	40.7	44.9	48.6	52.0	56.1	59.0
35	17.2	18.5	20.6	22.5	24.8	27.8	30.2	32.3	34.3	36.5	38.9	41.8	46.1	49.8	53.2	57.3	60.3
36	17.9	19.2	21.3	23.3	25.6	28.7	31.1	33.3	35.3	37.5	39.9	42.9	47.2	51.0	54.4	58.6	61.6
37	18.6	20.0	22.1	24.1	26.5	29.6	32.1	34.2	36.3	38.5	41.0	44.0	48.4	52.2	55.7	59.9	62.9
38	19.3	20.7	22.9	24.9	27.3	30.5	33.0	35.2	37.3	39.6	42.0	45.1	49.5	53.4	56.9	61.2	64.2
39	20.0	21.4	23.7	25.7	28.2	31.4	33.9	36.2	38.3	40.6	43.1	46.2	50.7	54.6	58.1	62.4	65.5
40	20.7	22.2	24.4	26.5	29.1	32.3	34.9	37.1	39.3	41.6	44.2	47.3	51.8	55.8	59.3	63.7	66.8
41	21.4	22.9	25.2	27.3	29.9	33.3	35.8	38.1	40.3	42.7	45.2	48.4	52.9	56.9	60.6	65.0	68.1
42	22.1	23.7	26.0	28.1	30.8	34.2	36.8	39.1	41.3	43.7	46.3	49.5	54.1	58.1	61.8	66.2	69.3
43	22.9	24.4	26.8	29.0	31.6	35.1	37.7	40.0	42.3	44.7	47.3	50.5	55.2	59.3	63.0	67.5	70.6
44	23.6	25.1	27.6	29.8	32.5	36.0	38.6	41.0	43.3	45.7	48.4	51.6	56.4	60.5	64.2	68.7	71.9
45	24.3	25.9	28.4	30.6	33.4	36.9	39.6	42.0	44.3	46.8	49.6	52.7	57.5	61.7	65.4	70.0	73.2
46	25.0	26.7	29.2	31.4	34.2	37.8	40.5	43.0	45.3	47.8	50.5	53.8	58.6	62.8	66.6	71.2	74.4
47	25.8	27.4	30.0	32.3	35.1	38.7	41.5	43.9	46.3	48.8	51.6	54.9	59.8	64.0	67.8	72.4	75.7
48	26.5	28.2	30.8	33.1	35.9	39.6	42.4	44.9	47.3	49.8	52.6	56.0	60.9	65.2	69.0	73.7	77.0
49	27.2	28.9	31.6	33.9	36.8	40.5	43.4	45.9	48.3	50.9	53.7	57.1	62.0	66.3	70.2	74.9	78.2
50	28.0	29.7	32.4	34.8	37.7	41.4	44.3	46.9	49.3	51.9	54.7	58.2	63.2	67.5	71.4	76.2	79.5

Table 4

[Continued]

Table 4. The critical values of chi-square function
(DF is degrees of freedom, and α is level of significance).

DF	α																
	0.995	0.99	0.975	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.025	0.01	0.005
51	28.7	30.5	33.2	35.6	38.6	42.4	45.3	47.8	50.3	52.9	55.8	59.2	64.3	68.7	72.6	77.4	80.7
52	29.5	31.2	34.0	36.4	39.4	43.3	46.2	48.8	51.3	53.9	56.8	60.3	65.4	69.8	73.8	78.6	82.0
53	30.2	32.0	34.8	37.3	40.3	44.2	47.2	49.8	52.3	55.0	57.9	61.4	66.5	71.0	75.0	79.8	83.3
54	31.0	32.8	35.6	38.1	41.2	45.1	48.1	50.8	53.3	56.0	58.9	62.5	67.7	72.2	76.2	81.1	84.5
55	31.7	33.6	36.4	39.0	42.1	46.0	49.1	51.7	54.3	57.0	60.0	63.6	68.8	73.3	77.4	82.3	85.7
56	32.5	34.3	37.2	39.8	42.9	47.0	50.0	52.7	55.3	58.0	61.0	64.7	69.9	74.5	78.6	83.5	87.0
57	33.2	35.1	38.0	40.6	43.8	47.9	51.0	53.7	56.3	59.1	62.1	65.7	71.0	75.6	79.8	84.7	88.2
58	34.0	35.9	38.8	41.5	44.7	48.8	51.9	54.7	57.3	60.1	63.1	66.8	72.2	76.8	80.9	86.0	89.5
59	34.8	36.7	39.7	42.3	45.6	49.7	52.9	55.6	58.3	61.1	64.2	67.9	73.3	77.9	82.1	87.2	90.7
60	35.5	37.5	40.5	43.2	46.5	50.6	53.8	56.6	59.3	62.1	65.2	69.0	74.4	79.1	83.3	88.4	92.0
61	36.3	38.3	41.3	44.0	47.3	51.6	54.8	57.6	60.3	63.2	66.3	70.0	75.5	80.2	84.5	89.6	93.2
62	37.1	39.1	42.1	44.9	48.2	52.5	55.7	58.6	61.3	64.2	67.3	71.1	76.6	81.4	85.7	90.8	94.4
63	37.8	39.9	43.0	45.7	49.1	53.4	56.7	59.6	62.3	65.2	68.4	72.2	77.7	82.5	86.8	92.0	95.6
64	38.6	40.6	43.8	46.6	50.0	54.3	57.6	60.5	63.3	66.2	69.4	73.3	78.9	83.7	88.0	93.2	96.9
65	39.4	41.4	44.6	47.4	50.9	55.3	58.6	61.5	64.3	67.2	70.5	74.4	80.0	84.8	89.2	94.4	98.1
66	40.2	42.2	45.4	48.3	51.8	56.2	59.5	62.5	65.3	68.3	71.5	75.4	81.1	86.0	90.3	95.6	99.3
67	40.9	43.0	46.3	49.2	52.7	57.1	60.5	63.5	66.3	69.3	72.6	76.5	82.2	87.1	91.5	96.8	100.6
68	41.7	43.8	47.1	50.0	53.5	58.0	61.4	64.4	67.3	70.3	73.6	77.6	83.3	88.3	92.7	98.0	101.8
69	42.5	44.6	47.9	50.9	54.4	59.0	62.4	65.4	68.3	71.3	74.6	78.6	84.4	89.4	93.9	99.2	103.0
70	43.3	45.4	48.8	51.7	55.3	59.9	63.3	66.4	69.3	72.4	75.7	79.7	85.5	90.5	95.0	100.4	104.2
71	44.1	46.2	49.6	52.6	56.2	60.8	64.3	67.4	70.3	73.4	76.7	80.8	86.6	91.7	96.2	101.6	105.4
72	44.8	47.1	50.4	53.5	57.1	61.8	65.3	68.4	71.3	74.4	77.8	81.9	87.7	92.8	97.4	102.8	106.6
73	45.6	47.9	51.3	54.3	58.0	62.7	66.2	69.3	72.3	75.4	78.8	82.9	88.8	93.9	98.5	104.0	107.9
74	46.4	48.7	52.1	55.2	58.9	63.6	67.2	70.3	73.3	76.4	79.9	84.0	90.0	95.1	99.7	105.2	109.1
75	47.2	49.5	52.9	56.1	59.8	64.6	68.1	71.3	74.3	77.5	80.9	85.1	91.1	96.2	100.8	106.4	110.3
76	48.0	50.3	53.8	56.9	60.7	65.5	69.1	72.3	75.3	78.5	82.0	86.1	92.2	97.4	102.0	107.6	111.5
77	48.8	51.1	54.6	57.8	61.6	66.4	70.0	73.2	76.3	79.5	83.0	87.2	93.3	98.5	103.2	108.8	112.7
78	49.6	51.9	55.5	58.7	62.5	67.3	71.0	74.2	77.3	80.5	84.0	88.3	94.4	99.6	104.3	110.0	113.9
79	50.4	52.7	56.3	59.5	63.4	68.3	72.0	75.2	78.3	81.5	85.1	89.3	95.5	100.7	105.5	111.1	115.1
80	51.2	53.5	57.2	60.4	64.3	69.2	72.9	76.2	79.3	82.6	86.1	90.4	96.6	101.9	106.6	112.3	116.3
81	52.0	54.4	58.0	61.3	65.2	70.1	73.9	77.2	80.3	83.6	87.2	91.5	97.7	103.0	107.8	113.5	117.5
82	52.8	55.2	58.8	62.1	66.1	71.1	74.8	78.1	81.3	84.6	88.2	92.5	98.8	104.1	108.9	114.7	118.7
83	53.6	56.0	59.7	63.0	67.0	72.0	75.8	79.1	82.3	85.6	89.2	93.6	99.9	105.3	110.1	115.9	119.9
84	54.4	56.8	60.5	63.9	67.9	72.9	76.8	80.1	83.3	86.6	90.3	94.7	101.0	106.4	111.2	117.1	121.1
85	55.2	57.6	61.4	64.7	68.8	73.9	77.7	81.1	84.3	87.7	91.3	95.7	102.1	107.5	112.4	118.2	122.3
86	56.0	58.5	62.2	65.6	69.7	74.8	78.7	82.1	85.3	88.7	92.4	96.8	103.2	108.6	113.5	119.4	123.5
87	56.8	59.3	63.1	66.5	70.6	75.7	79.6	83.0	86.3	89.7	93.4	97.9	104.3	109.8	114.7	120.6	124.7
88	57.6	60.1	63.9	67.4	71.5	76.7	80.6	84.0	87.3	90.7	94.4	98.9	105.4	110.9	115.8	121.8	125.9
89	58.4	60.9	64.8	68.2	72.4	77.6	81.6	85.0	88.3	91.7	95.5	100.0	106.5	112.0	117.0	122.9	127.1
90	59.2	61.8	65.6	69.1	73.3	78.6	82.5	86.0	89.3	92.8	96.5	101.1	107.6	113.1	118.1	124.1	128.3
91	60.0	62.6	66.5	70.0	74.2	79.5	83.5	87.0	90.3	93.8	97.6	102.1	108.7	114.3	119.3	125.3	129.5
92	60.8	63.4	67.4	70.9	75.1	80.4	84.4	88.0	91.3	94.8	98.6	103.2	109.8	115.4	120.4	126.5	130.7
93	61.6	64.2	68.2	71.8	76.0	81.4	85.4	88.9	92.3	95.8	99.6	104.2	110.9	116.5	121.6	127.6	131.9
94	62.4	65.1	69.1	72.6	76.9	82.3	86.4	89.9	93.3	96.8	100.7	105.3	111.9	117.6	122.7	128.8	133.1
95	63.2	65.9	69.9	73.5	77.8	83.2	87.3	90.9	94.3	97.9	101.7	106.4	113.0	118.8	123.9	130.0	134.2
96	64.1	66.7	70.8	74.4	78.7	84.2	88.3	91.9	95.3	98.9	102.8	107.4	114.1	119.9	125.0	131.1	135.4
97	64.9	67.6	71.6	75.3	79.6	85.1	89.2	92.9	96.3	99.9	103.8	108.5	115.2	121.0	126.1	132.3	136.6
98	65.7	68.4	72.5	76.2	80.5	86.1	90.2	93.8	97.3	100.9	104.8	109.5	116.3	122.1	127.3	133.5	137.8
99	66.5	69.2	73.4	77.0	81.4	87.0	91.2	94.8	98.3	101.9	105.9	110.6	117.4	123.2	128.4	134.6	139.0
100	67.3	70.1	74.2	77.9	82.4	87.9	92.1	95.8	99.3	102.9	106.9	111.7	118.5	124.3	129.6	135.8	140.2

Table 4

[Continued]

Table 4. The critical values of chi-square function
(DF is degrees of freedom, and α is level of significance).

DF	α											
	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
91	60.005	62.581	66.501	70.003	74.196	81.574	99.700	108.661	114.268	119.282	125.289	129.491
92	60.815	63.409	67.356	70.882	75.100	82.524	100.750	109.756	115.390	120.427	126.462	130.681
93	61.625	64.238	68.211	71.760	76.006	83.474	101.800	110.850	116.511	121.571	127.633	131.871
94	62.437	65.068	69.068	72.640	76.912	84.425	102.850	111.944	117.632	122.715	128.803	133.059
95	63.250	65.898	69.925	73.520	77.818	85.376	103.899	113.038	118.752	123.858	129.973	134.247
96	64.063	66.730	70.783	74.401	78.725	86.327	104.948	114.131	119.871	125.000	131.141	135.433
97	64.878	67.562	71.642	75.282	79.633	87.278	105.997	115.223	120.990	126.141	132.309	136.619
98	65.694	68.396	72.501	76.164	80.541	88.229	107.045	116.315	122.108	127.282	133.476	137.803
99	66.510	69.230	73.361	77.046	81.449	89.181	108.093	117.407	123.225	128.422	134.642	138.987
100	67.328	70.065	74.222	77.929	82.358	90.133	109.141	118.498	124.342	129.561	135.807	140.169
102	68.965	71.737	75.946	79.697	84.177	92.038	111.236	120.679	126.574	131.838	138.134	142.532
104	70.606	73.413	77.672	81.468	85.998	93.944	113.331	122.858	128.804	134.111	140.459	144.891
106	72.251	75.092	79.401	83.240	87.821	95.850	115.424	125.035	131.031	136.382	142.780	147.247
108	73.899	76.774	81.133	85.015	89.645	97.758	117.517	127.211	133.257	138.651	145.099	149.599
110	75.550	78.458	82.867	86.792	91.471	99.666	119.608	129.385	135.480	140.917	147.414	151.948
112	77.204	80.146	84.604	88.570	93.299	101.575	121.699	131.558	137.701	143.180	149.727	154.294
114	78.862	81.836	86.342	90.351	95.128	103.485	123.789	133.729	139.921	145.441	152.037	156.637
116	80.522	83.529	88.084	92.134	96.958	105.396	125.878	135.898	142.138	147.700	154.344	158.977
118	82.185	85.225	89.827	93.918	98.790	107.307	127.967	138.066	144.354	149.957	156.648	161.314
120	83.852	86.923	91.573	95.705	100.624	109.220	130.055	140.233	146.567	152.211	158.950	163.648
122	85.520	88.624	93.320	97.493	102.458	111.133	132.142	142.398	148.779	154.464	161.250	165.980
124	87.192	90.327	95.070	99.283	104.295	113.046	134.228	144.562	150.989	156.714	163.546	168.308
126	88.866	92.033	96.822	101.074	106.132	114.961	136.313	146.724	153.198	158.962	165.841	170.634
128	90.543	93.741	98.576	102.867	107.971	116.876	138.398	148.885	155.405	161.209	168.133	172.957
130	92.222	95.451	100.331	104.662	109.811	118.792	140.482	151.045	157.610	163.453	170.423	175.278
132	93.904	97.163	102.089	106.459	111.652	120.708	142.566	153.204	159.814	165.696	172.711	177.597
134	95.588	98.878	103.848	108.257	113.495	122.625	144.649	155.361	162.016	167.936	174.996	179.913
136	97.275	100.595	105.609	110.056	115.338	124.543	146.731	157.518	164.216	170.175	177.280	182.226
138	98.964	102.314	107.372	111.857	117.183	126.461	148.813	159.673	166.415	172.412	179.561	184.538
140	100.655	104.034	109.137	113.659	119.029	128.380	150.894	161.827	168.613	174.648	181.840	186.847
142	102.348	105.757	110.903	115.463	120.876	130.299	152.975	163.980	170.809	176.882	184.118	189.154
144	104.044	107.482	112.671	117.268	122.724	132.219	155.055	166.132	173.004	179.114	186.393	191.458
146	105.741	109.209	114.441	119.075	124.574	134.140	157.134	168.283	175.198	181.344	188.666	193.761
148	107.441	110.937	116.212	120.883	126.424	136.061	159.213	170.432	177.390	183.573	190.938	196.062
150	109.142	112.668	117.985	122.692	128.275	137.983	161.291	172.581	179.581	185.800	193.208	198.360
200	152.241	156.432	162.728	168.279	174.835	186.172	213.102	226.021	233.994	241.058	249.445	255.264
250	196.161	200.939	208.098	214.392	221.806	234.577	264.697	279.050	287.882	295.689	304.940	311.346
300	240.663	245.972	253.912	260.878	269.068	283.135	316.138	331.789	341.395	349.874	359.906	366.844
400	330.903	337.155	346.482	354.641	364.207	380.577	418.697	436.649	447.632	457.305	468.724	476.606
500	422.303	429.388	439.936	449.147	459.926	478.323	520.950	540.930	553.127	563.852	576.493	585.207
600	514.529	522.365	534.019	544.180	556.056	576.286	622.988	644.800	658.094	669.769	683.516	692.982
700	607.380	615.907	628.577	639.613	652.497	674.413	724.861	748.359	762.661	775.211	789.974	800.131
800	700.725	709.897	723.513	735.362	749.185	772.669	826.604	851.671	866.911	880.275	895.984	906.786
900	794.475	804.252	818.756	831.370	846.075	871.032	928.241	954.782	970.904	985.032	1001.630	1013.036
1000	888.564	898.912	914.257	927.594	943.133	969.484	1029.790	1057.724	1074.679	1089.531	1106.969	1118.948

The exact critical values for **DF** > 150 can be calculated as follows:

For $\alpha = 0.05$, $\chi_{0.05} = 1.0499\ \mathbf{DF} + 26.664$

For $\alpha = 0.025$, $\chi_{0.025} = 1.0623\ \mathbf{DF} + 30.192$

For $\alpha = 0.01$, $\chi_{0.010} = 1.0739\ \mathbf{DF} + 36.555$

For $\alpha = 0.005$, $\chi_{0.005} = 1.0781\ \mathbf{DF} + 43785$

Table 5

Table 5. Standard normal distribution graphs for proportions (α is level of significance).

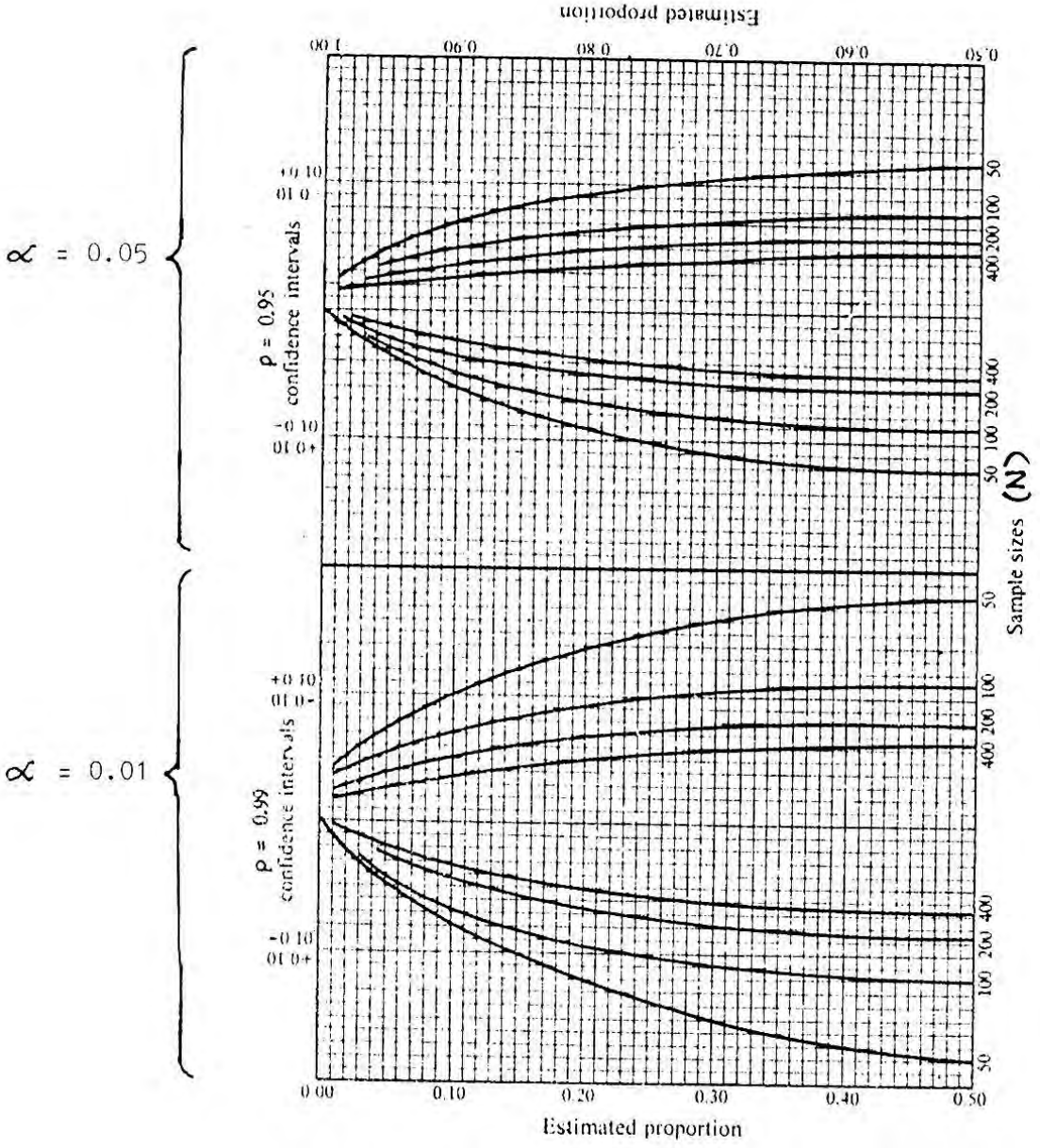


Table 6

Table 6. Critical values for the Cramer-von Mises normality test (ω^2 function); α is level of significance.

Omega-square

α =	0.10	0.05	0.025	0.001
w^2_α	0.104	0.126	0.148	0.178

Table 7

Table 7. Critical values for the Mann-Whitney U-test.

The table is divided into a series of boxes. Each column of boxes corresponds to a given number of specimens in sample 1 (N_1) each row of boxes to a given number of specimens in sample 2 (N_2). The test should be arranged so that N_1 is smaller than N_2 if the two are not equal. Within each box, there are four critical values of U, arranged thus:

number of tails
one two

$\alpha = 0.05$
$\alpha = 0.01$

N_2 \ N_1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
3	0 –																
4	0 –	– –	1 –	0 –													
5	1 –	0 –	2 –	1 0	4 1	2 0											
6	2 –	1 –	3 1	2 0	5 2	3 1	7 3	5 2									
7	2 0	1 –	4 1	3 0	6 3	5 2	8 4	6 3	11 6	8 4							
8	3 0	2 –	5 2	4 1	8 4	6 3	10 6	8 4	13 8	10 6	15 10	13 8					
9	3 1	2 0	6 3	4 1	9 5	7 3	12 7	10 5	15 9	12 7	18 11	15 9	21 14	17 11			
10	4 1	3 0	7 3	5 2	11 6	8 4	14 8	11 6	17 11	14 9	20 13	17 11	24 16	20 13	27 19	23 16	
11	5 1	3 0	8 4	6 2	12 7	9 5	16 9	13 7	19 12	16 10	23 15	19 13	27 18	23 16	31 22	26 18	
12	5 2	4 1	9 5	7 3	13 8	11 6	17 11	14 9	21 14	18 12	26 17	22 15	30 21	26 18	34 24	29 21	
13	6 2	4 1	10 5	8 3	15 9	12 7	19 12	16 10	24 16	20 13	28 20	24 17	33 23	28 20	37 27	33 24	
14	7 2	5 1	11 6	9 4	16 10	13 7	21 13	17 11	26 17	22 15	31 22	26 18	36 26	31 22	41 30	36 26	
15	7 3	5 2	12 7	10 5	18 11	14 8	23 15	19 12	28 19	24 16	33 24	29 20	39 28	34 24	44 33	39 29	
16	8 3	6 2	14 7	11 5	19 12	15 9	25 16	21 13	30 21	26 18	36 26	31 22	42 31	37 27	48 36	42 31	
17	9 4	6 2	15 8	11 6	20 13	17 10	26 18	22 15	33 23	28 19	39 28	34 24	45 33	39 29	51 38	45 34	
18	9 4	7 2	16 9	12 6	22 14	18 11	28 19	24 16	35 24	30 21	41 30	36 26	48 36	42 31	55 41	48 37	
19	10 4	7 3	17 9	13 7	23 15	19 12	30 20	25 17	32 26	28 22	44 32	38 28	51 38	45 33	58 44	52 39	

Table 8

Table 8. Critical values of Rao's K_0 for directional data.

Confidence semi-interval (semi-sector) $\Delta\gamma$	Significance level α	Required K_0
5°	0.10	354.6292
	0.05	504.0610
	0.01	870.6882
10°	0.10	88.7598
	0.05	126.1125
	0.01	217.9236
20°	0.10	22.1772
	0.05	31.5221
	0.01	54.4496

Table 9

Table 9. Special critical values for the Fisher t-test of Spearman's rank-correlation coefficient when $n \leq 8$ (α is level of significance).

n	$\alpha = 0.05$	$\alpha = 0.01$
≤ 4	—	—
5	1.000	—
6	0.886	1.000
7	0.750	0.893
8	0.714	0.857

Appendix

matrix algebra

From:
Stanley, L.T. (1973). *Practical Statistics for Petroleum Geologists.* Petroleum Publishing Company, Tulsa, Oklahoma.

Matrix methods

Nomenclature and General Application of Matrices

Though the nomenclature and application may seem rather complex at first, matrix procedures really are highly formalized and easy to follow.

An example of a matrix is

$$A = \begin{bmatrix} a_{11} & a_{12} \cdots a_{1n} \\ a_{21} & a_{22} \cdots a_{2n} \\ a_{m1} & a_{m2} \cdots a_{mn} \end{bmatrix}$$

The array has m rows and n columns, and would be called a matrix of $m \times n$ order.

For example, the two equations

$$a_{11}x + a_{12}y = k_1$$
$$a_{21}x + a_{21}y = k_2$$

are represented by the matrix

$$\begin{bmatrix} a_{11} & a_{12} & k_1 \\ a_{21} & a_{22} & k_2 \end{bmatrix}.$$

If there are n equations in n variables their matrix has n rows and n + 1 columns. The square array formed by a table of the n columns and n rows such as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

gives rise to a number called a *determinant* of the n^{th} order and expressed analogously by use of vertical bars, thus:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

In this case, where n = 2, the determinant is merely another way of writing the difference of the diagonal products. This is

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

The three determinants that form the elements of the example matrix are

$$D = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad D_x = \begin{vmatrix} k_1 & a_{12} \\ k_2 & a_{22} \end{vmatrix}, \quad D_y = \begin{vmatrix} a_{11} & k_1 \\ a_{21} & k_2 \end{vmatrix}$$

and

$$x = D_x/D; \; y = D_y/D$$

By orderly development of determinants the procedure of solving n simultaneous equations is reduced to a routine.

Matrices can be added, subtracted, multiplied, multiplied by a constant, transposed, and inverted.

Addition of two matrices is accomplished by adding corresponding elements; consequently addition is possible only for two matrices with the same number of rows and columns. For example,

$$\begin{bmatrix} 1 & 5 & 1 \\ 2 & 0 & 4 \\ 3 & 1 & 2 \end{bmatrix} + \begin{bmatrix} 4 & 2 & 0 \\ 1 & 3 & 1 \\ 2 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 5 & 7 & 1 \\ 3 & 3 & 5 \\ 5 & 1 & 4 \end{bmatrix}$$

Subtraction is defined in a similar fashion.

Multiplication of a matrix by any constant means multiplying every element in the matrix by the constant, e.g.,

$$4 \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 8 & 4 \\ 8 & 20 & 12 \end{bmatrix}$$

Given two matrices, the first with m_1 rows and n_1 columns and the second with m_2 rows and n_2 columns, their product is defined only if $n_1 = m_2$, i.e., the number of columns of the first equals the number of rows of the second. For two such matrices the elements of the product matrix are found by summing the element-by-element products of a row of the first matrix by a column of the second. For example.

$$A \times B = C$$

$$\begin{bmatrix} 4 & 1 & 3 \\ 2 & 2 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 5 \\ 2 & 5 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 9 & 28 \\ 7 & 21 \end{bmatrix}$$

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} = (4)(1) + (1)(2) + (3)(1) = 9$$
$$c_{12} = a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} = (4)(5) + (1)(5) + (3)(1) = 28$$
$$c_{21} = a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} = (2)(1) + (2)(2) + (1)(1) = 7$$
$$c_{22} = a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} = (2)(5) + (2)(5) + (1)(1) = 21$$

In matrix multiplication it is not necessarily true that AB = BA; order of multiplication usually makes a difference.

The transpose of a matrix, denoted by A′, is an m × n matrix whose rows are the columns of A and whose columns are the rows of A. If

$$A = \begin{bmatrix} 2 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

$$A' = \begin{bmatrix} 2 & 1 \\ 4 & 2 \\ 2 & 3 \end{bmatrix}$$

The inverse of a matrix is defined for square matrices only and is another matrix, A^{-1}, such that $A^{-1}A = I$ where the matrix I consists of ones down the main diagonal and zeros elsewhere, e.g., for a 4 × 4 matrix

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

The inverse exists only if the determinant of the matrix is nonzero.

Division of one matrix by another is an inverse operation of multiplication. Let

$$A \times B = G$$

Then multiplying the equation on the left by A^{-1}

$$A^{-1} \times A \times B = A^{-1}G$$

but $A^{-1}A = I$, the identity matrix—which plays the same role in matrix algebra that 1 plays in ordinary algebra—so

$$B = A^{-1}G$$

This means that the operation of dividing matrix G by matrix A to get matrix B actually consists of multiplying matrix G by the inverse matrix of A.

As mentioned before, the elements of a square matrix define the determinant of a matrix. It may be expressed as a polynomial by expanding the determinant by minors according to Cramer's rule. A minor of an element of a determinant is the determinant of one less order found by striking out the row and column containing the particular element. The cofactor A_{ij} of the element a_{ij} is obtained by multiplying $(-1)^{i+j}$ by the minor of a_{ij}.

Cramer's rule permits evaluation of a determinant by multiplying the

elements of any row or column by the corresponding cofactors and summing these products.

The element a^{ij} of the inverse A^{-1} of the matrix A is obtained by dividing the cofactor A_{ij} by the determinant $|A|$ of the matrix, that is,

$$a^{ij} = A_{ij}/|A|$$

For example, in order to find the inverse A^{-1} of the matrix

$$A = \begin{bmatrix} 0 & 3 & 2 \\ 1 & 2 & 4 \\ 3 & 0 & 2 \end{bmatrix}$$

it would be necessary to evaluate the determinant

$$|A| = \begin{vmatrix} 0 & 3 & 2 \\ 1 & 2 & 4 \\ 3 & 0 & 2 \end{vmatrix} = 0 \times \begin{vmatrix} 2 & 4 \\ 0 & 2 \end{vmatrix} - 3 \times \begin{vmatrix} 1 & 4 \\ 3 & 2 \end{vmatrix} + 2 \times \begin{vmatrix} 1 & 2 \\ 3 & 0 \end{vmatrix}$$

$$= 0 - 3(2 - 12) + 2(0 - 6) = 18$$

The cofactors are

$$A_{11} = (-1)^{1+1} = \begin{vmatrix} 2 & 4 \\ 0 & 2 \end{vmatrix} = 4$$

$$A_{12} = (-1)^{1+2} = \begin{vmatrix} 1 & 4 \\ 3 & 2 \end{vmatrix} = 10, \text{ etc.}$$

The complete matrix of cofactors is found to be

$$A_{ij} = \begin{bmatrix} 4 & 10 & -6 \\ -6 & -6 & 9 \\ 8 & 2 & -3 \end{bmatrix}$$

Upon interchanging rows and columns and dividing each element by A, the inverse matrix is found to be

$$A^{-1} = \begin{bmatrix} 4/18 & -6/18 & 8/18 \\ 10/18 & -6/18 & 2/18 \\ -6/18 & 9/18 & -3/18 \end{bmatrix}$$

It may be verified numerically that

$$A \times A^{-1} = A^{-1} \times A = I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

For example, the first element of the I matrix is

$$(0)(4/18) + (3)(10/18) + (2)(-6/18) = 1, \text{ etc.}$$

Least-squares regression—The matrix of normal equations

The least-squares principle is to minimize the sum of the squared deviations between the observed y values and those given by the regression equation. This is accomplished by taking the partial derivatives of

$$\text{SSD} = \sum_{i=1}^{N} (y_i - b_0 x_{0i} - b_1 x_{1i} - \cdots - b_p x_{pi})^2$$

with respect to b_0, b_1, b_2, etc. and equating each of these derivatives to zero. The resulting system of *normal equations* has the form

$$\begin{aligned}
b_0\Sigma x_0x_0 + b_1\Sigma x_0x_1 + b_2\Sigma x_0x_2 + \cdots + b_p\Sigma x_0x_p &= \Sigma x_0y \\
b_0\Sigma x_1x_0 + b_1\Sigma x_1x_1 + b_2\Sigma x_1x_2 + \cdots + b_p\Sigma x_1x_p &= \Sigma x_1y \\
b_0\Sigma x_2x_0 + b_1\Sigma x_2x_1 + b_2\Sigma x_2x_2 + \cdots + b_p\Sigma x_2x_p &= \Sigma x_2y \\
&\vdots \\
b_0\Sigma x_px_0 + b_1\Sigma x_px_1 + b_2\Sigma x_px_2 + \cdots + b_p\Sigma x_px_p &= \Sigma x_py
\end{aligned}$$

In these equations x_0 is a dummy variable that is always equal to one.

It can be seen that the normal equations are expressible in matrix form as

$$A \times B = G$$

where

$$A = \begin{bmatrix} \Sigma x_0^2 & \Sigma x_0x_1 & \Sigma x_0x_2 & \cdots & \Sigma x_0x_p \\ \Sigma x_1x_0 & \Sigma x_1^2 & \Sigma x_1x_2 & \cdots & \Sigma x_1x_p \\ \Sigma x_2x_0 & \Sigma x_2x_1 & \Sigma x_2^2 & \cdots & \Sigma x_2x_p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \Sigma x_px_0 & \Sigma x_px_1 & \Sigma x_px_2 & \cdots & \Sigma x_p^2 \end{bmatrix}$$

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_p \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} \Sigma y \\ \Sigma x_1y \\ \Sigma x_2y \\ \cdot \\ \cdot \\ \cdot \\ \Sigma x_py \end{bmatrix}$$

The matrix A is a square symmetric matrix, while the matrices B and G are single-column matrices. Note that matrix A is symmetric around its principal diagonal and the elements of the principal diagonal are of the form Σx_i^2. The element in the ith row and jth column is Σx_ix_j, where the summation is over the N observations in these two variables. Clearly $\Sigma x_ix_j = \Sigma x_jx_i$, that is, the matrix is symmetric.

Mathematicians have shown that the solution for the matrix of regression coefficients is $B = A^{-1}G$.

Designating

$$A^{-1} = C = \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0r} \\ c_{10} & c_{11} & \cdots & c_{1r} \\ \cdots & \cdots & \cdots & \cdots \\ c_{r0} & c_{r1} & \cdots & c_{rr} \end{bmatrix}$$

then

$$\begin{bmatrix} b_0 \\ b_1 \\ . \\ b_k \\ . \\ b_\nu \end{bmatrix} = \begin{bmatrix} c_{00}c_{01} & \cdots & c_{0\nu} \\ c_{10}c_{11} & \cdots & c_{1\nu} \\ \cdots & \cdots & \cdots \\ c_{k0}c_{k1} & \cdots & c_{k\nu} \\ \cdots & \cdots & \cdots \\ c_{\nu 0}c_{\nu 1} & \cdots & c_{\nu\nu} \end{bmatrix} \times \begin{bmatrix} \Sigma y \\ \Sigma x_1 y \\ . \\ \Sigma x_k y \\ . \\ \Sigma x_\nu y \end{bmatrix}$$

Hence, since two matrices are equal if, (and only if) corresponding elements are equal,

$$b_k = c_{k0}\Sigma y + c_{k1}\Sigma x_1 y + \cdots + c_{k\nu}\Sigma x_\nu y$$

Instead of inverting matrix A directly, as illustrated earlier, in order to obtain the individual elements c_{ij} of the C-matrix, it is more convenient to proceed indirectly. Since $A \times A^{-1} = I$ and $A \times C = I$,

$$\begin{bmatrix} \Sigma x_0^2 & \Sigma x_0 x_1 & \cdots & \Sigma x_0 x_\nu \\ \Sigma x_1 x_0 & \Sigma x_1^2 & \cdots & \Sigma x_1 x_\nu \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma x_\nu x_0 & \Sigma x_\nu x_1 & \cdots & \Sigma x_\nu^2 \end{bmatrix} \times \begin{bmatrix} c_{00}c_{01} & \cdots & c_{0\nu} \\ c_{10}c_{11} & \cdots & c_{1\nu} \\ \cdots & \cdots & \cdots \\ c_{\nu 0}c_{\nu 1} & \cdots & c_{\nu\nu} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

and, since two matrices are equal if, and only if, the corresponding elements are equal, it is immediately possible to write p sets of equations.

For example, the first set may be obtained by multiplying all the rows of the A matrix by the first column of the C matrix and setting the sums of these products in turn equal to the elements of the first column of the identity matrix, thus obtaining

$$\begin{aligned} c_{00}\Sigma x_0^2 &+ c_{10}\Sigma x_0 x_1 + \cdots + c_{\nu 0}\Sigma x_0 x_\nu = 1 \\ c_{00}\Sigma x_1 x_0 &+ c_{10}\Sigma x_1^2 + \cdots + c_{\nu 0}\Sigma x_1 x_\nu = 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ c_{00}\Sigma x_\nu x_0 &+ c_{10}\Sigma x_\nu x_1 + \cdots + c_{\nu 0}\Sigma x_\nu^2 = 0 \end{aligned}$$

In a similar manner the other p-1 sets of equations can be obtained. The solution of these sets of equations results in the elements c_{ij} of the inverse matrix $C = A^{-1}$.

The Abbreviated Doolittle method

Where a large number of equations is involved, matrix inversion would be a laborious process, but for several short-cut methods that have been devised. Probably the most well known is the *Abbreviated Doolittle technique*. Actually, to get the elements of B (the regression coefficient matrix) it is necessary to carry through only the *forward portion* of the Abbreviated Doolittle method.

Table B-1 shows symbolically the steps involved in the forward portion of the Abbreviated Doolittle method for a system of five normal equations. Lines (1) to (5) contain the elements a_{ij} of the A-matrix. The lower left-hand corner of the matrix is omitted because of its symmetry with the upper right-hand portion. The elements g_i are elements of the G-matrix. The j column is used to check the computations and will be discussed later.

TABLE B-1
TABLE FOR ABBREVIATED DOOLITTLE METHOD

Line							
(1)	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	g_1	j_1
(2)		a_{22}	a_{23}	a_{24}	a_{25}	g_2	j_2
(3)			a_{33}	a_{34}	a_{35}	g_3	j_3
(4)				a_{44}	a_{45}	g_4	j_4
(5)					a_{55}	g_5	j_5
(6)	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	$A_{1\nu}$	j_6
(7)		B_{12}	B_{13}	B_{14}	B_{15}	$B_{1\nu}$	j_7
(8)		A_{22}	A_{23}	A_{24}	A_{25}	$A_{2\nu}$	j_8
(9)			B_{23}	B_{24}	B_{25}	$B_{2\nu}$	j_9
(10)			A_{33}	A_{34}	A_{35}	$A_{3\nu}$	j_{10}
(11)				B_{34}	B_{35}	$B_{3\nu}$	j_{11}
(12)				A_{44}	A_{45}	$A_{4\nu}$	j_{12}
(13)					B_{45}	$B_{4\nu}$	j_{13}
(14)					A_{55}	$A_{5\nu}$	j_{14}
(15)						$B_{5\nu}$	j_{15}

The solution starts by the recopying of line (1) as line (6), that is $A_{11} = a_{11}$, $A_{12} = a_{12}$, . . . $A_{1\nu} = g_{1\nu}$. Line (7) is obtained from line (6) by dividing each number in line (6) by the leading element, A_{11}, that is

$$B_{12} = A_{12}/A_{11},\ B_{13} = A_{13}/A_{11},\ \text{etc.}$$

Line (8) is obtained by

$$\begin{aligned} A_{22} &= a_{22} - A_{12}B_{12} \\ A_{23} &= a_{23} - A_{12}B_{13} \\ A_{24} &= a_{24} - A_{12}B_{14} \\ A_{25} &= a_{25} - A_{12}B_{15} \\ A_{2\nu} &= g_2 - A_{12}B_{1\nu} \end{aligned}$$

Line (9) is obtained by

$$B_{23} = A_{23}/A_{22},\ B_{24} = A_{24}/A_{22},\ \text{etc.}$$

Line (10) is obtained by

$$\begin{aligned} A_{33} &= a_{33} - A_{13}B_{13} - A_{23}B_{23} \\ A_{34} &= a_{34} - A_{13}B_{14} - A_{23}B_{24} \\ A_{35} &= a_{35} - A_{13}B_{15} - A_{23}B_{25} \\ A_{3y} &= g_3 - A_{13}B_{1y} - A_{23}B_{2y} \end{aligned}$$

Line (11) is obtained by

$$B_{34} = A_{34}/A_{33},\ B_{35} = A_{35}/A_{33},\ \text{etc.}$$

Line (12) is obtained by

$$\begin{aligned} A_{44} &= a_{44} - A_{14}B_{14} - A_{24}B_{24} - A_{34}B_{34} \\ A_{45} &= a_{45} - A_{14}B_{15} - A_{24}B_{25} - A_{34}B_{35} \\ A_{4y} &= g_4 - A_{14}B_{1y} - A_{24}B_{2y} - A_{34}B_{3y} \end{aligned}$$

Line (13) is obtained by

$$B_{45} = A_{45}/A_{44},\ \text{etc.}$$

Line (14) is obtained by

$$\begin{aligned} A_{55} &= a_{55} - A_{15}B_{15} - A_{25}B_{25} - A_{35}B_{35} - A_{45}B_{45} \\ A_{5y} &= g_5 - A_{15}B_{1y} - A_{25}B_{2y} - A_{35}B_{3y} - A_{45}B_{4y} \end{aligned}$$

Line (15) is obtained by

$$B_{5y} = A_{5y}/A_{55}$$

When the point is reached where only a single B appears (here B_{5y}) the forward phase of the Doolittle is complete. The regression coefficients are now immediately available from

$$\begin{aligned} b_5 &= B_{5y} \\ b_4 &= B_{4y} - b_5B_{45} \\ b_3 &= B_{3y} - b_4B_{34} - b_5B_{35} \\ b_2 &= B_{2y} - b_3B_{23} - b_4B_{24} - b_5B_{25} \\ b_1 &= B_{1y} - b_2B_{12} - b_3B_{13} - b_4B_{14} - b_5B_{15} \end{aligned}$$

The j values are carried along as an arithmetic check on the rest of the work. The first six values are

$$\begin{aligned} j_1 &= a_{11} + a_{12} + a_{13} + a_{14} + a_{15} + g_1 \\ j_2 &= a_{12} + a_{22} + a_{23} + a_{24} + a_{25} + g_2 \\ j_3 &= a_{13} + a_{23} + a_{33} + a_{34} + a_{35} + g_3 \\ j_4 &= a_{14} + a_{24} + a_{34} + a_{44} + a_{45} + g_4 \\ j_5 &= a_{15} + a_{25} + a_{35} + a_{45} + a_{55} + g_5 \\ j_6 &= j_1 \end{aligned}$$

Subsequent j values are generated by operations analogous to those producing the A and B values, that is

$$\begin{aligned} j_7 &= j_6/A_{11} \\ j_8 &= j_2 - A_{12}j_7 \\ j_9 &= j_8/A_{22} \\ j_{10} &= j_3 - A_{13}j_7 - A_{23}j_9 \\ j_{11} &= j_{10}/A_{33} \\ j_{12} &= j_4 - A_{14}j_7 - A_{24}j_9 - A_{34}j_{11} \\ j_{13} &= j_{12}/A_{44} \\ j_{14} &= j_5 - A_{15}j_7 - A_{25}j_9 - A_{35}j_{11} - A_{45}j_{13} \\ j_{15} &= j_{14}/A_{55} \end{aligned}$$

If the j's are calculated in this way they serve as a check on the A's and B's as follows:

$$\begin{aligned} 1 + B_{12} + B_{13} + B_{14} + B_{15} + B_{1y} &= j_7 \\ A_{22} + A_{23} + A_{24} + A_{25} + A_{2y} &= j_8 \\ 1 + B_{23} + B_{24} + B_{25} + B_{2y} &= j_9 \\ A_{33} + A_{34} + A_{35} + A_{3y} &= j_{10} \\ 1 + B_{34} + B_{35} + B_{3y} &= j_{11} \\ A_{44} + A_{45} + A_{4y} &= j_{12} \\ 1 + B_{45} + B_{4y} &= j_{13} \\ A_{55} + A_{5y} &= j_{14} \\ 1 + B_{5y} &= j_{15} \end{aligned}$$

This checking system exposes immediately the errors made at each stage of the forward solution. A final over-all check on the b's can be made by substituting them back in the normal equations.

The elements of the inverse of A, denoted by c_{ij}, may now be found from the relations given below. Starting with c_{55}, the element of the lower right corner, work back towards c_{11}, as follows:

$$\begin{aligned} c_{55} &= 1/A_{55} \\ c_{45} &= -B_{45}c_{55} \\ c_{35} &= -B_{34}c_{45} - B_{35}c_{55} \\ c_{25} &= -B_{23}c_{35} - B_{24}c_{45} - B_{25}c_{55} \\ c_{15} &= -B_{12}c_{25} - B_{13}c_{35} - B_{14}c_{45} - B_{15}c_{55} \\ c_{44} &= 1/A_{44} - B_{45}c_{45} \\ c_{34} &= -B_{34}c_{44} - B_{35}c_{45} \\ c_{24} &= -B_{23}c_{34} - B_{24}c_{44} - B_{25}c_{45} \\ c_{14} &= -B_{12}c_{24} - B_{13}c_{34} - B_{14}c_{44} - B_{15}c_{45} \\ c_{33} &= 1/A_{33} - B_{34}c_{34} - B_{35}c_{35} \\ c_{23} &= -B_{23}c_{33} - B_{24}c_{34} - B_{25}c_{35} \\ c_{13} &= -B_{12}c_{23} - B_{13}c_{33} - B_{14}c_{34} - B_{15}c_{35} \\ c_{22} &= 1/A_{22} - B_{23}c_{23} - B_{24}c_{24} - B_{25}c_{25} \\ c_{12} &= -B_{12}c_{22} - B_{13}c_{23} - B_{14}c_{24} - B_{15}c_{25} \\ c_{11} &= 1/A_{11} - B_{12}c_{12} - B_{13}c_{13} - B_{14}c_{14} - B_{15}c_{15} \end{aligned}$$

This completes the calculation of the inverse matrix. For a symmetric matrix the inverse will also be symmetric so that the lower left corner need not be calculated. A final check should be made to see if $C \times A = 1$.