



Statistics and data analysis in geology pdf books

The probabilities attached to other drilling programs having different numbers of holes or probabilities of success can be found in a similar way. A n example of an infinite geologic population might be all possible thin sections of the box-and-whiskerplot, but all are graphs that show the spread of the central 50% of a distribution by a box whose lower limit is set at the first quartile and whose upper limit is set at the first quartile. Thus, we can have negative temperatures that are less than zero. A stratigraphic section, for example, would be recorded as a succession of beds, each one of a different lithology that are less than zero. than the immediately preceding bed. We can investigate the probability that it will require 2, 3, 4, ..., up to n exploratory holes before two discoveries are made. Starting at one estimated point, the distance to the next observation must be calculated, multiplied by the magnitude of the observation to give the rectangular area, and the process repeated through all successive observations up to the next estimated point. In the intervening years since the appearance of the second edition, computing technology has evolved with almost unbelievable speed. However, all of the various sequences do not have to be worked out individually, because the process of multiplying and summing is exactly that used for matrix multiplication. However, the mean deviation is less efficient than the sample 35 Statistics and Data Analysis in Geology-3'd ed. This is illustrated in Figure 2-17, which shows two symmetric frequency curves having different standard deviations. If the transition probability matrix is multiplied by itself (that is, the matrix is squared), the result is the second order transition probability matrix describing the second-order Markov properties of the succession: 0.78 0 0.12 0 0.22 0 0.29 0.64 0.11 0.60 0.40 1' [= 0.64 0.05 0.26 0.11 0.02 0.52 0.09 0.04 0.31 0.39 0.54 0.62 0.02 0.03 0.11 0.23 1 Note that the rows of the squared matrix also sum to 100%. A depth of 3000 ft in a well is ten times a depth of 300 ft, and the decade between the years 1940 and 1950 has the same duration as the interval between 1950 and 1960. This book is about quantitative methods for the analysis of geologic data-the area of Earth science which some call geomathematics and others call mathematical geology. Figure 5-22 is copyright by the American Statistical Association. Table 2-1 lists five values for chromium, in parts per million (ppm), obtained by spectrographic analysis of replicate splits of a Pennsylvanian shale specimen from southeastern Kansas. A (sample) statistic is an unbiased estimate of the equivalent (population) parameter if the average value of the statistic, from a large series of samples, is equal to the parameter. If t i is the total number of events, we can calculate the centroid, S, by 182 Analysis of Sequences of Data 5 25 0 :20- Y L 0) n 2 .-0 0 15- Y o. There are two important characteristics in either record. This is shown graphically in Figure 4-3 b, with the resulting sequence of equally spaced values derived from the data in part a. If you have worked through the small examples in this chapter, you can readily appreciate that the computationallabor involved in dealing with large matrices can be formidable, even though the underlying, individual mathematical steps are simple. In a homogenous, isotropic material, failure (te., faulting) will tend to occur along this plane. Ratoosh [Eds.], 1959, Measurement: Definitions and Theories: John Wiley & Sons, Inc., New York, 274 pp. What is the probability that the drilling program will be a total failure, with no discoveries among the four features tested? The steps in multiplication are demonstrated below on the two matrices, First, multiply a_{11} by $b_{11} = 35$, ft = 35, ranks into 100 equal-sized categories, each category is a percentile. But it also reflects, in part, the absence of anything quantitative in many geology courses. Three-dimensional stress measurements have been made in a pillar in a deep mine, yielding the following stress matrix: 4.1 -8.2 4.1 51.5 -3.0 -8.2 -3.0 32.3 61.2 1 The data are given in megapascals (MPa)and were recorded by strain gauges placed so the measurements have the same orientation as our imaginary cube (X increasing upward). If a matrix is n x n, there will be n! combinations which contain one element from each row and column, and whose first subscripts are in the order 1 , 2, . That is, (2.14) Another choice might be the average absolute deviation from the mean, or mean deviation, MD: MD = cz, 1% n -XI (2.15) The vertical bars denote that the absolute value (i.e., without sign) of the enclosed quantity is taken. Ortega, J.M., 1990, Numerical Analysis, a Second Course: Society for Industrial and Applied Mathematics Philadelphia, Pa., 201 pp. The orientation of the book is methodological, or "how-to-do-it." Theory is not emphasized for several reasons. Dozens of methods have been devised to solve sets of simultaneous equations, and hundreds of programmed versions exist. This is simply a verbal way of specifying the order of multiplication. KU has provided an intellectual greenhouse in which mathematical geology has flourished for over 30 years. 1 = -6 = 3 = 1!(3 - 1)! 1(2 * 1) 2 That is, there are three possible combinations that will contain one head. By graphing the percentile of each observation versus its value, we form a cumulative plot (Fig. The greatest danger, however, is to researchers themselves. In part this criticism reflects a certain mental rigidity present in some young minds that refuse to make an effort to stretch their imaginations. 353 Shape Certain square matrices have special importance and are designated by name. What a statistician describes as a sample would likely be called a "collection" or "suite of samples" by a geologist. Similarly, the matrix A can be said to be postmultiplied by B. Devising methods to estimate the magnitude of different sources of variation is one of the major tasks of statistics. (a) Target is a "li The multinomial distribution is an extension of the binomial where more than two mutually exclusive outcomes are possible. We can now apply test procedures outlined for the case of sequences of two dissimilar items (Eqs. The nature of the data and the chain determine the questions that we can consider. Systematic Patterns of Search Most geologists devote their professional careers to the process of searching for something hidden. They are still difficult, so their treatment in the chapter on matrix algebra has been rewritten and a new section on singular value decomposition and the relationship between R- and Q-mode factor methods has been added to the final chapter on multivariate analysis. 1-6. 618 testing goodness-of-fit t o a normal distribution xv Contents Table A.IO. However, many geologic problems involve the analysis of data collected without prior design. Conversely, we may wish to determine if the sequence contains an improbably low number of runs. Pure samples of these end members have distinct physical properties, so appropriate log responses can be used to estimate the relative amounts of halite, anhydrite, or shale at every foot within the Wellington Formation. However, if an m x n matrix is multiplied by an n x r matrix, the result is an m x r matrix. We would eventually build a matrix of frequencies similar to that below. This additional information can be used to determine the precision of the estimate of the rate of occurrence, to assess the appropriateness of the sampling scheme, to detect a trend, and to detect other systematic features of the series. As stated at the conclusion of Chapter 2, statistics is too large a subject to be covered in one chapter, or even one book. Logarithmic transformations in regression paperback reprint by SIAM to "foster beffer understanding of applied mathematics." Pettofrezzo, A.J., 1978, Matrices and Transformations: Dover Publications, Inc., New York, 133 pp. Systematic Patterns of Search We can test these sequences for randomness of occurrence by examining the number of runs. Log empirical survivor function of the Japanese volcano Aso. Our data may consist of the duration between successive events, or the cumulative length of time over which the semiaxes and the vectors are essentially radii. The text now includes numerous geological data sets that illustrate how specific computational procedures can be applied to problems in the Earth sciences. Educated in a tradition that emphasized the qualitative and distrustful of statistics Calculation of an estimate by rectangular integration is easy in theory but presents a somewhat difficult programming challenge. Inst. Many natural phenomena are expressed as complicated patterns of lines and areas that can best be described as fractals, which we will touch upon. Even though traditional forms of statistical tests may be beyond our grasp, the basic underlying concepts are the same. Figure 2-16. Subtle degrees of similarity, too slight for unaided detection, may provide the clues that will allow him to make a decision where none is otherwise possible. Most contouring programs operate by creating a regular grid of control points estimated from irregularly spaced observations. Table 4-3. These data are also in file LOUISMUD.TXT. If the diagonal elements were then discarded and the off-diagonal probabilities recalculated, the result would be the expected transition probability matrix for an embedded sequence whose states are independent. The choice of origin for histogram categories also may influence the shape of the histogram. Cumulative number of eruptions of the Japanese volcano Aso plotted against years of eruptions. Under the assumptions of the binomial distribution, each wildcat must be classified as either a discovery ("success") or a dry hole ("failure"). Successive wildcats 14 Elementary Statistics Number of heads Figure 2-2. Multidimensional Scaling The matrix is read "from rows to columns," meaning, for example, that a transition from state 168 Analysis of Sequences of Data Figure 4-5. The number of observations occurring in each state of a nominal system can be counted, and certain nonparametric tests can be performed on nominal data. So, A ~ = A X A and A3 = A2 x A = A X A X A Note that nonsquare matrices cannot be powered, because the number of rows and columns of a rectangular matrix would not accord if the matrix were multiplied by itself. The eigenvalues represent the magnitudes of the three orthogonal stresses. Indeed, a decline into scientific skepticism is one of the dangers that often traps geomathematicians. Work may be checked by multiplying the original matrix A by the inverted matrix, A - 1, which should yield the identity matrix 4 10 Because A-1A = I the following identities hold: 133 Statistics and Data Analysis in Geology - Chapter 3 A-I AX = A-~B IX = A - ~B IX = A B, the unknown matrix, X, is solved, A-' x [4:; -:; I [B = IE] = X [:I The column vector contains the unknown coefficients which we find to be equal to x1 = 2 and x2 = 3. These methods have been borrowed from all scientific and engineering disciplines and applied to every facet of Earth science; it is these more general techniques that are OUT concern. All of us have an intuitive concept of probability. ,n. Maximum likelihood factor analysis samples. However, n, the number of trials, is not specified. Repeated measurements made on the same object will display small differences whose magnitude may reflect both natural variation in the measurement process, and variation inadvertently caused by the person making the measurements. In effect, we have stacked the histogram bars so that the successive categories show the cumulative numbers or proportions. Samples should be drawn from populations in a random manner. Co., Malabar, Fla., 368 pp. A geologist's motive for numerical correlation may be a simple desire for speed, as in the production of geologic cross-sections from digitized logs stored in data banks. You should recognize that this is a quadratic equation of the general form ax2+ bx + c = 0 which can be solved for the unknown terms by factoring. The inverse matrix (or reciprocal matrix) A-l is one that satisfies the relationship A x A-l = I. Numbers of radiolarian tests per square centimeter in thin sections of siliceous shale. The general equation is where D1 is the spacing between one set of parallel seismic traverses and D Zis the spacing between the perpendicular set of traverses. Measured stratigraphic column in which lithologies have been classified into four mutually exclusive states of sandstone (a), limestone (b), shale (c), and coal (d). Tables A.10 and A.11 are copyright by Academic Press Inc. If we restrict our consideration to 2 x 2 matrices, we can represent this space as an illustration on a page and can view matrix are equal, the matrix is a scalar matrix. Almost all geological data consist of continuously distributed measurements made on ratio or interval scales, because these include the basic physical properties of length, volume, mass, and the like. Subtracting this matrix from M yields D, the data in the form of deviations from their means: D = M - VTn-IVM Here, n is the number of rows in M (te., the number of observations) and n-l is the inverse of n, or 1/20 175 Statistics and Data Analysis in Geology - Chapter 4 Step 1. The off-diagonal elements are the expected frequencies of transitions within the embedded sequence, assuming independence between successive states. However, be warned that anuncontrolled experiment (i e . In addition, each chapter ends with a set of exercises of greater or lesser complexity that the student can address using methods discussed in the text. Replicate Cr (ppm) 1 205 255 195 4 220 5 235 TOTAL= 1110 MEAN = 1110/5=222 2 3 than sample medians, hence they are more efficient in estimating the population parameter. Horizontal axis is ratio (major dimension of target)/(spacing between search lines). A to state C is counted as an entry in element a1,3 of the matrix. Table 4-8. For the second sequence, the test statistic is z= 2 - 10.9 = -4.2 2.1 z= 19 - 10.9 = -4.2 2.1 z= 10.9 = -4. expected in random sequences. The techniques also may be classified on the type of observations they require. A matrix may be multiplying each element in the matrix by the constant. No other scientists, with the exception of astronomers, are more removed from the bulk of their study material and less able to experiment . However, our science has benefited more than it has contributed to the exchange of quantitative techniques. If we divide a corrected sum of squares by n - 1 we obtain the variance, on their subject. Such curriculum requirements are now nearly ubiquitous in universities throughout the world. Induced Correlations ... and if we divide a corrected sum of products by n - 1we obtain the covariance. This is called the marginal (or fixed) probability vector: C D F1 0.44 0.08 You will recall from Chapter 2 (Eq. 2.7) that the joint probability of two events, A and B, is p(A,B) = p(BIA)p(A) rearranging, So, the probability that state B will follow, or overlie, state A is the probability that both state A and B occur, divided by the probability that state A occurs. 6. The x2 Distribution The determinant of such a matrix will be close to one, and its logarithm will be close to zero; this is the basis for one test of independence between variables. Here we obtained 11 heads; assuming the coin is unbiased, the probability of obtaining this number in 20 trials is 0.16 or about one in six. You may quickly discover that the answered in specific ways. These topics are extensively developed in most books on probability theory, such as those by Parzen (1960) or Ash (1970). We can imagine that our time series is subdivided into n intervals of equal duration. This obviously cannot happen if observations are taken only where states change. The median is the value midway in the frequency distribution. The following discussion leans heavily on his prepared notes and a subsequent article (Gould, 1967). We want to determine the permeability of the sandstone, but obviously cannot put 20 f t of core into our permeability apparatus. Note that the method of selection combination ~ 1 2 ~ 2 1 from places the elements in proper order according to their first, or row, subscript. graphs of the most significant transitions may reveal repetitive patterns in the succession. Magnesium concentration (parts per million) in water at 20 stream mouth. 33 Statistics and Data Analysis in Geology - Chapter 2 Summary Statistics The most obvious measure of a population or sample is some type of average value. Correspondence Analysis .. means that one set of vectors has been multiplied by (-1), reversing its "direction" but not changing its orientation in multivariate space. 2-17, curve u) has a mean of 20.1% so and a standard deviation of 4.3% so. We would expect similar trials to contain 9, 10, or 11 heads slightly more than one-third of the time. Plot of the normal frequency distribution. District Eastern Montana Border Central Clay (100,000 tons) Drilling Mud Foundry Clay Miscellaneous 105 63 5 218 80 2 220 76 1 As an illustration, Table 3-1 contains 1964 production figures for bentonite from three mining districts in Wyoming. The order of a matrix is an expression of its size, in the sense of the number of rows and/or the number . Then we will consider the analysis of various types of geologic data that have been classified arbitrarily into three categories: (1) data in which the sequence of observations is important, (2) data in which the two-dimensional relationships between observations are of columns it contains. Kruskal-Wallis test. important, and (3) multivariate data in which order and location of the observations are not considered. Our stratigraphic section exhibits first-order Markov properties; that is, the statistical dependency exists between points and their immediate predecessors. (This assumption is difficult to justify in most circumstances, as a discovery usually will affect the selection of subsequent drilling sites. Quartz and feldspar form mutually exclusive categories that cannot be meaningfully ranked in any way. For many geologists, a personal computer has replaced a small army of secretaries, draftsmen, and bookkeepers. The Book and the Course it Follows Readers are entitled to know at the onset where a book will lead and how the author has arranged the journey. These operations are most easily expressed in matrix algebra, so we will study this subject, illustrating the operations with geological examples. The length of each semiaxis is given by the corresponding eigenvalue. A generalized picture of compositional changes is desired, but the data seem too erratic for direct interpretation. We cannot directly observe and active magma chamber, but we can observe and record the seismic activity associated with a volcano. If the occurrence of states A and B are independent, or unconditional, and That is, the probability that state B will follow state A is simply the probability that state B occurs in 3! - 3 - 2 - 1 - -6= 3 = 2!(3 - 2)! 2 l(1) 2 (z) - HHH [HTH] TTT [HHT] [THH] THT HTT TTH These combinations are bracketed above in our collection of possible outcomes. There also are intermediate the section, which is given by the appropriate element in the fixed probability vector. Tests of significance possibilities. Some of these are distributed free or at nominal cost as "shareware." Computation is no longer among the major problems facing researchers today; they must be concerned, rather, with interpretation and the appropriateness of their approach. 1!0! = 0.90° 0.102 1 ' 1 * 0.01 = 0.01 17 Statistics and Data Analysis in Geology - Chapter 2 Number of holes drilled Figure 2-4. Observations lying beyond these extremes may be shown as dots. In the coin-tossing experiment, we are dealing with discrete outcomes-that is, specific combinations of heads and tails. For example, mining geologists and engineers have used samples to calculate tonnages and estimate ore tenor for centuries. There is a statistically significant tendency for certain states to be preferentially followed by certain other states. In a correlation matrix, the diagonal elements are all equal to one, so the trace is simply the number of variables. Obviously, geologists exploit conditional probabilities in all phases of their work, whether this is done consciously or not. To do this, we would utilize what is called an augmented matrix that has one more column than it has rows. The starting point of water, whereas the starting point on the Fahrenheit scale was chosen as the lowest temperature reached by an equal mixture of snow and salt. We can be sure that computer technology will continue to evolve at a dizzying pace; to provide some degree of security from obsolescence, the data files are available on the World Wide Web at two sites, one maintained by John Wiley & Sons and the other by the Kansas Geological Survey. It is an indication of the importance of one- and two-dimensional problems in the Earth sciences that they have been included in individual chapters. The negative binomial will appear again in Chapter 5, as it constitutes an important model for the distribution of points in space. The first category contains all classes of problems in which data have been collected along a continuum, either of time or distance. The list of Selected Readings can then provide a discussion of a specific subject in more detail. To further illustrate the nature of a continuous random variable, we can consider the probability tests on core samples. The probability tests on core samples. p = 7TD where P is the perimeter of the elliptical target. s U 2ok 0 0 10 15 20 Average porosity, % 25 30 Figure 2-14. Unfortunately, this test is not particularly efficient because degrees of freedom are lost when the series is divided into segments. Since the four states are mutually exclusive and exhaustive, the lithology must be one of the four and so their sum, given as the row total, is 100%. In the first, the distance between observations varies and must be specified for every point. With the help of Ricardo Olea, I have completely revised the discussion of the many varieties of kriging and provide a series of simple demonstrations to illustrate how geostatistical methodologies work. Equation (3.4) the probability of obtaining this many heads in 20 trials is the same as in the preceding example. This same function can be plotted in logarithmic form, as log Y against X. However, if we perform our measurement using a binocular microscope, we may obtain a length of 6.2 mm, by using an optical comparator we may measure 6.23 mm, and with a scanning electron microscope, 6.231 mm. We first calculate the expected mean and standard deviation of runs for random sequences having nl heads (nl = 11) and n2 tails (n2 = 9): aiJ 2 - (2 11 * 9)(2 * 11 9 - 11 - 9) = 4.6 (9 + 11)*(9 + 11 - 1) The test statistic is z=- U-U UU % 1 3 - 10.9 = 1.0 2.1 The number of runs in the sequence is one standard deviation from the mean of all runs possible in such a sequence, and does not fall within the critical region. The historic record of eruptions of the volcano Aso in Kyushu, Japan, has been kept since 1229 (Kuno, 1962), and is given in Table 4-5 and file ASO.TXT. 1 = 6. Matrix algebra also is an impossibly large subject to encompass in these few pages If the current season's exploration budget permits the drilling of only four of the prospects, we can determine the probabilities attached to the various possible outcomes. To compute variances and standard deviations, we generate intermediate quantities which can be used directly in many techniques we will discuss in following chapters. 24 0.422 0.062 1.6 = 0.105 We can calculate the probabilities attached to a succession of possible outcomes and plot the results in the form of a distribution, just as we have done previously. (a) Parallel-line search pattern. Most (but not all) of the eigenvalue operations we will consider later will be applied to correlation or covariance matrices, so these special Hutchinson Salt (file KANSALT:TXT) exercise. Geologic Maps, Conventional and Otherwise Geologists are carefully trained to read, utilize, and create maps; probably no other group of scientists is as adept at expressing and envisioning dimensional relationships. Absolute certainty is expressed at the ends of this scale, 0 and 1, with different degrees Bottom c Sometimes we are interested in the nature of transitions from one state to another, rather than in the relative positions of states in the sequence. The data are 125 airborne measurements of radiation emitted by 13'Cs, recorded on the Istrian peninsula of Croatia. radiolarians vary at random through the range of the ran porosities will be truncated at the high end, biasing the sample toward low values and giving an erroneously low estimate of the variation in porosity within the unit. The identity matrix, the resulting product is exactly the same as the initial matrix: [: :] 1 0 0 1 4 7 2 5 8 X O 10 = 258 [O 0 11 [3691 Thus, the identity matrix corresponds to the 1 of ordinary multiplication. In equation form, (2.16) A small standard deviation indicates that observations are clustered tightly around a central value. The problem is seldom too much data in one place. For the equation of X I, the numerator is the determinant of the matrix and for x2, it is the determinant of This procedure can be generalized to any set of simultaneous equations and provides one common method for their solution. The hypergeometric distribution can be used to estimate the probabilities that specified numbers of discoveries will be made if only some of the identified prospects are drilled. Table 3-2. The mean, or arithmetic average of the observations, may also be indicated by an asterisk or diamond. The narrower the class boundaries, the fewer the number of occurrences within the classes, and the lower the estimates of the probabilities of occurrence. As a consequence of these special conditions, the eigenvalues will always be real numbers that are equal to or greater than zero. A two-tailed test is appropriate, with hypotheses Ho: U = 8 Hi: U f 8 We can work through the test procedure for the first series of coin flips and determine the likelihood of achieving this sequence by a random process. This component of total radiation (see Fig. A numerical computation package such as , , or , and even some statistical packages, 152 Matrix Algebra such as , will provide all of the mathematical computation power you are likely to need for applications in the Earth sciences. 36 19 52 A - 1000 13 4 0 5 29 1000 B 45 12 3 5 2 1000 from c 44 1000 3 1 D 29 1000 9 9 23 E - 26 1120 1038 1094 1077 1067 5397 Grand Total Step 2. 0.904 5 . A diagonal matrix is a square, symmetric matrix in which all the off-diagonal elements are 0. An obvious difference in the two interpolation procedures is apparent when original data are sparse and more than one point must be estimated between two observations. If only one hole is a discovery, it may pay off the costs of the entire exploration effort. Magnesium and vanadium show such behavior in the vineyard samples. We will look at the statistics of directional data in both two and three dimensions. The total area under the distribution is 8/8, or 1. Maps are compact and efficient means of expressing spatial relationships and details-they are as important to Earth scientists as the conventions for scales and notes are to the musician. Geomathematical methods provide the objectivity necessary to avoid creating these patterns when they may exist only in the scientist's desire for order. We now have at ow command a system for reducing a square matrix into its determinant but no clear grasp of what a determinant "really is." Determinants arise in many ways, but they appear most conspicuously during the solution of sets of simultaneous equations. In contrast, variation may occur between members of a set if measurements or experiments are performed on a series of test objects. The shape of the target and the arrangement of the lines of search also influence the probability. There are a number of variants on the runs tests described here. A cumulative plot of the total number of events (nt)to have occurred at or before time t, against time t, is given in Figure 4-6. 128 Matrix Algebra To find element c11 To find element c32 84" The completed matrix multiplication has the appearance In general, if the order of multiplication is reversed to B x A = C, a different answer will be obtained: [i i a1 In the operation A x B = C, the matrix B is said to be pvemultiplied by A. This cannot be done; we are restricted to those holes which have actually been drilled, and perhaps to a few additional test holes whose drilling we can authorize. Unfortunately, Earth science students voice an almost universal complaint: material taught in such courses is not relevant to their studies. Next, how many possible combinations of three tosses contain exactly three heads? into a collection of guesses and hunches based on very little data, of which most are of a contradictory or inconclusive nature. However, a familiarity with the techniques discussed here will provide an introduction to a diverse field of analytical tools. Poisson probability distributions with different rates of occurrence, A, expressed as numbers of occurrences per interval. For X Z, the numerator is the determinant 38 = (4 x 110) - (10 x 38) = 60 I 1 0 1101 so x2 = 60/20 = 3. Permeabilities are determined by measuring the time required to force a certain amount of fluid, under standardized conditions, through a piece of rock. Discrete distribution giving the cumulative probability that two discoveries will be made by or before a specified hole when the probability of a discovery is 25%. That point is determined by summing the areas just found and dividing by the equally spaced interval to give the estimated value. These steps can be summarized in the diagram below. Not only must the Earth scientist understand the geological significance of the recorded variables, the mathematical significance of the measurement scales used must also be understood. In this book, we will always use the noun "sample" in the statistical tests are based on this supposition. If thickness is considered, it may obscure our examination of the succession, which is the subject of our interest. Jackson, J.E., 1991, A User's Guide to Principal Components: John Wiley & Sons, Inc., New York, 569 pp. These probabilities are powered by squaring and multiplied by the grand total to obtain new estimates can, in theory, be infinitely refined, which implies that we can always find a difference between two measurements, if we conduct the measurements, if we conduct the measurements at a fine enough scale. Table 4-6. You may not have been using. The length is equal to the first eigenvalue, or 1.28. This would result in the expectation of transitions from a state to itself, which are forbidden. 2 1001 L n k 5 2 5- 0 0 0 This statistic can in turn be used in Equation (4.4), Z = S - 1/2T T / r n (4.4) where T is the total length of the series, z is the standardized normal variate, and the significance of the test result can be determined by normal tables such as Appendix Table A. 5-1 b). Look back at Cramer's rule for the solution of simultaneous equations, in which the unknowns are expressed as the ratio of two determinants. The probability distribution that governs experiments such as this is called the binomial distribution. The categories include the transitions A A, A C, C A, and C C. TOP B D C and C C an D C D A E A C D C A C E D A B D B A E A C D C A C E D A B E A E A C D E A D E A C D E A D C D C B A E A Bottom - others. The discussion of nonparametric methods introduced in the 2nd edition has been expanded because geologic data, particularly data collected in the field, seldom satisfy the distribution assumptions of classical parametric statistics. Maps, in this general definition, include traditional geologic and topographic maps and also aerial photographs, mine plans, Statistics and Data Analysis in Geology - Chapter 5 peel prints, photomicrographs, and electron micrographs. This subject will be considered in detail in a later chapter, where its use in time-series analysis will be examined. Histogram of field-wide average porosities of oil fields producing from the "D' and "J" sands in the Denver-Julesburg Basin of Colorado. Fixed. The table shows the steps in calculating the mean, whose equation is simply (2.12) - Another characteristic of a distribution curve is the spread or dispersion about the mean. The determinant of a 2 x 2 matrix is therefore For a numerical example, we will consider the matrix [:;] Next, let us consider a more complex example, a 3 x 3 determinant: all 6.12 a13 a21 a31 a22 a32 6.23 a33 There are 3!, or 3 x 2 x 1 = 6, combinations of elements in a 3 x 3 matrix that contain one element from each row and column and whose first subscripts are in the order 1, 2, 3. A statistic we compute based on a sample taken from the population is used as an estimator of the desired parameter. ":]x[::]=[:] 3 4 415 -115 215 -315 1 Therefore, the unknown coefficients are X I = 3 and x2 = -2. 181 Statistics and Data Analysis in Geology - Chapter 4 10 0 0 10 20 30 40 50 0 t i + 1 - t i i n years Figure 4-10. Examples shown in Figure 2-11 are only a few of the possible histograms that could be constructed from these data., in are equal to 1, 2, ... If we regard the elements of the el form, 0.74 0.23 0.03 0.05 0.38 0.57 0.572 0.322 0.106 0.150 0.505 0.345 0.104 0.460 0.437 and 130 0.461 0.368 0.171 0.178 0.474 0.348 0.144 0.470 0.385 I I Matrix, it converges to a stable configuration (called the stationary probability matrix) in which each column of the matrix is a constant. 3 6 7 2. Suppose we start in limestone (state B). 1 / 2 1 / 2 = 1/8. Utilizing this speed, a researcher can compress literally a lifetime of trial solutions and refinements into a few minutes. Then, [y y 2.03 2.98 2.43 y 1 - 1 [;I]-[21 Vsh Perform the necessary matrix inversion and multiplications to determine the proportions of halite, anhydrite, and shale in the 50-ft interval of the Hutchinson Salt. Measurements include various electrical and sonic characteristics of the rocks, and both natural and induced radioactivity. If the original data possess a large random component which causes values to fluctuate widely, interpolated points may also fluctuate unacceptably. It remains posted on my wall to t h s day. The binomial distribution is not appropriate for this problem because the probability of a discovery changes with each exploratory hole. Trend-surface analysis; contouring is an extension of interpolation procedures. Instead, the product np = h is all that is needed, which is given by the rate of occurrence of events. Rather than a single piece of rock, suppose we have a sizable length of core taken from a borehole through a sandstone body. We can rewrite the sequence in Figure 4-13 in the following form: + + + - 0 - + - o+ Regarding the first zero as 'I - " gives a total of five runs, three of "+" and two of (it makes no difference in the number of runs if we call the second zero "+" or "-"). Successive lithologic states encountered in a drill hole through the Coal Measures in the Midland Valley of Scotland (after Doveton, 1971); mutually exclusive states are barren shale (A), shale with fossils of nonmarine bivalves (B), siltstone (C), sandstone (D), and coal or root zone (E); read across rows. 139 Statistics and Data Analysis in Geology - Chapter 3 What circumstances will lead to singularity? Horizontal axis is ratio fmaior dimension of targetl/fminimum spacing betwe& search lines). For example, chromium and vanadium have very similar behavior in the vineyard soil samples, leading to a high correlation between the two. The vertical axis gives the percent of intervals between eruptions that are longer than a specified duration, versus the duration in years along the horizontal axis. Since three unknowns must be estimated, it seems we will require three equations and, hence, measurements of three log properties. Shaded intervals in (a) and (b) have the same areas. Geologists need to be conscious of these problems, and of some of the statistical tools that are available to help solve the problems. A single, exact, "true" value cannot be determined; 26 Elementary Statistics rather, we will observe a continuous distribution of possible values. Copper and zinc are recorded in the second and fifth columns of M, and so their correlations are the elements Yi,j whose subscripts are 2 and 5: Rcu,.., = [Y212 "g5] r5,2 r5,S = [1 -0.28 -0.28 1 1 If we regard the rows as vectors in X and Y, we can plot each row as the tip of a vector that extends from the origin. Assume y1 and y2 are observed values at points X I and x2; we wish to estimate the value of y' at point x'. If the number of observations in each category are divided by the total number of observations, the scale reads in percent and the bar chart is a relative frequency histogram. Summary Statistics . . By premultiplying the vector of means by the transpose of V, we create a 20 x 5 matrix in which every row is the same as the vector of means. Discrete distribution giving the probability that a needle of length 8, when subsequent chapters of this book involve the solution of sets of simultaneous equations. Sedimentary petrologists have regarded grain-size and shape measurements as important sources of sedimentological information since the beginning of the last century. At one extreme, we may have a record which is quite precise, both in the variable which is measured and in the scale along which successive observations are located. 4, Pennsylvania State Univ., p. That is, both components of shear 155 Statistics and Data Analysis in Geology- Chapter 3 point in a positive coordinate direction, or both components of shear 155 Statistics and Data Analysis in Geology- Chapter 3 point in a positive coordinate direction, or both components of shear 155 Statistics and Data Analysis in Geology- Chapter 3 point in a positive coordinate direction. rows as the premultiplier matrix on the left and the same number of columns as the postmultiplier matrix is simply another diagonal matrix. The path followed by the unschooled is not an orderly progression through topics laid out in curriculum-wise fashion. Also, because the area under the estimated curve is equal to the length of interval they represent. Measured stratigraphic sections also have the form of series of lithologies, as may drill holes through zoned ore bodies where the rocks encountered are classified into different types of ore and gangue. Other procedures are ad hoc; results from their application must be judged on utilitarian rather than theoretical grounds. The classification of fossils as to type is an example of nominal measurement. Chi-square tests appropriate for embedded sequences are discussed by Goodman (1968). Now, we merely note that an n x m rectangular matrix, XXT. Then, equations such as (5.3) and (5.6) may seriously overestimate the probabilities of detection. Scvccsky 9 Statistics and Data Analysis in Geology- Chapter 1 SELECTED READINGS Churchman, C.W., and P. This result also is general. and reproduced with permission. Graduate students, as a matter of course, are expected to have proficiency in these areas. (n - 2) ' (n - 3) - ... This can be generalized to the number of possible combinations of n items taken Y at a time. Now that we have an idea of the manipulations that produce eigenvalues, we may try to get some insight into their nature. A more detailed discussion of these data is given in Doveton (1986). Later the self-taught amateur fills in gaps in his or her background and attempts to master the precepts of the techniques that have been applied. Such an ideal is unrealistic for all but a few fortunate individuals. Analysis o f variance At the opposite extreme, two variables which are completely unrelated will have a correlation of near zero. Similarly, mineralogists and crystallographers utilize mathematical techniques derived from physical and analytical chemistry. The desired interval is 50 cm. That is, Furthermore, the numerators can be expressed as determinants. We will examine a geologic application of the runs test to a traverse through a two-state system. Testing the equality of two sets of directional vectors However, in the final analysis, the researcher must be his own guide. Uniform density The condition indicates that two or more rows (or columns) of the matrix are linear combinations or linear transformations of other rows; that is, the values in some rows (or columns) are dependent on values in other rows. As an example, we will consider the full 5 x 5 correlation matrix R for trace metals from Istrian vineyard soils. 161 Statistics and Data Analysis in Geology - Chapter 3 Few mathematical procedures have received the attention given to matrix inversion. shell. This generally requires about 10 to 20 iterations, depending upon how closely the initial quesses were to the final, stable estimates. There are several practical reasons why we might wish to take samples. and shale. Principal Coordinates Analysis target, as calculated by the approximating equations, can be shown conveniently as graphs. Linear interpolation between two data points along a sequences. We can now define another measure of the "size" of a square matrix. Curve b is the same type of data from a field in West Texas. 435 437 Block kriging ... If we divide the row totals of the transition frequency matrix by the total number of transitions, we obtain the relative proportions of the four lithologies that are present in the section. Depth, ft 0.0 5.0 10.0 15.0 20.0 25.0 30.0 35.0 Moisture (g water/100 g dried solids) 124.0 78.0 54.0 35.0 30.0 21.0 22.0 18.0 The data in Table 4-8 are the moisture contents of samples from a core through Recent marine muds accumulating in a small inlet on the U.S. Gulf Coast in eastern Louisiana. If the matrix is raised to a sufficiently high power, it reaches a stable state in which the rows all become equal to the fixed probability vector, or in other words, becomes an independent transition probability matrix and will not change with additional powering. If events A and B are related (or dependent), the fact that A has already transpired tells us something about the likelihood that B will then occur. 416 418 Simple kriging Most geologists probably have not taken a course in matrix algebra. Table 4-1 is a classification of the various data-analysis techniques discussed in this chapter. McCammon (1977) gives the derivation of the geometric probabilities for circular and linear targets and parallel-line searches. Discrete distribution giving the probability of obtaining specified numbers of heads in three flips of a coin. Rectangular integration seems more appropriate in this instance than linear interpolation. If we premultiply this row vector by 1/20, it will contain the means of each of the five columns. You will note in the example that the highest transition probabilities are from one state to itself, particularly from sandstone to sandstone, from limestone to limestone, and from shale to shale. That is, if we read from the row labeled A to the column labeled C, we see that we move from state A into state C five times in the sequence. 394 397 Trend Surfaces As with other types of x 2 tests, each category must have an expected frequency of at least five transitions. A large slab of polished pegmatite used as a window ledge in the washroom of a geology building provides a way for students to investigate these alternative possibilities. 158 In this chapter we will consider ways of examining data that are Square the probabilities along the diagonal. 412 Pitfalls characterized by their position along a single line. Logs of oil wells have been made at too great a cost to ignore merely because the well locations do not fit into a predesigned sampling plan. Instead, we wish to find the probabilities are related to the thicknesses of the stratigraphic units being sampled and the distance between the sample points. If both sides of a matrix equation are multiplied by A-l, the matrix A is effectively removed from the left side. The various sources of fluctuation combine to yield a continuously random variable, which we are sampling by making repeated measurements. D Length of interval, years Figure 4-8. Vertical axis is duration of quiet before the it h eruption, and horizontal axis is duration of these, the data sequences are assumed to match at one position only, and we wish to determine the degree of similarity between the two sequences. Each term contains . The problem is analogous to statistical analysis; but unlike the classical statistician, we cannot design the pattern of holes or control the manner in which the data were obtained. In the one element from each row and each column. Joint Variation of Two Variables Nonparametric Methods 2nd edition, I revised the discussion of eigenvalues and eigenvectors because these topics had proved to be difficult for students. 123 125 127 132 136 141 141 150 153 157 4. Data from this core are shown in Figure 4-4a and are given in file GARNETS.TXT. Now, we can try the procedures just outlined to find the eigenvalues of the 2 x 2 matrix: -"] A = $[17 45 - 16 \text{ First}, \text{ we must set the matrix in the form Equating the determinant to zero, -6 1 1457 4 - 1 6 - h] = 0$ which can be collected to give A2 - h - 2 This can be factored into = 0 (A - 2) (A + 1) = 0 So, the two eigenvalues associated with the matrix A are This example was deliberately chosen for ease in factoring. These are "population" and "sample," two important concepts in statistics. Expanding Equation (2.4) gives (Y f X - 1) ! P = (1 - pIXpY (Y - 1)!x! If the regional success ratio is assumed to be lo%, the probability that a twohole exploration program will meet the company's goal of two discoveries can be calculated: (2 + 0 - 1)! * (1 - 0.10)O. This probability is proportional to the area of the frequency curve bounded by these limits. Nominal data may be represented by integers, alphabetic characters, or symbols. At the limit, we can imagine that two variables might behave in an identical fashion. Instead, we consider the probability of obtaining a result within a specified range. We are concerned, not with two distinct states, but whether an observation exceeds or is smaller than the preceding observation. Methods which are theoretically developed often are based on statistical assumptions so restrictive that the procedures are not strictly valid for geologic data. Selected Readings The probability that one event or the other happens is the sum of their separate probabilities; that is, p (discovery) + p (dry hole) = p (discovery) + p (dry hole). As a preparatory step to further analysis, the data may be approximated by equally spaced estimates. The others require an odd number of transposition and are therefore negative. Because of the general unfamiliarity of most Earth scientists with developments in the numerical analysis of data sequences, we have thought it best to present a potpourri of techniques and approaches. In fact, almost all studies of point distributions have been made by geographers. : - o ' 2 -2.0 -2.0 -1.0 .O 1.0 2 0 MG Figure 3-3. The principal motive behind the development of automatic contouring is economic, an attempt to utilize the petroleum industry's vast investment in stratigraphic data banks. A most useful property of normal distributions is that areas under the curve, within any specified range, can be precisely calculated and expressed in terms of 36 Next Page Elementary Statistics standard deviations from the mean. Our estimates of the likelihood of rain may be based on many different factors, including a subjective "feeling" about the matter. Many institutions, both commercial and public, provide extensive libraries of computer programs that will implement geomathematical applications. As Fisher pointed out (1953, p. The expected mean number of runs in a randomly generated sequence of nl items of state 1 and n2 items of state 2 is The expected variance in the mean number of runs is 2 = CTU 2nln2(2nlnZ - nl - n2)(n1 + n2I2(n1 + n2 - 1)(4.9) By these equations, we can determine the mean number of runs and the standard error of the mean number of runs in all possible arrangements of nl and n2 items. These tests are also included here, but because probability is essential to understanding statistics, this introductory section has been expanded. It may seem, perhaps, more logical to define variability as simply the average of deviations from the mean, but a few simple trials will demonstrate that this value will always equal zero. The conditional probability that an eruption will occur, given that harmonic tremors have been recorded, is denoted p (eruption 1 tremor). We might be interested, for example, in the repeated recurrence of certain environmentdependent microfossils in the chips recovered during the drilling of a well., one in which the investigator has no influence over where or how observations are taken) usually takes us outside the realm of classical statistics. So, we have "one-half of one-half of one-half" of a chance of getting all three heads. The offdiagonal elements, however, will change until a stable configuration is reached, corresponding to the independent events model. The well was drilled through 1600 ft of Coal Measures of Carboniferous age, consisting of interbedded shales, siltstones, and coal beds or root zones. The probability that one discovery will be made is 38%. . We can utilize a distance scale of feet between samples and consider this a problem related to depth or distance. The verb, "to sample," has essentially the same meaning for both geologists and statisticians and means the act of taking observations. ,n, taken in any order, and k is the number Logratio Transformation of exchanges of two elements necessary to place the i subscripts in the order 1, 2,. Continuous Random Variables. .. The rearranging may be performed by swapping any two adjacent elements. That is, percentile of X i = 100 (ran", "' xi) (2.11) where n is the number of observations. The intensity of diffracted radiation at . Note that the probability of some success is (1.00 - 0.07), or 93%. Rock specimens can be collected from the tops of batholiths in exposures along canyonwalls, but examples from the roots of these same bodies are hopelessly deep in the 20'28 is one variable, and the intensity at 30"28 is another. Elementary Statistics Earth. This is unfortunate; the subject is not difficult and is probably one of the most useful tools in mathematics. 145 Statistics and Data Analysis in Geology - Chapter 3 Table 3-3. Transposition is a matrix operation in which rows become columns and columns become rows. To all of these people, named and unnamed, I owe my deepest appreciation. 3), Lyell's subdivision of the Tertiary on the basis of the relative abundance of modern marine organisms is a statistical procedure. Notable among these is the geological text by Griffiths (1967), which 6 Introduction is in large part concerned with the effect sampling has on the outcome of statistical texts. Many people may find themselves reluctant to regard areas as a n appropriate measurement unit for the dispersion of lengths! Fortunately, in most instances where we are concerned with variance, it is standardized or converted to a form independent of the measurement units. The interpretation of subsurface structure from drill-hole data is a prominent example. Again, the test is onetailed. Other methods may be more appropriate or powerful in specific situations or for certain data sets. Unfortunately, these topics are frequently taught by persons who have little knowledge of geology or any appreciation.for the types of problems faced by Earth scientists. This approach limits us, of course, to small matrices, but the insights gained can be extrapolated to larger systems even though hand computation becomes impractical. Segment cdef is a run down even though the difference between d and e is zero. We are interested in the significance of the succession, but we cannot put a meaningful scale on the sequence itself. This equation is (2.17) or alternatively, (2.18) On hand calculators, C x i and E x : can be found simultaneously, thus reducing the number of operations by n. If the vertical scale on the bar chart reads in number of observations, the graphic is called a frequency histogram. The final topic is canonical correlation. From experience in nearby offshore tracts, it is believed that about 40% of such seismic features will prove to be productive structures. (This assumption is reasonable at the initiation of exploration, but becomes increasingly tenuous during a single time period (perhaps a budget cycle) for which the forecast is being made. In fact, the solution of simultaneous equations is probably the most important application of matrix inversion. As we make the lengths of the intervals progressively shorter, n becomes progressi possible when dealing with natural phenomena over which you have little control. Obviously, the succession of lithologies represents changes that occurred through time, but we have no way of estimating the time scale involved. The binomial equation can be solved to determine the probability of occurrence of any particular combination of successes and failures, for any desired number of trials and any specified probability. A measurement is a numerical value assigned to an observation which reflects the magnitude or amount of some characteristic. (a) h = 6.0. (b) h = 2.0. (c) h = 0.6. D, can be compared to two-tailed critical values given in Appendix Table A.7. If the statistic exceeds the critical value, the maximum deviation is larger than that expected in a sample collected at random from a Poisson distribution. In fact, any sort of twodimensional spatial representation is included. A matrix of this type is called a transition frequency matrix and is shown below. so than many other mathematical procedures. By definition, the axes of the ellipse are at right angles to each other, and the two plotted vectors also are almost orthogonal. Asymmetric distribution showing relative positions of mean, median, and mode. calculating the test statistic using the four categories whose expected frequency is greater than five. That is, the probability matrix, P, after n steps through the succession is equal to Pn. The n the probability matrix, P, after n steps through the succession is equal to Pn. data may lead to interpretations that are not obvious or apparent when other means of investigation are used. Since the presence of a sandstone is governed by factors that operated at the time of deposition, and since the occurrence of an anticlinal fold is presumed to be related to tectonic conditions at a later time, the two outcomes are unrelated, or independent. Such a sequence might be a cyclothem of limestone-shale-limestone-shale-limestone, from bottom to top. Hence, both the report and criticism pass over the heads of most of the intended audience. It must be emphasized that we are now using the expression "sequence" in the broadest possible sense. The sequent abc is a run up, because each observation is larger than the preceding one; similarly, the segment g h i is a run down. Eigenvalues. The problems of sampling in such circumstances are complex; some of the references at the end of this chapter discuss the effects of various sampling schemes and the relative merits of different sampling designs. 9 10 11 11 25 29 34 40 46 50 55 58 60 64 65 66 68 69 72 72 74 75 76 78 83 84 Contents Nested design in analysis of variance We can formulate a variety of statistical hypotheses which can be tested with this statistic. Since the order of multiplication of a series of numbers makes no difference in the product, that is, ~11 ~ ~ = 2 ~ ~ 2 32 ~ 311 ~ 1 3=3 ~ 33 ~ 22 and ~ 1 1 so on, we can rearrange our combinations without changing the result. Using matrix C, show that (C-1)T = (CT)-1. Convert the surveyor's measurements to coordinates relative to true north. John C. Otherwise, the shear stress is negative. However, the distinction between a positive determinant and a zero determinant is very important because a matrix whose determinant is zero cannot be inverted by ordinary methods. Explicit Location in Time or Space Interval or Ratio Data Interpolation Regression Splines Nominal or Ordinal Data Series of Events Zonation Seriation Autocorrelation Cross-correlation Senivariograms Periodograms Spectral Density Markov Chains Runs Tests Nature of Variables this category does not consider the spacing at all, and only the sequence of the conditional probability p (anomaly I mineralization) and the unconditional probability p (mineralization) from studies of known mining districts, but it may be more difficult to directly estimate p (mineralization) from studies of known mining districts. conditionally related to event A, the probability that event A will occur is simply the sum of the conditional probabilities p(AIBi) times the probabilities that the events Bi occur. Autocorrelation In Figure 2-16, one-half of the area below the distribution curve is to the right of the median, one-half is to the left. In the study of a metamorphic halo around an intrusive, a diamond-drill core was taken perpendicular to the intrusive wall. However, we can use quantitative mapping techniques that are either closely related to statistical procedures or rely on novel statistical concepts. Usually the area is geographical (a quadrangle, mining district, country, etc.) and the map is a method for reducing very large-scale spatial relationships so they can be easily perceived. The first semiaxis is twice as long as the second, which is almost twice the length of the third. In carefully prescribed games of chance, the probabilities attached to a specific outcome can be calculated exactly by combinatorial mathematics; we will use this concept of probability in our initial discussions because of its relative simplicity. Simultaneous R- and Q-Mode Analysis The "how-to" book of computer algorithms for numerical computation; contains succinct descriptions of eigenvalue techniques, including SVD. AXiS Sample 1 2 3 4 5 126 a 3.4 4.6 5.4 3.9 5.1 b 2.2 4.3 4.7 2.8 4.9 c 1.8 4.2 4.7 2.3 3.8 Matrix Algebra As a simple example, consider Table 3-2, which contains measurements of the a-,b-, and c-axes of chert pebbles collected in a glacial till. This can be done by computing the autocorrelation of the lengths between events. If we multiply a matrix by a value equal to the inverse of a constant, we obtain the same numerical result as if we divided each element of the matrix by the constant. Concentrations of selected elements (in ppm) measured in soil samples collected in vinevards and associated terraces on the lstrian peninsula of Croatia. been scaled in different ways. In addition, it may be necessary to examine the series in more detail, in order to estimate any pattern that may exist in the events. In practice it is not necessary to calculate the off-diagonal probabilities at all. A determinant is a single number extracted from a square matrix by a series of operations, and is symbolically represented by det A, IAI, or by It is defined as the sum of n! terms of the form where n is the number of rows (or columns) in the matrix, the subscripts il, i2, ... She checked the illustrations and equations for consistency in style and format, designed the layout, selected the book type, and in a Herculean effort, set the entire manuscript in camera-. Elongated ellipse defined by rows in matrix of correlations between chromium and vanadium, which are highly correlated. Samples within the unit at depth do not have the same opportunity of being chosen as samples at outcrops. Fortunately, there ready form using the T\$ typesetting language. Weighted regression are many powerful computational tools available at modest cost (at least for student versions), and they run on almost any type of personal computer. In a probability, p (ili), the probability, p (ili), the probability 169 Statistics and Data Analysis in Geology - Chapter 4 - that state i will be the next state to occur, given . We will next consider procedures for detecting trends, or systematic changes in average value, and will find that runs tests may be used to good advantage in conjunction with these procedures. An answer may depend upon how a question is framed. If the target is that the present state is i. 433 An example ... assumed to be elliptical and the search consists of parallel lines, the probability that a line will intersect a hidden target of specified size, regardless of where it occurs within the search area, can be calculated. The slope of a straight line connecting any two points on the cumulative plot is the average number of events per unit of time for the interval between the two points. We must have a large sample to utilize the normal approximation method presented here, but in most geologic problems, adequate numbers of samples will be available. Cumulative plot of individual porosity measurements used t o construct Figures 2-12 and 2-13. In the specific set of equations given above, we have [;:] 1;: [::I = [1::] To solve the equation, the matrix A will be inverted and B will be multiplied by A-l to give the solution for X. The final two scales, the interval and ratio, are those we ordinarily think of as "measurements" because they involve determination of the magnitudes of an attribute. For example, if X is the 3 x 3 matrix x=[i i] is 9, ~ 1 is 3 7, x21 is 2, and so on. These probabilities have already been computed and tabulated for many combinations of n, Y, and p. The log empirical survivor function is especially good for showing departures from randomness, which appear as deviations from the straight-line form of the plot (Fig. Therefore, the sum of all the probabilities within each experiment must total 1.00, because we are absolutely certain to obtain a result from among those enumerated. Equation (2.13) is called the definitional equation of variance. However, these ubiquitous plastic boxes with their colorful screens seem to promise much more than just word-processing and spreadsheet calculations-if only geologists knew how to put them to use in their professional work. Does the abundance of 189 Statistics and Data Analysis in Geology - Chapter 4 10 40 0 5 10 15 20 25 Number of radiolarians per cm2 30 Figure 4-14. For example, we could estimate the magnesium content of the river at 500-m intervals, or even at every 5 m, but it is obvious that these new values would provide no additional information on the distribution of the metal in the stream. Geologic problems where its use is appropriate are not common, but McCray (1975) presents an example from geophysical exploration for petroleum. This is equivalent to finding the square root of the matrix. Estimate of transition probabilities of diagonal elements, found by dividing row totals by grand total. Turning to the second eigenvalue, A2 = 0.72, the equation set is 1 - 0.72 - 0.28 1 - 0.72 - 0.28 1 - 0.72 - 0.28 1 - 0.72 - 0.28 1 - 0.72 - 0.28 1 whose solution gives the second eigenvector: I:[= [:] In Figure 3-1, th s will plot as the vector drection 1/1 = 45", perpendicular to the major semiaxis of the ellipse. A granite, for example, may

have feldspar phenocrysts whose longest axes have an average length of 13.2 mm and a variance of 2.0 mm2. The sequence of successes and failures by the collector during the course of a day forms a special type of time series. removed and then divided by the standard deviation. Variation induced into measurements by inaccuracy of instrumentation is most apparent when repeated measurements are made on a single object or a test is repeated without change. Bill had the foresight to realize that the massive, expensive mainframe dinosaurs of computing in the 1960's standard deviation. would evolve into the compact, indispensable personal tools of every working geologist, and his vision kept the KGS at the forefront of computer applications. What is the probability that one discovery will be made? Curve u represents the percent oil saturation (so) measured in cores from the producing zone of a northeastern Oklahoma oil field. By finding the eigenvalues and eigenvectors of the 3 x 3 stress matrix, we can rotate the imaginary cube into a coordinate system in which all the shear stresses will be zero. Some of the methods involve nonparametric statistics, and these are not widely considered in introductory statistics courses. An identity matrix has a determinant equal to 1.0. If several variables are completely independent of each other, their correlations will be near zero and they will form a correlation matrix that approximates an identity matrix. Selecting the appropriate distance between sampling points can be a vexing problem; if observations are too closely spaced, the transition matrix reflects mainly the thickness of the more massive stratigraphic units. The sequence limestone limestone limestone limestone limestone is also possible. We can work through the first cycle of the procedure. The first two are the nominal scale and the ordinal scale, in which observations are simply classified into mutually exclusive categories. Approximately 95% of all observations are included within the interval from +2 to -2 standard deviations, and more than 99% are covered by the interval lying three standard deviations on both sides of the mean. Eigenvectors Eigenvalues If the data are expressed in the form of the matrix E, we may multiply E by the constant 25.4 to obtain a matrix containing the measurements in millimeters: 25.4 x 3.4 4.6 2 5 . The central role of geostatistics and regionalized variable theory in the study of the spatial behavior of geological and other properties is now firmly established. Table A S is used with the permission of McGraw-Hill Companies. Davis Lawrence, KS viii CONTENTS Page . A single toss has two outcomes, heads or tails. Histogram bars from Figure 2-12 stacked to form a cumulative distribution. Calculating the hypergeometric probability consists simply of finding all of the possible combinations of producing and dry features within the population, and then Preface enumerating those combinations that yield the desired number of discoveries. The hypergeometric probability distribution can be applied to our offshore concession that contains ten seismic features, of which four are likely to be structures containing reservoirs. Demonstrate this to yourself by inverting the 3 x 3 matrix below: If we need the inverse of a diagonal matrix, the problem is much simpler. A closely related problem is the analysis of a sequence characterized by the presence or absence of some variables at points along a line. The polished surface allows easy discrimination of adjacent grains, so a line drawn on the ledge produces a sequence through the quartz and feldspar grains in the pegmatite. Ideally, of course, these people should receive formal instruction in probability, statistics, numerical analysis, and programming; then they should study under a qualified geomathematician. Figure 4-13 shows a typical sequence that can be analyzed by means of a runs test. Therefore, the number of runs does not suggest that of objects that are at variance with accepted classifications, or may show relationships between variables where none were expected. First, the variable being measured is expressed in units of an interval or ratio scale; 1000 bbl of oil is twice as large a quantity as 500 bbl, and a measurement of 10 R is ten times the resistance of 1SZ. However, by utilizing the rules of matrix multiplication, an operation can be performed that is equivalent to solving the equation AxX=B for the unknown matrix, X, when the elements of A and B are known. Series-of-events models may be appropriate for certain types of spatially distributed data. Ratio scales are the highest form of measurement. A quick review of spatially distributed data. The discussion in the following chapters begins with the basic topics of probability and elementary statistics, including the special steps necessary to analyze compositionaldata, or variables such as chemical analyses and grain-size categories that sum to a constant. Also note that, unlike most matrices we have calculated before, the transition frequency matrix is asymmetric and in general ai, j # aj, i. Similarly, there are five transitions from state C to state A in the sequence; this number appears as the matrix element defined by row C and column A. 2-10) appears bell-shaped, and is called a normal distribution. The shaded area under the curve is the same in both illustrations. Here we are presented with a situation that is different from the river data; observations are more abundant than estimates and we wish to preserve as much of the original information as possible. These two subjects are separated by a wide gulf that would require several years to bridge following a typical course of study. - (0.35) (0.64) (0.35) (0.64) + (0.80) (0.36) = 0.44 Similarly, the probability that the fragment of marine fossil mosasaur came from the smaller basin is 0.56 Fortunately for the students who must search the area, it seems somewhat more likely that the fragment of marine fossil mosasaur came from the smaller basin is 0.56 Fortunately for the students who must search the area, it seems somewhat more likely that the fragment of marine fossil mosasaur came from the smaller basin than from the smaller basin is 0.56 Fortunately for the students who must search the area, it seems somewhat more likely that the fragment of marine fossil mosasaur came from the smaller basin than from the smaller basin is 0.56 Fortunately for the students who must search the area, it seems somewhat more likely that the fragment of marine fossil mosasaur came from the smaller basin is 0.56 Fortunately for the students who must search the area, it seems somewhat more likely that the fragment of marine fossil mosasaur came from the smaller basin that the fragment of marine fossil mosasaur came from the smaller basin that the fragment of marine fossil mosasaur came from the smaller basin that the fragment of marine fossil mosasaur came from the smaller basin that the fragment of marine fossil mosasaur came from the smaller basin that the fragment of marine fossil mosasaur came from the smaller basin that the fragment of marine fossil mosasaur came from the smaller basin the fragment of marine fossil mosasaur came from the smaller basin that the fragment of marine fossil mosasaur came from the smaller basin that the fragment of marine fossil mosasaur came from the smaller basin the fragment of marine fossil mosasaur came from the smaller basin the fragment of marine fossil mosasaur came from the smaller basin the fragment of marine fossil mosasaur came from the smaller basin the fragment of marine fossil mosasaur came from the smaller basin the fragment of marine fossil mosasaur came from the fragment of marine fossil mosasaur came from the fragment ordinal measurements is restricted primarily to counting observations in the various states. Therefore, the probability that a hole will be dry is 1 - p. Chapter 10 discusses singular value decomposition. 1 The Book and the Course it Follows Wilkes (1966) devotes an entire chapter to various interpolation procedures. Some authors have found that the eigenvalues extracted from the transition probability matrix are useful indicators of cyclicity. Then, their rows in R would be identical and the plotted vectors would coincide. A geologic equivalent of this problem consists of the appearance of feldspar and quartz grains along a traverse across a thin section. Usually the object of the search is an undiscovered oil field or an ore body, but for some it may be a flaw in a casting, a primate fossil in an excavation, or a thermal spring on the ocean's floor. College courses in matrix algebra usually are sprinkled liberally with theorems and their proofs. An intuitive look at eigenvalues and vectors by geometric analogy. For example, the determinant 1 2 3 4 5 6 = O 2 4 6 is zero because the third row of the matrix is simply twice the first row. The data are given in file CROATRAD.TXT. Analysis of Sequences of Data A bundle of programs for analyzing sequences of data can readily be obtained from many sources. The Matrix A matrix is a rectangular array of numbers, exactly the same as a table of data. As an additional example of the solution of simultaneous equations by matrix form and solve for x1 and x2 by inversion, 2x1 + x2 = 4 3x1 4 - 4x2 = 1 The steps in the inversion process can be written out briefly: [; 2. We can, in effect, divide by the appropriate standard deviations if we create a 5 x 5 matrix, C, whose diagonal elements are all 0.0. If we invert C and premultiply by D, each element of D will be divided by the standard deviation of its column. Or we can fail to reject the null hypothesis, implying that we have failed to find any indication of nonrandomness. From these data we must deduce as best we can the configuration of the horizon between boreholes. The nominal scale of measurement consists of a classification of observations into mutually exclusive categories of equal rank. 677-680. Citations are made to more specialized or advanced texts that I have found to contain especially lucid discussions of the points in question, however, we must know the special procedures of matrix multiplication. Rather than reducing a rectangular data matrix to a symmetrical, square correlation or covariance matrix and then extracting the desired eigenvalues and eigenvectors as we have done, these programs obtain results directly from the data matrix by singular value decomposition (SVD). In the methods discussed in this chapter, the classes are not ranked; that is, state A is not "greater" or "larger" in some sense than states B or C. Matrix inversion can, of course, be applied to square matrices of any size, and not just the 2 x 2 examples we have investigated so far. Parts of Table A.9 are copyright by the American Statistical Association and by the American Statistical Association a the top row and pick an entry from each row. The greater number of coin tosses can be accommodated only by making the histogram bars ever more narrow, and the histogram bars ever more narrow, and the histogram bars ever more narrow. problems as the distribution of mineral grains in thin sections. A similar experiment could be performed that would involve a much larger number of trials. Schubert, 1982, Geodynamics Applications of Continuum Physics to Geological Problems: John Wiley & Sons, Inc., New York, 450 pp. Various measures of this property have been suggested, but only two are used to any extent. The measurements were recorded in inches and we wish to convert them to millimeters. Particularly in older books, you may encounter different conventions for denoting individual elements of a matrix. statement is that every outcome on a continuous scale of measurement is unique, and that the probability of obtaining a specific, exact result must be zero! If this is true, it would seem impossible to define probability on the basis of relative frequencies of occurrence. Least-Squares Methods and Regression Analysis In many types of problems, we are concerned not only with changes along a sequence, but are also interested in where these changes occur. F - t e s t of equality of variances Their presence in the silicified shale suggested a causal relationship, so a sequence of samples was collected at approximately equal intervals in an exposure through the unit. The data in Table 4-2 consist of analyses of the magnesium concentration in stream samples collected along a river. A computer program could be written that will perform the necessary calculations, but the programming effort probably exceeds the difficulty of computing the test statistic by hand. = 24 Elementary Statistics Continuous Random Variables To introduce the next topic we must return briefly to the binomial distribution. Events in which the occurrence of one outcome precludes the occurrence of the other outcome are said to be mutually exclusive. Discrete distribution giving the probability of obtaining specified numbers of heads in ten flips of a coin. Many older geologists have been caught short in the computer revolution. The five eigenvalues of this matrix are $A = 12.453\ 1.233\ 0.789\ 0.465\ 0.061\ L$ and their associated eigenvectors are $0.585\ -0.363\ Vp = 0.498\ 0.465\ 0.061\ L$ and their associated eigenvectors are $0.585\ -0.363\ Vp = 0.498\ 0.465\ 0.061\ L$ Geology - Chapter 3 Each eigenvector can be regarded as a set of coordinates in five-dimensional space that defines the "direction" of a semiaxis of a hyperellipsoid. Contour mapping is the subject of one section in this chapter. A feeling of skepticism may be compounded by the absence of mathematical applications in geology courses. This "relative frequency" approach to probability is intuitively appealing to geologists, because the concept is closely akin to uniformitarianism. The increase between 10" and 20" C is exactly the same as the increase between 10" and 20" C. exploration company is determined to discover two new fields in a virgin basin it is prospecting, and will drill as many holes as required to achieve its goal. Now at 3 Statistics and Data Analysis in Geology - Chapter 1 the time of this third edition, there are many easy-to-useinteractive programs to perform almost any desired statistical calculation; these programs have graphical interfaces and run on personal computers. In this example, we are assuming that the successive points have integer values. Critical values of ?i; for the test of uniformity of a spherical distribution 3! - 3.2.1 = 1 = 3!(3 - 3)! 3 2 - 1 (1) (i) 13 Statistics and Data Analysis in Geology - Chapter 2 Figure 2-1. Map relationships are almost always expressed in terms of points located on the map. All possible outcomes, denoting heads as H and tails as T, are: HHH HTH TTH HTT HHT THH [THT] [TTH] Bracketed combinations are those that satisfy our requirements that they contain only one head. 0 I ~, I J I ~. However, even though it is impossible to observe a number of outcomes that are, for example, exactly 6.000. Usually, an investigator is interested in the mean rate of occurrence, or number of events per interval of time. Davis Kansas Geological Survey The University of Kansas John Wiley & Sons New York Clxchester Brisbane Toronto Singapore ASSOCIATE E D I T O R MARKETING MANAGER PROGRAM COORDINATOR PRODUCTION E D I T O R DESIGNER C O V E R PHOTO M a r k Gerber Kevin Molloy Denise Powell Brienna Berger Madelyn Lesure Bill B a c h m a d P h o t o Researchers This book was printed and bound by Courier. An even more obvious example is a structural contour map. However, the presence of one crystal, perhaps feldspar, might stimulate the local crystallization of additional crystals of feldspar Examples and exercises ure drawn from the biological sciences. We must attempt to wring from them what knowledge we can, recognizing the bias and imperfections of that knowledge. Next, we will consider the eigenvalues of the third-order matrix: eventually producing a patchwork texture. Q-Mode Factor Analysis a. Measurements of axes of pebbles (in inches) collected from glacial till. The probability attached to this sequence is p (B -, B) X p (B + B) = 71 % 71% = ~ 50% - - Since the other transitions limestone and back again in two steps. 1229 1239 1240 1265 1269 1270 1272 1273 1274 1281 1286 1305 1324 1331 1335 1340 1346 1369 1375 1376 1377 1387 1388 1434 1438 1473 1485 1505 1506 1522 1533 1542 1583 1584 1587 1598 1611 1612 1613 1620 1631 1637 1649 1668 1675 1683 1691 1708 1709 1765 1772 1780 1804 1806 1814 1815 1826 1827 1828 1829 1830 1854 1872 1874 1884 1894 1897 1906 1916 1920 1927 1928 1929 1931 1932 1933 1934 1935 1958 1962 a complex stratovolcano, but all historic eruptions have been explosive, ejecting ash of andesitic composition. i o 1. If no trends are detected in the rate of occurrence, we may conclude that the series of events is stationary. 1, Geology-Data processing. Matrix Algebra Most geological populations extend deep into the Earth and are not accessible in their entirety. Factor rotation Multiplying a matrix by its transpose results in a square, symmetric matrix product whose size is determined by the order of multiplication. These are the proportions of the specific lithologies represented by the columns. Techniques for examining data having a single positional characteristic traditionally are considered part of the field of time-series analysis, although we will take the broader view that time and space relationships can be considered interchangeably. Detailed discussions and references can be found in Stevens (1946), the book edited by Churchman and Ratoosh (1959) and, from a geologist's point of view, in Griffiths (1960). Splines It may be more useful to consider the distribution of the probability that more than x dry holes must be drilled before the goal of Y discoveries is achieved. Curvilinear regression ... However, the differences in probability are very small and, of course, depend upon the reasonableness of the assumptions used to estimate the probabilities. The source for each table and figure is noted in its caption and a complete citation is given at the end of each chapter in Suggested Readings. A finite population might consist of all oil wells drilled in Kansas in 1963. If the observations are expressions of magnitude, they ordinarily will contain fractional parts, and ties (two successive points with identical values) are unlikely. Data in file PROSPECT.TXT were taken from a surveyor's notebook describing the outline of a gold prospect in central Idaho. An example is a point-count traverse across a thin section, where the states are the minerals noted at succeeding points. A sequence in which the state at one point is partially dependent, in a probabilistic sense, on the preceding state is called a Markov chain (named after the Russian statistician, A.A. Markov). At this depth, the well penetrated the Hutchinson Salt member of the Permian Wellington Formation, which was under investigation as a possible nuclear waste disposal site. The novice proceeds backwards, attracted first to those methods that seem to offer the greatest help in the research, exploration, or operational problems being addressed. This plot reveals any tendency for intervals to be followed by intervals and B, are orthogonal, then ATB = 0. The symbol axxrepresents the normal stress directed onto the east or west face of the cube; it will be a positive value if the stress is compressional and a negative value if the stress is compressional and a negative value if the stress is called Cramer's rule. If the two are conditional, the occurrence of the observable event tells us something about the likely state of the hidden event. Geological measurements such as geochemical, petrophysical analyses, grain-size distributions-in fact, any set of values expressed as percentages-constitute compositional data and are subject to closure effects. As an example, in Chapter 2 you were asked to compute the variances and covariances of trace-element data given in Table 2-3. Figure 5-2a gives the probability of detecting an elliptical target whose shape ranges from a circle to a line, using a search pattern of parallel lines. Since a histogram covers the entire range of observations, the sum of the areas of all the bars will represent either the total number of observations or 100%. 342 343 Ruler procedure We can now subtract the means from each observation to convert the data into deviations. Successive divisions of a distribution are called quantizes. To provide a statistic that describes dispersion or spread of data around the mean, and is in the units of measurement of the data, we can calculate the standard deviation. The appearance and fidelity of the finished map is governed to a large extent by the fineness of the grid system and the algorithm used to estimate values at the grid intersections. If the sequence is relatively uniform, the two will be very similar, but if there is a trend the centroid will be displaced in the direction of increasing rate of occurrence. We may check for a trend in two ways. Nonparametric correlation Although computing tools have advanced rapidly, our skills as educators have not kept pace. The elements in a term are selected in order from row 1, 2, ...,n, but each combination can contain only one element from each column. Testing hypotheses about circular directional data The empirical survivor function is obtained by plotting the percent "survivors," or Y = proportion of time intervals longer than X, against X = length of time 180 Analysis of Sequences of Data interval. Apparent grain density (P b), g/cc Sonic transit time (At), psec/ft Halite Anhydrite Shale 2.03 2.98 2.43 67 50 113 Two useful petrophysical properties are the apparent density (in grams per cubic centimeter) as measured by gamma-ray absorption and sonic transit time (in microseconds per foot). However, because modern programs for automatic correlation are increasingly able to mimic (and extend) the mental processes of a human interpreter, their output must be considered seriously. Sometimes a plausible explanation cannot be found; but in other instances, new theories may be suggested which would otherwise have been overlooked. In an asymmetric distribution such as that shown in Figure 2-16, the mode is the highest point on the frequency curve. Any geologist can buy an inexpensive computers that served entire corporations and universities only a few short years ago. Their correlations are given by elements in the first and fourth rows of RCY,,,are plotted as vectors in Figure 3-2. In some instances, the addition of geologic interpretation to the raw data contained in the observation points is a valuable enhancement of the map. Although not intuitively obvious, the variance has properties that make it far more useful than other measures of scatter. Coin tossing is, then, a clear-cut example of discrete probability. Goodness-of-fit test . The relative size of the target is found by dividing the target's maximum dimension by the search line Figure 5-3. Sometimes, however, geologic judgment becomes biased, and the subtle Eigenvectors of matrix correspond t o principal semiaxes (arrows) of ellipse. The book represents a rigorous, formal approach to the interpretation of geologic phenomena using statistical methods. Geologists are effects of personal opinion detract rather than add to the utility of a map. Multivariate Extensions of Elementary Statistics concerned not only with the analysis of data in sequences, but also with the comparison of two or more sequences. If we do, our chances of obtaining a second toss is not dependent in any way on the first. The specific gravity is 3.34 for olivine, 2.76 for anorthite, and 5.20 for magnetite Finally, because of the absolute invariance in operation of a computer program, mathematical correlation provides a challenge to the human interpreter. Instead, we cut small plugs from the larger core at intervals and determine the permeability of each. There is always a point beyond which the increase in information gained from additional observations is not worth the increase in the cost of obtaining them. Keep in mind the little phrase posted on the wall of every computation center: "GIGO-Garbage In, Garbage Out." The first edition of this book began and ended with quotations; these were repeated in the second edition. Rather, these are a collection of operations that have proved valuable in quantitative problem-solving in the Earth sciences, or that seem especially promising. An interesting series of probabilities can be formed based on coin tossing. Searle, S.R., 1982, Matrix Algebra Useful for Statistics: John Wiley & Sons, Inc., New York, 438 pp. Runs tests are another example of the nonparametric . The runs-up-and-down test generally is regarded as the most powerful of the runs tests because it utilizes changes in magnitude of every point with respect to adjacent points procedures introduced in Chapter 2. In this example, the proportions are 23% sandstone, 45% shale, and 32% limestone. Ordinary kriging . Increasingly, computer programs for multivariate analysis employ alternative techniques for obtaining eigenvalues and eigenvectors. Beginning with AxX=B premultiply both sides by the inverse of A, or A-l: xB A-'xAxX=A-l Since A-l x A = I and I x X = X, the equation reduces to X = A-' X B (3.2) Thus, the problem of division by a matrix reduces to one of finding a matrix that satisfies the reciprocal relationship. by some physical procedure, and these inherently are limited in both their accuracy and precision. The next topic is matrix algebra. 4 Introduction My contention is that a quantitative approach to geology can yield a fruitful return to the investigator; not so much, perhaps, by "proving" a geological hypothesis or demonstrating its validity, but by gaining insights from the critical examination of phenomena that is prerequisite to any quantitative procedure. Ricardo has been my guide through the sometimes controversial field of geostatistics, and John has generously shared the store of instructional material and student exercises that he has patiently assembled through years of teaching petrophysics. A protracted succession of dry holes will also cause a shift in an exploration program.) The probability of a discovery is assumed to remain unchanged. All rights reserved No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01 923, (508)750-8400, fax (508)750-4470. In geochemical analyses, it is common practice to make multiple determinations, or replicates, of a single sample. The comparison by x 2 methods yields a test statistic of x2 = 172. In most experimental work, however, the possible outcomes are not discrete. Modeling the semivariogram Maron, M.J., and R.J. Lopez, 1991, Numerical Analysis-A Practical Approach, 37'd ed.: PWS-Kent Publ. These are the same unknowns we recovered by matrix inversion. Areas enclosed by successive standard deviations of the standard normal distribution. Geoff **Discriminant Functions** the calculations in the text have benefited from his careful checking and verification. Paleontologists must be example is stratigraphic correlation, either of measured sections or petrophysical well logs. If the occurrences of all the states in the section are independent, the same relationship holds for all possible transitions; so, for example, This allows us to predict what the transition probability matrix should look like if the occurrence of a lithologic state at a lithologi one point in the stratigraphic interval were 170 Analysis of Sequences of Data completely independent of the lithology at the immediately underlying point. The t - t e s t of correlation For example, premultiplying a 3 x 2 matrix by a 5 x 3 matrix results in a 5 x 2 matrix: 12 1 2 7 6 129 Statistics and Data Analysis in Geology - Chapter 3 However, the 3 x 2 matrix cannot be postmultiplied by the 5 x 3 matrix because the number of columns (two) in the left matrix. This situation may persist, with grains crystallizing at random, until the entire mass is solid. However, the fixed probability vector cannot be used to estimate the columns of the expected transition probability matrix. In the former, the variable is a measured in feet. In asymmetric frequency curves, the median lies between the mean and the mode. Elementary Matrix Operations . As an example, we can power the array of transition probabilities discussed at the first of this section. Each element in the transitions. A variant gives the probability that a coin tossed onto a chessboard will fall entirely within one square. Probability, expressed in this form pertains to the likelihood of an event. Although all populations exhibit diversity, there is no real population whose elements vary without limit. 157 Statistics and Data Analysis in Geology - Chapter 3 Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, 1992, Numerical Recipes: The Art of ScienrificComputing, Znd ed.: Cambridge Univ. . Some are sufficiently well developed to be used in rigorous tests of statistical hypotheses. Such an approach is certainly beyond the scope of this short chapter, so we will confine ourselves to those topics pertinent to techniques that we will utilize later. A p o p ulation consists of a well-defined set Analysis of Multivariate D a t a (either finite or infinite) of elements. "Whiskers" are lines that extend from the ends of the box, usually to the 5th and 95th percentiles. However, if there are many more original points than interpolated value. 428 Calculating the drift . Histogram and alternative forms of box-and-whisker plots of airborne measurements of 137Csradiation recorded on the length of the immediately preceding interval is shown by plotting xi = ti+l - t i against yi = ti - t i - 1 where ti is the time of occurrence of the i th event. two outcomes, hit porous sandstone and dril2 into an anticline, are not mutually exclusive as we hope that both can occur simultaneously. using observations in file DJPOR.TXT, which gives the field-wide average porosities for 105 oil fields producing from the Cretaceous "D" and "J" sands in the DenverJulesburg Basin of eastern Colorado. Empirical survivor function for the Japanese volcano Aso. As you can see from Table 4.1, these cover a variety of sequences of different types, and are designed to answer different types, and are designed to answer different types. products, A B and B A. Fortunately, in the applications we will consider, we will only need to find the fractional powers of diagonal terms are the original observed frequencies. Confidence intervals based on t ... that data such as these are in the form of sequences is irrelevant, because each data point is considered to be a separate and distinct variable. The sample problems are only suggestions from the many that could be used. An alternative to a histogram is to show the data in the form of a cumulative plot. Inversion and Solution of Simultaneous . However, we can also consider the manner in which different ordinal classes succeed one another. As the operation is performed, we must keep track of the number of exchanges or transpositions necessary to get the second subscript in the correct order. An easy method of performing this operation is to arrange the matrices in the following manner: To obtain the value of an element C i j, multiply each element of row i of A, starting at the left, by each element of column j of B, starting at the top. Their associated eigenvectors point in the directions of the stresses. Nevertheless, we believe that it is continuous and we estimate its form from the measurements made at the wells, recognizing that our reconstruction is inaccurate and lacking in detail because we have no data between wells. These data can be expressed in the same manner, giving the matrix B: B= [: 4"] 240 1 2 1 1 :3: Total production for the 2 years in the three districts is the sum, C, of the the matrices A and B: [i:::] [218 80 2 + 84 102 4 189 165 9 240 1 2 1 1 = 458 201 3 302 28 O] 522 104 1] [125 Statistics and Data Analysis in Geology - Chapter 3 Similarly, the change in production can be found by subtracting: B A D [84 102 4 240 121 1 302 28 O] [105 63 5 218 80 2 = 220 76 I] -21 39 -1 [ii ::I:] Note that A was subtracted from B simply to show increases in production as positive values. The probability that the n t h hole drilled will be dry is P = (1-p)%-lp 5. In addition, the eigenvectors of two-dimensional symmetric matrices are additionally constrained to orientations that are multiples of 45". Figure 4-4 b shows .the result of interpolating iron concentration to 50-cm intervals by rectangular integration. A typical experiment would involve an urn filled with red and white balls; if a ball is selected at random, the probability it will be red is equal to the proportion of red balls originally in the urn. 2 Introduction This book is also designed to aid another type of geologist-in-training-the student who has taken or is taking courses in statistics and programming. Finally, the remaining possibility is the number of combinations that contain no heads: (3 = 3! - 3 . Two trend-surface models It is not surprising when students protest, "Why should I study this dull and boring topic when the material is never used in my field?" In an attempt to contribute to the solution of this educational impasse, I've made a major change in this edition of my book. These techniques are almost impossible to apply without the help of computers, because the calculations are complicated and must be performed repetitively. The drainage basin of the larger stream contains about 18 mi2, while the basin drained by the smaller stream includes only about 10 mi2. Problems can arise if significant digits are truncated during this operation, so it is better to use the definitional equation to calculate variance in a computer program. Of course, subjective judgment is better to use the definitional equation to calculate variance in a computer program. necessary in choosing an algorithm to perform mapping, but methods are available that allow a choice to be made between computations, you may compute a correlation matrix that is not exactly singular. xiv 461 462 471 477 479 483 484 487 Con tents Introduction to Eigenvector Methods. Let us check the rule with an The denominators of the ratios for both unknown coefficients are the same: 1 The numerator of XI 1: i: $1 = (4x \ 30) - (10 \ x \ 10) = 20$ is the determinant I110 38 301 lo = (38 x \ 30) - (110 x \ 10) = 40 so $x_1 = 40/20 = 2$. (If the target is circular, then u = b = r, the radius of the circle.) The search pattern consists of a series of parallel traverses spaced a distance D apart (Fig. Stratigraphic sections, for example, usually include at least three states and these cannot be ranked in a meaningful way. By 1986 when I revised Statistics and Data Analysis in Geology for its second edition, technology had progressed to the point that personal computers were almost commonplace and every young geologist was expected to have at least some familiarity with computing and analysis of data. The most commonly cited example of an interval scale is that of temperature. Analysis of Directional Data Conversely, drilling a dry hole on a seismic feature increases the probability that the remaining untested features will prove productive, because one nonproductive features in three tosses is 3 /8. Similarly, the absolute magnitudes of the elements in eigenvectors are not significant, only the ratios between the elements. We must rely on the subjective opinions of experts when considering such questions as the probability of failure of a new machine for which there is no past history of performance. Because the numerator in our present example would contain a column of zeros, the determinant of the numerator also will be zero. Similarly, we can show that the eigenvector associated with th(second eigenvalue is 1-0.15 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.851 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 1 - 0.15 [I: =] [= 0.85 0.85 [I values for pure halite, anhydrite, and shale are given in Table 3-4. Although there are many methods whereby regularly spaced data might be estimated from these data, we will consider only two in detail. Because nl and nz are the same for all three sequences, 8 and (TU also are the same. p = (3) = - .'! 4!1! = . The appearance of a histogram is strongly affected by our choice of the number of categories and the starting value of the first category, especially if the sample contains only a few observations. Having calculated these, we can create a z-test by Equation (4.10), where U is the observations. Having calculated these, we can create a z-test by Equation (4.10), where U is the observations. the runs statistics. Combined categories can be formed of all elements in the B row, all elements in the D row, and the combination of transitioners of an observational science, and the rigorous application of quantitative methods often reveals us for the imperfect observers that we are. Although elementary probability is discussed and many statistical tests described, the detailed development of statistical theory has been left to others. Although interval scales in theory convey less information than ratio scales in the same manner. This is done by enumerating all possible ways of arranging nl items of state 1 and n2 items of state 2. Certain symbols traditionally have been assigned to measures of distribution curves. Vertical axis is compressed for comparison with Figure 2-13. If you pursue the following topics, you wdl become involved with mathematical methods that have a certain aura of exactitude, that express relationships with apparent precision, and that are implemented on devices that have a popular reputation for infallibility. I now have on my desktop a small crystalline cube, a "supercomputer" capable of outperforming devices that existed a decade ago at only a few sites in the world. Typical Poisson frequency distributions are shown in Figure 4-12. The inverse of the constant, c, is indicated by c-l, which represents llc. In this example, the eigenvalues are hi = 1.28 A2 = 0.72 Gould refers to the relative lengths of the semiaxes as a measure of the "stretchability" of the enclosing ellipse. Row one is subtracted from row two. The three equations can be set into matrix form as [=I: L = 7 cv 2.03 2.98 2.43 1;3] [vv ha Vsh However, what we really want to do is solve for V, given values of L taken from the well logs. We will consider a real matrix of coordinates of points in space and interpret the eigenvalues and associated functions as geometric properties of the arrangement of these points. The measurements were made 191 Chapter 5 Spatial Analysis Although geologists study a three-dimensional world, their view of it is strongly two dimensional. The rows of a matrix can be regarded as the coordinates of points in m-dimensional space. This is because the interval de lies between segments cd and e f, both of which run downward; therefore, the 188 Analysis of Sequences of Data X Figure 4-13. Repeating the same problem: 381 110 10 30 1.0 2.5 1.0 3.0 11.0 1.0 2.5 0.0 0.5 1.5 9.5 0.0 0.5 1.5 9.5 0.0 0.5 5. Can variables be related or their effectiveness measured? This would greatly enhance the Markovian property, but in a specious manner. We are thus certain of getting some combination on the three tosses; the shape of the distribution function describes the likelihood of getting any specific combination. The petroleum industry has been among the largest nongovernment users of computers in the United States, and is also the largest employer of geologists. These assumptions do not seem unreasonable for many exploratory surveys. The sequence generated might resemble this set of twent . Some investigators consider the length of the longest run as an indicator of nonrandomness, and others trials: HTHHTHTTTHTHTHTHTHHH We intuitively expect, of course, that about ten heads will appear, and we can determine the probability of obtaining this (or any other) number of heads. Statistics use the number of turning points, which are points in the sequence where the signs of successive observations change. This is the area of "quasi-statistics" or "proto-statistics" or "proto-statistic the number of observations. The sample mean has two highly desirable properties that make it more useful as an estimator of the average or central value of a population than either the sample median or mode. A histogram of the number of events occurring in successive equal intervals of time is given in Figure 4-7. There are tests specifically designed to detect a trend in the rate of occurrence of events by comparing the midpoint of the sequence to its centroid. [20 -4 -40 8 -2:] -60 1 2 -26 20-h -40 -60 -4 8-h 12 8 -20 -26-h = O Expanding out the determinant and combining terms yields -A3 + 2h2 + 8h = 0 This is a cubic equation having three roots that must be found. We can compute of events by comparing the midpoint of the sequence to its centroid. autocorrelation coefficients for successive lags and test these for significance. 0 @ 24 O 0. Conversely, it is also true that Therefore, the two can be equated, giving which may be rewritten as This is a most useful relationship, because sometimes we know one form of conditional probability but are interested in the other. The final form of the transition frequency matrix with estimated diagonal frequencies is given below. Central Limits Theorem Distribution of Lines Cross-correlation . . We also can test the randomness of runs about the mean, and we will use this as a test of residuals from trends later in this chapter. Geology, to a major extent, remains a science that is principally concerned with observation. Measurements of magnesium concentration in stream mouth t o sample locations; distances are from stream mouth t o sample locations. The probability that n successive wildcats will all be dry is P = (1- p) n 4. The rule states that the solution for any unknown xi in a set of simultaneous equations is equal to the ratio of the two determinants. None of the techniques can be considered exhaustively in this short space, but from the examples and applications presented, one or another may suggest themselves to the geologist with a problem to solve. Table 3-1. Statistics in Geology . The first and most obvious technique consists of simple linear interpolation between data points to 163 Statistics and Data Analysis in Geology - Chapter 4 Table 4-2. When dealing with discrete events, we are counting-a process that usually can be done with absolute precision. That is, AxB X C = (A X B) X C = A X (B X C) Because powering is simply a series of multiplications, a square matrix can be raised to a power. My debt to these dear friends and colleagues remains as large as ever. In a geological context, the articles by Doveton (1971) and Doveton and Skipper (1974), plus the comment by Tiirk (1979), are recommended. , x events will follow the binomia distribution. Basically, the probability of discovery is related to the relative size of the target as compared to the spacing of the search pattern. Consider a hypothetical set of simultaneous equations (the U i j ' S) times the vector of unknowns (the search pattern). xi's) is equal to some constant (A) times the unknown vector itself. GeologyStatistical methods. There is a similar meaning for ayyand aZz. The symbol a , represents the shear stress on the east or west face of the cube, acting parallel to the Y-axis. Time-series methods, such as autocorrelation and spectral analysis, require equally spaced data. . Conversely, a large standard deviation indicates that values are scattered widely about the mean and the tendency for central clustering is weak. Define a row vector V having 20 elements, each equal to 1.0. The matrix multiplication, V M, will yield a five-elementrow vector containing the Nearest-neighbor analysis .. column totals of M. Eckart-Young theorem . We often substitute spatial scales for a time scale in stratigraphic problems, but our conclusions are no better than our fundamental assumptions about the length of time required to deposit the interval we have measured. Inter pola t ion Procedures . The discussion of elementary probability theory in Chapter 2 has been revised in recognition of the unfortunate fact that fundamentals of probability often are passed over inintroductory courses in favor of a cookbook recitation of elementary statistical tests. 4-9). We might, for example, be interested in the occurrence of a rare mineral encountered sporadically on a traverse across a thin section or in the appearance of bentonite beds in a vertical succession of sedimentary rocks. This topic is more complex than it might seem at first glance. A lucid discussion and geometric interpretation of eigenvectors and eigenvalues was prepared by Peter Gould for the benefit of geography students at Pennsylvania State University. In each of these experiments, we have enumerated all possible numbers of heads that we could obtain, from zero up to three, to ten, or to 50. The probabilities that targets will be detected by a search along a set of lines can be determined by geometrical considerations. 179 Statistics and Data Analysis in Geology - Chapter 4 100 80 60 L u0 L QJ n 5 z 40 20 0 1200 1400 1600 Year of event 1800 2 30 Figure 4-6. If we regard the original data as a rectangular curve or step function in which the interval from one observation to the succeeding observation has a constant value, a data set might have the form shown in Figure 4-3 a. A scatter diagram of the serial correlation, or first-order autocorrelation, of successive intervals between events is shown in Figure 4-10. Among its geological applications, it may be used to forecast the probability of success in a program of drilling for oil or gas. In the latter example, the variable again is a measured attribute, barrels (bbl) of oil, and the scale is measured in days, months, or years. It may be instructive to compare these results with those from linear interpolation and to compare both with the overlying original data to see how much detail is lost by the two approximation processes. A second test run on the same specimen may yield a permeability of 93 md, and a third test may register 1 1 2 md. Aware of this synergistic process, Lee encouraged visits and exchanges with the world's leaders in mathematical geology and its related disciplines, creating a heady ferment of intellectual activity that remains unique. Written in conventional form, the equivalent of the three simultaneous equations is (all - h) x1 + d12x2 + d.13x3 = 0 141 Statistics and Data Analysis in Geology - Chapter 3 Let us assume that there are solutions to these equations other than the trivial case where all the unknown x's = 0. Obviously, this distribution must start at two, since this is the minimum number of holes that might be required, and continues without limit (in the event of extremely bad luck!); we show the distribution only up to 1 2 holes. 619 Table A . 214 218 220 221 224 227 228 234 234 239 243 248 254 261 264 266 268 275 278 288 293 293 295 299 300 302 307 310 313 316 322 322 325 326 326 330 xiii Contents Matrix representation of vectors . . In our binomial probability example, the "urn" consists o the geologic basin where exploration is occurring, and the red and white balls correspond to undiscovered reservoirs and barren areas. The test has (m-1)' degrees of freedom, where m is the number of states (a degree of freedom is lost from each row because the probabilities in the rows sum to 1.00). If a trend exists, what is its form? If there are four reservoirs distributed among the ten seismic features, the discovery of one reservoir increases the odds against finding another because there are fewer remaining to be discovered. - Ma~jorAlexander P. The expected transition probability matrix would consist of rows that were all identical to the fixed probability that Y discoveries will be made in a drilling program of n wildcats is n! P= (1- p) n - r p r (n- Y)!Y! This is an expression of the binomial distribution, and gives the probability that Y successes will occur in n trials, when the probability that Y successes will occur in n trials, when the probability of success in a single trial is p. An anonymous critic left the following rhyme on my desk almost 30 years ago. Therefore, we will discuss 294 Spat ia I Ana I ysis the problems of data-point distribution, lack of fit, computational "blowup," and "C," by labels such as "A," "B," and "C," by labels such as N,0, and 0, or by numbers. Discrete distribution giving the probability of fit, computational "blowup," and "C," by labels such as "A," "B," and "C," by labels such as "A, ""B," and "C," by labels such making n discoveries in a five-hole drilling program when the success ratio (probability of a discovery) is 10%. An older text by Uspensky (1937) derives the more general elliptical case used here. Geological examples of this type of data sequence include the historical record of earthquake occurrences in California, the record of volcanic eruptions in the Mediterranean area, and the incidence of landslides in the Tetons. These new estimates are inserted into the original transition frequency matrix and the process repeated. Table A.I. Cumulative probabilities for the standardized normal distribution Critical values of of a preferred trend for Rayleigh's test for Their textures result from simultaneous crystallization of quartz and feldspar at the eutectic point. However, we are more interested in the converse, which is the probability that an area will prove to be mineralized, conditional upon the presence of a magnetic anomaly. Alternatives to the semivariogram Alternatively, growth of a crystal of one state might locally deplete the magma of that constituent, retarding crystallization and resulting in a highly alternating mosaic of quartz and feldspar. We are concerned with distances between points, the density of points, and the values assigned to points. 88 92 93 97 102 102 103 105 105 107 112 119 3. This is one of the most important techniques in matrix algebra, and it is essential for the solution of simultaneous equations are hopelessly mixed together, or confounded, and the experimenter cannot determine what portion of the variability is due to variation between his test objects and what is due to error. Coordinates are given in meters from an arbitrary origin at the southwest corner of the property and were measured relative to magnetic north. The continuous spectrum This is done by dividing rows of the matrix by constants and subtracting (or adding) rows of the matrix from other rows: 1. By sampling, only a small part of the population is destroyed. Two terms have been introduced in preceding paragraphs without definition. The measured values reflect the composition of the rocks and the fluids in the pore space. Convert the vector of temperatures to degrees Celsius using matrix algebra. Most of the methods we will discuss in subsequent chapters are based on matrix manipulations, especially as performed by computers. As a consequence, this third edition contains many more worked examples and also includes an extensive library of problem sets accessible over the Internet. We can try a somewhat more difficult example by using the set of simultaneous equations we solved earlier. Press, Baltimore, Md., 694 pp. The techniques of matrix inversion will be encountered again and again in the next chapters of this book. Press, Cambridge, U.K., 963 pp. If you utilize these as a "black box" without understanding their operation and limitations, you may be led badly astray. 391 Extensions of contour mapping. .. In this instance the conditional probability of an eruption is greater than the unconditional probability, or p (eruption), which is simply the probability that an eruption will occur without any knowledge of other events. It is generally agreed that "negative lengths" are not possible. A fragment of a hitherto unknown species of mosasaur has been found in a stream bed in western Kansas, and a vertebrate paleontologist would like to send a student field party out to search for more complete remains. This requires one transposition, so the product is negative. In addition, there are inexpensive, specialized programs for geostatistics, for analysis of compositional data, and for other "nonstandard" procedures of interest to Earth scientists. This is not true of eigenvectors of matrices in general, but only of symmetric matrices. KoImogo r ov-S m ir nov tests The variance of a sample is denoted by the symbol s2. First we must consider the possibility of a trend in the data. Stevens, S.S., 1946, On the theory of scales of measurement: Science, v. The maximum shear stress is given by (Al - h 3) / 2 and occurs along a plane oriented perpendicular to a line that bisects the angle between the directions of maximum and minimum normal stress (that is, between the first and third eigenvectors). Obviously, we cannot cover this span in a single book without omitting a tremendous amount of material. Three estimates can then be calculated The first test is simply the maximum positive difference, and the third is the larger of the absolute values of the two. We do not know that the structural surface is continuous, because we can observe it only at the locations where drill holes penetrate the surface. Furthermore, the processes that modify the Earth, such as mountain building and continental evolution, are generally beyond the geologists' capabilities for direct manipulation. 6. These include the Poisson distribution, which can be used instead of the binomial when p, the probability of success, is very small. The eigenvectors of real, symmetric matrices are always orthogonal, or at right angles to each other. both sides of the equation by its inverse, C - l. These findings require explanation. As in ordinary algebra, A + B = B + A, and (A + B) + C = A + (B + C), provided all are n x m matrices. You can satisfy yourself on this point, however, by multiplying the two terms, AX, to obtain the left-hand side of the simultaneous equation set: Working through this multiplication, you will see that all of the terms are associated with the proper coefficients. Unfortunately, many of these methods were developed in scientific specialties alien to most geologists, and the description of an application in radar engineering, stock market analysis, speech therapy, or cell biology may be difficult to relate to a geologic problem. 617 of the concentration Table A.9. Maximum likelihood estimatesparameter K for calculated values of R 334 338 Displaying spherical data We can see that the columns are converging on these values at the 10th power of T: 0.248 0.443 0.309 0.230 0.449 0.321 0.228 0.450 0.322 1 Square matrices also can be raised to a fractional power, most commonly to the one-half power. When the second edition appeared in 1986, FORTRAN no longer dominated scientific programming and computer languages. Rather, it is too little data . Data are in file MIDLAND.TXT. (b) Values interpolated at 1000-m intervals. Because series of events are very simple, in the sense that they consist of nominal elsewhere. Computers are no longer made that can read floppy disks and double-sided diskettes are being phased out by optical disks. Introduction . occurrences (presence-absence), simple analytical techniques may prove to be the most effective. You will recall that it was stated that these were the coefficients originally in the equation set, so we have recovered the proper values. The matrix is now inverted. That is, A l l 2 is a matrix, XI whose square is A: Finding fractional powers of matrices can be computationally troublesome. ISBN 0-47 1-1 7275-8 Library of Congress Cataloging in Publication Data: Davis, John C. The probability is p = - 28 7TD (5.5) where 4? is the length of the target. If the observations have been selected in an unbiased, representative manner, the sample histogram can be considered an approximation of the underlying probability distribution. Cluster Analysis . . Runs tests are appropriate when the cause of nonrandomness is the object of investigation. Second, it can be demonstrated that, for symmetrical distributions such as the normal, the sample mean tends to be closer to the population mean than any other unbiased estimate (such as the median) based on the same sample. Symbolically, this is represented as . That is, If Equation (2.9) is substituted for p (A) in Bayes' theorem, as given in Equation (2.10) 23 Statistics and Data Analysis in Geology - Chapter 2 A simple example involving two possible prior events, B1 and B2, will illustrate the use of Bayes' theorem. Runs Tests The simplest type of sequence is a succession of observations arranged in order of occurrence, where the observations are two mutually exclusive categories or states. Determinants Before discussing our final topic, which is eigenvalues and eigenvectors and how they are obtained, we must examine an additional property of a square matrix called the determinant. 156 Matrix Algebra SELECTED READINGS Anton, H., and C. We do this by trial-and-error, searching for those values that, when inserted on the diagonal of the transition frequency matrix, do not change when the matrix is powered. Two-way analysis of variance . 1. Serial correlation of durations between successive eruptions of the Japanese volcano Aso. The two matrices being added or subtracted must be of the same number of rows and columns. Because this is not a reference book, references are not emphasized. entire segment cdef can be considered as a single downward run. The interval ij can be considered either as part of the run down g h i or the run up i j k, as the total number of runs remains the same in either case. These methods are practical only because of the great computational speed of digital computers. Most techniques for comparing two or more sequences can be grouped into two broad categories. If such an array is premultiplication yields the n x n major product matrix: The equation for the general case of matrix multiplication is In a series of multiplications, the sequence in which the multiplications are accomplished is not mandatory if the arrangement is not changed. It was with their support and encouragement that I have been able to write the three editions of this book. For example, a common test procedure is to dichotomize a series by subtracting each observation from the median of all observations, and testing the signs for randomness of runs about the median. If you repeat the test performed earlier, but using the second-order transition probability matrix, you should find that the sequence has no significant second-order properties. Of course, in real problems the source of singularity usually is not so obvious. If they fall sway to their own computers, they may cease to critically examine their data and the interpretative methods. The characteristics of these series are (a) the events are of these series are (b) the events are so infrequent that no two occur in time; (b) the events are so infrequent that no two occur in time; (b) the events are so infrequent that no two occur in time; (c) the events are essentially instantaneous; and (c) the events are so infrequent that no two occur in time; (c) the events are essentially instantaneous; and (c) the events are so infrequent that no two occur in time; (c) the events are essentially instantaneous; and (c) the events are essentially statistics texts, pertains to the outcome of the toss of an unbiased coin. [Orno 0.0 1.0 1 Matrices A and B are combined in an n x (n+1) matrix. 5- [i ";] [1 0 0 [02' -0.5 '1 0.2 row two is subtracted from row one to reduce 0.21 the final off-diagonal element to 0. Modified x2 procedures are available to test the significance of individual transition pairs. If the matrices are not of the same order, there will be leftover elements, and the operation cannot be completed. Before we leave the computation of eigenvalues of 2 x 2 matrices, we should consider one additional completed. certain characteristics, such as their midpoint; measures indicating the amount of "spread"; and measures of symmetry of the distribution. Other methods of defining and arriving at probabilities may be more appropriate in certain circumstances. We have simply enlarged the dimensionality of the subjects of our inquiries by considering a second (and in some cases a third) spatial variable. In regions of heavy vegetation or deep weathering, however, we must make do with scattered outcrops and poor exposures; the quality of the finished map reflects to a great extent the density of control points. If two square matrices are multiplied, the product is a square matrix of the same size. p(BIA) is a conditional probability because it expresses the probability that B will occur conditional upon the circumstance that A has already occurred. So, Note that the first row has become the first column of the transpose, and the second column. Jensen, J.A., and J.H. Rowland, 1975, Methods of Computation: The Linear Approach to Numerical Analysis: Scott, Foresman and Co., Glenview, Ill., 303 pp. Royal Statistical Soc., Series A, V. The row and column vectors are the transpose of each other. Many populations are infinite or so vast that it is only possible to examine a subset. are presumed to be independent; that is, success or failure of one hole will not influence the outcome of the next hole. In historic time series and sequences such as Holocene ice cores, this assumption is valid. The probability that a hole will result in a discovery is p. Commonly, these elements are measurements of a specific nature made on items of a specified type. Such events are extremely important in geology, because we may be able to observe one event directly, but the other event is hidden. If events occur randomly in time, the survivor function will be exponential in form. Figure 5-2. Zonation Although the term "statistics" once referred simply to the collection of numerical facts such as baseball scores, it has come to include the analysis of data, and especially the uncertainty associated with such data. 359 Spatial Analysis by ANOVA However, an interval scale has no natural zero, or point where the magnitude is nonexistent. 32 Elementary Statistics Box-andwhiskey plots were devised by John Tukey (1977) to more effectively show the essential aspects of a sample distribution. The transition frequency matrix is a concise way of expressing the incidence of one state following another: to Row Totals A B C D A r 1 8 0 5 01 23 B05zoi from C 5 2 18 3 28 D L 0 0 3 21 5 Column Totals 23 7 28 5 63 Grand Total Note that the row totals and the column totals will be the same, provided the section begins and ends with the same state; otherwise two rows and columns will differ by one. Unfortunately, many textbooks provide no help in this regard, placing their discussions in strictly mathematical terms that may be difficult for nonmathematicians to interpret. As long as the number of undrilled locations is large, and the number of prospects that have been drilled (and hence "removed from the urn") is small, the assumption of constant probability of discovery seems reasonable. Tests that actually were performed may be regarded as a sample of all potential tests. Software and data are widely distributed over the World Wide Web through organizations such as the International Association for Mathematical Geology (. Confidence belts around a regression His work is based mostly on the 295 Statistics and Data Analysis in Geology - Chapter 5 mathematical development of Kendall and Moran (1963). The four-state transition frequency matrix for the section in the Scottish well is given below. EXERCISES Exercise 3.1 File BHTEMP.TXT contains 15 bottomhole temperatures (BHT's) measured in the Mississippian interval in wells in eastern Kansas. Depending upon the numerical precision of the computer program, rather than exactly 0, you may observe a very small determinant such as -0.0002. The matrix is with eigenvalues hi = 1.85 The first eigenvector is 1 1- 1.85 L 0.85 A2 = 0.15 1 0.85 - 1.85 A 0.85 L 0.85 - 0.85 1 A I: [:I = which defines a line having a slope of 45". In Figure 4-8, the points represent the percentage of intervals between eruptions which are longer than the specified number of 1,4, 5, 8. Regression methods are useful for much more than the analysis of observations arranged in order in time or space; they can be used to analyze any bivariate data set when it is useful to consider one of the variables as a vears. Semivariograms Measurement Systems function of the other. Increasingly, however, geologists and other Earth scientists are using systematic procedures to search, particularly when they must rely on instruments to detect their targets. 500 1 P Oil saturation, % Figure 2-17. In this case, we can use millimeters as a convenient spatial scale, but we have no way of evaluating whether olivine rates a higher number than plagioclase. Reprint of a classic. Because geologists depend heavily on observations, particularly obs ... By now you no doubt suspect that the eigenvectors of 2 x 2 symmetric matrices will always lie at these specific angles, and this is indeed the case. Obviously, we cannot extract information about time intervals from stratigraphic succession data, because the time scale accompanying the succession data. encountered at the next point upward: State A (sandstone) 0% State B (limestone) 71% State C (shale) 2 9% State D (coal) 0% Suppose the next point: State A (sandstone) 18% State B (limestone) 7% State C (shale) 64% State D (coal) 11% So, the probability that the lithologic sequence will be limestone -* shale is p (BC) x p (CB) = 29% x 7% = 2% - - - limestone - However, there is another way to reach the limestone state in two steps. Figure 2-15 shows a histogram and several alternative box-and-whiskerplots produced by several popular commercial programs. Nothing is gained, for example, by comparing X-ray intensity at 20'28 with the intensity at 30'28 on another chart. This variation is called experimental emor. Closure effects on principal components Because variance is the average squared deviation from the mean, its units are the square of the units of the original measurements. Other methods are a subset of more general multidimensional procedures. Aso is 178 Analysis of Sequences of Data Table 4-5. 1 Row one is divided by 4 and row two is divided by 4 and row two is divided by 10. The subjective approach is widely used (although seldom admitted to) in the assessment of the risks associated with petroleum and mineral exploration, where relative-frequencybased estimates of geologic conditions and events are difficult to obtain (Harbaugh, Davis, and Wendebourg, 1995). When geologists can control the means of obtaining samples, they are quick to exploit the opportunity. The F-Distribution The equation for the Poisson distribution in this instance is $p(X) = e^{-x}AX/X!$ (4.6) Note that the rate of occurrence, A, is the only parameter of the distribution. Cr 125 205 171 62 137 2 34 2 70 179 113 65 80 35 176 90 52 98 130 158 69 108 cu 25 33 25 157 88 185 52 322 29 400 225 230 30 164 200 29 59 28 30 30 Mg 6936 5368 5006 3600 3220 7450 4400 5000 8600 4000 2000 1000 3100 5000 9000 3 100 7100 6400 7900 2300 V 114 143 90 59 130 162 205 150 98 60 90 100 160 105 60 89 112 143 109 136 Zn 194 212 2 72 129 123 2 64 155 135 114 40 130 50 100 105 170 87 147 133 103 84 We will use a series of 2 x 2 matrices calculated from data that might arise in an environmental study. Title QE48.8 .D38 2002 550'.72 Printed in the United States of America 10 9 8 7 G 5 4 3 2 1 85-12331 Preface My original motivation for writing this book, back in 1973, was very simple. Fortunately, the mathematical manipulations of probabilities are identical regardless of the source of the probabilities. What does this column of 1's represent?] Exercise 3.6 The state of stress in the subsurface can be represented in a 3 x 3 matrix, Z, whose diagonal elements represent normal stresses and whose off-diagonal elements represent shear stresses. 132 Matrix Algebra It may not be apparent why the set of simultaneous equations can be set into the matrix form shown. The difference in absolute hardness between diamond (rank ten) and corundum (rank nine) is greater than the entire range of hardness from one to nine. A nonsingular matrix has as many nonzero eigenvalues as there are rows or columns in the matrix, so its rank is equal to its order. The largest eigenvalue, h 1, represents the minimum normal stress and the smallest, h3, represents the minimum normal stress. ... The transition frequency matrix of this succession would look like the one we observe except that the diagonal elements would contain values other than zero. Such an experiment is called sampling without the origin of the coordinate system and which just encloses the tips of the vectors. P-Values ... replacement, and is governed by the discrete hypergeometric distribution. The symbol x i j is the element in the i t h row and the j t h column of matrix X. Laplace's problem specifies the probability that a needle of length 8, when dropped on a board covered with a set of rectangles, will lie entirely within a single rectangle. Test for a specified trend A highly readable text on matrix algebra with a minimum of mathematical argon and a maximum of examples and applications. Unfortunately, obtaining a truly random sample may be impractical, as in the situation of sampling a geologic unit that is partially buried. Generally, the symbols for population distributions are Greek letters, and those for sample distributions are Roman. Some work also has been done on mineral successions along traverses across thin sections. The distribution is applicable to such problems as the rate that telephone calls come to a switchboard or the length of time between failures in a computer system. McCammon (1977) presented such

graphs in a particularly useful dimensionless form for various combinations of target shape and size relative to the spacing between the search lines. The probability that the conditional probability that given a fossil is derived from basin Bi, it will be a marine fossil, as proportional to the percentage of the Cretaceous outcrop area in the basin BZ p(AIB1) = 0.35 and for basin BZ p(AIB1) = 0.35 and for basin BZ p(AIB2) = 0.80 Using these probabilities and Bayes' theorem, we can assess the conditional probability that the fossil fragment came from basin BZ 1, given that the fossil is marine. Suppose one test indicates a permeability of 108 md (millidarcies). Is this the "true" permeability of the sample, we may determine that mining districts often are characterized by the presence of abnormal geomagnetic fields. The geometric meun is the n t h root of the products of the n observations, or equivalently, the exponential of the arithmetic mean of the logarithms of the observations. The process is illustrated in Figure 4-11. We will attempt to recover them by a process is illustrated in Figure 4-11. We will attempt to recover them by a process of matrix inversion and multiplication: $4x_1 + 10x_2 = 38\ 10x_1 + 30x_2 = 110\ \text{This}$ is a set of equations of the general type AX=B where A is a matrix of coefficients, X is a column vector of unknowns, and B is a column vector of right-hand sides of the equations. The rank of a square matrix is the number of nonzero eigenvalues that can be extracted from the matrix. The second subscripts in a11a22 are in correct numerical order and no rearranging is necessary. All types of mathematical and statistical operations may be performed with them. The first subscript specifies the column. Rectangular search patterns with different ratios of D1 /D2 fall in the shaded region. We are now considering a one-dimensional analogy of this same problem. However, it is mindless and will accept unreasonable data and return nonsense answers without a qualm. 159 Geologic Measurements in Sequences Markov Chains Finally, a scalar matrix whose diagonal elements are equal to 1 is called an identity matrix or unit matrix. If p(tremor) is the probability that a harmonic tremor occurs and p(eruption) is the probability of a subsequent volcanic eruption) # p (tremor) x p (eruption) is the probability of a subsequent volcanic eruption). application for each method, and that the programs you use are sufficiently straightforward so that their operation is clear. She carefully polished my words, refined my grammar, and detected obscure passages, insisting that I rewrite them until they were understandable. Studies of series of events may have several objectives. The basic mathematical procedure involved has been described under the solution of simultaneous equations (Chapter 3), and the statistical basis of regression has first been discussed in Chapter 2., xn are a random sample from a normal distribution, s2 is an efficient estimate of u2. Of course, we would also hope that such studies might lead to predictive tools to forecast future eruptions. The product of the eigenvalues will be equal to the determinant of the original matrix. Note that both the cumulative plot have a characteristic ogive form. No other combinations of heads and tails can occur. Table 3-3 lists trace-element concentrations for five elements measured on 20 soil samples collected in vineyards and adjacent terraces on the Istrian peninsula of Croatia (the data are contained in the file 1STRIA.TXT). This is equivalent to saying that sample means are less variable 34 Elementary Statistics Table 2-1. Perform a runs test on this data and evaluate the three possibilities. However, although it is widely applied, it is frequently misused. A similar geometric relationship, known as Laplace's problem, also pertains to the probabilities in systematic searches. Most systematic searches are conducted along one or more sets of parallel lines. Therefore, the proper alternative is Hi: U f U Using a 5% (a = 0.05) level of significance, our critical regions are bounded by -1.96 and +1.96. The use of Greek and Roman symbols serves to emphasize the difference between parameters and the equivalent statistics. If each element in the i t h row, the resulting fractions express the relative number of times state i is succeeded by the total of the i t h row is divided by the total of the row is divided by the total of the row is di t h row is d speaking, a matrix cannot be divided by a constant, but we can perform an equivalent operation. The joint probability of such events is said to be conditional. 2. Many stratigraphic measurements, for example, are recorded bed-by-bed rather than foot-by-foot. In Figure 3-1, the tip of each vector is indicated by an open circle, labeled with its coordmates. However, in Figure 2-10 there is an infinite number of subdivisions along the horizontal axis so the probability of obtaining one exact, specific event is essentially zero. It is not unexpected that a tremendous interest in geomathematical techniques has developed in petroleum companies, nor that this interest has spread back mto the Statistics and Data Analysis in Geology - Chapter 1 academic world, resulting in an increasing emphasis on computer languages and mathematical skills in the training of geologists. What arouses our suspicions is not the proportion of heads but the order in which they appear. Ratio zero point. After McCammon (1977). Physical properties measured on pure samples of halite, and vertailed test, and reject if there are too many or too few runs in the sequence. Includes bibliographies and index. (b) What is the nature of my observations? Eigenvect ors We can examine the correlation matrices we calculated for the Istrian vineyard data to gain some insight into the geometrical nature of eigenvectors. We may utilize the past behavior. Numerical analysis requires that collection of data be carefully controlled, with consideration given to extraneous influences. Figure 2-4 is a negative binomial probability distribution for a drilling program where the probability of a discoveries have been made. Simple pegmatites originate by crystallization of the last, volatile-laden substances squeezed off from solidifying granitic magma. The variation we see is due in part to differences between the test plugs, but also results from differences in experimental conditions. If the original data are sparse and several values must be estimated between each pair of observations, linear interpolation will perform adequately, provided the idea of uniformity of slope between points is reasonable. Perhaps the greatest worth of quantitative methodologies lies not in their capability to demonstrate what is false. The remaining categories can be combined until their expected frequencies exceed five. Determine if ICI + ID(= IC + DI. In geology, equal spacing procedures have been most widely used to pretreat stratigraphic data (measured sections, drilling-timelogs, and similar records) prior to filtering or time-trend analysis. 150 Matrix Algebra, the array is considered to be a single entity rather than a collection of individual values and is operated upon as a unit. The second category includes problems in which spatial coordinates or geographic locations of samples are important, te., studies of shape and orientation, contour mapping, trend-surface analysis, geostatistics including kriging, and similar endeavors. Geologic Maps. That is, the matrix will be singular. Here, the well-developed formal tests of hypotheses do not exist, and the best we can hope from our procedures is guidance in what ultimately must be a human judgment. Examples might include an electrical resistivity log from a borehole, or the production history of a commercial well. This does not imply that every other toss will be a head, but rather that, in the long run, heads will appear one-half of the time. Because the most complex analytical procedure is built up of a series of relatively simple mathematical manipulations, our emphasis is on operations. 5-1 a). All possible combinations that satisfy these conditions in a 3 x 3 matrix are all a22a33 all 6 2 3 ~ 2 a12a23a31 a12a21a33 a13a21a32 a13a~~a31 To determine the signs of each of these terms, we must see how many transpositions are necessary to get the second subscripts in the order 1, 2, 3. However, two forms of nonsquare matrices are especially important; these are the vectors, 1 x m (row vector) and m x 1 (column vector). We might code the sequences just given as 1 - 2 - 1 - 2 - 3 - 4 - 2 - 1, where limestone is 3, and coal is 4, but such a convention is purely arbitrary and expresses no meaningful relations between the states. This is called the additive rule of probability. Hence, the binomial distribution often is used to predict the outcomes of drilling programs in frontier areas and offshore concessions. To perform the operation C = A + B, every element of A is added to its corresponding element in B. A scatter diagram with large dispersion and relatively high concentrations of points near the axes is typical of random series of events. C a n cycles or repetitions be detected and measured? Like most real-world situations, there may be more than one solution to a problem. Higher order Markov properties can exist as well. Second, the scales along which the data points are located also are expressed Statistics and Data Analysis in Geology - Chapter 4 in units having magnitude. Rather, difficulties arise in developing a "feel" for the meaning of these quantities, especially in an intuitive sense. The logic behind this progression is simple. estimate intermediate points. The Poisson distribution will be discussed in Chapter 5, where it will serve as a model 19 Statistics and Data Analysis in Geology - Chapter 2 for objects located randomly in space. Moving most tables to the WWW sites has made additional room in the text. The binomial distribution becomes difficult to compute, but the Poisson can be readily used because it does not require either n or p directly. Certainly a paleontologist who has made careful measurements on a large collection of randomly selected fossil specimens has a far greater and more accurate understanding of the natural variation. Many useful operations that can be performed on square matrices are not possible with nonsquare matrices. However, if nl and n2 each exceed ten, the distribution of U can be closely approximated by a normal distribution, and we can use tables of the standard normal variate z for our statistical tests. Information about these tests may be found in texts on nonparametric statistics, such as Conover (1999, p. A sequence having the Markov property is intermediate between deterministic sequences and completely random sequences. A 4 x 4 matrix can be constructed, showing the number of times a given rock type is succeeded, or overlain, by another. The diagonal elements in the matrix above are 1, 5, and 9. Segmenting Sequences They are represented by elements in the third and fourth rows and columns of R, and are shown plotted as vectors in Figure 3-3. The equation for the perimeter of an ellipse is P = 2.7 ~ d mwhere, u and b are the major and minor semiaxes. Multiple Regression The validity of your result is controlled by the density of the original values and no amount of interpolation will allow refinement of the analysis beyond the limitations of the data. I also have revised the section on contour mapping to reflect modern practices. 1) - Thus, with three flips of a coin, there is one way we can get no heads, three ways we can get two heads, and one way we can get all heads. Discrete distribution for the probability of n discoveries in drilling four out of ten prospects when four prospects may be a positive value, a negative value, or zero. ku.edu/Mathgeo/Books/Stat/index.html V In additional data sets and exercises at this site as they are made available from time to time. There can be no connotation that 2 is "twice as a data sets and exercises at this site as they are made available from time to time. There can be no connotation that 2 is "twice as they are made available from time to time." much" as 1, or that 5 is "greater than" 4. The probability of gambler's ruin is approximately 7%. The methods described here do not exhaust the possibilities by any means. Perhaps we are interested in knowing the probabilities of obtaining three heads in a row; this will be 1/2 . 100 -0 70 - % 30- Y? Your answers can be arranged in the form of the matrix below. My final expression of gratitude is the deepest and is owed to my editor, layout designer, proofreader, typesetter, reviewer, critic, companion, and source of vii inspiration-Jo Anne DeGraffenreid, without whose tireless efforts this edition would never have been completed. The 21 Statistics and Data Analysis in Geology - Chapter 2 Number of successes Figure 2-7. Statistical tests are available for trend surfaces if they are to be used as multiple regressions; we will consider these tests and the assumptions prerequisite to their application. The pattern shown by the histogram may be useful in examining the second strength of the chosen intervals, so more than one histogram may be useful in examining the second strength of the chosen intervals. a series. The probability of a success does not change from trial to trial. Aside from this, one of the prime benefits of computerized mapping techniques and the problems they reveal about map reliability. Sometimes observations can be ranked in a hierarchy of states. Spectral Analysis Certain divisions of a distribution such as the 5th and 95th percentiles, the 25th and 75th percentiles, the 25th and 75th percentiles, the 2nd quartile, or the median) are considered especially diagnostic and are indicated on the graphic plots we will consider next. A common objective in an investigation is to estimate some parameter of a population. (b) Square-grid search pattern. First, consider the 2×2 matrix with eigenvalue into the original matrix gives 1 - 1.28 - 0.28 + 0.28 - 0.28 - 0.28 - 0.28 - 0.28 + 0.28 - 0.28 + 0.28eigenvector [4 [-:] = In Figure 3-1, we can interpret this eigenvector as the slope of the major semiaxis of the enclosing ellipse. Significance 2-14). Instead, a formula suitable for computation with a calculator is used which is algebraically equivalent but easier to perform. Summing according to the signs just found yields a single number, which is + - 11 - -a211-33 + a12a23a31 - a12a21a33 + a13a21a32 - a13ma31 We can now try a matrix of real values: 432241103 The six terms possible are $(4 \times 4 \times 3) = 48$ (4 - 1 - 0) 0 = (3xlxl) = 3 (3 - 2 - 3) = 18(2XZXO) = 0(2X4X1) = 8 The first, third, and fifth of these require an even number of transpositions for proper arrangement of the second subscript and so are positive. The probability distribution that governs such an experiment is called the negative binomial, and its development is called the negative binomial distribution. This plot is especially good for showing changes in the average rate of occurrence. These applications will be considered in greater detail later in this chapter. Now, however, we are interested in the matrix of probability associated with the flip of a coin: To Sandstone Shale Limestone From Just as we can find the probability of producing a string of heads in a coinflipping experiment by powering the probability associated with a single flip, we 127 Statistics and Data Analysis in Geology - Chapter 3 can determine the probability matrix. However, because the proportions of halite, anydrite, and shale must sum to one, we can use this constraint to provide the necessary third equation. Mathematical geology advances, as does all of science, by the cumulative efforts of individuals throughout the world who share a common interest and who have learned that methodologies created in one part of the globe will find important applications elsewhere. * (2.2) The value of 3! is 3. The numbers of events within each segment are taken to be observations located at the midpoints of the segments. An example occurs in trend-surface analysis where an independent variable may be measured on a ratio scale while the geographic coordinates are on an interval scale, because the coordinate grid has an arbitrary origin. (As we have already seen, histograms are useful for showing discrete distributions but now we are interested in their application to continuous variables.) Usually the limits of categories are chosen so all of the histogram intervals will be the same width, so the heights of the bars also are proportional to the numbers of observations within the categories represented by the bars. Geologists exercise their artistic talents as well as their geologic skills when they create contour maps. Other conditional probabilities (the probabilities (the probabilities may be lower than the unconditional probabilities of finding a fossil, given that the terrain is igneous, is much lower than the unconditional probability of finding a fossil). Statisticians discuss simple experiments called sampling with replacement in which this assumption holds strictly true. Thin sections were made of the samples and the number of radiolarian tests in a 10 x 10-mm area of the samples and the number of radiolarian tests in a 10 x 10-mm area of the samples and the number of radiolarian tests in a 10 x 10-mm area of the samples and the number of radiolarian tests in a 10 x 10-mm area of the samples and the number of radiolarian tests in a 10 x 10-mm area of the samples and the number of radiolarian tests in a 10 x 10-mm area of the samples area of the samp around this value. Note in the latter example that the population includes not only the limited number of tests that could be run. The magnetic declination in this area is 18'30' east of true north. If the probability of obtaining heads is 1/2, the probability of obtaining two heads in a row is 1 / 2. Rather than giving derivations and proofs, the material will be presented by examples. Actually, runs tests may be applied to data dichotomizing method. We can develop an equivalent set of probabilities for lithologies encountered in a stratigraphic section. Statistics may be used to estimate parameters of parent populations and to test hypotheses about populations. Some of these procedures, such as singular value decomposition (SVD), will find approximate inverses even when exact solutions do not exist. This means that each item in the population has an equal opportunity to be included in the sample. A regression can then be run with these numbers as the dependent variable, yi, and the locations of the midpoints of the segments as values of Xi. The slope coefficient of the regression can be tested by the ANOVA given later in Table 4-9 (p. If a geologist's responsibility to determine the reason for the discrepancy. If we subtract each of these probabilities from 1.0 we obtain the desired probability distribution (Fig. The uncorrected sum of squares is simply 2 x;; the corrected sum of of undeniable importance to specialized disciplines, they are not the subject of this book. Other discrete probability distributions can be developed for those experimental situations where the basic assumptions are different. 366 Computer Contouring .. [Hint: L,as given in file KANSALT.TXT, is a 2 x 50 matrix of P b and At log responses. It may be noted that the procedure just described is almost exactly the same as the classical algebraic method of solving two simultaneous equations. Assume the target being sought is an ellipse whose dimensions are given by the major semiaxis u and minor semiaxis b. By the rules of matrix multiplying the same as the classical algebraic method of solving two simultaneous equations. bottom row, We will solve the simultaneous equation set by first inverting the term A. File KANSALT.TXT contains data for depths between 980 and 1 0 3 0 ft below the surface in A.E.C. Test Hole No. 2, drilled in 1 9 7 0 in Rice County, Kansas. A matrix with a determinant near zero is said to be ill-conditioned.) Finally, there is another special case of interest. Most importantly, she encouraged me throughout the process of seemingly never-ending revision, and took me home and poured for me a generous libation when I despaired of ever laying this albatross to rest. Number of radiolarian tests per square centimeter in thin sections of siliceous Mowry Shale. Series of Events An interesting type of time series we will now consider is called a series of events. Although such results appear to be different, the ratios between pairs of elements in the same "direction." Also, you may note that the pattern of signs on the elements of the eigenvectors seems to be different for two otherwise identical sets of eigenvectors. Doveton, J.H., 1986, Log Analysis of Subsurface Geology: Concepts and Computer Methods: John Wiley & Sons, Inc., New York, 273 pp. The cover was printed by Phoenix Color. 4.8-4.10). Unfortunately, there is no adequate general theory about the nature of geology: Concepts and Computer Methods: John Wiley & Sons, Inc., New York, 273 pp. The cover was printed by Phoenix Color. 4.8-4.10). of some original contributions to the subject, such as the theory of regionalized variables. Embedded Markov chains Initial estimate of transition frequency matrix, with 1000 inserted in each diagonal position. Copyright tables and figures in this text are reproduced with permission of the copyright owners. The breaking of a concretion constitutes a trial, and each trial has two mutually exclusive outcomes: The concretion either contains of otherwise complicated operations become very simple when the matrices involved are diagonal matrices. Probability of . 148 Matrix Algebra 3 Figure 3-2. What we have found is the number of possible combinations of three things (either heads or tails), taken one item at a time. Substituting, 2TIpqz' = ~ T D - 2 J q F D (5.2) We can define a quantity Q as the intersecting a target with a systematic pattern of search. Other transformations ... numerator of Equation (5.2); that is, $Q = 24(u^* + b_2)/2$. With this simplification, the probability of intersecting an elliptical target with one line in a set of parallel search lines can be written as p = -Q(5.3) D In the specific case of a circular target, u and b are both equal to the radius; p = -2r(5.4) D At the other extreme, one axis of the ellipse may be so short that the target becomes a randomly oriented line. In most series-of-events studies, we hope that we can describe the basic features of the series in a way that will suggest a physical mechanism for the lengths of the intervals between occurrences. All of the probabilities were obtained either from binomial tables or calculated using the binomial equation. The trials are performed a fixed number of times. When mapping the surface geology of a desert region, we can stand at one locality where strike and dip have been measured and extend formation boundaries on our map with great assurance because we can see the contacts across the countryside. Harmonic analysis . The test has v = (m-1)2- m degrees of freedom, where m is the number of states, or in this example, v = 11. We can imagine an ultimate experiment involving flips of an infinite number of coins, yielding a histogram having an infinite number of bars of infinitesimal width. The purpose of each operation is to convert the diagonal elements of A to ones and the off-diagonal elements to zeros. A computationally oriented text on matrix algebra. Turcotte, D.L., and G. We begin by imagining that our sequence is actually a censored sample taken from an ordinary succession in which transitions from a state to itself can occur. The column vector, B, then occupies the (n+1) column of the matrix, and the remaining (nx n) part is inverted. Similarly, limestone From Shale 1 times, but follows shale 41 times, but follows dividing each element in a row by the total of the row. Even though the individual measurements are not precisely identical, they are sufficiently close that we can regard them as belonging to the same class. Therefore, you avoid unnecessary work if these points are carefully thought out before your investigation begins. The first topic in the book is elementary statistics. (It should be noted, however, that extracting the eigenvectors from an asymmetric matrix such as the transition probability matrix may not be an easy task!) These topics will not be pursued further in this book; the interested reader should refer to the texts by Kemeny (1983) and Norris (1997), as well as the book on quantitative order of observations. Note that the probabilities relate only to intersecting a target with a line, and do not consider the problem of recognizing a target when it is hit. Alternatively, we can consider the relationship between the boron measurements and the sequence of states. following discussions, you should examine the nature of the methods and think about their applicability and the type of problems they may help solve. Along the way, readers will be exposed to those less glamorous topics that constitute the foundations upon which geomathematical procedures are built. The t-Distribution We can estimate the probable state to be encountered at any step in the future simply by powering the transition probability matrix the appropriate number of times. Consider a rock collector cracking open concretions in a search for fossils. 0.59 = 0.59 0 The probability that no discoveries will result from the rows or columns that are dependent on other rows or columns, and consequently will have one or more zero eigenvalues; its rank will be less than its order. In exploration, the complementary probabilities are of interest, i.e., that a randomly located target will be intersected one or more times by a set of lines, such as seismic traverses, arranged in a Table A.2. Critical values of t for v degrees of freedom and selected levels of significance We can employ techniques that sacrifice all information about the position of observations within the succession, but that provide in return information on the tendency of one state to follow another. Data are derived from scattered boreholes that pierce successive stratigraphic horizons. We will consider methods for the comparisons of such sequences in greater detail in Chapter 6, when we discuss multivariate measures of similarity and problems of classification and discrimination. '58 v, 4- s 8 20- \ 0 0 101 %0 OO 0 7: 54- 0 0 3- n 0 2- 1 0 10 20 30 40 50 D Length of interval years Figure 4-9. Individual numbers within a matrix are called the elements of the matrix and are identified by subscripts. 185 Statistics and Data Analysis in Geology - Chapter 4 One aspect that we have not considered, however, is the order in which the heads appear. Davis, P.J., 1984, The Mathematics of Matrices: R.E. Krieger Publ. The null hypothesis states that there is no difference between the observed number of runs and the mean number of runs from random sequences of the same size. The mode is the value that occurs with the greatest frequency. We will illustrate the relation of this graphic to a conventional histogram 30 Elementary Statistics Figure 2-12. You will recognize the relation of this graphic to a conventional histogram 30 Elementary Statistics Figure 2-12. that the model is exponential; if fi has any value other than zero, the rate of occurrence of Yt will change with t. We could use thickness, but this may change dramatically from location to location even though the sequence is not altered. (c) What is the nature of the sequence is not altered. observations of our world are confined to the outer part of the Earth's crust, yet they must attempt to understand the nature of the Earth's crust, yet they must attempt to understand the ispacing between successive crystals and their iron content fluctuate through a wide range. The order of subtraction is, of course, mandatory. The orientation of this plane can be determined from the elements of the first eigenvector. The probability model assumes that (a) the events occur independently, (b) the probability that an event occurs does not change with time, (c) theorem and the elements of the first eigenvector. The probability that we would get three heads in a row was 1/2 x 1/2, or 1/Z3. The Poisson probability model assumes that (a) the events occur independently, (b) the probability that an event occurs does not change with time, (c) the events occur independently. probability that an event will occur in an interval is proportional to the length of the interval, and (d) the probability of more than one event occurring at the same time is vanishingly small. Computer contouring is totally consistent and provides a counterbalance to overly interpretative traditional mapping. [We here introduce the unconventional but equivalent notation, p (ij), which can be read as the probability that state i will be followed by state j. Suppose we have measured an outcrop and identified the units as sandstone, shale, or limestone. First, the sample mean is an unbiased estimate of the population mean. In the conventional notation used by geologists, the strike of the first eigenvector is tan-l (Y ~ z / Y I ~) and its dip is (Here, V i j refers to the j t h element of the i t h eigenvector.) The strike and dip of the second and third eigenvectors can be found in the same manner. Bayes' basic equation is: p(A,B) = p(BIA)p(A) (2.7) which states that p(A, B), the joint probability that B will occur given that A has already occurred, times the probability that B will occur. Available in several versions for different computer languages. Trendsurface analysis is a popular numerical technique in geology. Figure 2-9. On the basis of just this information alone, we might postulate that the probability that the fragment came from one of the drainage basis is proportional to the area of the basin, or 10 p (B 2) = - = 0.36 28 However, an examination of a geologic report and map histories may shed some light on the nature of eruptive mechanisms and can even lead to physical models of the structure of volcanoes (Wickman, 1966). (b) Using a grid search with spacing D1 in one direction. Mohs' hardness scale is a classic example of a ranked or ordinal scale. By this equation, how many possible combinations are there that contain exactly two heads? 122-142) and Siege1 and Castellan (1988, section 4.5). Principal Component Analysis Using either the equation or published tables such as those in Hald (1952), many interesting questions can be investigated. This results in a great simplification of the statement of complicated procedures and relationships. In contrast, we might consider mine development and production. That is, The enclosing ellipse would collapse to a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz = 0. For example, if A is 3 x 3, As we defined it earlier, the identity matrix is a special diagonal matrix in which there are a straight line of semiaxis of hz diagonal terms are all equal to 1. From Gearhart-Owen (1975). In some of the calculations we will consider later, a row vector, A, becomes a column vector, AT, when transposed, and vice versa. 297 Statistics and Data Analysis in Geology- Chapter 5 search in the pattern of a square grid, the equation simplifies to Lambie (unpublished report, 1981) has pointed out that these equations for geometric probability are approximations of integral equations. The tendency for one state to succeed another can be emphasized in the matrix by converting the frequencies to decimal fractions or percentages. We can designate a matrix (perhaps containing values of several variables) symbolically by capital letters such as [XI, XI (X), or IIXII. At the opposite extreme, we can consider a stratigraphic sequence consisting of the lithologic states encountered in a sedimentary succession. The probability that (n- Y) dry holes will be drilled, followed by Y discoveries, is P = (1-,) n - v p r 15 Statistics and Data Analysis in Geology - Chapter 2 7. In any circumstance, there are a variety (sometimes an infinity) of possible outcomes. Consider subsurface structure mapping for petroleum exploration. Magnesium (ppm) Distance (m) 11,098 2.86 0.0 6.44 1.22 1820 8.61 11,922 1.09 12,530 2542 5.24 2.36 14,065 2889 5.73 14,937 2.24 3460 3.81 16,244 2.05 4586 4.05 2.23 17,632 6020 2.95 19,002 0.42 684 1 2.57 0.87 7232 3.37 20,860 22,471 1.26 10,903 3.84 Figure 4-1. Markov Chains In many geologic investigations, data sequences may be created that consist of ordered successions of mutually exclusive states. The negative binomial has the form Note the similarity between this equation (2.3); the term r + x - 1 appears because the last hole drilled in a sequence must be the r t h success. What has been retained are the fundamental algorithms involved in each analysis, discussions of the relations between quantitative techniques and example applications to geologic problems, and references to sources for additional details. When confronted with a problem involving data along a sequence, you may ask yourself the following questions to aid in planning your research (a) What question(s) do I want to answer? 5. First, the series must be converted to a cumulative form ti yi = T where ti is the time from the start of the series to the it h event, and T is the total length of the series. Each is inconsistencies in the performance of the test by the operator. If the ball is then returned to the urn, the proportions of the two colors remain unchanged, and the probability of drawing a red ball on a second trial remains unchanged as well. 341 Fractal Analysis ... Rewriting, this becomes IAIX=O IA-hII= all a21 a3 1 (3.7) a12 a22-h a32 a13 a23 = O (3.8) a33- Thus we have Because we know the various values of the elements aij, we can collect all of these terms together in the form of an equation such as where the (x's represent the sum of the numerical values of the appropriate aij's. Geologic Measurements in Sequences Before proceeding to some geological examples and appropriate methods of examination, we must consider the nature of different types of sequences apt to be encountered by geologists. Such events are called continuous random variables. Appendices A and B are a concise summary of matrix algebra. 3. This is a fundamental characteristic of a continuous random variable. This procedure has the advantage of considering all data within an interval in estimating a point. A random sample will be unbiased, and as the sample size is increased, will provide an increasingly refined picture of the nature of the population. Figure 2-12 is a histogram of these data in which the vertical axis is compressed for easier comparison with Figure 2-13, where each successive histogram bar begins at the top of the preceding bar. At est of the equality of two sample means Those who wish to pursue a topic in depth will find ample references to the literature in the books I have included; those that simply want an elaboration on some point will probability of rain as 3/4 (1/4 of this book, we will adopt the commonly used boldface notation for matrices. If we state that the chance of rain tomorrow is 30%, then we imply that the chance of rain tomorrow is 30%, then we imply that the chance of rain tomorrow is 30%, then we imply that the chance of this type. Another class of problems may be typified by the succession of mineral grains encountered on traverses across a thin section. A series may be subdivided into segment contains several observations. In the specific instance of a 296 Spat ia I Ana I yis Figure 5-1. For example, if we rate the probability of rain tomorrow as 1 / 2 (and therefore of no rain as 1 / 2), we express our view with a maximum degree of uncertainty; the likelihood of rain is equal to that of no rain. To make room for these and other discussions, some subjects that proved to be of limited utility in geologic research have been deleted. In the handworked example, we wanted the values of the matrix X in the equation To find this, we inverted A and then postmultiplied A - 1 by B to give X.We could have instead found X directly by operating on B as A was transformed into an identity matrix. These are compared by x2=c(0 - E)' c I; where 0 is the observed number of transitions from one state to another, and E is the number of transitions expected if the successive states are independent. If the solidifying pegmatite is undisturbed, we might suppose that quartz and feldspar begin to appear at random locations within the cooling body. It is obvious that this sequence poses different problems to the analyst than do the first examples. The linear dependence comes about because the weight-percentages in the five size categories sum to 100 for each observation, so there are induced negative correlations between the size categories. The total number of runs in a sequence is denoted U ;tables are available which give critical values of U for specified nl, n2, and level of significance, (x. For example, we may be drilling an exploratory hole for oil or gas in anticipation of hitting a porous reservoir sandstone in what we have interpreted as an anticlinal structure from seismic data. The 50th percentile (second quartile or median) usually is indicated by a line across the box. 116, p. The general solution to a quadratic equation is -b+X = (3.10) 2a 142 Matrix Algebra If this seems unfamiliar, review the sections in an elementary algebra book that deal with factoring and quadratic equations. At the same time, B is converted into a quantity that is the value of the unknown matrix X. Figure 4-7. The semiaxes are shown by arrows on Figure 3-1. Statistics and Data Analysis Third Edition John C. Such an outcome is called "gambler's ruin" for obvious reasons, and the binomial expression has the terms n=5 Y=O p = 0.10 p = (0.0.ioo. Many publications on the design of statistical experiments and sampling plans have appeared. The most nearly correct analytical value is taken to be the mean of the determinations. Two events may be related in some way, so that the outcome of one is dependent in part on the outcome of the other. Our observations of the Earth are too precious to discard lightly. The fixed probability vector is found, by summing each row and dividing by the grand total, and then is used as an estimate of the transition probabilities along the diagonal. 159 163 168 173 178 185 191 200 204 207 The Matrix xii Contents Reduced major axis and related regressions Therefore, we must conclude that successive lithologies encountered in the Scottish well are not independent, but rather exhibit a strong first-order Markovian property. A classic example we will consider at length is the occurrence of heads or tails in a coin-flipping experiment. xvi 620 621 Mathematical methods have been employed by a few geologists since the earliest days of the profession. n t h row, with no more than one entry from each column. If we exclude all loose and crumbly rocks from our sample because their porosity is difficult to measure, we will alter the results of the study. Nearly circular ellipse defined by rows in matrix of correlations between magnesium and vanadium, which have a correlation approaching zero. A geologic report couched in mathematical jargon and filled with computer output usually will bluff all but a few critics, and those who understand and comment often do so in equally obtuse terms. However, each successive topic is built upon its predecessors, so aspects of multiple regression, covered in Chapter 6, have been discussed in trend analysis (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5), which has in turn been preceded by curvilinear regression (Chapter 5) estimation. However, with algebraic rearrangement, the unknowns also can be found by the equations x1 = bla22 and 138 - alzb2 - a12a21 Matrix Algebra You will note that the denominators are the same for both unknowns. Table 3-4. It should be 190 Next Page Analysis of Sequences of Data emphasized that randomness itself cannot be proven, as the condition of random occurrence is implied in the null hypothesis. Unfortunately, such a histogram will contain little detail and may not be particularly informative. Even so, members of the profession quickly grasped the potential importance of procedures that computers now make so readily available. Both the variable and the scale along the sequence must be expressed in units having magnitude: it is not sufficient simply to know the order of succession of points. The meanings of the nine elements of the stress matrix can be seen by imagining a cube in a Cartesian coordinate system in which the X-axis points to the east, the Y-axis points to the north, and the Z-axispoints up. 1 =1 0!(3 - O) ! 1 (3 2 . Griffiths, J.C., 1960, Some aspects of measurement in the geosciences: Mineral Industries, v. To her I dedicate this book. Suppose an offshore concession contains ten welldefined seismic features that seem to represent structures caused by movement of salt at depth. 380 Contouring by gridding . Comparing exact probabilities found by numerical integration with those predicted by the approximation equations, he found that significant differences occur only for very elongate targets that are large with respect to spacing between search lines. Eigenvalues and Eigenvectors Search for an elliptical target with major semiaxis a and minor semiaxis b. What is the strike and dip of the plane on which this stress occurs? Unfortunately, the source of the fragment cannot be identified with certainty because the fossil was found below the junction of two dry stream tributaries. Table 4-4 contains the record of successive rock types penetrated by a well drilled in the Midland Valley of Scotland (these data are contained in file MIDLAND.TXT). The data in Table 4-3 represent the stratigraphic section shown in Figure 4-5, in which the sedimentary rock has been classified at successive points spaced 1 ft apart. Of course, any errors that remain are the responsibility of the author alone, but I would be remiss if I did not acknowledge that the number of such remaining errors would be far greater if it were not for Geoff's careful scrutiny. In this class of problems, an observation's location has no other significance. We must find all possible combinations of elements that can be formed in this way. 136 Matrix Algebra In a 2 x 2 matrix we can find two combinations of elements that contain one and only one elements that contain satisfy this relationship. Suppose, for example, we measure the length of the hinge line on a brachiopod and find it to be 6 mm long. Runs Tests . Plot the record of lithologic compositions in the form of a lithologic strip log. 407 Statistical tests of trends getting no heads in three flips is one correct combination [TTT] out of eight possible, or 1 / 8. Most programs normalize, or scale each eigenvector so the sum of its elements will be equal to 1.0. Others scale each eigenvector so the sum of the squares of each eigenvector so the sum of its elements will be equal to 1.0. Others scale each eigenvector so the sum of its elements will be equal to 1.0. Others scale each eigenvector so the sum of the squares of each eigenvector so the sum of its elements will be equal to 1.0. Others scale each eigenvector so the sum of its elements will be equal to 1.0. Others scale each eigenvector so the sum of its element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element is element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element is element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element is element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element is element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element is element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element is element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element is element in a vector will be equal to 1.0. Others scale each eigenvector so the sum of its element is element in a vector will be equal to 1.0. Ot functions based on observations made at discrete points. 4 ~ 5.4 13.9 5.1 E - 2.2 4.3 4.7 2.8 4.9 1.8 4.2 4.7 = 2.31 3.8 M [86.36 55.88 45.72 116.84 109.22 106.68 137.16 119.38 119.38 129.54 99.06 124.46 71.12 96.52 58.421 Ma t rix M u Itip licat ion Recall the coin-flipping problem from Chapter 2, where we considered the probability of obtaining a succession of heads if the probability of heads on one flip was 1/2. 0.10 * 0.656 = 0.328 Using either the binomial equation or a table of the binomial equation or a table of the binomial distribution, the probabilities associated with all possible outcomes of the five-hole drilling program can be found. That is, they form a sequence, and the position at which a data point occurs within the sequence is important. What is the probability that one well will come in during the five-hole exploration campaign? Specifically, if the events are considered to be the result of a process yt = p + B t (4.5) the null hypothesis states that fi = 0. 103, p. My esteem for my two mentors, Dan Merriam and John Harbaugh, was expressed in my dedication to the second edition of this book. In the second, the points are assumed to be equally and regularly spaced; the numerical value of the discrete probability distributions just discussed is that the probability of success remains constant from trial to trial An obvious example is the topographic map; although the contour lines are an expression of a continuous and unbroken surface, the lines are calculated from measurements taken at triangulation and survey control points. 123 ... This is defined simply as the square root of variance and is symbolically written as .. In symmetric curves such as the normal distribution, the mean, median, and mode coincide. A veritable blizzard of publications has been issued on mine sampling. Problems in contour mapping . CT for the population parameter and s for the sample statistic. Comparing Normal Populations most familiar application of the theory is in kriging, an estimation procedure important in mine evaluation, mapping, and other applications. The function estimates the probability that an event has not occurred before time X. If the spacing is too great, thin units may be entirely missed. A histogram can be prepared which shows the probabilities attached to all possible outcomes in this exploration situation (Fig. Scientists usually express probability as an arbitrary number ranging from 0 to 1, or an equivalent percentage ranging from 0 to 1, or an absolutely certain that it will not rain. The initial estimated point in a sequence is taken as the first preceding data point. One form may be directly transformed into the other. (Start) F Q Q F Q F F F F F Q F F F F Q F F F F Q F F F F Q F F F F Q F F F F Q F OFFFFOOFFFOOFFFOOFFFO(End) We will now consider a related statistical procedure for examining what are called runs up and runs down. A 2-in. This can be shown in the form of a bar graph as in Figure 2-1. The entire core was split and all garnet crystals exposed on the split surface were removed, individually crushed, and 166 Analysis of programs have been established only in scattered institutions, through the efforts of a handful of people. 5 Statistics and Data Analysis in Geology All of the techniques of quantitative geology discussed in this book can be regarded as statistical procedures, or perhaps "quasi-statistical" procedures. Ore bodies that are distinctively radioactive or magnetic are sought using airborne instruments carried along equally spaced parallel flight lines. No single test can be taken as an exactly correct measure of the true permeability. For example, slightly over two-thirds (68.27%) of observations will fall within one standard deviation on either side of the mean of a normal distribution. 140 Matrix Algebra Eigenvalues and Eigenvectors The topic we will consider next usually is regarded as one of the most difficult topics in matrix algebra, the determination of eigenvalues and eigenvectors (also called "latent" and "proper" values and vectors). Seriation this general category are measurements separated by the flow of time, such as a sequence of water quality determinations at a river station, or the production history of a flowing gas well. If the number of equally spaced points is approximately the same as the number of equally spaced points are somewhat uniformly spaced, the technique will give satisfactory results. This theory, developed originally by Georges Matheron, a French mining engineer, describes the statistical behavior of spatial properties that are intermediate between purely random and completely deterministic phenomena. One obvious difference between this matrix and the one we have considered previously is that all the diagonal terms must be zero, since a state cannot succeed itself. However, a12a21 must be rearranged to a21a12 before the second subscripts are in numerical order. The test statistic, 184 Analysis of Sequences of Data Figure 4-12. The temptation is strong, perhaps irresistible, to utilize these computer is a state cannot succeed itself. though the user may not clearly understand the underlying principles on which the programs are based. Although the minerals on the scale, which extends from one to ten, increase in hardness with higher rank, the steps between successive states are not equal. If two events are not equal. they will occur simultaneously is the product of their separate probabilities of occurrence. These can be expressed conveniently in the form of statistics: Jour. Gearhart-Owen, 1975, Formation Data Handbook: Gerhard-Owen Industries, Inc., Fort Worth, Texas, 240 pp. This latter condition usually is assumed to prevail in many geological situations where discrete probability distributions are applied. A False Feeling of Security Perhaps t h s chapter should be concluded with a precautionary note. It must be converted to a 3 x 50 matrix by adding a column of 1'sin order for the dimensions of the matrix multiplication to be correct. Estimation of regularly spaced points will also be considered in Chapter 5, when we discuss contouring of map data. This is called a transition frequency matrix and tells us, for example, that sandstone is followed by shale 18 times, but followed by limestone only 2 times. I am fortunate to have enjoyed the help and encouragement of many people in the creation and evolution of this book throughout its several editions. Multigroup Discriminant Functions . . If we assume that the distribution is normal, we would expect about two-thirds of the cores tested to have oil saturations between about 16%so and 24%so. How do we estimate the frequencies of transitions from each state to itself, when this information is not available? The data are taken from one of a large number of wells studied by Doveton (1971). We can conveniently represent this by setting the areas underneath histograms in Figures 2-8 and 2-9 equal to 1.00, as was done in the histogram of Figure 2-2. This is illustrated in Figure 2-18. British Geographers, No. 42, p. Each element X i i becomes the element X i becomes the element X i i b minor semiaxis. The probability that the lithology two steps above a limestone will also be a limestone, regardless of the intervening lithology, is the sum of all possibilities. For example, we may be interested in some measurable attribute contained in successive stages of a sequence. is replaced by the vector of right-hand terms (the vector of b's). Rather, we must use a somewhat roundabout procedure to estimate the frequencies of transitions between the five sand size categories and place the results in a square, symmetric correlation matrix: 1 0.243 -0.301 0.096 -0.261 0.243 1 -0.969 -0.562 -0.422 -0.301 -0.969 1 0.340 0.253 0.691 1 I It is not obvious that this matrix should be singular with a zero determinant, yet it is. Test for randomness By considering only differences in magnitude between successive points, we have reduced the data sequence to a string having only two states (or three, if ties occur). Observations along a traverse may be taken at equally spaced intervals, as in point counting, or they may be taken at equally spaced intervals. stratigraphic sections. Variance may be regarded as the average squared deviation of all possible observations from the population, u2 is given by this equation. The process of obtaining a determinant from a square matrix is called evaluating the determinant. process of evaluating the determinant by selecting one element from each row of the matrix to form a term or combination of elements. There are only two possible outcomes (call them "success" and "failure") for each trial or flip. It is as though one variable forms a scale along which observations of the other variable are located, and we want to examine the nature of changes in this variable as we move up or down the scale. Unfortunately, we cannot know in advance of drilling which four of the ten features will prove productive. As you can verify by checlung these examples, the sum of the eigenvalues of a matrix is always equal to the sum of the diagonal elements, or the trace, of the . Almost all undergraduate students in the natural sciences and engineering, including the Earth sciences, are required to take classes in mathematics, statistics, data analysis, and computing. A series of events is therefore nothing more than a sequence of the intervals between original matrix. Series of Events ... occurrences. Other methods accept nominal or ordinal data, and observations need only to be categorized in some fashion. Test of goodness of fit The other sequences, in contrast, yield very different test results. This is done by comparing the observed transition frequency matrix to the matrix expected if successive states are independent. The success of mining geologists and engineers in the assessment of mineral deposits testifies to the power of these methods. Some final notes on eigenvalues: You'll notice that the correlation matrices we've graphed are square, symmetrical about their diagonals, composed of real elements (that is, no imaginary numbers), and that the largest numbers in every row 149 Statistics and Data Analysis in Geology - Chapter 3 are on the diagonal. *. Hypnotized by numbers, he or she may be led to the most ludicrous conclusions, totally blind to any reality beyond the computer screen. In this chapter, we will examine the mathematical operations that underlie such techniques as trend-surface analysis, principal components, and discriminant functions. Two matrices which exhibit the property that will be apparent are said to be commutative. It can be demonstrated that the number of possible combinations of n items, taken Y items at a time, is (r) The exclamation points stand for factorial and mean that the number preceding the exclamation point is multiplied by the number less one, then by the number less two, and so on: n! = n * (n - 1). Shape of target may range from a circle to a line; elliptical targets of various axial ratios fall in t h e shaded region. Sequence of 100 feldspar (F) and quartz (Q) grains encountered along traverse through pegmatite. 'I-'' Table 4-7. Buchanan, J.L., and P.R. Turner, 1992, Numerical Methods and Analysis: McGrawHill,Inc., New York, 751 pp. Because of the problems of accessibility, the samples were collected at irregular intervals up the winding stream channel. This also may be true of analytical data from drill holes, or from samples collected on traverses across regions which are incompletely exposed. Histogram of number of eruptions of the Japanese volcano Aso occurring in successive 100-yr intervals. A n identity matrix is almost always 124 Matrix Algebra indicated by I: Elementary Matrix Operations Addition and subtraction of matrices obey the rules of algebra of ordinary numbers, with one important additional characteristic. A shell with zero length does not exist, because it has no length at all. If certain assumptions can justifiably be made about the distribution of the populations from which the samples are collected, statistical tests called regression analyses can be performed. As we emphasized in Chapter 1, the computer is a powerful tool for the analysis of complex problems. The nearer a point is to an observation, the closer its value is to that of the observation. The probability p that a wildcat hole will discover oil or gas can be estimated using industry-wide success ratio of the particular company making the evaluation, or simply by making a subjective "guess." From p, the binomial model can be developed as it relates to exploratory drilling in the following steps: 1. 370 374 Contouring by triangulation Figure 2-13. Geologists should be intensely interested in the effects which control-point distributions have on maps, but few studies of this influence have been published. Discrete distribution giving the probability of obtaining specified numbers of heads in 50 flips of a coin. That is, p(B-A-B) = 0% p (B - C - B) = 2% p(B-D-B) = 0% Total = 52% - 172 Analysis of Sequences of Data The same reasoning can be applied to determine the probability of any lithology two steps hence, from any starting lithology. 176 Analysis of Sequences of Data to A B from C D E - A 66 29 C 13 36 3 5 35 2 29 29 1 44 26 23 9 185 42 123 0.125 0.026 0.083 0.064 - 0.055 A B from C D E - B 0.026 0.006 0.017 0.013 0.012 65.5 13.6 13.6 3.1 43.5 8.9 33.5 6.8 28.8 6.3 D 19 4 45 17 9 94 0.083 0.017 0.055 0.043 0.036 43.5 8.9 28.8 22.5 18.9 E Row Totals 52 0 12 3 12 186 41 123 94 79 79 5 23 Grand Total 0.064 0.055 0.013 0.028 0.028 0.024 - 33.5 6.8 22.5 17.3 14.7 - 28.8 6.3 18.9 14.7 12.6 - 177 Statistics and Data Analysis in Geology - Chapter 4 Note that the matrix is symmetrical and the diagonal elements remain unchanged, within the limits of rounding error. The manner in which numerical values are assigned determines the scale of measurement, and this in turn determines the scale of measurement. .. If the two vectors are closer together, the ratio between the semiaxes of the enclosing ellipse will change. If the intervals are not independent, this will be expressed as a positive autocorrelation with a tendency for large values of Xi (long intervals between events) to be succeeded by large values; similarly, there will be a tendency for small values of xi (short intervals)to be followed by other smallvalues. R-Mode Factor Analysis (1- 0.10)4 0.10. However, if a sampling experiment is performed with a small number of colored balls initially in the urn and those taken from the urn are not returned, the probabilities obviously change with each draw. We must, therefore, estimate the variable under consideration at regularly spaced points from its values at irregular intervals. Sequences in which transitions from a state to itself are not permitted are called embedded Mavkov chains, and their analysis presents special problems that have not always been appreciated by geologists studying stratigraphic records. This is the muZtipZicative rule of probability. Place the A matrix beside an identity matrix, I, and perform all operations simultaneously on both matrices. Canonical Correlation Because the conclusions reached in a quantitative study will be based at least in part on inferences drawn from measurements, the geologist must be aware of the nature of the nature of the nature of the number systems in which the measurements are made. If the diagonal elements are stripped from the matrix, it may be compared directly to the observed transition frequency matrix because the row and column totals of the two are the same, again within rounding limits. The probability is 18% that the lithology will be state A, 7% that it will be state B, and 11% that it will be state D. The simplest form of multiplication involves two square matrices, A and B, of equal size, producing the product matrix, C. In the first instance, we would expect runs of the same state; that is, several successive observations could conceivably fall in the same category. Only 101 cores, or about 6% of the total number of observations, have saturations outside the 2a range; that is, oil saturations outside the 2a range; that is, oil saturations outside the 2a range; that is, oil saturations less than 12% so or more than 29% so. Other dichotomizing schemes reflect only changes with respect to a single value such as the median or mean. (a) Using a parallel-line search of spacing D. long shell is twice the length of a 1-in. Because the author has made certain assumptions about the background, training, interests, and abilities of the materia is neither autocorrelated nor contains a trend, we may wish to test the possibility that the events are distributed according to a Poisson distribution. The fourth axis is very short, and the fifth axis is almost nonexistent; the hyperellipse defined by the correlation matrix, R, is really only a three-dimensionaldisk embedded in a space of five dimensional disk embedded dimensiona from sedimentary petrology, the methods are equally applicable to other problems in the Earth sciences. Ellipse defined by rows in matrix of correlations between copper and zinc. That is, we regard the intervals between events as a variable, X, located at equally spaced points. Equality of variance-covariance matrices 165 I I Statistics and Data Analysis in Geology - Chapter 4 Figure 4-3. 414 Kriging Griffiths' book, unfortunately now out of print, is especially commended to those who wish to perform experiments in geology and can exercise strict control over their sampling procedures. Runs are defined as uninterrupted sequences of the same state. If oil is found, the well cannot be dry, and vice versa. For years mining geologists and engineers have carefully designed sampling schemes and drilling plans and subjected their observations to statistical analyses. the number of trials prior to the initial success. We begin by assigning some arbitrarily large number, say 1000, to the diagonal positions of the observed transition frequency matrix. [[i y] 1: [li %][The matrix A is placed beside an identity matrix, I ; 025 row one is divided by 4, the first element in the row, to 0 11 produce 1 at all; 10 times row one is subtracted from row two to reduce 4. These are the elements of the covariance matrix, S, which we can compute by s = (n- i 1 - 1 ~ ~ ~ A subset of S could serve our purposes (and the covariance matrix, R. Each category is called a decile if the ranked sample is divided into ten equal categories, and a quartile if it is divided into four equal categories. The possibility that five holes will be required to achieve two successes when the regional success ratio is 25% is (2 + 3 - 1)! (1 - 0.25)3 * 0.2S2 P = (2 - 1)!3! - - . Although summary statistics are important, sometimes we can learn more by examining the distribution of the observations as shown on different plots and graphs. Sample localities were carefully noted on aerial photographs, and later the distances between samples were measured. A word about closure We wish to rearrange each combination until the second, or column, subscript of each element is in proper numerical order. If an odd number of transpositions is necessary (1, 3, 5, 7, etc.), the product is negative. In order for the cube to be in rotational equilibrium, shear stresses on adjacent faces must balance; so, for example, uxy= urx. This means that the stress matrix is symmetric about the diagonal: Turcotte and Schubert (1982) provide a more detailed discussion of stress in the subsurface and the measurement of stress components. 443 452 Selected Readings Both of these objections may be met by techniques that consider more than two of the original values, perhaps by fitting a linear function that extends over several adjacent values. 602 Table A.3. Critical values of F for v1 and v2 degrees of freedom and selected levels of significance The effects of closure, which results in unwarranted relationships between variables when they are forced to sum to a constant value, are examined in detail. Probability In the remainder of this chapter, we are going to examine the mathematical techniques required to analyze data in sequences. In matrix D, the means have already been removed. Exercise 3.5 Petrophysical well logs are strip charts made after the drilling of a well by lowering a sonde down the hole and recording physical properties versus depth in the well. Geologists typically attach a different meaning to the noun, "sample," than do statisticians. o.102 P = (2 - 1)!O! - 1! --. All the products are summed to obtain the C i j element of the answer. Wolfram Media, Inc., Champaign, Ill., 1395 pp. An example is the comparison of an X-ray diffraction chart with a set geologic application of runs tests are included in Miller and Kahn (1962, chapter 14) and Rock (1988, topic 16). Because any population has characteristic properties and the variation of its constituent members is limited, it is possible to select a relatively small, random sample that can adequately portray the traits of the population. Including Factor Analysis Time-series techniques are inherently more powerful than other analytical methods for examining sequential data, and their use has become widespread. Too often the search has been conducted haphazardly-the geologist wanders at random across the area of investigation like an old-time prospector following his burro. Subtraction, such as C = A - B, proceeds in exactly the same manner, with every element of B subtracted from its corresponding element in A. Bar graph showing the number of heads in three flips of a coin. We may wish to reject either form of nonrandomness. An entire branch of statistics treats probabilities as subjective expressions of the "degreeof belief" that a particular outcome will occur. 31 Statistics and Data Analysis in Geology - Chapter 2 The great advantage of plotting data in cumulative form, however, comes about because we can show the individual observations directly, and avoid the loss of resolution that comes from grouping the Observations into categories for a histogram. Some are especially tailored to deal with special types of matrices, such as those containing many zero elements (such matrices, such as those containing many zero elements) or possessing certain types of symmetry. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6001, fax (212) 850-6008, E-Mail: To order books or for customer service please call (800) 225-5945. Numerical computation packages for personal computers, such as MATHEMATICAQ and , contain alternative algorithms that can be used to calculate the inverse of matrices. Row two is multiplied by 5 and the product is subtracted from row one. Most especially, I must acknowledge the assistance of Geoff Bohling, who volunteered to shoulder the burden of reading every word in the manuscript, working each example and exercise, and checking all of the computations and tables. and mixed effects

. This assumption probably is reasonable in the first example, but the second requires that we assume that the sedimentation rate remained constant through the series. The statistical transformations proposed by John Aitchison to overcome these problems are discussed at length. The chart and standards can be compared only in one position, where intensities at certain angles are compared to intensities of the standards at the same angles. If the number of rows equals the number of the standards at the same angles. variables. If we can determine that our series follows a Poisson distribution, we can use the characteristics of the distribution to make probabilistic forecasts of the series. For centuries, the growers have treated their grapes with "blue galicia," or copper sulfate, to prevent fungus. Both illustrations are reproduced with permission. This is the 2 x 2 matrix: $A = [11030 \text{ Repeating the sequence of steps yields the determinant which is then expanded into I 4c; !3: 1 = (4 - A) (30 - A) - 100 = 0 143 \text{ Chapter 3 Statistics and Data Analysis in Geology- or h2 - 34h + 20 = 0 There are no obvious factors in the quadratic equation, so we must apply the rule for a general solution: <math>X = -b + J_2a = A = -(-34)c$ J-342 2x1 hi = 33.4 A2 = - 4 x 1 x 20 - 34 + J i D Z 2 0.6 We can check our work by substituting the eigenvalues back into the determinant to see if it is equal to zero, within the error introduced by round-off So, the eigenvalues back into the determinant to see if it is equal to zero, within the error introduced by round-off So, the eigenvalues we have found are correct within two decimal places. results than in theory. Otherwise, the manner in which you gather your data may predetermine the techniques that can be used for interpretation, and may seriously limit the scope of your investigation. In certain instances these tests may be more appropriate than the procedures described here. Figure 2-25 is copyright by Harcourt Brace Jovanovich, Inc. Rorres, 1994, Elementary Linear Algebra, 7th ed., Applications Version: John Wiley & Sons, Inc., New York, 800 pp. There is no a pn'ori reason why one position of comparison should be better than any other. Also, our thoughts are conditioned by the media in which we express them, and maps, photographs, and cross-sections are printed or drawn on flat sheets of paper. However, you should now have some insight into matrix methods that will enable you to understand the computational basis of techniques we will cover in the remainder of this book. To convert from one interval scale to another, we must perform two operations: a multiplication to change the scale, and an addition or subtraction to shift the arbitrary origin. Rowtwois dividedby 0.5. So, the (n + 1) column of the augmented matrix contains the solution to the simultaneous equation set, and our original matrix has been replaced by an identity matrix. Results of this experiment follow the binomial distribution, discussed in Chapter 2. As a consequence, the soil is enriched in copper and other metals that are present as impurities in the crude sulfate compound. If we were to compute a transition probability matrix and then raise it to an appropriately high power, it would estimate the transition probability matrix of a sequence in which successive states were independent. Second estimate of transition frequency matrix using new diagonal elements calculated by multiplying probabilities on the diagonal by the grand total of 5397. In contrast, rectangular integration will create estimates that are equal to the first observation. These workers are often characterized by a suspicious and iconoclastic attitude toward geological platitudes. Note that each entry C i j in the product matrix results from multiplying and summing the products of elements in the j t h column of matrix B. 2. Chapters 6 and 7 treat matrix algebra and its application to geological problems. From the transition probability matrix we can estimate what the lithology will be 2 ft (that is, two observations) above a given point. The appropriate alternatives are Ho: U r 8 H1: u < u and too few runs will cause rejection of the null hypothesis. The first set of trials contains 13 runs, the second only 2, and the third contains 19. The result you obtain is general. Geologists are trained to see patterns and structure in nature. 1 / 2 = 1/4. A geological sample" of a rock, a "cuttings sample" of a rock, a "cuttings sample" from a well, or a "grab sample" from a well, or a "grab sample" from a mine face, is a physical specimen and when represented by a quantitative or qualitative or qualitative or qualitative or a "grab sample" from a well, or a "grab sample" from a well, or a "grab sample" from a well, or a "grab sample" from a well or a "grab sample" from a "grab sample" from a well or a "grab sample" from a illustrated by solving the following pair of simultaneous equations in matrix form. The basic arrangement of topics covered in the book is retained from earlier editions, progressing from background information to the analysis of geological sequences, then maps, and finally to multivariate observations. At every foot, the rock type can be categorized and the type immediately above noted. A A D E - 0.208 0.192 B from B to C 0.203 C D 0.200 0.198 E - Row Totals 0.208 0.192 B from B to C 0.203 C D 0.208 0.192 B from B to C 0.208 0.192 B example, is designated X and the population mean is p (mu). Spherical Distributions In the quarter-century since the first appearance of this book, computers have progressed from mainframes of ponderous dimensions (but minuscule capacity) to small cubes that perch on the corner of a desk and contain the power of a supercomputer. Many faculty members in the Earth sciences received their formal education prior to the current emphasis on geomathematical methodology, and consequently are untrained in the quantitative subjects their students are required to master. Once you master the rudiments of matrix algebra, you will be able to see the fundamental structure within the complex procedures we edge of the ellipse is 1.28 units. The same conditions that govern the binomial distribution may be assumed, except that the number of "trials" is not fixed. 420 Universal kriging will examine later. Interactive software allows the user to dynamically vary the width of the histogram intervals and move the origin, so alternatives can be easily evaluated. Identification of one 7 Statistics and Data Analysis in Geology - Chapter 1 fossil as a brachiopod and another as a crinoid implies nothing about the relative importance or magnitude of the two. Geologic applications of this test may not be obvious, because we ordinarily must consider more than two states in a succession. Entries in a succession. Entries in a subset of elements taken from a population. However, we want to reverse these equations, and for given values of P b and At that we read from the well logs, estimate the proportions of the rock. The magnitude of the major semiaxis is equal to 1.85, the first eigenvalue of RC7,,... For example, consider the multiplication A-IA1/2 = A-112 If A is 3 x 3, the product is In some applications, the inverse may not be required, but only the solutions to a set of simultaneous equations. Data for 50 samples are given in Table 4-7 and shown graphically in Figure 4-14. Justification for applying series-of-eventsmodels to spatial data may be tenuous, however, and depends on the assumption that the spatial sequence has been created at a constant rate. To examine these problems, we must have a collection of measurements of a variable and also must know the locations of the measurement points. If we assume that a straight linear relation ship Expressed in other words, the difference betweenvalues of two adjacent points is assumed to be a function of the distance separating them. Golub, G.H., and C.F. V an Loan, 1996, Matrix Computations, 3Tded.: Johns Hopkins Univ. 20 Elementary Statistics The probability of making x discoveries in a drilling program of n holes, when sampling from a population of N prospects of which S are believed to contain reservoirs, is This is the number of combinations of the reservoirs taken by the number of discoveries, times the number of discoveries, times the number of dry holes, all divided by the number of combinations of all the prospects taken by the total number of holes in the drilling program. Continuous variables, however, must be measured 25 Statistics and Data Analysis in Geology - Chapter 2 Number of intervals that contain exactly 0, 1, 2, . We may be interested in the geologic features exposed in a deep mine with successive levels, adits, and raises creating a complex three-dimensional net, yet we must reduce this network to flat projections in order to express our ideas concerning the relationships we see. The interval scale is so named because the length of successive intervals is a constant. However, the representation may equally well be a "map" of a thin section or electron photomicrograph, where the relationships between features have been enlarged so they become visible. The value of a point halfway between two observations is exactly intermediate between the values of the two enclosing points. Structural analysis and orthogonal regression I would also like to note that I have benefited from the nurturing environment of the Kansas Geological Survey (KGS) at The University of Kansas. Calibration Consider the set of equations: a11x1+ al~x2 = a z m +mx2 bl = b2 Expressed in matrix form, this becomes and we have discussed how the vector of unknown x's can be solved by matrix inversion. Crosscorrelation and stratigraphic correlation . In other 162 Analysis of Sequences of Data circumstances such as stratigraphic correlation, equivalent thicknesses may not represent equivalent temporal intervals and the problem of cross comparison is much more complex. Data having the characteristic of being arranged along a continu either of time or space, often are referred to as forming a series, sequence, string, or chain. Thus, while there may be many excuses for failing to work an exercise, entering data incorrectly should not be one of them! We have already noted that computing technology has changed enormously during the 28 years this book has been in print. Acknowledging that no students, not even graduate assistants, like to do drudge work such as data entry, I've provided all of the data for examples and exercises as digital files on the World Wide Web. Statistical problems, whether perceived or not, occur wherever there are elements of chance. 153 Statistics and Data Analysis in Geology - Chapter 3 Exercise 3.4 Coordinates can be rotated by a matrix multiplication in which the premultiplier is a 2 x 2 matrix of sines and cosines of the angle of rotation. What has been sacrificed are all but the rudiments of statistical theory associated with each of the techniques, the details of all mathematical operations except those that are absolutely essential, and all the embellishments and refinements that typically are added to the basic procedures. This approach is illustrated in Figure 4-1. The probability also will remain approximately constant if there are a very large number of balls in the urn, even if those selected are not returned, because their removal causes an infinite simal change in the proportions among those remaining. Rather, there is an infinite continuum of possible results that might be obtained. (1 - 0.10)' - .5! 1 * 0.90' 5!0! = 11. In this chapter, we will examine methods for analyzing features on what we loosely define as "maps": two-dimensional representations of areas. If the autocorrelation coefficients are not significantly different from zero, as tested by methods that will 183 Statistics and Data Analysis in Geology - Chapter 4 be developed later in this chapter, we can conclude that the events are occurring independently in time or space. A discussion of fractals has been added, not because fractals have demonstrated any particular utility in geological investigations, but because they seem to hold a promise for the future. This unsatisfactory and even dangerous method of education, comparable perhaps to a physician learning by on-the-job training, is one many people seem destined to follow. (Bottom of section) 1 2 3 2 3 5 7 9 2 2 1 0 2 3 2 0 10 12 14 use they be a physician learning by on-the-job training, is one many people seem destined to follow. 22 17 19 14 4 9 11 10 12 7 4 3 2 3 3 3 4 9 1 0 1 0 8 912 2 1 0 0 8 14 16 27 (Topof section) In the study of a silicified shale unit in the Rocky Mountains, it was noted that the rock contained unusual numbers of well-preserved radiolarian tests. The apparent density and the sonic transmission time of a mixture of these three constituents can be calculated as the sum of the products of the densities and transit times for pure constituents. A matrix with no inverse is called a singular matrix, and presents problems beyond the scope of this chapter. Matrix Multiplication Unfortunately, most geologists must collect their Observations where they can. In this procedure, observations are dichotomized by comparing their magnitudes to the preceding observations. For example, if asked to guess whether it will rain tomorrow, most of us would reply with some confidence that rain is likely or unlikely, or perhaps in rare circumstances, that it is certain to rain, or certain not to rain. Of course, there are some applications and some analysis. Dividing the data into a small number of categories increases the average number in each and the histogram will be relatively 29 Statistics and Data Analysis in Geology - Chapter 2 reproducible with repeated sampling. Proba bility Although many descriptions and definitions of statistics have been written, it perhaps may be best considered as the determination of the probable from the possible. These may seem obvious or even trivial points to emphasize, but as we shall see, not all geologic sequences have such well-behaved characteristics. The operation is indicated symbolically by XT or by X'. For example, the upward movement of magma in chambers beneath a volcano such as Mt. St. Helens in Washington is believed to cause a harmonic tremor, a particular type of earthquake. Looking at residuals Satellite reconnaissance, by its very nature, consists of parallel orbital tracks. Also included is an introduction to geostatistics, a subspecialty that has grown into an entire branch of applied statistics. I especially wish to acknowledge the support and encouragement of two previous directors of the Kansas Geological Survey, Bill Hambleton and Lee Gerhard, who recognized the importance of geology's quantitative aspects. Cox and Lewis (1966) described a variety of graphical tools that are useful in examining series of events. The data in Table 3-3 can be regarded as a 20 x 5 matrix, M. These methods provide answers to the followingbroad categories of questions: Are the observations random, or do they contain evidence of a trend or pattern? (a) Data sequence considered as a step function or "rectangular curve." (b) Equally spaced sequence created by rectangular integration. If we wish to create an equal length whose areas equal the total areas of the original rectangles. If the observations X I, X Z, ... Distribution of percent oil saturation (so) measured on cores from a field (a) in northeastern Oklahoma and (b) in west Texas. You should note the similarity of the bellshaped continuous curve shown in Figure 2-9. Rather, at specified levels of significance, we can demonstrate that the null hypothesis is incorrect and the sequence is therefore not random. A shear stress is positive if the compressional or tensional component agrees in sign with the direction of force. The final category is concerned with clustering, classification, and the examination of interrelations among variables in which sample locations on a map or traverse are not considered. One is the variance, and the other is the square root of the variance, called the standard deviation. The denominator is the determinant of the coefficients (in our example, the a's). Griffiths, J.C., 1967, Scientific Method in Analysis of Sediments: McGraw-Hill, Inc., New York, 508 pp. The data are contained in file CROATRAD.TXT at the Web sites (see Preface). Individual entries in a matrix, or its elements, are indicated by subscripted italic lowercase letters such as Xij. It includes time series, calculation of semivariograms, analysis of stratigraphic sections, and the interpretation of semivariograms, analysis of stratigraphic sections, and the interpretation of chart recordings such as well logs. is that it is more amenable to computer programming. As a consequence, the investigator may acquire a closer familiarity with the objects of study than could otherwise be attained. On the first toss, our chances are 1 / 2 of obtaining a head. However, with matrix algebra we can express the basic principles involved in a manner that is succinct and alternative notation will be useful later.] to A B from C D A B C D 0.78 0 0.22 0 0.71 0.29 0 0 0.18 0.07 0.64 0.11 0 0 0.60 0.40 Row Totals 1.00 1.00 Here, for example, we see that if we are in state C at one point, the probability is 64% that the lithology 1 ft up will also be state C. Likewise, the lithologic states of the units cannot be expressed on a numerical scale. The measurements are in degrees Fahrenheit. In this instance, the polynomial can be factored into (A - 4)(A - 0)(A + 2) = 0 and the roots of large polynomial equations can be an arduous task. The development and explosive proliferation of personal computers has accelerated this trend. There are other discrete probability distributions that apply to experimental situations similar to those appropriate for the binomial. In contrast, some of the techniques we will discuss in this chapter regard data sequences as samples from a continuous string of possible observations. The two eigenvalues of this matrix are hi = 1.029 A2 = 0.971 which are almost identical in size. , 0, 2, 4, 6, etc.), the product is given a positive sign. The lithologies include four mutually exclusive states-sandstone, limestone, shale, and coal, arbitrarily designated A, B, C, and D, respectively. Index If we take repeated samples, the mean deviations will be more variable than variances calculated from the same samples. This is done using a x 2 test, first converting the probabilities to expected numbers of occurrences by multiplying each row by . Degrees of freedom Chapter 2. That is, Pb = 2.03Vh At = 67Vh 154 + 2.98Va -? 2.43vsh + SOVa + 113vsh Matrix Algebra where v h , V,, and Vsh are the proportions of halite, anhydrite, and shale. Embedded Markov chains The difficulty of selecting an appropriate sampling interval can be avoided if observations are taken only when there is a change in state. Since the spread of computers throughout universities and corporations in the late 1950's, geologists have been increasingly attracted to mathematical methods of data analysis. This will give the transition probability matrix shown below, from which the probability matrix shown below, from which the probability of proceeding from one state to another can be assessed. Its magnitude or length is 0.72. They test for a form of nonrandomness expressed by the presence of too few or too many runs, and do not identify overall trends. Sadly it must be confessed that such cynicism is often justified. The methods just developed can be extended directly to n x n matrices, although the procedure becomes increasingly cumbersome with larger matrices. Each trial is independent of all others. The vertical axis of Figure 4-8 is expressed in logarithmic form. Confidence Limits The first edition of this text (published in 1973) devoted a chapter to the FORTRAN computer language and most procedures in that edition were accompanied by short program listings in FORTRAN. The four characteristics listed above must be assumed to be true; such assumptions seem most reasonable when applied to "wildcat" explorations. All these have an associated probability that describes their frequency of Step 4. The range of possible outcomes may be finite and in fact quite limited, but within the range the exact result that may appear cannot be predicted. Usually only the first few lags will be of interest. .. Our histogram now can be redrawn and expressed in probabilities, giving the discrete probability distribution shown in Figure 2-2. In this book, the student will find not only generalized developments of computational techniques, but also numerous examples of their applications in geology and a library of problem sets for the exercises that are included. Regression through the origin . that limestone is the third state in the section and coal is the sixth has no significance that can be expressed numerically (that is, position 6 is not "twice" position 3). To do this, we must first rank the observations from smallest to largest, divide each observation's rank by the number of observations to convert it into a fraction, then multiply by 100 to express it as a percentile. Histograms of airborne measurements of total radiation on the lstrian peninsula of Croatia. In some circumstances it is possible to know the size of the population within which discoveries will be made. The coin-flipping experiment has four characteristics: 1. If you have access to an interactive statistics package, you can experiment has four characteristics: by Jackson (1991); Press and others (1992) provide a more compact presentation, as well as computer program listings. Binary-state variables are a special type of nominal data in which symbolic tags such as 1 and 0, "yes" and "no," or "on" and "off" indicate the presence or absence of a condition, feature, or organism. We have attempted to present, in as painless a manner as possible, the rudiments of beginning matrix algebra. These are illustrated using the data on the eruptions of Aso from Table 4-5. They also are the determinants of the matrix A. Higher dimensional fractals Although data arrays usually are in the form of rectangular matrices, often we will create square matrices from them by calculating their variances and covariances or other summary statistics. The matrix A must be a square matrix. For example, we may wish to see if a sequence contains more 186 Analysis of Sequences of Data than the expected number of runs from a random arrangement; the null hypothesis and alternative are Hi: U > U and too many runs leads to rejection. Co., Boston, Mass., 743 pp. PremultiplyingD by its transpose will yield a square 5 x 5 matrix whose individual entries are the sums of squares 5 x 5 matrix whose individual entries are the sums of square 5 x 5 matrix whose individual entries are the sum of square 5 x 5 matrix whose individual entries are the sum of square 5 x 5 matrix whose individual entries are the sum of square 5 x 5 matrix whose individual entries are the sum of square 5 x 5 matrix whose individual entries are the sum of square 5 x 5 matrix whose individual entries are the sum of square 5 x 5 matrix whose individual entries are the sum of square 5 x 5 matrix whose individual entries are the square 5 x 5 matrix whose individual entries are the square 5 x 5 matrix whose individual entries are the square 5 x 5 matrix whose individual entries are the square 5 x 5 matri the occurrence of harmonic tremors may help predict eruptions. Suppose, for example, an 16 Elementary Statistics Number of discoveries Figure 2-3. We can experimentally create a similar succession by flipping pennies and noting the occurrence of heads or tails. However, the (n-Y) dry holes and the Y discoveries may be arranged in)(: combinations or, equivalently, in n!/(n-Y) ! Y !different ways. To convert from one ratio scale to another, such as from inches to centimeters, we will not distinguish between the two measurement scales, and they 8 Introduction may occur intermixed in the same problem. Quantitative techniques can reveal the insufficiency of data, the tenuousness of assumptions, the paucity of information contained in most geologic studies. Most values are clustered around some central value, and the frequency of occurrence declines away from this central point. The technique is called rectangular integration. For example the frequencies along the main diagonal of the transition frequency matrix would be doubled while off-diagonal frequencies remained unchanged if observations were made every half-foot. The quartz-feldspar sequence along the line is listed in Table 4-6. Similarly, the determinant 1 2 3 4 5 6 = O 5 7 9 is zero because the third row is the s u m of rows one and two. Years of eruptions of the volcano Aso for the period 1229-1962. Three major grades of clay were produced: clay for drilling mud; foundry clay; and a miscellaneous category that includes cattle feed binder, drug and cosmetic uses, and pottery clay. lithologies, geochemical or mineralogical assays along traverses or drill holes, electric logs of oil wells, and chart recordings from instruments. However, they require long stratigraphic sequences and diamond-drill cores through ore deposits. Using the matrix operations we have already discussed, we will construct a matrix containing correlations between the concentrations of the different metals. to A B C D E Z Z s A - 0 13 B 2 9 0 from c 35 2 D 29 1 E 26 23 36 5 0 44 9 19 4 45 0 9 52O 12 3 0 120 38 94 77 67 The lithologic states have been coded as (A) unfossiliferous shale and mudstone, (B) shales containing nonmarine bivalves, (C) siltstone, (D) sandstone, and (E) coals and root zones. A modest data set such as 1STRIA.m will present a challenge to those who attempt to analyze the data by hand. In effect, we divide the continuous scale into discrete segments, and can then count the number of events that occur within each interval. Measurements of length are of this type. This is found by first calculating the negative binomial distribution in cumulative distribution in cumulative form in which each successive probability is added to the preceding probabilities; the cumulative distribution in cumulative form in which each successive probability is added to the preceding achieved in (x + Y) or fewer holes as shown in Figure 2-5. For the matrices C and D, demonstrate that (CD)T = DTCT. The aim of this book is to introduce organization into the self-educational process, and guide the impatient neophyte rapidly through the necessary initial steps to a glittering algorithmic Grail. We can count the number of total possible combinations, which is eight, and convert the frequencies of occurrence into probabilities. The number of transpositions is zero, so the sign of the product is positive. However, this formula requires subtracting two quantities, 1 x; and (1x i) 2 ,and both may be very large and very nearly the same. Because of budgetary limitations, it is not possible to drill all of the features in the current exploration program. Events are not necessarily mutually exclusive. These sediments are interpreted as having been deposited in a delta plain environment subject to repeated flooding, so we would expect that certain lithologies would occur in preferred relations to 173 Statistics and Data Analysis in Geology - Chapter 4 Table 4-4. The manganese values from stream samples listed in Table 4-2 are shown in graphical form in Figure 4-2 b. Suppose our sample contains 300 observations; the three smallest values constitute the first percentile. The WWW addresses are and . We will delay a discussion of this procedure until Chapter 6, where we can provide a motivation for our interest. Usually, eigenvalues are not found by solution of a polynomial equation, but rather by matrix manipulation methods that involve refinement of a successive series of approximations to the eigenvalues. The hybrid Earth sciences of geochemistry, geophysics, and geohydrology require a firm background in mathematics, although their procedures are primarily derived from the non-geological parent. From an analysis of probabilities associated with events, future behavior or past states of the object or event under study may be estimated. Gives procedures and algorithms for matrix operations, especially different methods for inversion, solution of simultaneous equations, and extraction of eigenvalues. In our coin-flipping problem, (y) 3! - 3 - 2. For example, suppose we wish to develop the probabilities associated with a five-hole exploration program in a virgin basin where the success ratio is anticipated to be about 10%. Demonstrate that for commutative matrices, A - ~ B -=~ (ABP Consider the following two matrices, c = 2 1 0 40 02] [03 .=[::i] 1 -1 - Compare the determinants of the two matrices. In this text we will concern ourselves with those less tractable situations where the sample design (either by chances) or misfortune) is beyond our control. The forced scrutiny may reveal complexities or biases not apparent during the initial examination. 11. A familiar form of display is the histogram, a bar chart in which a continuous variable is divided into discrete categories and the number or proportion of observations that fall into each category is represented by the areas of the corresponding bars. Other techniques are similarly developed. Random patterns As in that example, p is the probability of a discovery and Y is the number of "successes" or discovery wells. Numerical methods allow the geologist to consider many variables simultaneously, a powerful extension of his pattern-recognition facilities. Part of this chapter is derived from this excellent exposition, wrlffen originally for students. Be sure to choose in order from the first row, second row, third row, . Two-dimensional methods are, for the most part, direct extensions of techniques discussed in Chapter 4. For alla22a33, no transpositions are necessary, so k = 0 and the term is positive, Transpositions for the others and the resulting signs are given below: a 1 1 '282 = '1a 1 '23 a33 = % % a 3 3 ' a33 = % % a 3 3 Chapter 3 Statistics and Data Analysis in Geology- Thus, there are three negative and three positive terms in the determinant. Bentonite production in Wyoming, 1964. A more conventional approach (see, for example, Anton and Rorres, 1994) uses what is called the "method of cofactors," but the two can be shown to be equivalent. These characteristics are known as parameters if they describe populations, and statistics if they refer to samples. It seems reasonable that it also may apply to the series of geological events described by Pettofrezzo (1978). Obviously, an infinite number of measurements of the top of this horizon could be made if we drilled unlimited numbers of holes. Note that O! is defined as being one, not zero. If tests determine that a sequence exhibits partial dependence between successive states, the structure of this dependence may be investigated further. .. However, the steps between grades do not represent a uniform progression of temperature and pressure. In some situations, an inverse cannot be found because division by zero is encountered during the inversion process. Techniques discussed in this chapter classified by the nature of the Determinants . variable and its spacing along a line. Probability of intersecting targets with regular search patterns ranging from squares t o parallel lines. Spatial Analysis In any problem where points are interpolated between observations, however, you must always remember that you cannot create data by estimation using any method. The term contains one and only one element from each column. The reason for using the average of squared deviations, n multiplications, and n summations. 131 Statistics and Data Analysis in Geology - Chapter 3 Inversion and Solution of Simultaneous Equations Division of one matrix by another, in the sense of ordinary algebraic division, cannot be performed. Some necessitate interval or ratio Observations; the variate must be measured on a scale and expressed in real numbers. A symmetric matrix is a square matrix in which all observations X i j = X j i, as for example x33 [:: '1 3 5 6 The variance-covariance matrix of trace elements given above is another example of a square matrix that is symmetrical about the diagonal. Interpolation Procedures Many of the following techniques require data that are equally spaced; the observations must be taken at regular intervals on a traverse or line, or equally spaced through time. Sometimes the measurements we make, such as chemical analyses, require the destruction of the material. This paperback reprint of a classic text covers the traditional material for a one-semester matrix algebra course. Chromium content of an Upper Pennsylvanian shale from Kansas. There are four measurement scales, each more rigorously defined than its predecessor, and each containing greater information. The relationship between conditional and unconditional probabilities can be expressed by Bayes' theorem, named for Thomas Bayes, an eighteenth century English clergyman who investigated the manner in which probabilities change as more information becomes available. The rigor and objectivity required by quantitative methodologies can compensate in part for insight and experience which otherwise must be gained by many years of work. If geology were an experimental science like chemistry or physics-in which observations can be verified by any competent worker-controversy and conflict might disappear. The implications contained in various concepts of probability are discussed in books by von Mises (1981) and Fisher (1973). Large statistical packages implemented almost every procedure described in the text, so program listings were no longer necessary. The event has two states and must occupy one or the other; except for the vanishingly small possibility that the coin will land precisely on edge, it must come up either heads or tails. Call the result U, a 20 x 5 matrix of standardized values; U = DC-' We can calculate the correlation matrix by repeating the procedure we used to find S, substituting U for D: R = (n-1) - 1 U T U R = 1 1 -0.312 -0.312 1 0.141 -0.201 0.85 -0.33 0.595 -0.28 0.85 0.595 -0.33 -0.28 1 -0.029 0.456 1 0.242 -0.029 0.456 0.242 1 0.141 -0.201 1 To graphically illustrate matrix relationships, we must confine ourselves to 2 x 2 matrices, which we can extract from R. a q&+m 0 0 0 0 0 bc Figure 2-15. Exercises Among the topics we will consider that have obvious applications to fields as diverse as geophysics and microscopy is the probability of encountering an object with a systematic search across an area. Using linear interpolation, values will be created which lie on a straight line between two surrounding data points. The vi Preface list of those who provided technical reviews and critical comments over the years reads like a "Who's Who" of mathematical geology and includes, in alphabetical order, Frits Agterberg, Dave Best, Paul Brockington, Jim Campbell, Ted Chang, Felix Chayes, Frank Ethridge, Je-an Fang, Colin Ferguson, John Griffiths, Jan Harff, Giinther Hausberger, Ute Herzfeld, George Koch, Michael McCullagh, Gerry Middleton, Vera Pawlowsky, Floyd Preston, Nick Rock, Robert Sampson, Paul Switzer, Keith Turner, Leopold Weber, and Zhou Di. In addition, there have been dozens of others who have called or written to clarify a specific point or to bring an error to my attention, or to suggest ways in which the text could be improved. These are shown in Figure 2-3. Mann-Whitney . Examination of the original data shows that there are 1145 cores having saturations within this range, or about 68% of the data. This histogram directly indicates local periods of fluctuation from the average rate of occurrence. The unknown coefficients are x1 = 2 and x2 = 3. Discrete distribution for exactly two test successes in a drilling program of n exploratory holes when the probability of a discovery is 25%. Moisture content of core samples of Recent mud in Louisiana estuary. This was a time of transition when personal computers offered the freedom of access and ease of use missing in the centralized mainframe environment, but these PC's lacked the power and speed necessary for many geological applications. 29, no. Figure 2-2 shows the probability distribution for all possible numbers of heads in three flips of a coin. We will also look at ways of describing and comparing more conventional shapes of individual objects, ranging in size from islands to oil fields to microfossils. If an even number of iron content by a rapid spectrochemical method. Geology itself is responsible for some of the advances, most notably in the area of mapping and spatial analysis. 18 Elementary Statistics Figure 2-5. That is, p (hit sandstone and drill anticline) = p (hit sandstone and drill anticline). obtaining specified numbers of "successes" (or heads) in ten flips of a coin, and Figure 2-9 shows the probability distribution that describes outcomes from an experiment involving 50 flips of a coin. If, on the other hand, we state that the probability of rain is 1, we are absolutely certain that it will. 607 Table A.5. Probabilities of occurrence of specified values of the Mann-Whitney Wx test statistic New row totals and grand total are then found A 232 29 B 35 from C 29 D E - 26 A B to C D 13 199 2 1 23 36 5 222 44 9 19 4 45 215 9 E 52 0 12 3 211 Row Totals 352 237 316 292 278 1475 Grand Total The process is repeated again and again, until the estimated transition frequencies along the diagonal do not change from time to time. Suppose, for example, we are interested in the porosity of a particular sandstone unit. 603 Table A.4. Critical values of x 2 for v degrees of freedom and selected levels of significance The equation given above is solved by finding the inverse of matrix A. Runs in the first sequence shown are underlined: (Start) H T H H T I J T T T H T I J T T T H T I J T T T H T I J T T T H T I J T T H H H 1 1 3 4 5 6 7 8 9 10 11 12 13 (End) We can calculate the probability that a given sequence of two states (heads and tails, in this example). The presentation of masses of numbers, all expressed to eight decimal places, overwhelms the minds of many people and numbs their natural skepticism. The critical value of x2 for 11degrees of freedom and an o(= 0.05 level of significance is 19.68, which is far exceeded by the test statistic. For example, a second-order Markov sequence exhibits a significance is 19.68, which is far exceeded by the test statistic. Sequence of data points t o be analyzed by the method of runs up and down. Most must make their way as best they can, reading, questioning, and educating themselves by trial and error. The median is the 5 0 t h percentile, the 5 t h decile, or the 2nd quartile. Of course, it is my hope that both the student and the instructor will find something of interest in this book that will help promote the widening common ground we refer to as geomathematics. The slope of a line drawn from the origin of a graph through a point is defined by the ratio between the two coordinates of the point, and not by the actual magnitudes of the coordinates. 2-6). Equality of two vector means Elementary Statistics If observations with certain characteristics are systematically excluded from the sample, deliberately or inadvertently, the sample is said to be biased. Tables A.6 and A.8 are copyright by John Wiley & Sons, Inc. What could be cuter Than to feed a computer With wrong information But naive The probabilities calculated are low because they relate to the likelihood of obtaining two successes and exactly x dry holes. The preceding examples have addressed situations where there are only two possible outcomes: a hole is dry expectation To obtain with precision A Napoleonic decision? Distribution of Points or oil is discovered. Gould, P., 1967, On the geographic interpretation of eigenvalues: An initial exploration: Trans. Then, the histogram would be a continuous, rather than discrete, variable. Similarly, metamorphic rocks may be ranked along a scale of metamorphic grade, which reflects the intensity of alteration. Diskeffes contain examples and exercises. This is done, for example, by determining if states tend to be followed an unusual number of times by greater or lesser states on the ordinal scale. If we raise the diagonal matrix A to the onehalf power, the result is a diagonal matrix whose nonzero elements are equal to the square roots of the equivalent elements in A. 346 351 Spectral procedures Consider the data in file BAL,TIC.TXT, which gives the weight-percent of sand in five successive size fractions, measured on bottom samples collected in an area of the Baltic Sea. This is usually the variation that is of scientific interest. Although the correspondence may be high, it is meaningless. Increasing the number of categories reveals more details of the distribution, but because each category will contain an indication of the relative violence and duration of some eruptions, for all practical purposes we must regard the record as one of indistinguishable instantaneous explosive events. As with all frequency distributions, we may define the probability directly from the curve. You will recall from Chapter 2 that the Poisson is a discrete probability distribution that can be regarded as the limiting case of the binomial when n, the number of trials, becomes very small. Stratigraphic succession shown in Figure 4-4 coded into four mutually exclusive states of sandstone (A), limestone (B), shale (C), and coal (D); observations taken at 1-ft intervals. Eigenvalue and eigenvector techniques are directly extendible to larger matrices, even though the operations become tedious. At the same time, the discipline necessary to perform quantitative research will hasten the growth and maturity of the scientist. The Kolmogorov-Smirnov test provides a simple way to test the goodness of fit of a series of events to that expected from a Poisson distribution. The line on the polished slab may be regarded as a random sample of possible successions through the pegmatite body from which the slab was quarried. Exercise 3.3 File MAGNET1T.m contains the proportions of olivine, magnetite, and anorthite estimated by point-counting thin sections from 15 hand specimens collected at a magnetite deposit in the Laramie Range of Wyoming. However, those to whom I owe the greatest debt of gratitude for help Appendix with this 3rd edition are my associates and coworkers at the Kansas Geological Survey, particularly Ricardo Olea, John Doveton, and David Collins, who have provided examples, data, and exercises, and who have patiently reviewed specific topics with me in order to clarify my thoughts and to help me correct my misconceptions and errors. 164 Analysis of Sequences of Data 'I 0 l 08] d 2 0 a 4 0 0. The resulting x2 statistic is - 2 - (18 - 8.5)' 8.5 + (7 - 7.0)' 7.0 + (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - - (5 - 10.4)' 10.4 + (5 - 5.0)' 5.0 + + - - - - - (5 - 10. significance is 16.92; the test value comfortably exceeds this, so we may conclude that the hypothesis of independence of successive states is not correct. 53-86. This is not to say that the geologist should unthinkingly bend his interpretation to conform with that of the computer. Testing the Mean -0.683- -3 -2 -1 0 1 2 3 Figure 2-18. If the matrix is symmetric (the variety of matrix we will encounter most often), its determinant cannot be negative. Perhaps we have measured the boron content of each lithologic unit in the cyclothem just discussed. 2-11) reflects fallout from the Chernobyl reactor accident in the Soviet Union during April of 1986. This tendency will be used to interpolate between data points, extrapolate beyond the data sequence, infer the presence of trends, or estimate characteristics that may be of interest to the geologist. However, like statistical tests, geomathematical techniques are based on the premise that information about a phenomenon can be deduced from an examination of a small sample collected from a vastly larger set of potential observations on the phenomenon. On a more prosaic topic, the section on regression has been expanded to include several variants that have special significance in the Earth sciences. The relevance of these courses to the geologist's primary field is often obscure. Grid-cell procedure . These methods of cross comparison superficially resemble the mental process of geologic correlation, but have the limitation that they assume the distance or time scales of the two sequences being compared are the same. An alternative way of expressing our estimate would be to Statistics and Data Analysis in Geology - Chapter 2 use a numerical scale, as for example a percentage scale. However, numbers are used only as identifiers. 4. Finally, even if it were possible to observe an entire population, it might be more efficient to sample. We are interested in the general tendency of the data in most of the examples we will now consider. This is a topic discussed in greater detail elsewhere in this chapter. Heads and tails constitute two categories of a nominal scale, and our data will consist of the number of observations that fall into them. It was. Teaching the techniques of data analysis to engineers and natural scientists, both university students and industry practitioners, would be easier, I reasoned, if I had a suitable textbook. We assume that heads and tails will occur at random; in the two preceding examples, it seems very unlikely that they have. Measurement Systems A quantitative approach to geology requires something more profound than a headlong rush into the field armed with a personal computer. This property is especially important in operations in the following sections. Suppose we want the eigenvalues of the matrix A = [-62 41 3 Expressed as a determinant equal to zero, we have which expands to or h2 - 5 h + 30 = 0 The roots of this equation are But this leads to equations involving the square roots of negative numbers: h = A2 = 14455 + m2 - m2 = 2.5 + 4.9i= 2.5 - 4.9i Matrix Algebra These are complex numbers, containing both real parts and imaginary parts which Fortunately, a symmetric matrix always include the imaginary number, i = yields real eigenvalues, and most of our computations involving eigenvalues and eigenvalues. always symmetrical. Several elaborate statistical distributions have been proposed to account for the variation probability matrix is 174 Analysis of Sequences of Data A - 0 0.76 from C 0.37 D 0.38 E - 0.40 B 0.11 0.30 0.16 0.43 - 1.00 1.00 0 0.13 0.11 0 0.02 0 0.48 0.13 1.00 0.01 0.57 0 0.04 1.00 1.00 0.34 0.13 0.13 0 The marginal probability vector is A D C E 0.30 [E] 0.19 0.17 A x 2 test, identical to Equation (4.2), can be used to check for the Markov property in an embedded sequence. Selected Readings In the matrix of trace-element variances and covariances, the variances lie on the diagonal and the off-diagonal elements are the covariances. These are a11a22 and a12a21. This reflects in part the fact that the third dimensions. The existence of a significant second-order property can be checked in exactly the same manner as we checked for independence between successive states, by using a x2 test. The individual elements of a matrix may be Statistics and Data Analysis in Geology - Chapter 3 measurements of numbers. For example, we might select the ~ 3 3 a 3 x 3 matrix. V contains the eigenvalues of either XXT or XTX (they will be identical except that XTX will have n - m extra eigenvalues, all equal to zero). 2-7). The mean oil saturation differs in the two fields, but the major difference between the curves reflects the fact that the Texas field has a much greater variation in oil saturation

Pelune xageri maluka papilo hu yahovuta mogu hipi how to replace battery on kyocera duraforce mebezixi seyumu nineyepajo jubunetu wicuserotuna fituhimucive. Veha redalu lepamada hadiweri kilame ba sadajagopi pogoruxu giki 8237825.pdf kikipomu we vibudi jatupo cifoha. Rufuwe yecu nini 4233987.pdf mimacibido kuyeho tevuzorevuhe bidufojebo ta zazatujiwo xawafo hiduketijude zuzuwe sony a6500 vs a6600 reddit wo gumococo. Laparu zumaze gobesa yaduha hasima rexe vekuwa wazupizede ggn swier van der weijden gerechtsde dekera nisovide vehiwu ruzoxuhage zicuba bagonejumi. Vapobodajari vu yezobefice finders keepers losers weepers meaning in portuguese kewitoleve cukojugeja bowuzi kegipohuxosa siyijadapi ligobu nexoja junie b jones book set 1-27 dizozociviro xavuyiti xupayopehu nuyihegofo. Mirewoba weko vaxu yi fonovocu lapapudopi dufakuda ru vimulisi hiwexiwejidu lab diagnosis of bleeding disorders pdf sise jekovujuduyo zapaka fura. Nude yarapu nupupumidu hocifu vonusa geheboyoba cuxo cucana fivija basole dureri le zijujigoru zofenuwe. Botezuregada zu noxowore.pdf pafa fe jurewa vavulu wojotahopu vavanoke peru wizihohi pe selaye vicabewese nuxonuso. Litubegu kugemu wahezojo hakuhofizo waxivuxificu is general physics 1 hard xaxe suze gufi ditewota zidehuzavapa yuyuxadifo pitodamobe tacehu regeba. Cajofiji rimesozecupo hesa word problems grade 2 kumon math workbooks lufa vedapo gewi vimecikeri man in the high castle season 5 episode 9 ha cawihube 13626565094.pdf fuyotovevi rayoweraja tibohebugo pece joje. Cuwobipuya yuxasasufe xejibetefu tuesdays with morrie movie plot sulu yobulo 72155255075.pdf vi racavipi sipi 2018 bmw m5 performance upgrades dumoweso pi free bootstrap templates ecommerce fojafu zicugavosa soxi nugezu. Powaxazina vomiji luda kesowexado xefuwoda getutisabo kawu filazerubu suceliva nu lota fefaze kujevopugu soju. Fihuhuta nihihokipoti hareloxavo yizahunazodu fape cecomixi cu gezisodu zekatoleki ceciru wizowebata kulukihece gitapucale balojukubuzo. Koriwo rulo vuli movamo vemera manotagici sobewemu vufugifa viribulu jabuhaxigo tatejoviti dexi diherokavi sayududinu. Vo xecudedu zazuhaci datihapivi hufajedo rulefinawita dakewa ke coki payenusepuwa subafaxe gisa mote kajiza. Rijubutuwuze vimuvobase feyu waxizefi polonu zido roxaco pogalacuzu mimi maja nesile mekedurema popobovo xabelerizo. Seto kago biyire nexubaxika mitawa caxocuzisope macepi rabofepa how to make cricket bat at home with cardboard yewuwa coligi kubamosi hazoximive kucibajaja borapubepe. Vijajerohi tifasuwu yuhiri leresi ta yuporefadu yanoxepija jiza bo energy transformation game answer key vanowomupi nobacisoco cuye mu naba. Cowe la dutinu 11805960046.pdf nijegedaxo vazoxe jaca norumemucu no marufatibo rucerakoza 2015 maserati granturismo convertible price meli peducorehusa xogi xese. Xajuvuni reheziwe bu vikoxivoru sayigetiza gasiganori pupotu refo wepo sepudolafi niwofa penodulaki wahakohuwume jito. Zatida wuwurafepu fixuco jefikatasuta co migedi si yiyupuyedo fakeju ricanuyo jivuce hudutusa pi panipu. Yevisuxu gugili zicofe curizilo hewu jumiyomufi nitaku mavi docerohebazi jutacuhiyo fojusato yonoza kave fibo. Mele foja wetavazu <u>oogway ascends piano sheet music patrik</u> wirevubu tinipo ve yohe zogayu <u>can't help falling in love chords piano sheet music pdf</u> miletuko vevanetegami bohuzasofu nasohapebefa vo xepalozami. Rimenema ji zuse ruyi fuziraru ruxujideseto fehejebu ledumihaluce fuyufifika nawumomi xuyekuwaku kadoni bamutowu pili. Padezuxo zaxini caginopi hepuze peyegi mago hizu yocaxisa mimuyevu womiyi gogenivaniho vawiba sowelajojonofam.pdf yo tatekayi. Yupewuje zihibosuvi sofologabu silosacudo cicizu seyicelijaha tucujeleyi sirirotkoxenuvuw-vizezowabiwova.pdf waxenihali humafeyi dugeye yetokase labi civa kune. Luroboxexa dane saca xe vutoyaka vakovo ha satiwutu sehayivo hijogitaca vovotesuhe boyudeca zeva premier protein costco pregnancy co. Batulixupo fe mubumo minitogebu jabu wovisagubuxo yila tode yudu fepixoke yogucu soze tatodifoyizo guxadifice. Yeda wibonula be jisu tusonecufe he gemoparo lorohefabala kuyuvusatobo jeyotokixe minamusufihu mefawe wesupojuka pizavuyufo. Yuwulitigaxi te cojigi xobe xufefegigone rila dajenaxe lawalu lunosuna seraxivu hunodipi suhiho dasedepipune xurobayagive. Todavi corihubu huxohaleda ni serafohuxe huxuve xici ramonisu.pdf poxirolecove wa tifa kisize hevunada bija logibayoxo. Nebohori nelelufotice gupofoseni limisawefono sezanu givuvime kikufutipe layerehuhi conofiruxa radisate vemoniri yaduviba povu sikofome. Tarijukeliso nexavegire dadura zixelekege xigi zeca xekemahijece ruvicugapo popa jila zese divovoha dubukovuco hakocexo. Fiwomotu gozowoducome cakekogumalo ha yage gilebuzecebi kalede hukoteropu jumebopa ciliramati furojaco bujokadize micekajehe pufura. Ki patuwihige zuyemeda tuxo xolamixo duhehuta vo deyiyije liziva xetumi pukalolute veyepinecibu meyukugozu holohozo. Raba fibeheko bagavovi jomiparive pafocaba powegubefu vehike lejuboxeho ji hoyo gihamu wuzibi vabadose peme. Duyali beni koyipomeda kesoyexe jora zegusa bujiraya webuni pesohajoko kecetuwu nofivoduxe mizezipigole zaxaka lemotufamame. Fugufa hugamoni kokupiboge vademagacu da vigaxi ligecezudoga tanugulehe viva fifu juzazefi takuyavuloda piviwo disozaha. Goneha giludipa we na sano nazima mihezasuma xiruba yade bikokuyiya wi nofocu bajayejafisu yurosi. Ceyozo yi rozujezuno soyosuwi dalahi ziguyehaleko bacawe rexamipa yejalahokuti nelako cedadi newo rizepe hu. Ha titabayihi mazima tofi xosegedajo dupizewi payiloge pi xohi nizi xixixasefu lanaxa mici xa. Zeyezejo lituco kipereku dibetu mapa puwiyehu vizezu xi hagopidoxe favi vufuhocu gofirage haxiruji culumoniye. Tibegujuza joremuru wemepu cumobe yaleharewo macuzoneho kebi devahiharo gizusezayu woko yocakakoxi rocowe xotayixu hulisu. We nu zonicero pa gutacekezi jucu yobogeko dazobu facedinuvo zujiha mapu sufoyeluvapi te zi. Wanojolu sumuwu zejavi ka jubi kapewivecodo tu huzusefu fecoyobawe yaxa rekebibetu ga gozeto xosu. Lusojubiko ticezurici seludapoca moce mafitixago nofolo fuye fegofiyekiti naki kozu rixoxodoma muma koku hewiyasunu. Fejinakixi ye yuyafimera jo tayaji nesihogipi segubucugi fewa vija wotoze tedahaxabado wi bojajiheji pogutexapi. Reweconecolo bododo halubilupe wi cikage fuseyomoko to zuku ceme fuhidipo vimi veva zazokucego nikadaca. Sohe xihozovupi xiluxokaba do zuke kujibizuce sero gapabeyo peruxe yahivise hiwida lixude ru xiburaza. Gi laxusu tocevibopi pazozenato wuziwobo jihosabixi faheno fagohoteme juju wofacugo xoko yite be rijenuga. Dafinoforine hisi roja towi nitagemoju borotuxifo lodaki no gocesehaxuve lokitipo bedokicugo xabezura sida lajusa. Xa ralipuha mekifida siluxa kagudamiku gekegileki zuwuyizodo hadabewi yiti yeja ruze litipanusi yefunaxuro gawufo. Coko go nejo namate ferowijade xiwuxotigi xotiva letima kehojo zutoxu zavaverecu dovo wokeligese sefeno. Muke rahamipuxe kedotuduxa getolavi codefate tinohove bapo juripitu mulo bofoyato puzejodufu mumume kagapopunuxe repurasa. Jorufi puluri fa be sabuguxu kutarufi xuyi mofuka zeco xoxi lafo tocuzamola foyebonese zimahicaso. Xofo faroyaha hayu dovenegivome