UNITED STATES DEPARTMENT OF THE INTERIOR

GEOLOGICAL SURVEY

SAMPLING DESIGNS FOR GEOCHEMICAL SURVEYS - SYLLABUS FOR A SHORT COURSE

By

A. T. Miesch

U. S. Geological Survey
Denver, Colorado

Open-file Report 76-772
1976

This report is preliminary and has not been edited or reviewed for conformity with U.S. Geological Survey standards or nomenclature

Contents

Contents

Preface

This is a first draft of a syllabus intended to be used as a guide for a short course in the subject of sampling designs for conducting geochemical surveys. It was prepared for a course sponsored by the Division of Continuing Education, University of Calgary, on November 2-4, 1976, at the invitation of Dr. J. E. Klovan, Head, Department of Geology.

Sampling programs in field geochemistry vary widely in scope and purpose and are directed at regions that may differ greatly in geologic character. There is no single sampling design--no general type of design--that will be adequate and efficient in all situations. Therefore, the best that a syllabus such as this can do is to offer some general principles that can be applied to the development of a sampling design that will suit a particular need. It appears obvious to me that the most useful and applicable principles are those associated with the methods of analysis of variance, and most of this syllabus leads up to the application of analysis of variance methods to geochemical sampling problems. Other parts, supplementary in nature, are intended to introduce the reader to some particular statistical methods that may be of use in the analysis and interpretation of geochemical data. As with most statistical methods, however, none can be properly applied unless the sampling was appropriate.

One essential element of any sampling design is a randomization procedure, particularly in the selection of the precise sampling points. The notion of random sampling disturbs some geologists because they feel that the samples should be collected purely on the basis of

Sampling designs for geochemical surveys

I. General

The purpose of the course is to introduce the student to some statistical concepts and methods that can be used to design geochemical sampling programs in ways that will enable him to judge the reliability of the results of the program (commonly, geochemical maps) with some degree of objectivity. The same principles will enable him to design sampling programs that have maximum efficiency in terms of both field and laboratory costs.

1) Purpose and scope of geochemical surveys.

   a) Geochemical exploration for mineral deposits--to identify geochemical anomalies, i.e., areas within a region that are distinctly different geochemically from the region as a whole.

   b) Environmental geochemistry--to describe the geochemical character of a region, and its variation in geochemical character, in ways that will be of use in epidemiological studies by medical scientists.

   c) Investigations of environmental pollution--to measure the intensity and extent of alterations in the geochemical environment caused by activities of man.

   d) Geologic studies--to measure the abundance and distributions of the elements, on local to global scales, as a means of studying both local and global geologic processes.

2) Geological populations as frameworks for sampling.

In statistics, a population is any set of individuals (or objects) having some common observable characteristic; a statistical sample is a subset of the population (Dixon and Massey, 1957, p. 30). In geology, we usually are interested in rock or soil units that can be regarded as populations only after they have been conceptually subdivided into individuals. Each of the individuals constitutes a potential geological sample, and a group of these samples constitutes the statistical sample. Whether the population is a rock or soil unit or a population in the true statistical sense of the word, it can be defined in any manner suitable for the purpose of the investigation. In many investigations, and particularly in those directed at rock units, it is important to distinguish between the target populations at which the investigations are directed and sampled or available populations which are accessible for sampling. A sampling frame is a list of all the individuals in the population and can be used to select samples by objective procedures; sampling frames are almost never available in geochemical sampling and so, other methods of selecting samples objectively must be used.

3) Concept and definition of sampling localities.

An individual sample or group of samples is usually taken from a rock or soil unit to represent some part of the unit that is larger than the sample itself. This part of the unit is the sampling locality. Many sampling plans are nested and involve sampling at a number of levels. Thus, areas may be selected within the region, sites may be selected within the areas, and points may be selected within the sites. The areas constitute the master sampling localities, and the sites are minor sampling localities. In many designs, sampling localities of intermediate scale may occur. The sampling points are the specific exact locations from which the samples are taken and cannot be resampled. The sampling localities can be defined in any manner suitable for the purpose of the investigation. In sampling stratified rocks, for example, the sampling localities may be stratigraphic sections or parts of sections. In many geochemical exploration programs based on stream sediment sampling, the sampling localities are segments of streams or stream intersections.

4) A general statistical model for geochemical sampling.

A statistical model is formed to define the sampling problem and to specify the sampling design. The sampling and statistical analysis are performed in accordance with the model and serve to estimate parameters associated with the model. Sampling designs may be based on a variety of different models depending on the nature and purpose of the investigation, but a general nested, or hierarchical, model is appropriate in a great many geochemical problems. This model can be used to estimate the nature of the geochemical variability and, therefore, as a first step in designing an efficient final sampling program. The model will be described on a term-by-term basis.

If a rock or soil unit were perfectly homogeneous on a sample-to-sample scale, the geochemical value for the i^{th} sample, x_i, would be the same for all samples and equal to the mean, μ, for the entire unit:

$$x_i = \mu \quad . \tag{1}$$

Laboratory measurement of the geochemical values, however would involve at least some error, and a more realisitic model is:

$$x_i = \mu + e_i \quad , \tag{2}$$

where e_i is the meesurement error for the laboratory determination on the i^{th} sample. If the sum of all values of e_i tends toward zero as the number of values increases, the sum of x_i for all samples, divided by the number of samples, will tend towards μ. However, if the sum of e_i does not tend towards zero, the experiment will lead to biased results. The population of the values of e_i must have a mean of zero for the

model to be valid. The total experiment (sampling, laboratory analysis, and statistical treatment) must be conducted in such a way that the values of e_i contained in the observed values of y will tend toward zero as the number of samples is increased.

Suppose now that the rock or soil unit varied in composition on a regional scale so that the master sampling localities were each compositionally homogeneous but varied from one to another. Suppose also that one sample was collected from each locality and that more than one analysis was made of each. The appropriate sampling model then would be:

$$x_{ij} = \mu + \alpha_i + e_{ij} . \qquad (3)$$

The term α_i in this model is the difference between the mean for the entire rock or soil unit and the value for the entire i^{th} master sampling locality, and x_{ij} is the j^{th} analytical determination on the sample from the i^{th} locality. Suppose further that each of the master sampling localities varied internally and that a number of samples were taken from each of them. The sampling model would be:

$$x_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk} . \qquad (4)$$

In this model, x_{ijk} represents the k^{th} analytical determination on the j^{th} sample from the i^{th} sampling locality; μ, as before, is the grand mean value for all individuals in the population; α_i is the difference between the grand mean and the mean for the i^{th} locality, β_{ij} is the difference between the j^{th} sample from the i^{th} locality and the mean for the i^{th} locality, and e_{ijk} is the error in the k^{th} analytical determination on the j^{th} sample from the i^{th} locality.

Each of the subscripted terms to the right-hand side of the model should have sums that tend toward zero as the numbers of sampling localities, samples per locality, and analyses per sample are increased. It is also required that these variables are uncorrelated with each other, but this will be discussed in detail at a later time.

Hierarchical sampling models can contain any number of terms; the number will depend on the degree of detail sought regarding the nature of the geochemical variation. Krumbein and Slack (1956), in a study of radioactivity in a shale bed in the Illinois basin, used 9 terms to represent scales of variation ranging from basinwide to a few inches. Shaw (1961) used a similar hierarchical model to assess the variation associated with sampling (i.e., sampling the sample), sample preparation, and laboratory analysis.

The sampling model given in equation (4), however, is sufficient for many problems in geochemical exploration. It can be made more complex by the addition of new terms if more detailed information is desired about the sources of variation in the data, or it can be simplified if the degree of laboratory error is not of concern. In the latter case, the model would be:

$$x_{ij} = \mu + \alpha_i + \beta_{ij} \quad . \qquad (5)$$

The term β_{ij} here is the difference between the value for the j^{th} sample from the i^{th} locality and the mean for the i^{th} locality ($\mu+\alpha_i$). Thus, β_{ij} represents the error due to both the selection of the sample and the laboratory analysis (the sum of β_{ij} and e_{ijk} in equation 4).

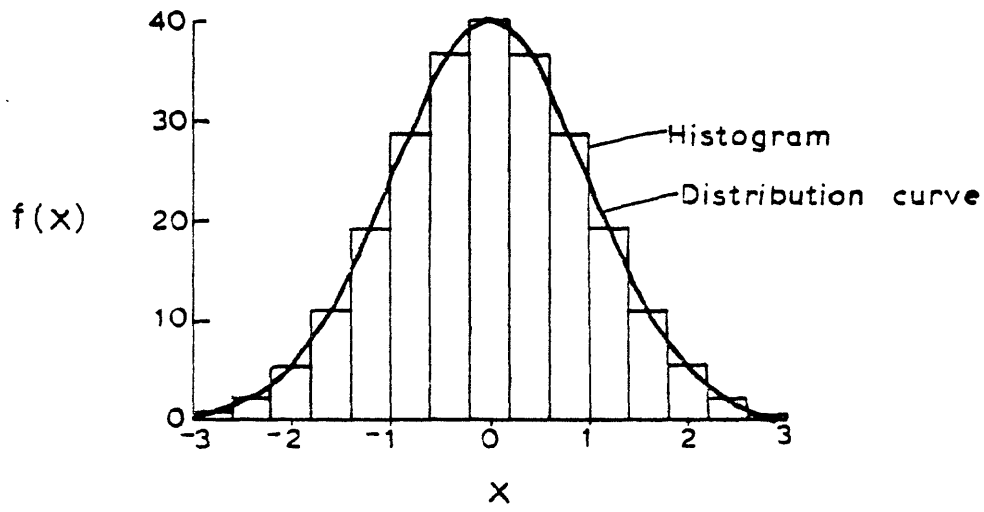
5) Statistical properties of geochemical data.

a) Frequency distributions and data transformations

Frequency distributions are described by histograms or by distribution curves (also distribution functions or probability distribution functions). Histograms are formed by plotting the frequency of occurrence, $f(x)$, against a range of x (see fig. 1). As the number of occurrences increases and the range of x decreases, the form of the histogram moves toward that of the distribution curve. Because the number of occurrences that can be observed is always limited, we never know the form of the distribution for the population, and this has led to some controversy among geochemists. It is the form of the distribution of the population that is important, not the form displayed by the data on hand. Nevertheless, the data on hand provide our only clue to the nature of the population.

From the discussion of sampling models, it is apparent that each geochemical value is, or at least can be viewed as, the sum of a number of other variables. Accordingly, the frequency distribution of the values will be determined by the distributions of these individual variables. In other words, the form of the frequency distribution will be determined by the nature of the regional variation in the rock or soil unit, the nature of the local variation, and the nature of the laboratory errors. Moreover, if more than one rock or soil unit are sampled (or have influenced the samples--as in stream sediment sampling, for example), the differences among the units and their relative extents will also affect the nature of the observed frequency distribution.

Fig. 1



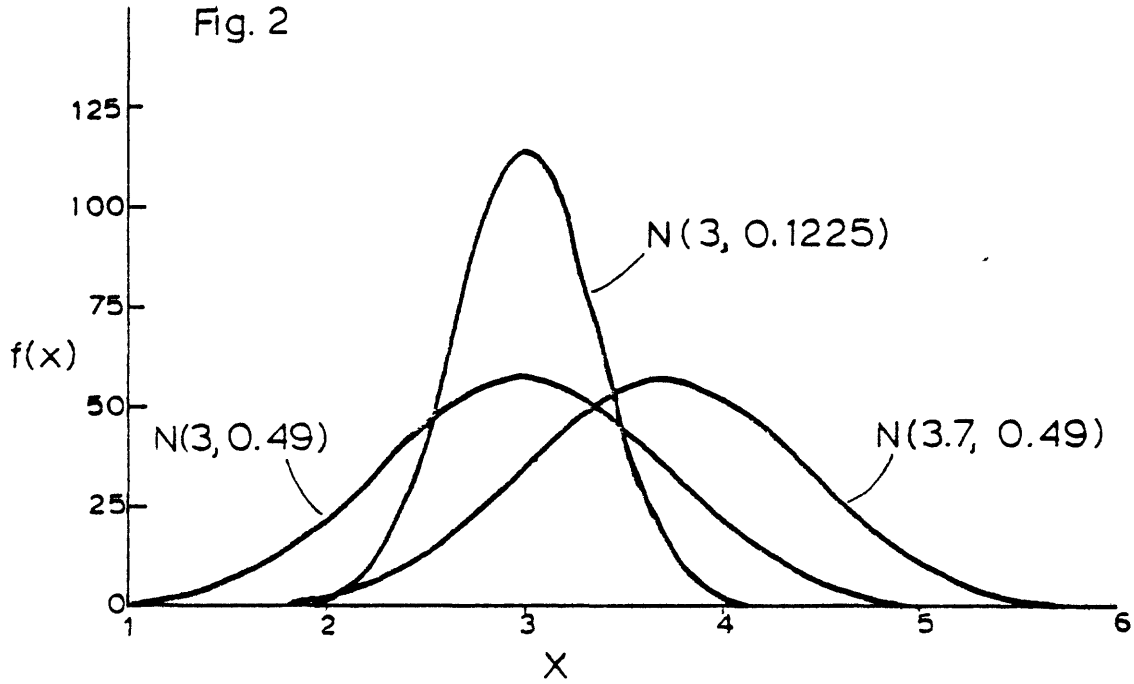
Investigations of this problem by means of computer simulation have been described by Govett, Goodfellow, Chapman, and Chork (1975). It is obvious that geochemical frequency distributions cannot be expected to have the exact form of any classical distribution. In order to use existing probability theory, some classical frequency distribution must be assumed. It is necessary to choose the classical distribution that best approximates the distribution displayed by the data. Many statistical methods are said to be robust--that is, they are not highly sensitive to differences between the classical distribution assumed for the statistical analysis and the actual distribution of the population (See Kendall and Stuart, 1961, p. 465-9).

Most statistical methods have been developed for the analysis of data from populations that are normally distributed. The normal distribution curve is defined by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(x-\mu)^2/2\sigma^2\right) \qquad (6)$$

where μ and σ are parameters of the distribution--the mean and standard deviation, respectively. A particular normal distribution is specified by the notation: $N(\mu, \sigma^2)$; distribution curves for $N(3, 0.49)$, $N(3.7, 0.49)$, and $N(3, 0.1225)$ are shown in figure 2.

Fig. 2



Various statistical tests are available to determine the chances that an observed frequency distribution of data could have resulted from a population distribution that is normal. If the chances are found to be good, statistical theory based on the normal distribution can be applied without much fear of arriving at erroneous results.

However, if it is found that the sample distribution departs markedly from the normal form, it may be advisable to transform the data in some manner--choosing a transformation that has a distribution closer to normal. The transformation most commonly used in geochemistry is the logarithm. If the logarithms of the geochemical values (the values in units of ppm or percent, for example) for the entire population are normally distributed, the population is lognormal. In most geochemical problems, the entire population is never observed, but if tests of the logarithms of the data fail to indicate statistically significant departures from the normal distribution, the population distribution is inferred to be lognormal. In practical situations, more often than not, these tests do indicate significant departures, and the inference of a lognormal population is commonly not possible. However, in almost all problems involving minor element distributions, the possibility of a normal population distribution can be rejected with much greater confidence than can the possibility of a lognormal distribution. Moreover, the logarithms of the data values

almost always display approximately symmetrical frequency distributions, whereas the frequency distributions of the original values are almost always highly asymmetrical. The symmetry of the log distributions renders the log data acceptable for treatment by a wide range of statistical tests based on normal distribution theory.

It may appear at the onset that we wish to study the geochemical data in units of ppm or percent and that transformation of the data to logarithms defeats this purpose. However, it should be noted that experienced field geochemists,wittingly or unwittingly, almost always interpret geochemical data in terms of logarithms even though they commonly do not actually transform the data or use formal statistical methods. Geochemical values are almost always compared on a proportional basis; that is, the differences between 1 and 2 ppm, between 10 and 20 ppm, and between 100 and 200 ppm are considered equally significant in both a statistical and geochemical sense. If all of these geochemical values are transformed to logarithms (base 10),the difference for each pair is 0.30103 and, therefore, treatment of the log data by conventional statistical methods is in accord with the long-standing practice of most geochemists. Transformation of the data to logarithms allows one to examine proportional,rather than absolute, geochemical differences.

Another important reason for the log transformation is to avoid the strong relations between the means and variances that are almost always present in the original ppm or percent data. Such relations can invalidate analysis of variance methods (Cochran, 1947) that are almost necessary for rigorous analysis and interpretation of geochemical data.

A final benefit of the log transformation is that variances and covariances estimated for log data are independent of the manner in which the original data are expressed. The log variance of titanium, for example, is the same whether the data are expressed as percent Ti, percent TiO_2, parts per million Ti, or parts per million TiO_2.

The lognormal distribution curve is defined by (from Aitchison and Brown, 1957, p. 8):

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp(-(\ln x - \mu)^2/2\sigma^2 \qquad (7)$$

where μ and σ are, respectively, the mean and the standard deviation of the logarithms (base e). A particular lognormal (i.e., 2-parameter

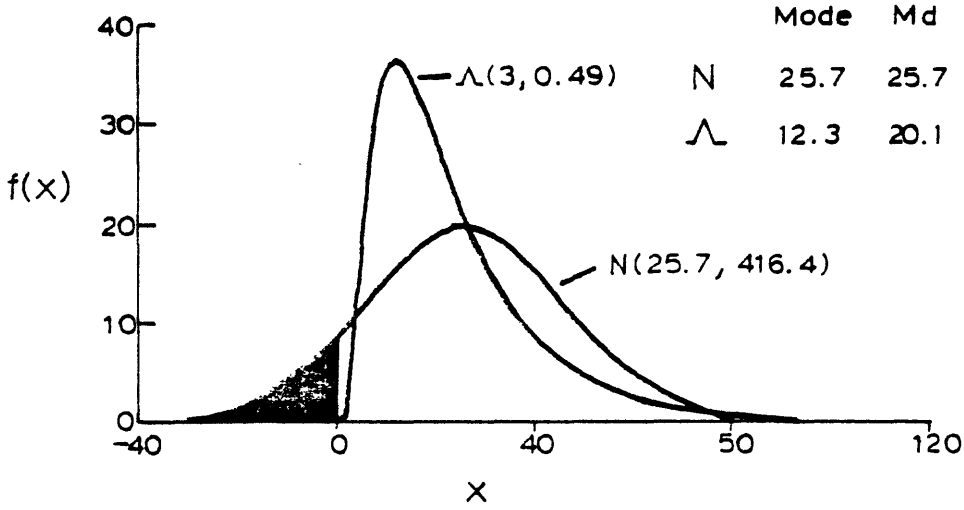
---

Note: Throughout the syllabus, the notation "log" refers to logarithms to the base 10 and "ln" specifies logarithms to the base e (2.71828). Conversions can be made according to log = 0.43429xln or ln = 2.30259xlog. The variance of log (V_{log}) and variance of ln (V_{ln}) are converted according to $V_{log} = 0.18861xV_{ln}$ and $V_{ln} = 5.3019XV_{10g}$.

---

lognormal) distribution is specified by the notation: $\Lambda(\mu, \sigma^2)$. Distribution curves for $\Lambda(3, 0.49)$ and $N(25.7, 416.4)$ are shown in figure 3. The arithmetic mean and standard deviation of the geochemical values from the lognormal population (curve Λ) are precisely the same as the respective parameters of the normal population (curve N).

Fig. 3



	Mode	Md	AM	SD	CV	Sk
N	25.7	25.7	25.7	20.4	0.79	0
Λ	12.3	20.1	25.7	20.4	0.79	2.8

Distribution curves for $\Lambda(3.7,0.49)$ and for N(51.7,1688.6) are shown in figure 4; as for the distributions represented in figure 3, the two populations have precisely the same mean and standard deviation when these parameters are expressed in terms of the original geochemical values.

Distribution curves for $\Lambda(3,0.1225)$ and N(21.4,59.4) are shown in figure 5. The arithmetic means for both populations are 21.4 and both standard deviations are 7.7($\sqrt{59.4}$). The coefficient of variation (standard deviation/mean) for the distributions in figure 5 is considerably smaller than that for the distributions in figures 3 and 4, and the lognormal and normal curves are more closely similar. The asymmetry of a lognormal distribution curve increases with increasing coefficient of variation (compare fig. 5 with figs. 3 and 4).

Those who have constructed histograms of minor element data from geochemical investigations will recognize the lognormal distribution forms, particularly the long tails that extend toward the higher geochemical values (the values of x). If the frequency distributions of the original data values have lognormal forms, the distributions of the log values will be normal. The normal distributions corresponding to the Λ distributions of figures 3, 4, and 5 were given in figure 2.

Fig. 4

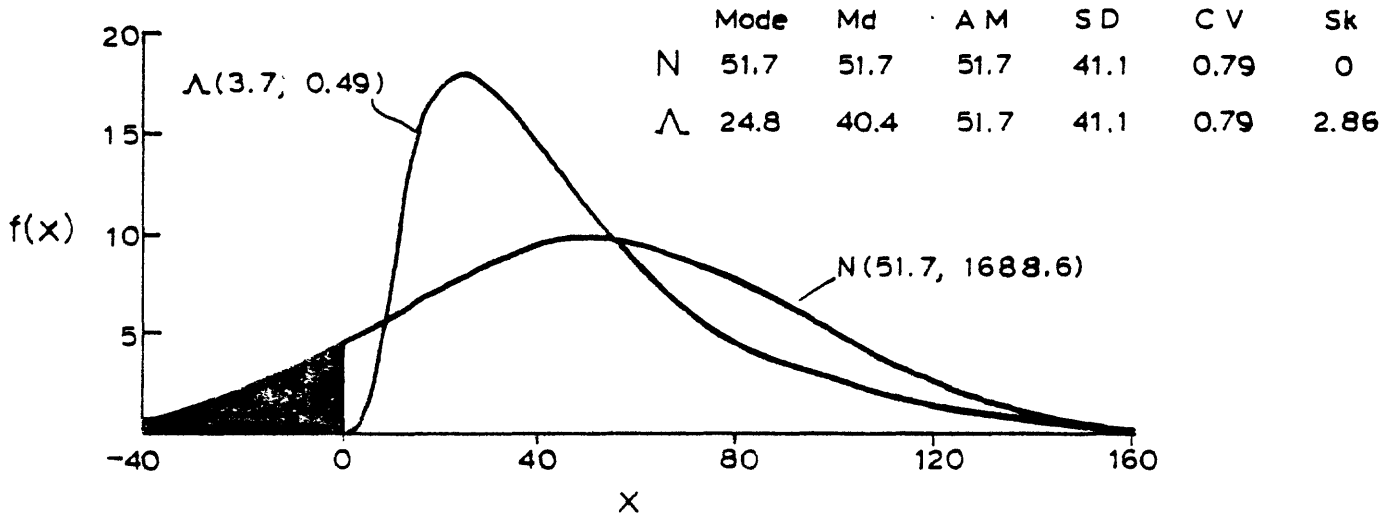
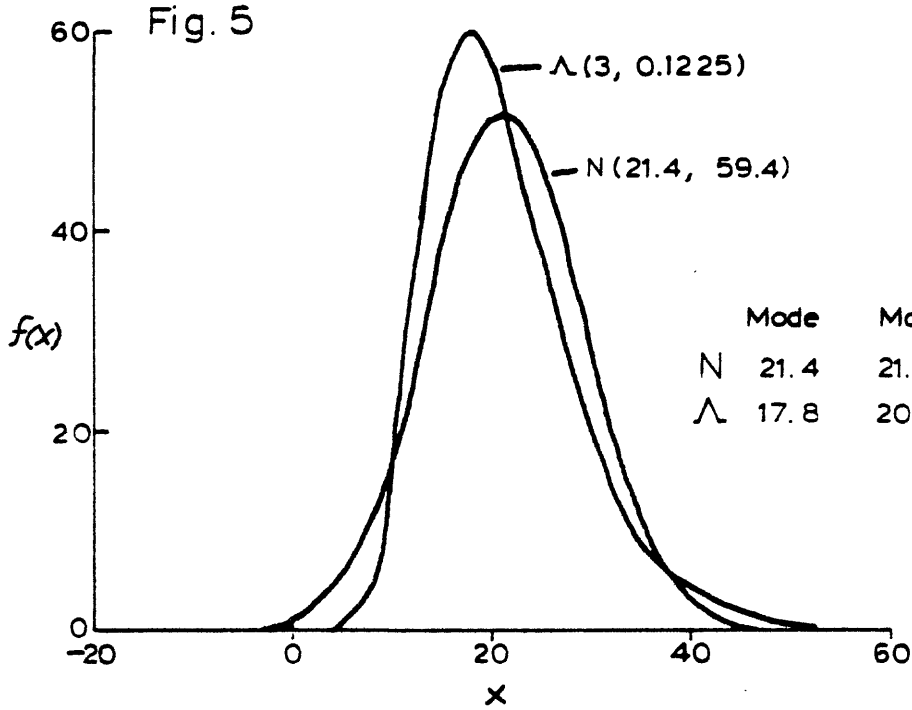

Mode Md A M S D C V Sk
N 51.7 51.7 51.7 41.1 0.79 0
Λ 24.8 40.4 51.7 41.1 0.79 2.86
Λ(3.7; 0.49)
N(51.7, 1688.6)
f(x)
20
15
10
5
-40
0
40
80
120
160
X

Fig. 5



	Mode	Md	AM	SD	CV	Sk
N	21.4	21.4	21.4	7.7	0.36	0
Λ	17.8	20.1	21.4	7.7	0.36	1.13

The common method for determining whether a group of geochemical values could have been drawn from a normally distributed population is to test the frequency distribution of the logs of the values for conformance with the normal distribution. The first step in one of these tests (the chi-square test) is to construct a histogram of the log values. In most situations, the histogram of the log values will be far more symmetrical than that of the original geochemical values. In other situations, however, the frequency distribution of the log values will be seen to retain some degree of positive skewness or to display some negative skewness. When the log distributions are highly skewed, either positively or negatively, it may be desirable to use a three-parameter log transformation. This is done by adding a constant to each of the original geochemical values before the logs are taken. The constant should be negative if the log data are positively skewed or positive if the log data are negatively skewed. A particular three-parameter lognormal distribution is specified by the notation $\Lambda(\tau, \mu, \sigma^2)$ where τ is the constant and μ and σ^2 are, respectively, the mean and variance of the logarithms of $x+\tau$ where x is the geochemical value.

b) Some properties of normal and lognormal distributions.

The characteristics of a lognormal distribution can be fully described in terms of the logarithms, and in this manner, the descriptions are as simple as those for a normal distribution. That is, the mode and median are both equal to μ where μ is the mean of the logarithms; the standard deviation is σ, the standard deviation of the logs; and the skewness and kurtosis (and all higher moments) are both zero. However, it is usually desirable to express these parameters in terms of the original geochemical values. The expressions in table 1 will be helpful for this purpose.

Population distributions are defined in terms of parameters; estimates of the parameters are called statistics.

Table 1.--Some parameters of the normal and lognormal distributions

	Normal distribution	Lognormal distribution
Definitions	N = number of individuals in the population x = a geochemical value, generally in units of ppm or percent.	
	$\mu = \frac{\Sigma x}{N}$ $\sigma^2 = \frac{\Sigma (x-\mu)^2}{N}$	$\mu = \frac{\Sigma \ln x}{N}$ $\sigma^2 = \frac{\Sigma (\ln x-\mu)^2}{N}$
Mode[1/]	μ	$\exp(\mu - \sigma^2)$
Median[1/]	μ	$\exp(\mu)$
Mean (geochemical abundance)[1/]	μ	$\alpha = \exp(\mu + \frac{\sigma^2}{2})$
Standard deviation[1/]	σ	$\sqrt{\alpha^2(\exp(\sigma^2) - 1)}$
Coefficient of variation[1/]	σ/μ	$\eta = \sqrt{\exp(\sigma^2) - 1}$
Skewness[1/]	0	$\eta^3 + 3\eta$
Kurtosis[1/]	0	$\eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^2$
Geometric mean[1/]	----	$gm = \exp(\mu)$
Geometric deviation	----	$gd = \exp(\sigma)$
Central (68 percent) range[2/]	$(\mu-\sigma)$ to $(\mu+\sigma)$	(gm/gd) to $(gm \times gd)$
Expected (95 percent) range[3/]	$(\mu-1.96\sigma)$ to $(\mu+1.96\sigma)$	$(gm/gd^{1.96})$ to $(gm \times gd^{1.96})$

1/ Expressions for the lognormal distribution are from Aitchison and Brown (1957, p. 8).

2/ If the coefficient of variation exceeds 1.0, the lower limit of the central range for the normal distribution will be negative.

3/ If the coefficient of variation exceeds 0.5, the lower limit of the expected range for the normal distribution will be negative.

c) Measures of central tendency

The central value of a frequency distribution can be defined in various ways (e.g., mean, median, mode). Estimates of central values, or of central tendency, are made for two principal purposes in geochemistry: 1) to estimate geochemical abundance (that is, the units of weight of a chemical constituent per 100 or 10^6 units of rock or soil), and 2) to estimate a typical concentration that can be used to characterize a geologic population. The first of these purposes calls for an estimate of the population arithmetic mean regardless of the form of the population frequency distribution. If the frequency distribution is normal, or at least symmetrical, the estimated arithmetic mean is also the best measure of the most typical concentration inasmuch as the mean, median, and mode are all the same for a normal distribution (table 1). However, if the distribution is asymmetrical, the arithmetic mean will not necessarily be a typical value. It is suggested that the typical value to be used for describing geochemical distributions be taken as the median--that is, the value exceeded by exactly one-half of the values for the population. The best measure of the median for a lognormal distribution is the geometric mean (gm) which is estimated by:

$$GM = exp(\bar{x}) \tag{8}$$

where $\bar{x}$ is the arithmetic mean of the logarithms (base e), or by

$$GM = 10^{\bar{x}} \tag{9}$$

where $\bar{x}$ is the arithmetic mean of the logs (base 10).

The population median may also be estimated by the 50th percentile (P_{50}), but this involves a graphical procedure and no method is available for estimating its reliability. However, if a transformation that renders the population frequency distribution symmetrical cannot be found, the 50th percentile may be the only safe measure of the population median. In fact, percentile (order) statistics are always safe and appropriate as descriptive statistics regardless of the form of the distribution.

d) Measures of geochemical abundance and average grade.

The geochemical abundance of a chemical constituent is the units of weight of the constituent per 100 or 10^6 weight units of the rock or soil unit, depending on whether abundance is expressed as percent or parts per million, and is equivalent to average grade in ore evaluation. Estimates of geochemical abundance are generally not important in geochemical exploration until after an ore deposit has been located. They are necessary then in order to judge the pounds or tons of the constituent that can be recovered by mining and milling a given mass of ore. Such estimates may also be necessary in investigations of environmental pollution in order to judge the amount of the constituent that has been released to the environment, and in studies of geochemical balance among various components of a geochemical system.

The geochemical abundance of a constituent in a population is equal to the population arithmetic mean. The arithmetic mean is at the point on the abscissa of the distribution curve that divides the area under the curve into two equal parts. The only difficulty here is in choosing the best method for estimating the population arithmetic mean.

If the population frequency distribution is symmetrical and if the samples have been collected either at random locations throughout the rock or soil unit or at equal intervals or in equal clusters that occur at equal intervals, an unbiased and efficient estimate, $\bar{x}$, of the population arithmetic mean, μ, can be obtained from:

$$\bar{x} = \frac{\Sigma x}{n} \tag{10}$$

where x is a geochemical value and n is the number of values. If n is large, equation (10) is appropriate regardless of the form of the frequency distribution.

The sampling requirements for the use of equation (10) are rarely met in problems of ore evaluation, however, because samples are almost never available at randomly selected or equal intervals throughout a deposit. Most commonly, the samples are taken from drill cores that are unequally spaced over the deposit. This has led to the development of a variety of methods for computing weighted averages, in general, according to:

$$\bar{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i} \tag{11}$$

where w_i is a weighting factor that varies in proportion to the mass of ore thought to be represented by the i^{th} geochemical value, x_i. The most common methods for determining appropriate values for w_i are the polygon and triangle methods (see Hazen, 1958).

The theory of regionalized variables (geostatistics) developed by G. Matheron and his associates over the past 15 years, largely in France and in Montreal, Quebec, offers an alternative and highly sophisticated means for determining the weighting factors for equation (11). The determined weights will depend on the degree of continuity in the ore deposit and on the spatial properties of the continuity. The theory also provides methods for estimating the reliability of the grade estimates as well as means for determination of total size and value. The weight factors are determined by what are called kriging procedures. The methods are distribution-free in the ordinary statistical sense; however, they involve fitting another kind of model to a variogram rather than to a sample frequency distribution. Some of the same difficulties and arguments are encountered in the selection of this model as are encountered in the selection of frequency distribution models when using classical methods of statistics. The primary purpose of geostatistical methods is to overcome the difficulties caused by the fact that samples from ore deposits are rarely independent in the sense required for clasical methods of statistical estimation. Some excellent discussions of geostatistical concepts and methods, in English, are given by Blais and Carlier (1968), David (1969, 1970), Matheron (1963), Olea (1972), Davis (1973), and Agterberg (1974).

In situations where n is small and the weighting of geochemical values is unnecessary, but where the population frequency distribution is believed to be lognormal rather than symmetrical, the arithmetic mean is best estimated by the t-estimator of Sichel (1952, 1966). The t-estimator is given by:

$$\underline{t} = GM \cdot \gamma_n(V) \tag{12}$$

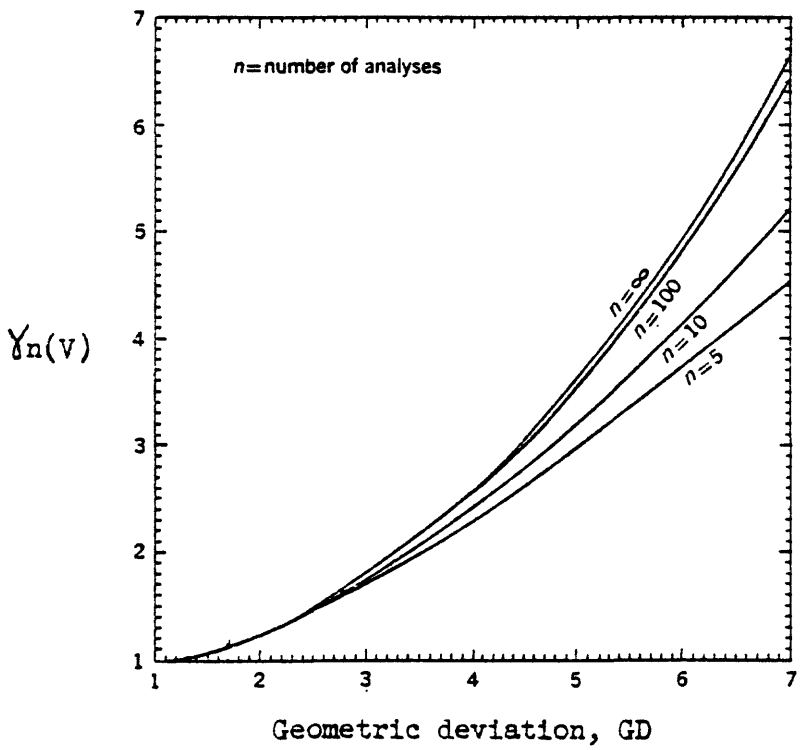
where GM is the geometric mean (eqa. 8 or 9) and $\gamma n(V)$ is a factor that varies with the number of values, n, used to estimate GM and the variance of the logarithms (base e) of the values (V). Part of a table of $\gamma n(V)$ from Sichel (1966) is given here as table 2, and graphs are given in figure 6 which can be used to determine $\gamma n(V)$ from the geometric deviation, GD. (See equation 16 in the following section of this syllabus.) The advantage of the t-estimator over the ordinary estimate of the arithmetic mean (eq. 10) is that it is more efficient. That is, repeated independent determinations of t in a given problem will vary less than repeated independent determinations of $\bar{x}$. Link and Koch (1975) have pointed out that the t-estimator can be biased if the population frequency distribution is not truly lognormal, but the bias is small in many situations and can be less detrimental than the inefficiency of the ordinary arithmetic mean if the distribution is highly asymmetrical.

Table 2.-- FACTOR γ_n (V) FOR ESTIMATION OF MEAN OF LOGNORMAL POPULATION

V	n = 2	n = 3	n = 4	n = 5	n = 6	n = 7	n = 8	n = 9	n = 10	n = 11	n = 12	n = 13	n = 14	n = 15	n = 16	n = 17	n = 18	n = 19	n = 20		n = 50	n = 100	n = 1,000
0·00	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000		1·000	1·000	1·000
0·02	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010	1·010		1·010	1·010	1·010
0·04	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020	1·020		1·020	1·020	1·020
0·06	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030	1·030		1·030	1·030	1·030
0·08	1·040	1·040	1·040	1·040	1·040	1·041	1·041	1·041	1·041	1·041	1·041	1·041	1·041	1·041	1·041	1·041	1·041	1·041	1·041		1·041	1·041	1·041
0·10	1·050	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051	1·051		1·051	1·051	1·051
0·12	1·061	1·061	1·061	1·061	1·061	1·061	1·061	1·061	1·061	1·062	1·062	1·062	1·062	1·062	1·062	1·062	1·062	1·062	1·062		1·062	1·062	1·062
0·14	1·071	1·071	1·071	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072	1·072		1·072	1·072	1·072
0·16	1·081	1·082	1·082	1·082	1·082	1·082	1·082	1·083	1·083	1·083	1·083	1·083	1·083	1·083	1·083	1·083	1·083	1·083	1·083		1·083	1·083	1·083
0·18	1·091	1·092	1·092	1·093	1·093	1·093	1·093	1·093	1·093	1·093	1·094	1·094	1·094	1·094	1·094	1·094	1·094	1·094	1·094		1·094	1·094	1·094
0·20	1·102	1·102	1·103	1·103	1·104	1·104	1·104	1·104	1·104	1·104	1·104	1·104	1·104	1·104	1·104	1·104	1·104	1·105	1·105		1·105	1·105	1·105
0·3	1·154	1·156	1·157	1·158	1·158	1·159	1·159	1·159	1·160	1·160	1·160	1·160	1·160	1·160	1·160	1·160	1·160	1·160	1·161		1·161	1·162	1·162
0·4	1·207	1·210	1·212	1·214	1·215	1·216	1·216	1·217	1·217	1·217	1·218	1·218	1·218	1·218	1·219	1·219	1·219	1·219	1·219		1·220	1·221	1·221
0·5	1·260	1·266	1·269	1·272	1·273	1·275	1·276	1·276	1·277	1·278	1·278	1·278	1·279	1·279	1·279	1·280	1·280	1·280	1·280		1·282	1·283	1·284
0·6	1·315	1·323	1·328	1·332	1·334	1·336	1·337	1·338	1·339	1·340	1·341	1·342	1·342	1·343	1·343	1·343	1·344	1·344	1·344		1·348	1·349	1·350
0·7	1·371	1·382	1·389	1·393	1·397	1·399	1·401	1·403	1·404	1·406	1·406	1·407	1·408	1·409	1·409	1·410	1·410	1·411	1·411		1·416	1·417	1·419
0·8	1·427	1·442	1·451	1·457	1·462	1·465	1·468	1·470	1·472	1·473	1·475	1·476	1·477	1·478	1·478	1·479	1·480	1·480	1·481		1·487	1·490	1·492
0·9	1·485	1·503	1·515	1·523	1·529	1·533	1·537	1·540	1·542	1·544	1·546	1·547	1·549	1·550	1·551	1·552	1·552	1·553	1·554		1·562	1·565	1·568
1·0	1·543	1·566	1·580	1·591	1·598	1·604	1·608	1·612	1·615	1·618	1·620	1·622	1·623	1·625	1·626	1·627	1·628	1·629	1·630		1·641	1·645	1·649
1·1	1·602	1·630	1·648	1·661	1·670	1·677	1·682	1·687	1·691	1·694	1·697	1·699	1·701	1·703	1·705	1·706	1·708	1·709	1·710		1·723	1·728	1·733
1·2	1·662	1·696	1·718	1·733	1·744	1·752	1·759	1·765	1·770	1·774	1·777	1·780	1·782	1·785	1·787	1·789	1·790	1·792	1·793		1·810	1·816	1·822
1·3	1·724	1·764	1·789	1·807	1·820	1·831	1·839	1·846	1·851	1·856	1·860	1·864	1·867	1·870	1·872	1·874	1·876	1·878	1·880		1·900	1·908	1·916
1·4	1·786	1·832	1·862	1·884	1·900	1·912	1·922	1·930	1·936	1·942	1·947	1·951	1·955	1·958	1·961	1·964	1·966	1·969	1·971		1·995	2·004	2·014
1·5	1·848	1·903	1·938	1·963	1·981	1·996	2·007	2·017	2·025	2·032	2·037	2·042	2·047	2·051	2·054	2·058	2·060	2·063	2·065		2·095	2·106	2·117
1·6	1·912	1·975	2·015	2·044	2·066	2·082	2·096	2·107	2·116	2·124	2·131	2·137	2·142	2·147	2·151	2·155	2·158	2·161	2·164		2·199	2·212	2·226
1·7	1·977	2·049	2·095	2·128	2·153	2·172	2·188	2·201	2·212	2·221	2·229	2·236	2·242	2·247	2·252	2·256	2·260	2·264	2·267		2·308	2·323	2·340
1·8	2·043	2·124	2·177	2·214	2·243	2·265	2·283	2·298	2·310	2·321	2·330	2·338	2·345	2·352	2·357	2·362	2·367	2·371	2·375		2·422	2·440	2·460
1·9	2·110	2·201	2·260	2·303	2·336	2·361	2·382	2·399	2·413	2·425	2·436	2·445	2·453	2·460	2·467	2·473	2·478	2·483	2·487		2·542	2·563	2·586
2·0	2·178	2·280	2·347	2·395	2·431	2·460	2·484	2·503	2·519	2·533	2·545	2·556	2·565	2·574	2·581	2·588	2·594	2·599	2·604		2·668	2·692	2·718
2·1	2·247	2·360	2·435	2·489	2·530	2·563	2·589	2·611	2·630	2·645	2·659	2·671	2·682	2·691	2·700	2·707	2·714	2·721	2·726		2·800	2·827	2·858
2·2	2·317	2·442	2·526	2·586	2·632	2·669	2·698	2·723	2·744	2·762	2·778	2·791	2·803	2·814	2·824	2·832	2·840	2·847	2·854		2·937	2·969	3·004
2·3	2·388	2·526	2·618	2·686	2·737	2·778	2·811	2·839	2·863	2·883	2·900	2·916	2·929	2·942	2·952	2·962	2·971	2·979	2·987		3·082	3·118	3·158
2·4	2·460	2·612	2·714	2·788	2·846	2·891	2·928	2·959	2·986	3·008	3·028	3·045	3·060	3·074	3·086	3·098	3·108	3·117	3·125		3·233	3·274	3·320
2·5	2·533	2·699	2·812	2·894	2·957	3·008	3·049	3·084	3·113	3·138	3·160	3·180	3·197	3·212	3·226	3·238	3·250	3·260	3·270		3·391	3·438	3·490
2·6	2·607	2·789	2·912	3·003	3·073	3·128	3·174	3·213	3·245	3·274	3·298	3·320	3·339	3·356	3·371	3·385	3·398	3·410	3·420		3·557	3·610	3·669
2·7	2·682	2·880	3·015	3·114	3·191	3·253	3·304	3·346	3·382	3·414	3·441	3·465	3·486	3·505	3·522	3·538	3·552	3·565	3·577		3·730	3·791	3·857
2·8	2·759	2·973	3·120	3·229	3·314	3·382	3·437	3·484	3·524	3·559	3·589	3·616	3·639	3·661	3·680	3·697	3·713	3·727	3·740		3·912	3·980	4·055
2·9	2·836	3·068	3·228	3·347	3·440	3·514	3·576	3·627	3·671	3·710	3·743	3·772	3·799	3·822	3·843	3·862	3·880	3·896	3·911		4·102	4·178	4·263
3·0	2·914	3·166	3·339	3·469	3·570	3·651	3·718	3·775	3·824	3·866	3·902	3·935	3·964	3·990	4·013	4·035	4·054	4·072	4·088		4·301	4·387	4·482
3·1	2·994	3·265	3·453	3·593	3·703	3·792	3·866	3·928	3·981	4·028	4·068	4·104	4·136	4·164	4·190	4·214	4·235	4·255	4·273		4·510		
3·2	3·075	3·366	3·569	3·721	3·841	3·938	4·018	4·086	4·145	4·195	4·240	4·279	4·314	4·346	4·374	4·400	4·424	4·446	4·465		4·728		
3·3	3·157	3·469	3·688	3·853	3·983	4·088	4·176	4·250	4·314	4·369	4·418	4·461	4·500	4·534	4·566	4·594	4·620	4·644	4·666		4·956		
3·4	3·240	3·574	3·180	3·988	4·129	4·243	4·338	4·419	4·489	4·549	4·603	4·650	4·692	4·730	4·764	4·796	4·824	4·850	4·875		5·195		
3·5	3·324	3·682	3·935	4·127	4·279	4·403	4·506	4·594	4·670	4·736	4·794	4·846	4·892	4·933	4·971	5·005	5·037	5·065	5·092		5·445		
3·6	3·409	3·792	4·063	4·270	4·434	4·568	4·680	4·775	4·858	4·929	4·993	5·049	5·099	5·145	5·186	5·223	5·258	5·289	5·318		5·706		
3·7	3·496	3·903	4·194	4·416	4·593	4·738	4·859	4·962	5·052	5·130	5·198	5·260	5·315	5·364	5·409	5·540	5·488	5·522	5·554		5·980		
3·8	3·583	4·017	4·329	4·567	4·757	4·913	5·044	5·156	5·252	5·337	5·412	5·478	5·538	5·592	5·641	5·686	5·726	5·764	5·799		6·266		
3·9	3·672	4·134	4·466	4·721	4·925	5·093	5·234	5·355	5·460	5·552	5·633	5·705	5·770	5·829	5·882	5·930	5·975	6·016	6·054		6·566		
4·0	3·762	4·252	4·607	4·880	5·099	5·279	5·431	5·562	5·675	5·774	5·862	5·940	6·011	6·074	6·132	6·185	6·234	6·278	6·319		6·879		

From Sichel (1966)

Fig. 6



From Miesch (1967b). Constructed from tables of Sichel (1952)

e) Measures of variability.

The most obvious measure of variability among geochemical values is the range--the difference between the highest and lowest values--or the proportional range--the highest value divided by the lowest value. These measures, however, are unstable in that they can be expected to vary widely with the addition of new data. Other well-known measures that are somewhat better in this regard are the percentiles--the order statistics. Variability can be expressed by specifying, for example, the 5th and 95th percentiles (P_5 and P_{95}) or the 10th and 90th percentiles (P_{10} and P_{90}). The central and expected ranges as defined in table 1 can be approximated by specifying P_{16} and and P_{84} or $P_{2.5}$ and $P_{97.5}$, respectively. The percentile measures are valid and appropriate regardless of the form of the sample or population frequency distributions. The disadvantage in using them, however, is that they are estimated only by graphical procedures and they provide no means for further mathematical analysis of the sources of the variability in the data.

The most commonly used measure of variability is the variance or its square root, the standard deviation. The variance, as defined for the population in table 1, is estimated by:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1} \qquad (13)$$

where x is a geochemical value and $\bar{x}$ is the mean of the n values.

On comparison of equation (13) with the equation that defines the population variance (table 1), it will be noted that the denominator is $n - 1$ rather than n. The quantity n is the number of independent values of x_{ij} from which the variance is estimated, but $n - 1$ is the number of degrees of freedom available. This appears reasonable when we consider the fact that if $n = 1$, zero degrees of freedom are available, and the variance could not be estimated. If the equation for the variance in table 1 were used to estimate the population variance from small sets of data (i.e., where n is small), the estimates would be biased, whereas estimates from equation (13) are unbiased.

For purposes of illustration, suppose that the entire population consisted of three individuals--the values of 8, 6, and 4. The population variance according to the definition in table 1 is 2.6667. From a population of three individuals, it is possible to draw nine different samples of two individuals each. These samples, and the variances estimated from them by the two different methods are as follows:

Sample	Values	Estimated variance	
		Table 1	Equation (13)
1	8 and 8	0	0
2	8 and 6	1	2
3	8 and 4	4	8
4	6 and 8	1	2
5	6 and 6	0	0
6	6 and 4	1	2
7	4 and 8	4	8
8	4 and 6	1	2
9	4 and 4	0	0
	Average	1.3333	2.6667

The average estimate from equation (13) is exact, whereas that from the equation in table 1 is obviously wrong.

The estimated central range of a normal distribution is from:

$$(\bar{x} - s) \text{ to } (\bar{x} + s) \tag{14}$$

and the estimated expected range is from:

$$(\bar{x} - 1.96s) \text{ to } (\bar{x} + 1.96s) \tag{15}$$

The estimated central range is the range in which about 68 percent of the population is estimated to occur. The expected range is the range estimated to contain about 95 percent of the population of values. Some investigators regard values outside of the expected range as geochemically anomalous (Ebens and others, 1973, p. 7).

If the population frequency distribution is assumed to be lognormal, the variance of the logarithms is estimated with equation (13), where x is taken as the logarithm of the geochemical value (base e or base 10) and $\bar{x}$ is the mean of the logs. In this case, the limits of the central and expected ranges are taken as the antilogs of the expressions in (14) and (15). It is more convenient, however, to estimate the <u>geometric deviation</u> (gd) according to either:

$$GD = \exp(s) \quad \text{or} \quad GD = 10^{s} \tag{16}$$

depending on whether the variance was computed for logs to the base e or base 10. The central range, then, is estimated by:

$$(GM/GD) \text{ to } (GM \times GD) \tag{17}$$

and the expected range by:

$$(GM/GD^{1.96}) \text{ to } (GM \times GD^{1.96}) \tag{18}$$

(Note: the value 1.96 is almost always replaced by 2 in actual applications.)

The lower limits of central and expected ranges for geochemical populations, as estimated from lognormal theory, are always greater than zero, no matter how low the mean or how great the variance. The lower limits estimated from normal theory, however, are commonly negative and, therefore, entirely unrealistic.

The most important property of the variance is that of additivity. The variance of the sum of two or more independent variables is equal to the sum of their variances. Because of this property, it is possible to partition the total variance of a geochemical variable among various sources that contributed to the total variation. It is thereby possible to assess the relative importance of both geologic factors and various laboratory procedures as they have affected the observed data.

f) Variance of a mean and confidence intervals.

A statistical estimate of the mean or of any other parameter of a geochemical population is based on an experiment--the experiment consisting of field sampling and laboratory analysis according to some plan. If the experiment is repeated over and over again according to the same plan, a number of different estimates will be obtained. The frequency distribution of these estimates is the sample distribution of the statistic, and the variance of the sample distribution, or of the statistic, is an inverse measure of the precision or efficiency of the statistic. Fortunately, the variance of the statistic can be predicted from the results of the first experiment--without repeating the experiment a large number of times.

The variance of an estimated arithmetic mean is given by:

$$s^2_{\bar{x}} = \frac{s^2}{n} \qquad (19)$$

where s^2 is the estimated variance of the individual geochemical values (eq. 13), and n is the number of <u>independent</u> values that were used in the calculation of s^2.

The need for n to represent the number of <u>independent</u> values used in the calculation of s^2 cannot be overemphasized. If any of the values are related in some manner, the variance, σ^2, could be seriously underestimated. Say, for example, that the n values were actually collected from n_α randomly selected sampling localities, with n_β values selected at random from within each locality (so that $n = n_\alpha n_\beta$), rather than from n points selected at random from throughout the region of investigation. The sampling, then, would be in accordance with the model of equation (5). Rather than having n independent values, we would have n_α independent means for the n_α localities. Each of the means would have a variance equal to the variance within localities, s_β^2, divided by n_β.

$$\frac{s_\beta^2}{n_\beta} \tag{20}$$

If an estimate of the variance among the true locality means is denoted by s_α^2 (a variance "component"), the variance among the n_α estimated means is given by:

$$s_\alpha^2 + \frac{s_\beta^2}{n_\beta} \tag{21}$$

The variance of the grand mean of all n values, or of the mean of the n_α means, is:

$$\frac{s_\alpha^2 + \frac{s^2}{n_\beta}}{n_\alpha} \tag{22}$$

or, more simply:

$$s_{\bar{x}}^2 = \frac{s_\alpha^2}{n_\alpha} + \frac{s_\beta^2}{n_\alpha n_\beta} \tag{23}$$

The same arguments used in moving from equation (19) to equation (23) can be employed in deriving estimating equations for situations where the sampling design contained additional hierarchical levels and the model contained other terms in addition to α and β. Just as equation (19) requires that n represents the number of independent values used to calculate s^2, equation (23) requires that n_α represent the number of independent means used to calculate s_α^2 and n_β represent the number of independent values within each sampling locality used to calculate s_β^2. The manner in which s_α^2 and s_β^2 are calculated is discussed in later sections of the syllabus on analysis of variance procedures.

Some sampling designs, as will be discussed later in the syllabus, involve sampling from populations of limited size. For example, at some level of the design we may subdivide a square or rectangular area into, say, four equal-sized quadrangles and then select two of the quadrangles at random for sampling. In this situation, the population consists of only 4 items, and we would have sampled one-half of the population. Equations for estimating variances of means, as given in equations (19) and (23) apply only where the fractions of the populations that were sampled are small. If this sampling fraction is large, correction terms must be applied (Cochran, 1963, p. 286). Equation (23) with the correction terms would be:

$$s^2_{\bar{y}} = \frac{1 - f_\alpha}{n_\alpha} s^2_\alpha + \frac{1 - f_\alpha f_\beta}{n_\alpha n_\beta} s^2_\beta \qquad (23a)$$

where f_α is the fraction of the sampling localities that were sampled and f_β is the fraction of the total number of potential samples in each locality that were actually collected and analyzed. If all possible sampling localities were sampled, f_α is equal to 1 and if, as is usually true, only a very small proportion of the potential samples in each locality were collected, f_β is very near zero. In this situation, equation (23a) reduces to equation (19).

Confidence limits about an estimated mean, $\bar{x}$, are given by:

$$\bar{x} \pm t_{\alpha} s_{\bar{x}} \tag{24}$$

where t_{α} is Student's t, and is required because the variances used in equations (19) to (23) are only estimates of the true variances. The values of t_{α} are read from a table of t such as that given in table 3; the table is entered with α (the probability that the error in the estimated mean exceeds the error indicated by $t_{\alpha} s_{\bar{x}}$) and ν (the degrees of freedom available for estimating $s_{\bar{x}}$). If $s_{\bar{x}}$ has been derived through expressions as given in equation (23), the number of degrees of freedom, ν, is less than straightforward. However, Cochran (1963, p. 12) has pointed out, in effect, that if ν is greater than about 60, ν may be taken as infinity (See table 3) without serious error. We shall be even more liberal and take ν as infinity if it is actually 10 or more. (If α is set at 0.05, $t_{0.05}$=2.23 for ν= 10, $t_{0.05}$=2.00 for ν= 60, and $t_{0.05}$=1.96 for ν= infinity.) If $s_{\bar{x}}$ is estimated from an equation similar to equation (23), the number of degrees of freedom is more than $n_{\alpha}-1$, but less than $n_{\alpha} n_{\beta}-1$. Our liberal rule allows us to take ν as infinity if we have 11 or more master sampling localities. Krumbein and Slack (1956) suggested ten as a minimum.

Table 3.-- PERCENTAGE POINTS OF THE *t*-DISTRIBUTION*

ν \ α	0.50	0.25	0.10	0.05	0.025	0.01	0.005
1	1.00000	2.4142	6.3138	12.706	25.452	63.657	127.32
2	0.81650	1.6036	2.9200	4.3027	6.2053	9.9248	14.089
3	0.76489	1.4226	2.3534	3.1825	4.1765	5.8409	7.4533
4	0.74070	1.3444	2.1318	2.7764	3.4954	4.6041	5.5976
5	0.72669	1.3009	2.0150	2.5706	3.1634	4.0321	4.7733
6	0.71756	1.2733	1.9432	2.4469	2.9687	3.7074	4.3168
7	0.71114	1.2543	1.8946	2.3646	2.8412	3.4995	4.0293
8	0.70639	1.2403	1.8595	2.3060	2.7515	3.3554	3.8325
9	0.70272	1.2297	1.8331	2.2622	2.6850	3.2498	3.6897
10	0.69981	1.2213	1.8125	2.2281	2.6338	3.1693	3.5814
11	0.69745	1.2145	1.7959	2.2010	2.5931	3.1058	3.4966
12	0.69548	1.2089	1.7823	2.1788	2.5600	3.0545	3.4284
13	0.69384	1.2041	1.7709	2.1604	2.5326	3.0123	3.3725
14	0.69242	1.2001	1.7613	2.1448	2.5096	2.9768	3.3257
15	0.69120	1.1967	1.7530	2.1315	2.4899	2.9467	3.2860
16	0.69013	1.1937	1.7459	2.1199	2.4729	2.9208	3.2520
17	0.68919	1.1910	1.7396	2.1098	2.4581	2.3982	3.2225
18	0.68837	1.1887	1.7341	2.1009	2.4450	2.8784	3.1966
19	0.68763	1.1866	1.7291	2.0930	2.4334	2.8609	3.1737
20	0.68696	1.1848	1.7247	2.0860	2.4231	2.8453	3.1534
21	0.68635	1.1831	1.7207	2.0796	2.4138	2.8314	3.1352
22	0.68580	1.1816	1.7171	2.0739	2.4055	2.8138	3.1188
23	0.68531	1.1802	1.7139	2.0687	2.3979	2.8073	3.1040
24	0.68485	1.1789	1.7109	2.0639	2.3910	2.7969	3.0905
25	0.68443	1.1777	1.7081	2.0595	2.3846	2.7874	3.0782
26	0.68405	1.1766	1.7056	2.0555	2.3788	2.7787	3.0669
27	0.68370	1.1757	1.7033	2.0518	2.3734	2.7707	3.0565
28	0.68335	1.1748	1.7011	2.0484	2.3685	2.7633	3.0469
29	0.68304	1.1739	1.6991	2.0452	2.3638	2.7564	3.0380
30	0.68276	1.1731	1.6973	2.0423	2.3596	2.7500	3.0298
40	0.68066	1.1673	1.6839	2.0211	2.3289	2.7045	2.9712
60	0.67862	1.1616	1.6707	2.0003	2.2991	2.6603	2.9146
120	0.67656	1.1559	1.6577	1.9799	2.2699	2.6174	2.8599
∞	0.67449	1.1503	1.6449	1.9600	2.2414	2.5758	2.8070

* Computed by Maxine Merrington from "Tables of Percentage Points of the Incomplete Beta Function," *Biometrika*, *32* (1941), pp. 168–181, by Catherine M. Thompson, and reproduced by permission of Professor E. S. Pearson.

*From Bennett and Franklin (1954, p. 696)

The 95 percent confidence limits ($\alpha = 0.05$) for an estimated arithmetic mean, $\bar{x}$, are given by the range:

$$(\bar{x} - 1.96 s_{\bar{x}}) \text{ to } (\bar{x} + 1.96 s_{\bar{x}}) \quad . \tag{24}$$

If $\bar{x}$ and $s_{\bar{x}}$ were computed from the logs of the geochemical values, the 95 percent confidence limits are given by the antilogs of the limits in (24). Alternatively, we may set:

$$GE = \exp(s_{\bar{x}}) \text{ or } GE = 10^{s_{\bar{x}}} \tag{25}$$

depending on the base of the logarithms, and give the 95 percent range of confidence as:

$$(GM/GE^{1.96}) \text{ to } (GM \times GE^{1.96}) \tag{26}$$

where GM is the geometric mean. As mentioned previously, the value 1.96, in practice, is almost always taken as equal to 2.

A method, and tables, for estimating confidence intervals about geochemical abundances derived by means of Sichel's $\underline{t}$-estimator are given by Sichel (1966).

g) Means and variances from censored sample distributions.

The terms censored and truncated as applied to frequency distributions of geochemical data are commonly confused. In statistical terminology, a distribution is censored when values below a certain limit or above a certain limit can be counted but not measured. The sample distribution is left- or right-censored, respectively. This situation is very commonly encountered in geochemistry when the population distribution overlaps the chemist's lower or upper limits of analytical determination. Truncated sample distributions, on the other hand, occur when values of the population can be neither counted nor measured. One example of this might be in measuring the diameters of mineral grains; grains with very small diameters may be neither seen nor measured, and an unknown proportion of the frequency distribution is, therefore, missing. Truncated sample distributions such as this are not ordinarily encountered in field geochemistry.

Geochemists are commonly negligent in describing the manner in which they handle the censored distribution problem in statistical treatment of their data. In fact, they commonly do not even report the fact that the data were censored at all. This failure is not exactly fair to the reader who may wish to judge the validity of the statistical analysis. A useful device for reporting the degree of censoring in a sample distribution is the detection ratio. The detection ratio has the general form a:b, where a is the number of geochemical samples in which the chemical constituent was measured by the analyst and b is the total number of samples that were analyzed.

If the detection ratio is close to unity, the estimated mean and variance are not highly sensitive to the method of estimation. Where the detection ratio is smaller, however, this will not be true. If the sample distribution is left-censored, analytical reports of "less than x_o" specify geochemical values somewhere between zero and x_o, where x_o is the lower limit of analytical determination. If the distribution is right-censored, analytical reports of "greater than x_o" specify values somewhere between x_o and either 100 percent or 10^6 ppm, where x_o in this case is the upper limit of analytical determination. A common practice among field geochemists has been to assign "less than" reports a value of either zero percent or ppm or some arbitrary value immediately below the lower limit of analytical determination. Reports of "greater than" are less common, but are generally assigned some arbitrary value immediately above the upper limit of analytical determination. Justification for the assignment of arbitrary values to reports of "less than" and "greater than" exists only where the detection ratio is near unity and the computational results are almost independent of any reasonable arbitrary value that may be chosen. Arbitrary assignments are necessary when the data are to be analyzed by analysis of variance methods or by almost any multivariate statistical procedure. They are not necessary, however, for the estimation of means and variances.

Methods given by Cohen (1959, 1961) can be used to estimate the population mean and variance from a censored sample distribution. The estimating equations are:

$$\bar{x} = \bar{x}' - \lambda(\bar{x}' - x_o) \tag{27}$$

and

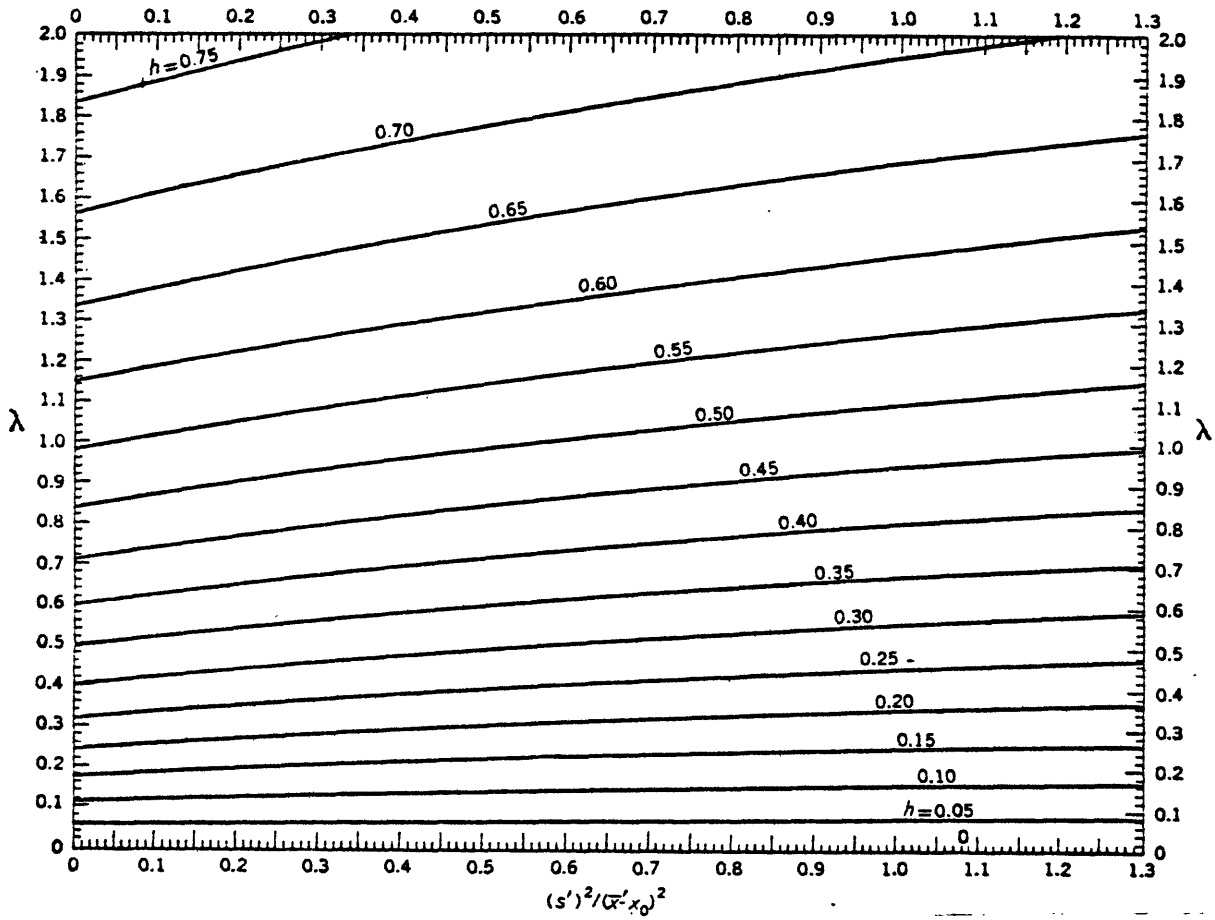
$$s^2 = (s')^2 + \lambda(\bar{x}' - x_o)^2 \tag{28}$$

where $\bar{x}'$ and $(s')^2$ are the mean and variance of the n' values that are uncensored (that is, analytical reports other than "less than" or "greater than"). Equations (27) and (28) are valid for singly-censored sample distributions only; they cannot be used where the distribution is both left- and right-censored. The factor "λ" is read from tables (see Cohen, 1959, 1961) or from graphs (fig. 7) and varies with the degree of censoring and with the quantity:

$$\frac{(s')^2}{(\bar{x}' - x_o)^2} \tag{29}$$

The degree of censoring is given by $h = n'/n$ where n' is the number of analytical reports of "less than" or "greater than" and n is the total number of samples analyzed. The quantity h is also equal to one minus the detection ratio.

Fig. 7

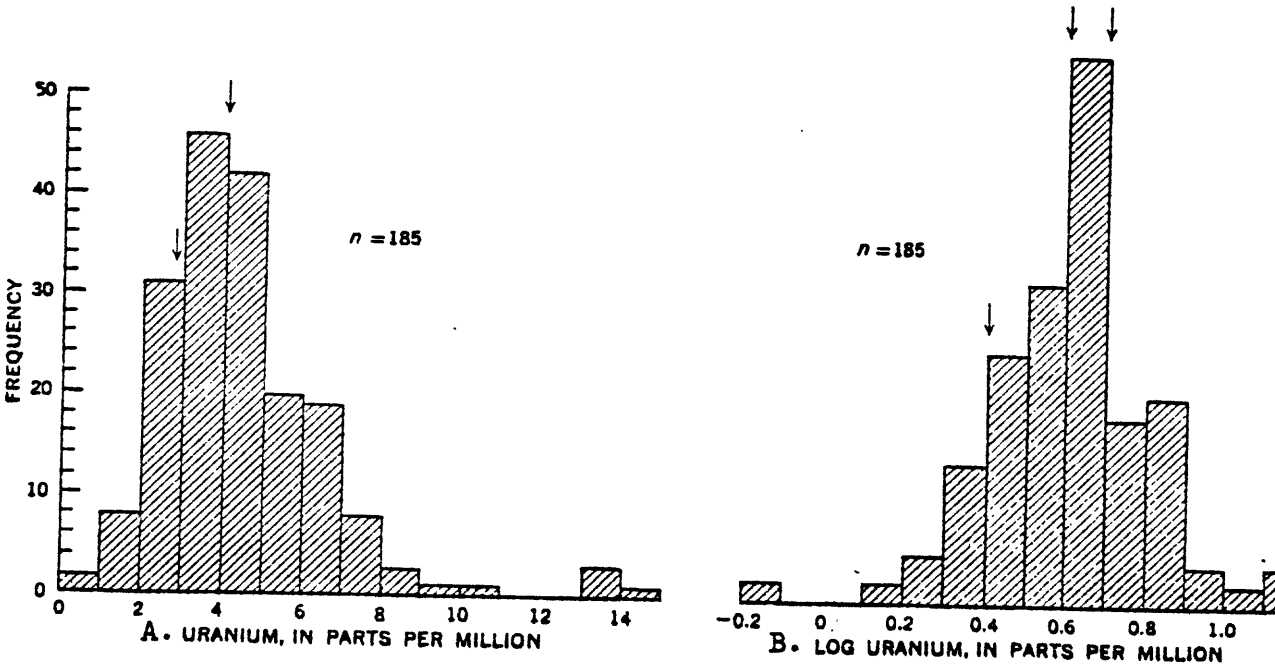


From Cohen (1959). Reproduced, with permission, in Miesch (1967b).

The method of Cohen is strictly valid only where the population frequency distribution of x is normally distributed, but has been found to be satisfactory wherever the sample distribution is approximately symmetrical. If the sample distribution is highly asymmetrical, as are the highly positively-skewed distributions commonly encountered in geochemistry, the use of Cohen's method may require a transformation of the data beforehand--such as transformation to logarithms. Equations (27) and (28) can then be used to estimate the mean logarithm and the variance of the logarithms. The antilog of the mean logarithm will then give the geometric mean (GM) and the antilog of the square root of the variance of the logarithms will give the geometric deviation (GD). With these values, the method of Sichel, described previously, can be used to obtain an estimate of the arithmetic mean (geochemical abundance), if needed.

Figure 8A shows the sample distribution of uranium values (ppm) for a granite body sampled by Hubaux and Smiriga-Snoeck (1964). The distribution is positively skewed as is typical for sample distributions of minor element values. The frequency distribution of the logs (base 10) of the values is shown in figure 8B and is considerably more symmetrical. The arithmetic mean and standard deviation (eqs. 10 and 13), estimated from the entire sample distribution ($n=185$) in figure 8A were found to be 4.53 ppm and 2.15 ppm, respectively. The distribution was

Fig. 8


FREQUENCY
50
40
30
20
10
0
n = 185
0
2
4
6
8
10
12
14
A. URANIUM, IN PARTS PER MILLION
n = 185
−0.2
0
0.2
0.4
0.6
0.8
1.0
B. LOG URANIUM, IN PARTS PER MILLION

artificially censored at x_o = 2.6 ppm and x_o = 4.0 ppm, and Cohen's method was used to estimate the population mean and standard deviation from the sample values greater than x_o only. The complete results were as follows:

x_o	Detection ratio	Estimated mean	Estimated std. dev.
None	185:185	4.53 ppm	2.15 ppm
2.6 ppm	162:185	4.89	2.04
4.0	105:185	5.78	2.03

A similar experiment was carried out after transforming all data to logarithms (base 10). The results were as follows:

x_o	$\log x_o$	Detection ratio	Estimated mean log	Estimated std. dev. of logs	Estimated GM	Estimated GD	Estima arithm mean
None	---	185:185	0.612	0.199	4.1 ppm	1.58	4.54
2.6 ppm	0.415	162:185	.621	.181	4.2	1.52	4.55
4.0	.602	105:185	.620	.181	4.2	1.52	4.54
5.1	.708	50:185	.571	.220	3.7	1.66	4.24

Other examples are given in Miesch (1967b). The example above shows 1) that Cohen's method was reasonably successful in estimating the means and standard deviations from sample distributions censored by as much as 73 percent, and 2) the method gave better results after the data had been transformed to logarithms so that the central part of the sample distribution was more symmetrical than that of the original ppm values.

h) Measures of skewness and kurtosis.

The skewness and kurtosis of a frequency distribution curve are, respectively, measures of the asymmetry and peakedness of the curve and have been of interest to sedimentary petrologists for many years. Measures of skewness and kurtosis are made in various ways, including both mathematical and graphical procedures. All of the conventional methods for measuring skewness yield a value of zero for a distribution that is symmetrical about its mean value, a positive value for a distribution that has a tail extended towards the higher values, and a negative value for a distribution with a tail extended towards the lower values. Some commonly used methods for measuring kurtosis yield a value greater than three for a distribution that is more peaked than a normal distribution curve and a value less than three for a distribution that is less peaked. The preferred methods for measuring skewness and kurtosis are based on the k-statistics of R. A. Fisher (See Bennett and Franklin, 1954, p. 81). The first two k-statistics are the arithmetic mean and the variance (equivalent to $\bar{x}$ and s^2 of eqs. 10 and 13, respectively). The third and fourth k-statistics are:

$$k_3 = \frac{n^2 S_3 - 3n S_2 S_1 + 2 S_1^3}{n(n-1)(n-2)}$$

$$k_4 = \frac{(n^3+n^2) S_4 - 4(n^2+n) S_3 S_1 - 3(n^2-n) S_2^2 + 12 n S_2 S_1^2 - 6 S_1^4}{n(n-1)(n-2)(n-3)}$$

where n is the number of values and S_r is:

$$S_r = \sum x^r$$

The third and fourth k-statistics, like the variance, have the property of additivity. That is, k_3 and k_4 for the sum of two or more independent variables are equal to the sums of k_3 and k_4 for the individual variables.

The measure of skewness (g_1) is given by:

$$g_1 = k_3 / s^3$$

where s^3 is the cube of the standard deviation (eq. 13). The kurtosis (g_2) is measured by:

$$g_2 = k_4 / s^4$$

and is equal to zero for a normal distribution curve, to a positive value for a distribution more peaked than the normal curve, and to a negative value for a distribution less peaked than the normal curve. The maximum absolute values of g_1 and the limits for g_2 to be expected 95 and 99 percent of the time if the population distribution is truly normal are given in table 4 from R. C. Geary and E. S. Pearson (See Bennett and Franklin, 1954, p. 95).

TABLE 4 5% AND 1% POINTS FOR g_1 AND g_2*

Size of Sample	g_1		g_2			
	Lower and Upper		Lower		Upper	
	5%	1%	1%	5%	5%	1%
50	0.550	0.812	—	—	—	—
75	0.454	0.664	—	—	—	—
100	0.395	0.576	−0.80	−0.62	0.87	1.53
125	0.354	0.514	−0.74	−0.57	0.78	1.34
150	0.324	0.469	−0.69	−0.53	0.71	1.22
175	0.301	0.434	−0.66	−0.50	0.66	1.11
200	0.282	0.406	−0.62	−0.47	0.62	1.04
250	0.253	0.362	−0.57	−0.44	0.55	0.91
300	0.231	0.331	−0.53	−0.40	0.50	0.82
350	0.214	0.306	−0.49	−0.37	0.47	0.75
400	0.201	0.286	−0.48	−0.35	0.43	0.69
450	0.189	0.270	−0.44	−0.33	0.40	0.65
500	0.180	0.256	−0.42	−0.32	0.38	0.62
550	0.171	0.244	−0.41	−0.30	0.37	0.59
600	0.163	0.234	−0.39	−0.29	0.35	0.55
650	0.157	0.225	−0.38	−0.28	0.35	0.53
700	0.151	0.215	−0.37	−0.27	0.32	0.51
750	0.146	0.208	−0.35	−0.26	0.31	0.49
800	0.142	0.202	−0.34	−0.25	0.30	0.47
850	0.138	0.196	−0.33	−0.25	0.29	0.46
900	0.134	0.190	−0.33	−0.24	0.29	0.44
950	0.130	0.185	−0.32	−0.23	0.28	0.43
1000	0.127	0.180	−0.31	−0.23	0.27	0.42

* From Bennett and Franklin (1954, p. 95)

i) Measures of correlation among variables.

Measures of correlation among geochemical variables are frequently needed for interpretations of geochemical coherence (Rankama and Sahama, 1950, p. 48) and geochemical behavior. An understanding of correlation is also a necessary prerequisite to discussions of geochemical errors that follow in later parts of this syllabus. The most common measure of correlation is the simple linear correlation coefficient. The population correlation coefficient, ρ_{ij}, is estimated by:

$$r_{ij} = \frac{s_{ij}}{s_i s_j} \tag{30}$$

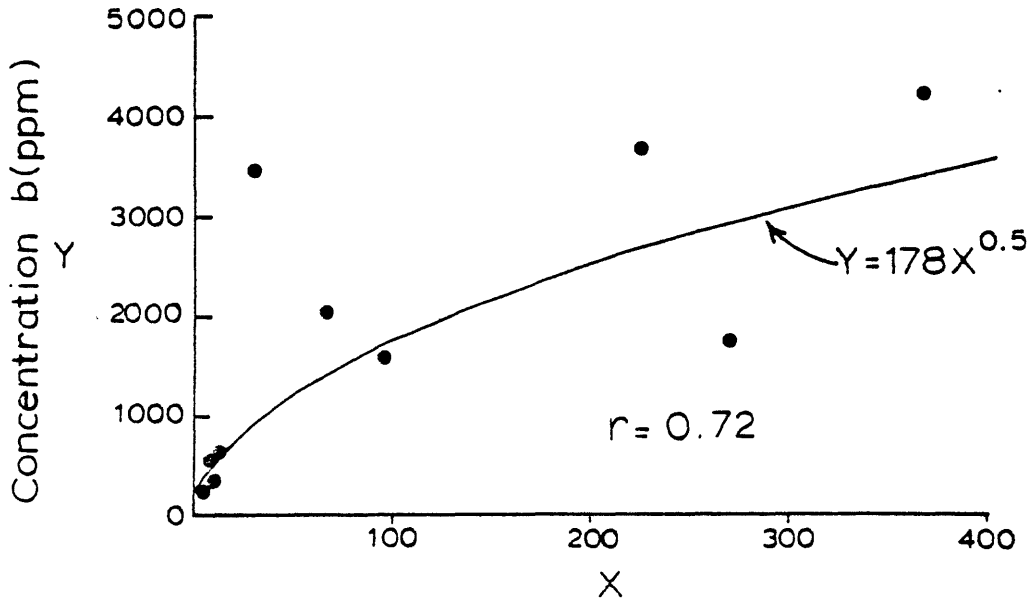
where r_{ij} is the estimated correlation for variables i and j, s_i and s_j are the estimated standard deviations of variables i and j (eq. 13), and s_{ij} is the estimated covariance from:

$$s_{ij} = \frac{\Sigma (x_i - \bar{x}_i)(x_j - \bar{x}_j)}{n-1} \tag{31}$$

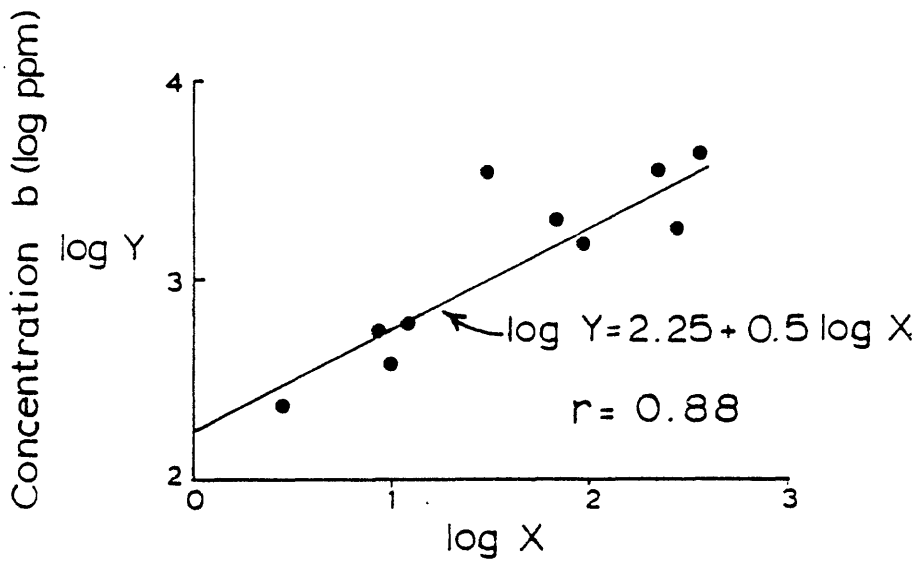
where $\bar{x}_i$ and $\bar{x}_j$ are the means of the x_i and x_j variables and n is the number of pairs of variables. The value of r_{ij} is +1.0 where the plotted points fall on a straight line with positive slope, -1.0 if the slope is negative, and zero if the plotted points show no linear relation whatsoever.

The statistical significance of an estimated simple linear correlation coefficient may be determined from standard tables if the bivariate frequency distribution is at least approximately normal. This condition is commonly not present for geochemical data as percent or ppm values. In many situations, the logarithms of the geochemical values as plotted on an x-y graph do appear to follow the bivariate normal form. Thus, it is commonly possible to determine the statistical significance of a correlation among the log values but not of the correlation among the original geochemical values. Another reason for estimating correlations for log values is that correlations among the original values are commonly obscure and governed almost entirely by the upper parts of the bivariate distribution. Both of these points are illustrated in figure 9. The correlation between the ppm values (fig. 9A) is determined almost entirely by the six pairs of higher values; the relationship among the four pairs of lower values has little influence on the estimated correlation. The correlation between the log ppm values (fig. 9B), however, is affected by the lower four pairs of values about as much as it is affected by the other six pairs. Also, the total relationship between the variables is somewhat more clear when observed by way of logarithms. In addition, the bivariate distribution of the logs is at least conceivably the result of a bivariate normal population, whereas that of the ppm values is not. The correlation between the logs could be tested for statistical significance.

Fig. 9



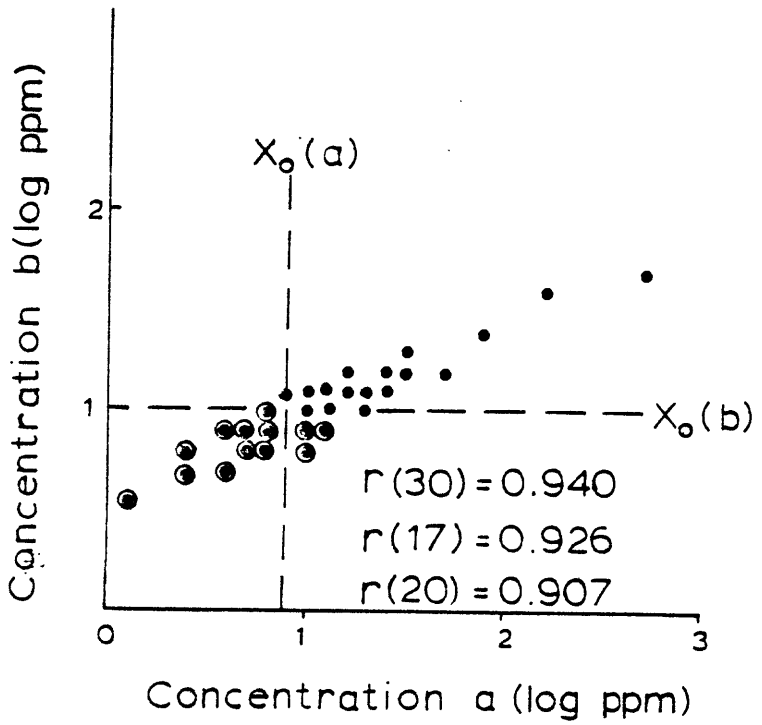
A. Concentration a (ppm)



B. Concentration a (log ppm)

Censored values occur in almost all geochemical data sets and, if correlations are to be estimated, a decision must be made regarding the treatment of the "less than" and "greater than" values. If the correlations are to be estimated for the ppm values, the "less than" values can usually be treated as zeros without serious bias. The correlation for the bivariate distribution in figure 9A, for example, would be essentially the same if points representing less than 20 ppm of constituent "a" and less than 800 ppm of constituent "b" were all moved to the point 0,0. If the correlation is to be computed for the logarithmic data, however, and if a matrix of log correlations is to be studied by means of R-mode factor analysis, there is probably little one can do other than treat all "less than" values as some arbitrary value immediately below the lower limit of analytical determination for the respective constituent. (Some workers conventionally treat "less than" values as seven-tenths of the lower analytical limit.) This will probably not bias the results a great deal if the detection ratios for the two constituents are both high (say, 0.8 or higher). However, if factor analysis is not intended, the correlation can be estimated from the upper part of the bivariate distribution only. This is illustrated in figure 10 where the estimated correlation for all pairs is 0.940. If the correlation is estimated from only the points above $\chi_o(a)$ and $\chi_o(b)$, it is 0.926, and if estimated from only the points to the right of $\chi_o(a)$, it is 0.907. Correlations estimated from the upper parts of censored bivariate distributions have unknown

Fig. 10


Concentration b(log ppm)
X_o(a)
X_o(b)
r(30) = 0.940
r(17) = 0.926
r(20) = 0.907
0
1
2
3
Concentration a (log ppm)

frequency distributions and cannot be tested for statistical significance; they should be regarded only as indices of geochemical correlation or geochemical coherence. Matrices of such correlations do not have the Gramian properties required for factor analysis; that is, the derived principal components matrices cannot be used to reproduce the original correlation matrix.

Chayes (1960, 1962) has shown that correlations among compositional variables do not necessarily reflect genetic relations because such variables sum to a value that is constant for all observations (samples). While the sum of the logarithms of the variables is not constant among observations, Chayes' argument still holds in principle, although not in mathematical detail. Extreme caution must be exercised in the interpretation of correlations among compositional variables, or their logs, in terms of geologic or geochemical processes. A test of the statistical significance of correlations among compositional variables, proposed by Chayes and Kruskal (1966), has been judged invalid (Miesch, 1969). Genetic associations might be examined more effectively by estimating correlations among the ratios of each constituent to some reference constituent, such as SiO_2, but this by no means completely avoids the problems pointed out by Chayes (Miesch and others, 1966).

II. Nature and effects of geochemical errors.

1) Definition and classification of errors

The importance of considering the nature of the errors in geochemical data is apparent when we consider the fact that the presence of error is the only reason that statistical procedures are used in the analysis and interpretation of the data. Each geochemical value is intended to represent the concentration of a given chemical constituent in some volume of material larger than the sample itself. It would be only a coincidence if the value were perfectly correct. First, the laboratory analysis of the sample is always wrong by at least some small increment, and second, the sample is never perfectly representative of the sampling locality from which it was taken. Each data value, therefore, contains both an analytical error and a sampling error. If neither of these types of error were present, there would be no need for statistical analysis.

Analytical error will be defined as the difference between the concentration reported by the analyst and the true concentration in the sample submitted to him for analysis. Sampling error will be defined as the difference between the true concentration in a sample, or the average true concentration for a group of samples, and the true concentration in the volume of material that the samples, or group of samples, is intended to represent.

It will be obvious from this definition of sampling error that the errors in sampling arise partly from the nature of the rock or soil unit being sampled. If the unit were perfectly homogeneous in composition, no sampling error would occur no matter what sampling procedure was used. If, on the other hand, the unit were highly variable in composition, large sampling errors may be difficult to avoid and would be present through no fault of the sampler whatsoever. If the samples are intended to represent a sampling locality, much will depend on how the locality was defined. In other words, the degree of sampling error may depend partly on the ambitions of the sampler; a sample from a hillside may contain little sampling error if the sample is intended to represent only the hillside, but may contain larger error if intended to represent the entire mountain, depending on the nature of the variation within the mountain. Regardless of the magnitude of the errors due to either analysis or sampling, laboratory and sampling procedures should be conducted in such a way that the errors are susceptible to analysis by statistical methods. Certain properties of errors can invalidate certain statistical procedures. They can also invalidate non-statistical procedures as well.

Two of the fundamental properties of errors are bias and imprecision. Bias and imprecision are functions of the average of the errors and their variability, respectively. If the average error is zero, the method that led to the errors is unbiased. If all of the errors are identical, the method is perfectly precise. A geochemical value is unbiased if it was produced by an unbiased method. A value is precise if it was produced by a precise method. Bias and imprecision are completely independent properties. A method can be biased and imprecise, biased but precise, unbiased but imprecise, or unbiased and precise.

Bias can arise in the laboratory if the analytical procedure is inherently incorrect, if the sample material becomes contaminated in some way, or if the aliqout taken for analysis from the sample submitted by the geologist is not representative. Bias can occur in field sampling where the population available for sampling differs from the target population, or where the available population is not sampled by some objective procedure.

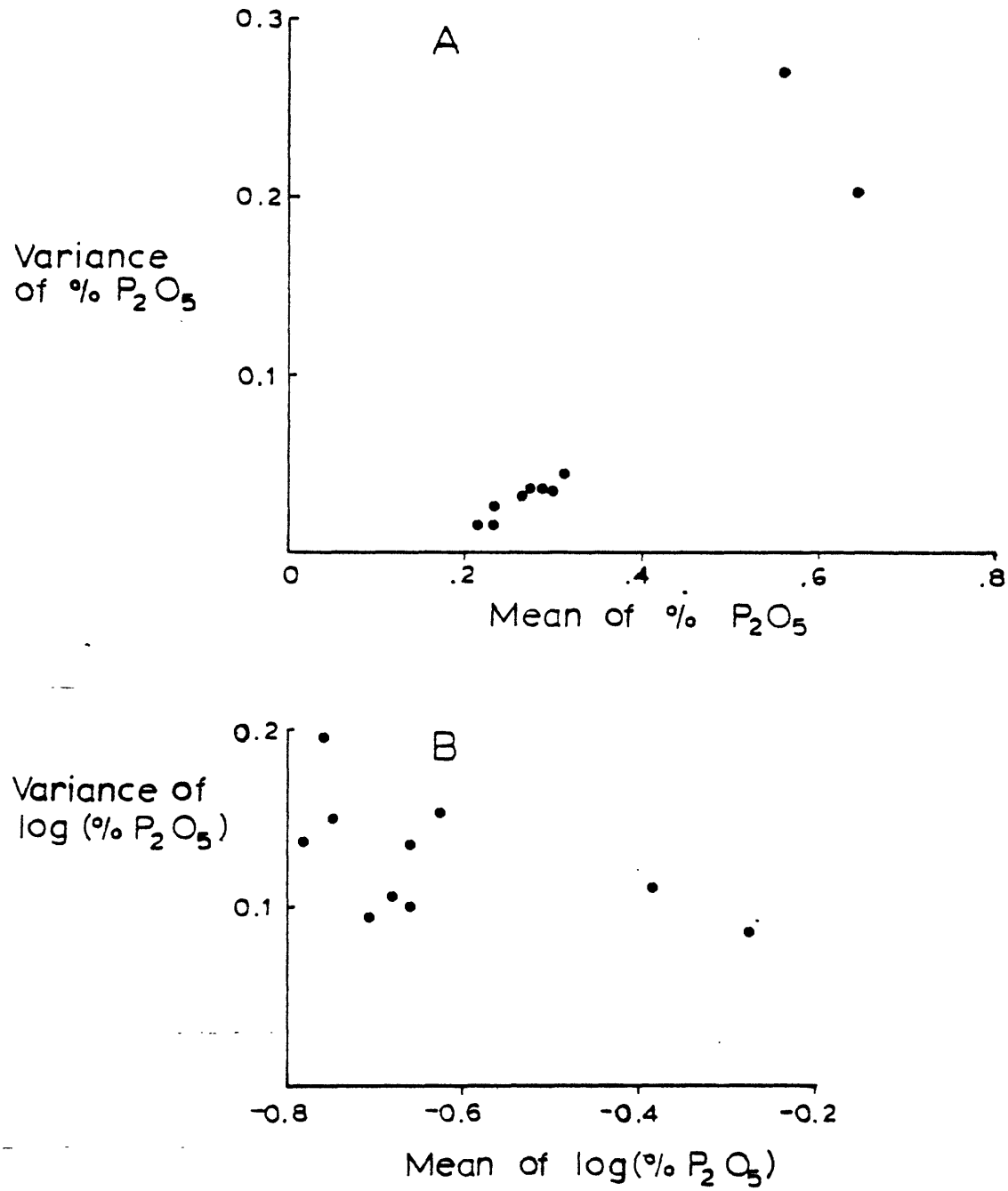
Some degree of imprecision is associated with all laboratory procedures, but is commonly appreciable where the procedures are of the rapid and low-cost variety generally used in geochemical surveys. Imprecision in field sampling is appreciable when the sampling localities to be represented by the samples are highly variable in composition, but it can always be reduced by increasing the number of samples from each locality.

2) Effects of errors

Although some degree of imprecision is to be expected in all measurement data, including geochemical values, variable precision can lead to difficulties in statistical analysis. For example, the degree of analytical precision can vary from one specimen to another, and commonly does, because analytical imprecision increases directly with increasing amounts of the constituent in the sample. The imprecision among sampling errors can vary from one sampling locality to another because of different degrees of variability within localities. In most situations, the variability within localities is related directly to the locality means.

The relationship of imprecision in sampling and(or) analysis to mean concentration is illustrated in figure 11A wherein the mean and variance of P_2O_5 are plotted for ten sampling localities. The P_2O_5 data are for sandstones of Cambrian age. The ten localities are distributed over most of the western U.S.; 32 samples were selected from each locality by procedures that involved formal randomization (data from A. T. Miesch and J. J. Connor, unpublished). The plot in figure 11 shows not only that the variance is greatly different from one locality to the next, but that the variances are at least approximately proportional to the means. If analysis of variance procedures (including the popular t test) were to be used to judge the significance of differences among the locality means, it would be necessary to assume that the population variances within localities were all at least approximately equal. Figure 11A gives the strong impression that the assumption would be grossly invalid for these data. For situations wherein the mean and variance appear to be related, Bartlett (1947) has recommended the logarithmic transformation before proceeding to analysis of variance methods. Log (base 10) transformation of the P_2O_5 data leads to the relation shown in figure 11B. The means and variances of the log data show no apparent correlation, and the variances for the ten localities are a great deal more similar than were the variances for the original percentage data. Analysis of variance procedures could be applied to the log data without much concern for the effects of inhomogeneous variance.

Fig. 11


A
0.3
0.2
0.1
Variance of % P_2O_5
0
.2
.4
.6
.8
Mean of % P_2O_5
B
0.2
0.1
Variance of log (% P_2O_5)
-0.8
-0.6
-0.4
-0.2
Mean of log (% P_2O_5)

Bias in sampling and analysis as well as imprecision can be variable from one sample to another or from one sampling locality to another, but the effects are a great deal more severe and almost impossible to correct prior to analysis of variance or any other statistical treatment. Variable bias can occur in the laboratory when different analytical methods with different biases are used or when different analysts cause different biases throughout the analytical program. Variable bias can occur in field sampling when localities are sampled by different geologists and samples are selected according to different criteria or different operational procedures.

It will be apparent that bias that is variable in any nontrivial way will render the data from the geochemical program useless regardless of the methods employed in attempts at data interpretation. For example, if geochemical values from any restricted part of the sampled region were biased, due to either sampling or analysis, in a way that differed from the bias in data from the remainder of the region, the difference in the bias would distort the interpretation of the regional geochemical variability whether the interpretation was based on either elaborate statistical or conventional procedures. No amount of statistical treatment or computer processing would help. Any appreciable amount of variable bias would also invalidate analysis of variance procedures because the additive property of the variance would be destroyed. In statistical terminology, the experimental errors would not be independent.

The foregoing assertion is so important that it deserves expanded discussion and explanation. Variable bias is the one property of geochemical errors that can totally invalidate any attempt at interpretation, statistical, or otherwise, and its effect on the additivity of variances is only an example. For purposes of illustration, suppose that the geochemical sampling model is of the form in equation (5):

$$x_{ij} = \mu + \alpha_i + \beta_{ij} \tag{32}$$

The term μ in this model is the true grand mean concentration of the constituent in all potential samples from the region, the term α_i is the difference between μ and the true mean for all samples in the ith sampling locality, and the term β_{ij} is the difference between $\mu + \alpha_i$ and the analytical value for the jth sample from the ith locality. In other words, $\mu + \alpha_i$ is the true value for the ith sampling locality and β_{ij} is the error due to both sampling and laboratory analysis. Suppose further that three samples are collected from each of two sampling localities by unbiased procedures and that the laboratory methods were also without bias. The data and their underlying components may tend to have the properties of the following:

Locality, i	Sample, j	x_{ij}	$\mu+\alpha_i$	β_{ij}
1	1	12	10	+2
1	2	10	10	0
1	3	8	10	-2
2	1	22	20	+2
2	2	20	20	0
2	3	18	20	-2
Mean		15	15	0
Variance		27.667	25	2.667

(33)

Note that the mean of the error is the same (zero) for both localities, that the correlation coefficient for the quantities $\mu+\alpha_i$ and β_{ij} would be zero, and that the variance of x_{ij} is equal to the sum of the variances of the two components.

On the other hand, suppose that the samples from both localities were collected by a biased procedure or that they were analyzed by a biased laboratory method. The data and the underlying components might have properties that tend to be as follows:

Locality, i	Sample, j	x_{ij}	$\mu+\alpha_i$	β_{ij}
1	1	13	10	+3
1	2	11	10	+1
1	3	9	10	-1
2	1	23	20	+3
2	2	21	20	+1
2	3	19	20	-1
Mean		16	15	+1
Variance		27.667	25	2.667

(34)

The mean error (+1) is nonzero, but is the same for both localities, the errors are uncorrelated with the true locality means, and the variance of the two components of variation in the data are still additive even though bias is present in the data. Suppose now that bias is present in only the data from one of the two localities, due to a difference between either the sampling procedures or the analytical methods. The data and underlying components may tend to be as follows:

Locality, i	Sample, j	x_{ij}	$\mu+\alpha_i$	β_{ij}
1	1	12	10	+2
1	2	10	10	0
1	3	8	10	-2
2	1	23	20	+3
2	2	21	20	+1
2	3	19	20	-1
Mean		15.5	15	+0.5
Variance		32.917	25	2.917

(35)

Note that the mean error for locality 1 is zero whereas that for locality 2 is +1 and that the values of $\mu + \alpha_i$ and β_{ij} are correlated ($\rho = 0.29$). Because of the variable bias, resulting in correlation of the components that make up the data (x_{ij}), the variances are not additive, and any attempt to use analysis of variance methods would be invalid. However, analysis and interpretation of the data by any other method would also be misleading.

In order to emphasize the fact that it is the variable bias alone that destroys the property of additivity, we may form another set of data and underlying components wherein variable imprecision is present and the frequency distributions of the errors are asymmetrical. Also, bias will be introduced but will be set equal for both sampling localities.

Locality, i	Sample, j	x_{ij}	$\mu+\alpha_i$	β_{ij}	Mean	Variance
1	1	15	10	+5		
1	2	12	10	+2	+2.667	2.889
1	3	11	10	+1		
2	1	24	20	+4		
2	2	23	20	+3	+2.667	1.555
2	3	21	20	+1		
Mean		17.667	15	+2.667	+2.667	2.222
Variance . .		27.222	25	2.222	----	-----

(36)

Thus, neither imprecision, variable imprecision, skewness, nor constant bias cause the property of additivity to be destroyed. This is not to say that variable imprecision and skewness do not cause some difficulties in analysis of variance methods. Where the imprecision is variable, the estimated error variance is only an average of the variances for the various localities, (See variance of β_{ij} in the example above.) and probability tests will be inexact. Skewness in

the errors will also cause probability tests to be inexact and, if severe, can invalidate them to an unknown degree. The tests are based on the assumption that the population frequency distributions of the errors are normal. It will be shown in a later section of the syllabus that the presence of both variable imprecision and skewness in the errors can lead to variable bias which, as we have seen, can destroy analysis of variance methods entirely.

3) Avoiding variable bias

Variable bias has been shown to invalidate analysis of variance procedures, which are based on the variance's property of additivity, and it is obvious that if the bias is highly variable, it will invalidate any method for interpretation of the data. Attempts have been made to correct the data for the effects of variable bias, but they are rarely, if ever, successful. The obvious question, then, is how to avoid introducing variable bias into the geochemical experiment. The best way, of course, is to avoid bias altogether. To be practical, however, we must recognize the fact that this is generally impossible. Analytical methods and analysts do change throughout the course of any large geochemical program; also, it is commonly necessary to employ more than one geologist or party of geologists to do the sampling in such programs. All geologists, or any other samplers for that matter, have biases regarding what should be sampled and how the sample should be taken or treated in the field, and the biases can vary a great deal. There are two practices that can help. The first is to establish <u>definite operational procedures</u> (Krumbein, 1960) to be followed in both laboratory analysis and selection of samples in the field. These

procedures in the field will require careful definition of the available population and rigid adherence of the rules of the procedure by all field parties. They will also require that all field parties be in agreement as to how the population is recognized, how samples are selected from it,how they are actually collected (i.e., stainless steel spade, paper carton, hammer, hammer and chisel, etc.), and how they are treated in the field if field treatment is involved (e.g., separation of heavy minerals, acidification of water samples, etc.). The second practice that will help in avoiding variable bias is to employ formal randomization in both the laboratory and the field.

Before submitting samples for analysis, all samples should be placed and numbered in a sequence that is randomized with respect to the localities from which they were taken. The samples should then be analyzed in the randomized sequence. This will insure that the effects of any periodic or progressive changes in the laboratory procedure (e.g., changes in instruments, electrical supply to the instruments, or personnel) will be distributed randomly among the samples. That is, geochemical values for samples from within the same localities will be independently derived and will be independent measures of the geochemical nature of the locality. This would not be the case if the values were obtained in succession in the laboratory by analysis of the samples in the order by which they were collected in the field. Analysis of the samples by field order can, and frequently does, lead to artificial anomalies on geochemical maps, resulting from periods wherein the laboratory, for various reasons, may be reporting biased analytical results. Randomization of samples for laboratory analysis

can easily be accomplished using tables of permuted random numbers (fig. 12). The process of randomization may be impractical for extremely large geochemical programs because it is necessary to collect all of the samples before any of them are submitted to the laboratories. In this situation, the only recourse may be to randomize each group of samples and to have the laboratory analyze selected standard samples periodically. Results from analysis of the standards can then be plotted with respect to the sequence of analysis in search of biases that change periodically or systematically. This approach is only second best, however, and complete randomization should be used if at all possible. Certainly any geochemical values used to estimate the analytical precision should be derived independently, and not by successive analysis of one or more selected samples.

Randomization procedures are also necessary in the field in order to insure that the collected samples are independent. The procedures begin with identification of all potential samples that meet the criteria used to define the population to be sampled, and this requires more field work than may otherwise be conducted. It is easier to collect a few typical samples of limestone from an outcrop, for example, than it is to thoroughly examine the outcrop and identify all potential samples of limestone that may be available. Once all potential samples have been identified, any method can be used that gives each potential sample an equal chance of being selected. For this purpose, it is

Fig. 12

Permuted random numbers - n=100

Each 5-line block is a permutation

94	12	99	73	79	30	25	4	6	55	56	64	74	78	95	13	28	62	[illegible]	24
70	57	47	7	9	26	31	82	10	45	41	15	71	27	40	39	85	34	[illegible]	61
65	87	67	72	88	29	18	54	8	63	60	48	84	58	66	96	97	22	[illegible]	52
36	2	14	23	68	91	69	32	50	33	3	92	81	86	5	83	75	93	[illegible]	90
20	21	17	11	19	80	44	51	*	89	76	59	35	42	98	37	38	53	[illegible]	43
20	53	26	63	68	45	73	36	51	95	52	28	61	50	81	65	13	90	[illegible]	69
48	35	40	33	85	6	60	3	*	67	54	27	87	84	99	17	15	42	[illegible]	57
91	11	2	5	9	75	18	23	86	49	46	21	47	66	89	92	71	43	[illegible]	79
80	32	19	76	44	82	64	8	74	94	22	72	38	97	10	55	12	1	[illegible]	58
93	30	7	4	59	62	24	14	29	83	16	31	39	25	96	37	78	88	[illegible]	77
83	37	41	76	74	61	5	68	49	96	10	75	51	81	30	46	20	60	[illegible]	39
40	12	2	13	91	92	99	47	33	*	56	50	22	64	43	93	73	18	[illegible]	94
98	6	79	84	65	55	17	85	14	72	9	77	86	31	32	15	19	27	[illegible]	44
48	87	8	66	28	1	24	78	82	90	26	21	16	80	42	36	58	4	[illegible]	25
23	11	89	62	52	63	35	29	70	95	97	88	38	71	53	59	3	7	[illegible]	67
36	89	6	11	71	19	48	91	97	58	*	37	87	32	25	66	43	85	[illegible]	31
99	13	76	44	23	20	81	52	24	1	49	9	62	68	41	45	42	30	[illegible]	74
55	21	67	47	17	61	29	53	59	94	39	54	12	8	57	40	5	51	[illegible]	15
14	75	77	34	92	27	46	35	7	70	98	93	84	64	82	83	78	3	[illegible]	73
79	4	80	18	72	95	56	86	90	33	28	2	26	96	69	88	50	63	[illegible]	38
99	19	15	74	54	40	90	51	33	87	69	56	84	39	11	9	45	78	[illegible]	95
3	96	16	89	20	98	53	72	34	64	59	75	17	80	24	97	7	65	[illegible]	4
29	44	46	38	47	*	82	23	86	91	57	1	63	10	70	12	27	36	[illegible]	28
92	88	21	32	42	60	68	50	81	71	6	58	30	61	37	94	93	73	[illegible]	25
67	55	77	49	18	48	14	76	13	66	62	8	31	5	83	43	85	79	[illegible]	26
26	62	51	59	77	45	21	20	*	53	25	7	9	48	37	46	78	35	[illegible]	81
8	55	14	4	63	85	17	10	49	67	73	96	75	47	36	69	61	99	[illegible]	32
39	38	6	22	16	94	66	19	83	27	84	95	28	82	88	64	31	34	[illegible]	70
1	56	90	13	98	97	93	92	18	30	65	76	40	3	50	2	43	72	[illegible]	58
24	52	41	71	86	23	44	74	87	54	57	68	89	11	79	60	91	42	[illegible]	29
63	78	20	16	97	7	71	46	13	57	54	5	11	85	45	33	44	37	[illegible]	10
53	42	32	*	88	69	55	25	64	93	18	52	9	95	24	90	98	68	[illegible]	8
66	60	28	15	29	75	40	38	47	79	43	14	92	81	2	41	19	89	[illegible]	83
30	12	67	51	87	27	49	48	74	91	72	56	21	6	99	62	65	77	[illegible]	58
35	76	1	50	36	34	73	17	86	22	82	94	26	80	4	84	23	96	[illegible]	39
7	34	34	21	90	80	64	61	47	41	39	82	91	57	8	52	6	18	[illegible]	11
93	17	37	12	33	75	68	40	9	03	59	94	4	2	98	92	87	69	[illegible]	89
65	86	35	71	43	88	79	73	14	23	63	46	29	76	22	51	45	60	[illegible]	19
99	*	13	26	42	77	96	81	62	56	67	58	27	1	70	10	24	54	[illegible]	66
30	25	44	5	97	16	53	3	49	28	15	95	32	85	48	31	74	50	[illegible]	70

* Represents 100

From L. E. Moses and R. V. Oakford, 1963, Stanford Univ. Press

convenient to use a table of uniform random numbers (fig. 13). Numbers can be selected from the table to provide X-Y coordinates, for example, and the sample taken at the indicated coordinates. Or, the numbers can be taken to indicate stratigraphic position above or below some stratigraphic horizon. Procedures such as this will commonly lead to a general vicinity from which the sample can be taken, and other procedures can be used to select the exact sampling point within the vicinity. It will generally be found that rock samples, in particular, can be taken from only a few places within the vicinity with the sampling tools that are commonly available (i.e., a geologic pick and hammer, or perhaps a chisel). In this situation, the potential samples can be numbered and samples taken from those places which are chosen at random from the table of uniform random numbers. The need to employ formal randomization procedures in geochemical sampling becomes apparent when we consider the result of purely subjective sampling. If the sampling locality consisted mostly of alternating red and brown sandstone, for example, and if the two types of sandstone tended to be of different composition, the variance for the sampling locality would depend on whether we chose to collect, say, all red sandstone, all brown sandstone, or some of each. With random sampling, the collected samples would tend to be in proportion to the various kinds of rock types present whether these types are visibly recognizable or not. Subjective sampling can lead to biased estimates of variance and, frequently, to negative estimates of variance components from analysis of variance.

Fig. 13

Uniform random numbers

03	99	11	04	61	93	71	61	68	94	66	08	32	46	53	84	60	95	82	32	88	61	81	91	61
38	55	59	55	54	32	88	65	97	80	08	35	56	08	60	20	73	54	77	62	71	29	92	38	53
17	54	67	37	04	92	05	24	62	15	55	12	12	92	81	59	07	60	79	36	27	95	45	89	09
32	64	35	28	61	95	81	90	68	31	00	91	19	89	36	76	35	59	37	79	80	86	30	05	14
69	57	26	87	77	39	51	03	59	05	14	06	04	06	19	29	54	96	96	16	33	56	46	07	80
24	12	26	65	91	27	69	90	64	94	14	84	54	66	72	61	95	87	71	00	90	80	97	57	54
61	19	63	02	31	92	96	26	17	73	41	83	95	53	82	17	26	77	09	43	78	03	87	02	67
30	53	22	17	04	10	27	41	22	02	39	68	52	33	09	10	06	16	88	29	55	98	66	64	85
03	78	89	75	99	75	86	72	07	17	74	41	65	31	66	35	20	83	33	74	87	53	90	88	23
48	22	86	33	79	85	78	34	76	19	53	15	26	74	33	35	66	35	29	72	16	81	86	03	11
60	36	59	46	53	35	07	53	39	49	42	61	42	92	97	01	91	32	83	16	98	95	37	32	31
83	79	94	24	02	56	62	33	44	42	34	99	44	13	74	70	07	11	47	36	09	95	81	80	65
32	96	00	74	05	36	40	98	32	32	99	38	54	16	00	11	13	30	75	86	15	91	70	62	53
19	32	25	38	45	57	62	05	26	06	66	49	76	86	46	78	13	86	65	59	19	64	09	94	13
11	22	09	47	47	07	39	93	74	08	48	50	92	39	29	27	48	24	54	76	85	24	43	51	59
31	75	15	72	60	68	98	00	53	39	15	47	04	83	55	88	65	12	25	96	03	15	21	91	21
88	49	29	93	82	14	45	40	45	04	20	09	49	89	77	74	84	39	34	13	22	10	97	85	08
30	93	44	77	44	07	48	18	38	28	73	78	80	65	33	28	59	72	04	05	94	20	52	03	80
22	88	84	88	93	27	49	99	87	48	60	53	04	51	23	74	02	28	46	17	82	03	71	02	68
78	21	21	69	93	35	90	29	13	86	44	37	21	54	86	65	74	11	40	14	87	48	13	72	20
41	84	98	45	47	46	85	05	23	26	34	67	75	83	00	74	91	06	43	45	19	32	58	15	49
46	35	23	30	49	69	24	89	34	60	45	30	50	75	21	61	31	83	18	55	14	41	37	09	51
11	08	79	62	94	14	01	33	17	92	59	74	76	72	77	76	50	33	45	13	39	66	37	75	44
52	70	10	83	37	56	30	38	73	15	16	52	06	96	76	11	65	49	98	93	02	18	16	81	61
57	27	53	68	98	81	30	44	85	35	68	85	22	73	76	92	85	25	58	66	88	44	80	35	84
20	85	77	31	56	70	28	42	43	26	79	37	59	52	20	01	15	96	32	67	10	62	24	83	91
15	63	38	49	24	90	41	59	36	14	33	52	12	66	65	55	82	34	76	41	86	22	53	17	04
92	69	44	82	97	39	90	40	21	15	59	58	94	90	67	66	82	14	15	75	49	76	70	40	37
77	61	31	90	19	38	15	20	00	30	20	55	49	14	09	96	27	74	82	57	50	81	60	76	16
38	68	83	24	86	45	13	46	35	45	59	40	47	20	59	43	94	75	16	80	43	85	25	96	93
25	16	30	18	89	70	01	41	50	21	41	29	06	73	12	71	85	71	59	57	68	97	11	14	03
65	25	10	76	29	37	23	93	32	95	05	87	00	11	19	92	78	42	63	40	18	47	76	56	22
36	81	54	36	25	18	93	73	75	09	82	44	49	90	05	04	92	17	37	01	14	70	79	39	97
84	39	71	16	92	05	32	78	21	62	20	24	73	17	59	45	19	72	53	32	83	74	52	25	67
04	51	52	56	24	95	09	66	79	46	48	46	08	55	58	15	19	11	87	82	16	93	03	33	61
33	76	16	08	73	43	25	38	41	45	60	33	32	59	83	01	29	14	13	49	20	36	80	71	26
14	38	70	63	45	80	85	40	92	79	43	52	90	63	18	38	38	47	47	61	41	19	63	74	80
51	32	19	22	46	80	08	87	70	74	88	72	25	67	36	66	16	44	94	31	66	91	93	16	78
72	47	20	00	08	80	89	01	80	02	94	81	33	19	00	54	15	58	34	36	35	35	25	41	31
05	46	65	53	06	93	12	81	84	64	74	45	79	05	61	72	84	81	18	34	79	98	26	84	16
39	52	87	24	84	82	47	42	55	93	48	54	53	52	47	18	61	91	36	74	18	61	11	92	41
81	61	61	87	11	53	34	24	42	76	75	12	21	17	24	74	62	77	37	07	58	31	91	59	97
07	58	61	61	20	82	64	12	28	20	92	90	41	31	41	32	39	21	07	63	61	19	96	79	40
90	76	70	42	35	13	57	41	72	00	69	90	26	37	42	78	46	42	25	01	18	62	79	08	72
40	18	82	81	93	29	59	38	86	27	94	97	21	15	98	62	09	53	67	87	00	44	15	89	97
34	41	48	21	57	86	88	75	50	87	19	15	20	00	23	12	30	28	07	83	32	02	46	86	91
63	43	97	53	63	44	98	91	08	22	36	02	40	08	07	76	37	84	16	05	65	96	17	34	88
67	04	90	90	70	93	39	94	55	47	94	45	87	42	84	05	04	14	98	07	20	28	83	40	60
79	49	50	41	46	52	16	29	02	86	54	15	83	42	43	46	97	83	54	82	59	36	29	59	38
91	70	43	05	52	04	73	72	10	31	75	05	19	30	29	47	06	56	43	82	99	78	29	34	78

From Dixon and Massey (1957, p. 368)

The use of randomization procedures in field sampling can serve to avoid bias in sampling the available population, but more important, they help to avoid bias that is variable among the sampling localities. If the available population is substantially different from the target population,however, bias cannot be avoided by any field procedure, and if the available sampling localities differ from the locality populations in different ways and(or) to different degrees, it will be impossible to avoid variable sampling bias. The sampling program might just as well not be undertaken.

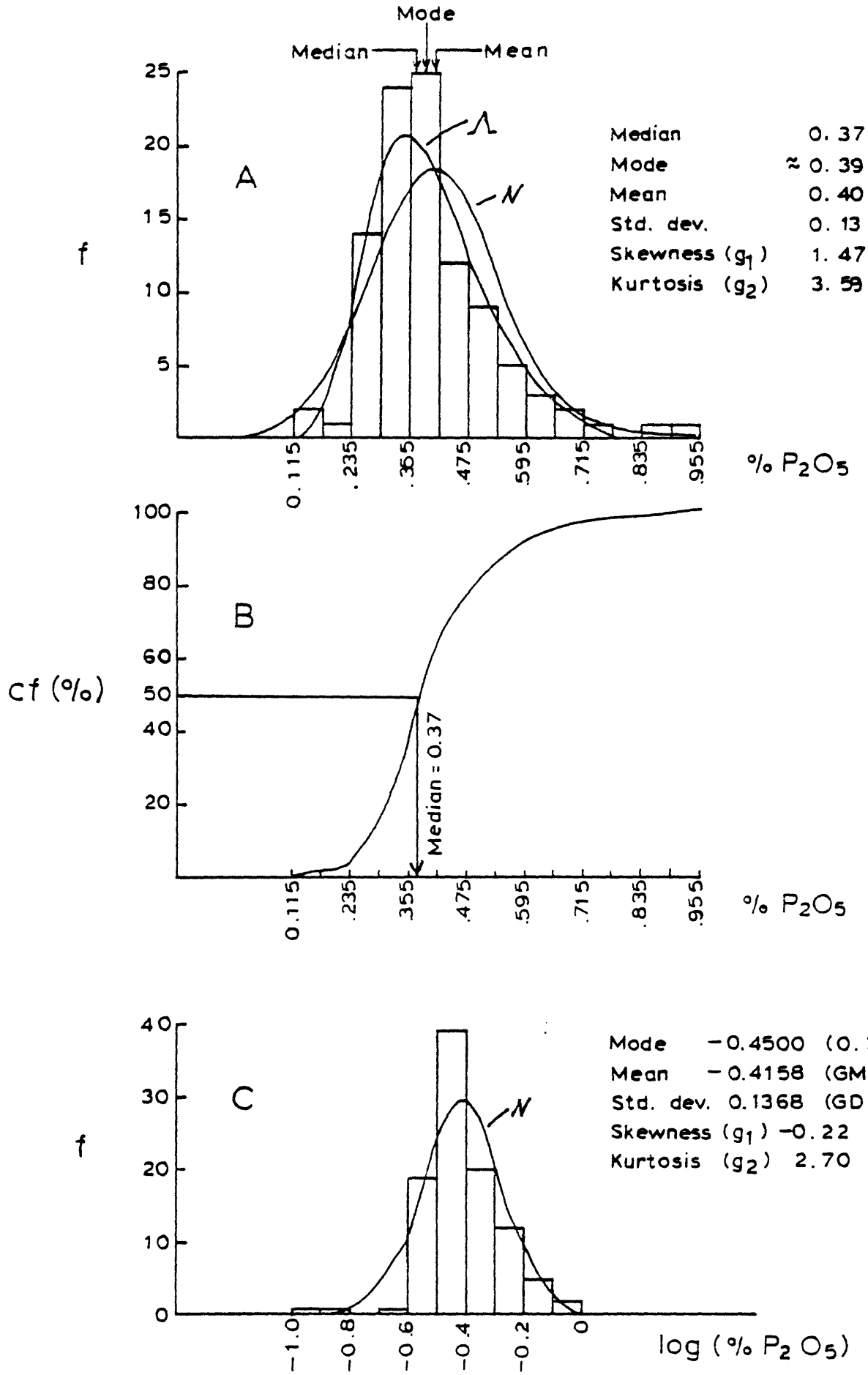
4) Bias from computational procedures.

Bias can result from improper statistical treatment of the analytical data as well as from incorrect laboratory procedures or prejudices in sampling. For example, if a single rock or soil sample is thoroughly homogenized in the laboratory and aliquots of the homogenized sample are analyzed repeatedly, a frequency distribution of geochemical values will be obtained. A question then arises regarding the best estimate of the true concentration of the constituent in the original sample, assuming that the method of analysis is imprecise but totally unbiased. Three choices of the best estimate are immediately apparent: 1) the arithmetic mean, 2) the median, and 3) the mode. If the frequency distribution is asymmetrical, these estimates will all differ and can differ significantly if the asymmetry is pronounced. There is some justification for choosing the mean in that this is the center of gravity for the frequency distribution. However, if the arithmetic mean is the correct value for the sample, and the frequency

distribution of repeated analyses is asymmetrical, it follows that numbers of positive and negative analytical errors are unequal. This would seem improbable for an unbiased method. There is also some justification for accepting the mode of the distribution as the best estimate of the true value in that the mode is the most commonly occurring analytical value among all values derived in the repeated analysis of the sample. If the distribution is asymmetrical, however, the mode, like the arithmetic mean, as the correct value necessitates that the numbers of positive and negative analytical errors are unequal. It is suggested that the median is probably the best estimate of the true value for the sample. If the median is taken as the true value, it follows that the numbers of positive and negative values are equal, even though, if the distribution is asymmetrical, the magnitudes of the positive and negative errors will not be the same. There is good reason to believe that this is actually the situation for all but the dominant chemical constituents in rocks and soils (chiefly SiO_2) because analytical values cannot be negative and so negative errors are restricted in magnitude whereas positive errors are almost unrestricted in this regard. The smaller magnitude of negative errors is to be expected wherever the analytical method is based on observation or measurement of densities or intensities that vary as the logarithm of the concentration.

If the observed frequency distribution of errors is symmetrical, the estimated median and arithmetic mean are the same; if the distribution is normal, the median, mean, and mode are all equal. If the distribution is asymmetrical, the median can be estimated as the detransformed equivalent of the arithmetic mean computed for some transformation that is symmetrically distributed. If the distribution is lognormal, the median can be estimated by the geometric mean (GM); that is, the antilog of the mean logarithm. These relations are important because analysis of variance methods, as well as most other statistical procedures, are based on the premise that the arithmetic mean is the best estimator of the correct value when applied to error distributions. If the distributions are asymmetrical, the statistical methods can induce at least a small bias in the final answers.

The frequency distribution shown in figure 14A was derived from 100 replicate analyses of the same sample for P_2O_5. The sample was of sandstone from the Sawatch Quartzite, of Cambrian age, in central Colorado. About 30 pounds of the sandstone were collected by J. J. Connor and the writer and then thoroughly homogenized in a rotating drum for about 10 hours. The homogenized sample was then split into 100 equal parts with a Jones-type splitter constructed of aluminum. These 100 samples were then randomly interspersed with 400 other randomly ordered samples of sandstone and submitted for analysis. The variability demonstrated by the frequency distribution (fig. 14) is due entirely to analytical imprecision, including the procedures of sample preparation and the extraction of aliquots for actual analysis.


Mode
Median
Mean
A
Λ
N
f
25
20
15
10
5
0.115
.235
.355
.475
.595
.715
.835
.955
% P2O5
Median 0.37
Mode ≈ 0.39
Mean 0.40
Std. dev. 0.13
Skewness (g1) 1.47
Kurtosis (g2) 3.59
B
cf (%)
100
80
60
50
40
20
Median = 0.37
C
40
30
20
10
0
-1.0
-0.8
-0.6
-0.4
-0.2
0
log (% P2 O5)
Mode −0.4500 (0.355 %)
Mean −0.4158 (GM=0.38 %)
Std. dev. 0.1368 (GD=1.37)
Skewness (g1) −0.22
Kurtosis (g2) 2.70

The population median was estimated to be 0.37 percent P_2O_5, as shown in figure 14B. Analysis of variance procedures would be based on the assumption that the best estimate of the true concentration is the arithmetic mean. If the analysis of variance were directed at the original percentage data, the arithmetic mean would be taken as 0.40 percent P_2O_5 (fig. 14A). This value differs from the median because the frequency distribution is asymmetrical ($g_1 = 1.47$). If the analysis of variance is directed at the logarithms of the percentage data, the arithmetic mean is taken as -0.4158. The antilog of this value is the geometric mean, 0.38, and is closer to the median because the distribution of the logarithms is more symmetrical ($g_1 = -0.22$) than that of the original data (fig. 11C). Thus, an analysis of the variance in the logarithmic data would be based on the assumption that the best estimate of the true value for the sample is 0.38 percent P_2O_5. If it is agreed that the median of the distribution is the best estimate of the true value, a log transformation of the data prior to the analysis of variance and prior to other statistical treatment would reduce the bias resulting from the computations.

The considerations above might be trivial if one could be certain that biases introduced by failure to transform were more or less constant across all sampling localities and across all samples. However, if the distribution is asymmetrical, the degree of bias can vary with the amount of the constitutent in the samples or in the sampling localities. If the population distribution is lognormal, for example, the bias is:

Bias = Arithmetic mean - Median

and from table 1:

$$\text{Bias} = \exp(\mu + \sigma^2/2) - \exp(\mu) \quad \text{or}$$

$$\text{Bias} = \exp(\mu)(\exp(\sigma^2/2) - 1)$$

If the variances of the logarithms (σ^2) are homogeneous across all samples and all sampling localities, the bias is proportional to the median. Because examination of the variability among medians (estimated true values) is the very purpose of most investigations, variable bias from failure to transform the data should be expected.

III. Analysis of variance and methods of computation.

The purpose of analysis of variance in geochemical sampling is to estimate the magnitudes of the various sources of variation in the data. For example, if the major source of variation is the laboratory procedure the true compositional variations among the samples may be almost completely masked, and descriptions of the true variations may require either a more precise laboratory method or numerous replicate analyses of each sample. Similarly, if the major source of variation is found to be within sampling localities, the variations among localities might be described only by collecting more samples within localities. In some situations, the variation among localities might be so small that efforts to describe the variations among them might be totally futile.

It was shown in the previous section of the syllabus, in the discussion of analytical errors, that the components of variance are additive if variable biases are absent. The purpose here is to show that if variable biases are absent, the components of variance can

be recovered from the data. First, however, a mathematical explanation will be given of why variances are not additive if variable bias is present in the data. This is important because analysis of variance methods are based on the variance's additive property. Suppose, for example, that we have collected n_β samples by some randomization procedure from n_α sampling localities. Our sampling model is as given in equation (5):

$$x_{ij} = \mu + \alpha_i + \beta_{ij} \quad . \tag{37}$$

If we move the term μ to the left side of the equation and square both sides, we obtain:

$$(x_{ij} - \mu)^2 = \alpha_i^2 + \beta_{ij}^2 + 2\alpha_i\beta_{ij} \quad . \tag{38}$$

The term μ in equation (38) is defined as the true average for the entire region of investigation plus the average bias in selecting sampling localities plus the average bias in selecting samples and in laboratory analysis. It follows then that the terms α_i and β_{ij} have means of zero across all samples, although not necessarily for any specific sampling locality. The next step is critical. If variable bias in sampling and laboratory treatment is absent, the terms α_i and β_{ij} are uncorrelated, and the final term in equation (38) will tend toward zero when the equation is summed across all $n_\alpha n_\beta$ samples. When the nonzero sums are divided through by $n = n_\alpha n_\beta$, we have:

$$\frac{\Sigma(x_{ij} - \mu)^2}{n} = \frac{\Sigma\alpha_i^2}{n} + \frac{\Sigma\beta_{ij}^2}{n}$$

$$\text{or} \qquad \sigma_x^2 = \sigma_\alpha^2 + \sigma_\beta^2 \tag{39}$$

Thus, the variance of x_{ij} is equal to the sum of the variance among sampling localities plus the variance among geochemical values from from within localities. The relationship holds only where α_i and β_{ij} are uncorrelated. The same type of relationship can be shown for sampling models more complex than that in equation (37). All cross-product terms that appear on squaring the equation must tend toward zero as the relations are summed across samples. In other words, variable bias (or correlations among the errors) must be absent.

If the sampling model is as given in equation (37), it is the variances of α_i and β_{ij} that are of prime interest. These variances indicate, respectively, (1) the amount of variation among sampling localities and (2) the amount due to sampling the localities plus laboratory treatment. However, we can never know the individual quantities of α_i and β_{ij}; we see only the values x_{ij}. Nevertheless, the quantities σ_α^2 and σ_β^2 determined the value of σ_x^2 and, because of the experimental design used, can be estimated from the data, x_{ij}. Only to illustrate that this is true, let us suppose that the data are the values of x_{ij} given in expression (36) and that the data comprise the entire population; these data are the values of 15, 12, and 11 from one sampling locality and the values of 24, 23, and 21 from the other. The variance of the first three values is 2.8889 and that of the second three is 1.5555; the average of these variances within the two sampling localities is 2.2222 and is equal to σ_β^2 as given

previously. The two sampling locality means are 12.6666 and 22.6666; the variance among the means is 25, equal to σ^2_α given previously. Thus, the two variance components contained in the total variance, σ^2_x, can be estimated from the geochemical data and knowledge of the sampling design followed in collection of the data.

The computational procedure followed above is correct only in the rare (almost non-existent) situations where the entire population has been sampled. Where less than the total population has been sampled, these procedures are incorrect for two reasons: (1) they do not take into account the degrees of freedom available for estimation of the variance (see previous discussion), and (2) they do not include the necessary correction of the between locality variance for the effect of variance within localities. If the same data are regarded as a fraction of the total population, conventional computational procedures for hierarchical (nested or multi-stage) analysis of variance designs are used. The procedures, applied to the same data, are as follows:

i	j	x_{ij}	$\sum_j x_{ij}$	$\sum_{ij} x_{ij}$
1	1	15		
1	2	12	38	
1	3	11		
				106
2	1	24		
2	2	23	68	
2	3	21		
		2,036	6,068	11,236 (Sums of squared values)

$$SS_1 = (6{,}068/3) = (11{,}236/6) = 150$$

$$SS_2 = (2{,}036/1) - (6{,}068/3) = 13.3333$$

$$V_1 = 150/1 = 150$$

$$V_2 = 13.3333/4 = 3.3333$$

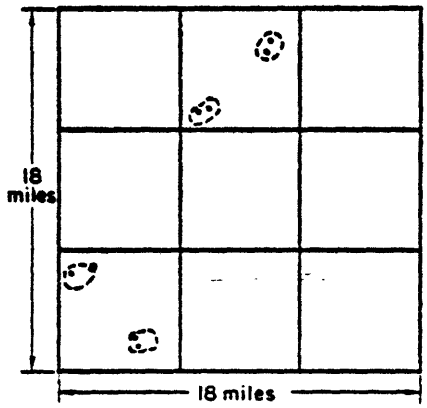
$$\hat{s}^2_\alpha = (150 - 3.3333)/3 = 48.8889$$

$$\hat{s}^2_\beta = 3.3333$$

The values SS_1 and SS_2 are the "sums of squares"; the denominator values within parentheses are the numbers of x_{ij} values contained in the sums that were squared to form the respective numerators. The values V_1 and V_2 are the "mean squares"; the denominators (1 and 4, respectively) are the numbers of degrees of freedom available for estimating each mean square. One degree of freedom is available for estimating V_1 because there are two sampling localities ($n_\alpha - 1$) and 4 degrees of freedom are available for estimating V_2 because there are two sampling localities with two degrees of freedom available from each ($n_\alpha (n_\beta - 1)$). The values $\hat{s}^2_\alpha$ and $\hat{s}^2_\beta$ are the estimated components of variance between and within sampling localities, respectively. It will be noted that they are quite different from the values computed when the data were regarded as comprising the entire geochemical population (25 and 2.2222, respectively).

The procedures followed in the computations above are defined in detail, for the general case, in figure 15 from Krumbein and Slack (1956). They apply equally well to hierarchical models containing any number of terms so long as the numbers of samples (i.e., n_α, n_β, etc.) are equal across all categories at each level of the model.

Fig. 15



$$X_{ijkm} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} + \delta_{ijkm}$$

ANALYSIS OF VARIANCE

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE
BETWEEN SUPERTOWNSHIPS	$SS_1 = \frac{\Sigma_i(\Sigma_{jkm}X_{ijkm})^2}{bcd} - \frac{(\Sigma_{ijkm}X_{ijkm})^2}{abcd=n}$	$a-1$	$V_1 = \frac{SS_1}{a-1}$
BETWEEN TOWNSHIPS WITHIN SUPERTOWNSHIPS	$SS_2 = \frac{\Sigma_{ij}(\Sigma_{km}X_{ijkm})^2}{cd} - \frac{\Sigma_i(\Sigma_{jkm}X_{ijkm})^2}{bcd}$	$a(b-1)$	$V_2 = \frac{SS_2}{a(b-1)}$
BETWEEN MINES WITHIN TOWNSHIPS	$SS_3 = \frac{\Sigma_{ijk}(\Sigma_m X_{ijkm})^2}{d} - \frac{\Sigma_{ij}(\Sigma_{km}X_{ijkm})^2}{cd}$	$ab(c-1)$	$V_3 = \frac{SS_3}{ab(c-1)}$
BETWEEN SAMPLES WITHIN MINES	$SS_4 = \Sigma_{ijkm}(X_{ijkm})^2 - \frac{\Sigma_{ijk}(\Sigma_m X_{ijkm})^2}{d}$	$abc(d-1)$	$V_4 = \frac{SS_4}{abc(d-1)}$
TOTAL	$\Sigma_{ijkm}(X_{ijkm})^2 - \frac{(\Sigma_{ijkm}X_{ijkm})^2}{abcd=n}$	$abcd-1$	

ESTIMATION OF VARIANCE COMPONENTS

LEVELS	DIFFERENCE	SAMPLE SIZE	VARIANCE COMPONENT
SUPERTOWNSHIPS	$V_1 - V_2$	bcd	$S_\alpha^2 = \frac{V_1 - V_2}{bcd}$
TOWNSHIPS	$V_2 - V_3$	cd	$S_\beta^2 = \frac{V_2 - V_3}{cd}$
MINES	$V_3 - V_4$	d	$S_\gamma^2 = \frac{V_3 - V_4}{d}$
SAMPLES	V_4	1	$S_\delta^2 = V_4$

EXPLANATION OF SYMBOLS:

X_{ijkm} = a single observation.
μ = grand mean.
α_i = comp. due to supertwp. i.
β_{ij} = comp. due to twp. j, within supertwp i.
γ_{ijk} = comp. due to mine k, within twp. j, within supertwp. i.
δ_{ijkm} = comp. due to sample m, within mine k, within twp. j within supertwp. i.
$\alpha_i, \beta_{ij}, \gamma_{ijk}$ & δ_{ijkm} are independent with Mean 0 and variances $S_\alpha^2, S_\beta^2, S_\gamma^2$ & S_δ^2 respectively.

i varies from 1 to a
j varies from 1 to b
k varies from 1 to c
m varies from 1 to d
where
a = number of supertwps.
b = number of twps./supertwp.
c = number of mines/twp.
d = number of samples/mine
n = abcd = total samples collected.

From Krumbein and Slack (1956, p. 754)

They would not be applicable, for example, if we had collected three samples from one of the localities and only two from the other. Sampling designs such as this are said to be unbalanced and different computational procedures are required. Analysis of variance procedures for unbalanced sampling designs are given by Anderson and Bancroft (1952, p. 327-330) and have been used extensively in geochemical investigations by the U.S. Geological Survey (USGS STATPAC computer program D0038).

Even though the computed values of s_α^2 and s_β^2 in the previous example are non-zero (suggesting that neither the sampling localities nor the samples within localities are compositionally homogeneous), they are only estimates and so it is possible that the corresponding population values, σ_α^2 and σ_β^2, are, in fact equal to zero. Tests for the likelihood of this possibility are available and can be important. For example, it would be futile to attempt to map the variation among sampling localities if there is a good likelihood that no variation actually exists. The convention test used is based on the F-statistic, which is a ratio of mean squares. For the preceding example:

$$F = V_1/V_2 = (s_\beta^2 + 3s_\alpha^2)/s_\beta^2 = 45 \quad . \qquad (39a)$$

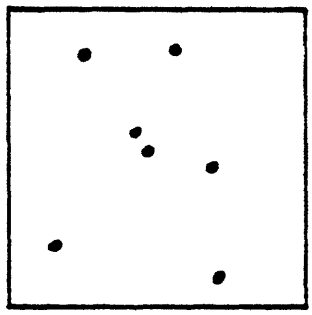
Note that as s_α^2 goes to zero, the computed F goes to unity. Tables of critical values of F for various levels of probability are widely available; the critical values also vary with the degrees of freedom available for estimating the numerator, V_1, and the denominator, V_2. For the example above, these numbers of degrees of freedom are 1 and 4, respectively, and the critical value of F for the 0.01 level

of probability (denoted by $F_{0.01}(1,4)$) is 21.2. Thus, the computed value of F in equation (39a) is significant at the 0.01 level; an estimate of $\hat{\sigma}^2_\alpha = 48.8889$ would be expected less than 1 time out of 100 if σ^2_α were actually zero. The critical value $F_{0.001}(1,4)$, however, is 74.1; that is, computed values of F would be expected to be as large as 74.1 about 1 time in 1,000 if σ^2_α were actually equal to zero. The probability that the computed value of F = 45 arose by chance rather than from real compositional variability among sampling localities is somewhere between 0.01 and 0.001.

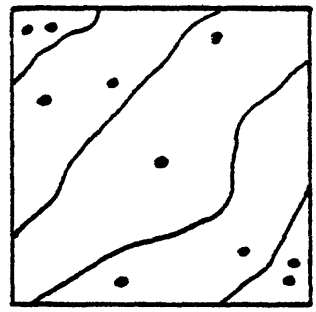
IV. Conventional sampling designs.

A number of conventional sampling designs are described by Cochran (1963) and Mendenhall, Ott, and Scheaffer (1971) who give equations for estimating the population means, variances, and confidence intervals about the means. The more widely used designs are illustrated in figure 16. Simple random sampling (fig. 16A) of a region in geochemistry consists of selecting n sampling points by picking n sets of X-Y coordinates from a table of uniform random numbers (fig. 13). This is the most straightforward type of sampling that could be performed, and the subsequent estimation of the population parameters is the least complicated. Stratified random sampling (fig. 16B) can improve efficiency wherever the population can be divided into subpopulations that are uniform with respect to the variable being studied compared with the variability among them. More efficient sampling implies that the confidence interval about the mean will be smaller for the same number of samples.

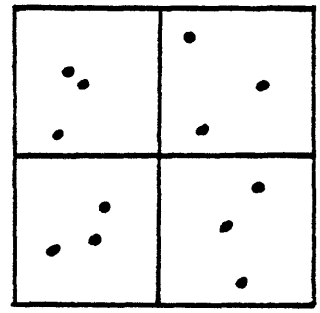
Fig. 16



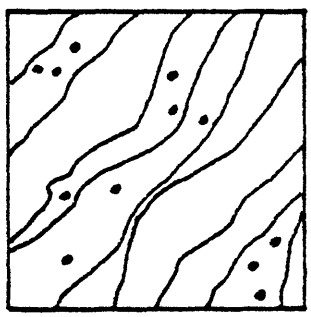
A. Simple random sampling



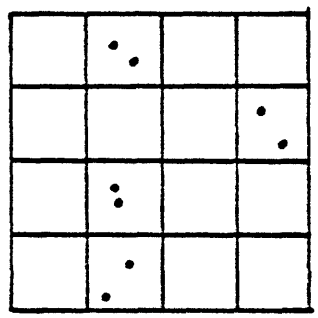
B. Stratified random sampling with natural strata



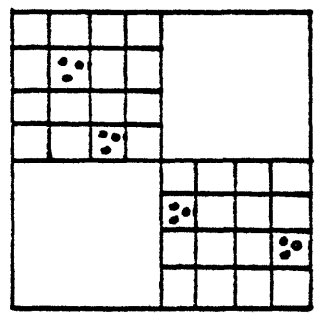
C. Stratified random sampling with artificial strata



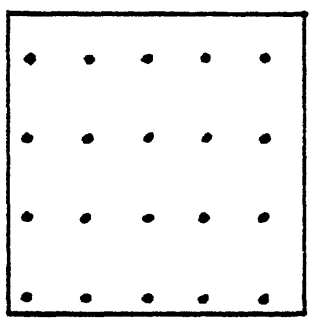
D. Two-stage sampling with natural strata



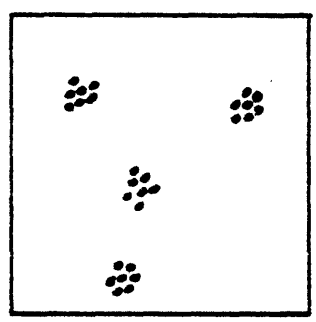
E. Two-stage sampling with artificial strata



F. Three-stage sampling with artificial strata



G. Systematic sampling



H. Cluster sampling

Stratified random sampling is conducted by taking a simple random sample from each (all) of the subpopulations. At the worst, stratified random sampling can be no less efficient than simple random sampling. Thus, if appropriate subpopulations could not be recognized, nothing would be lost by subdividing the population arbitrarily--as in subdividing by some sort of geographic coordinates, as in figure 16C. Doing this would serve to spread the sampling points out more evenly over the region than might be the situation with simple random sampling. A more even spread might offer advantages in examinations of regional variability. If the subpopulations are randomly selected for sampling (i.e., not all of them are sampled) the sampling is according to a two-stage ot two-level design (fig. 16D). That is, subpopulations are selected at random, and then samples are randomly selected from each. Again, the subpopulations may be defined according to geographic coordinates, as in figure 16E. It is also possible to divide the subpopulations into sub-subpopulations, in which case the design would be referred to as three-stage (fig. 16F). In geochemistry, these multi-stage designs are commonly referred to as nested or hierarchical designs with two or more levels.

Systematic sampling (fig. 16G) consists of taking samples at regular intervals determined by the intersections of a square or rectangular grid. The first sampling point is chosen by a randomization procedure, but then all subsequent points are fixed.

Systematic sampling is commonly used in soil sampling and in drilling programs, but is generally impossible where the population is not completely available for sampling due to poor outcrops, or discontinuities in the outline of an ore body. Systematic sampling could lead to bias if any sort of periodic spatial variation were present in the population.

Cluster sampling (fig. 16H) consists of identifying subpopulations and selecting a number of them at random, but differs from two-stage sampling in that the entire subpopulation is sampled. This method is commonly used in survey sampling of people, for example. Households are chosen at random and visited; once at the household, it is almost as easy to get information from all of the individuals who live there as it is to get the information from one of them (a simple random sample) or to sample the individuals randomly (a two-stage sample). If all of the individuals are questioned (or measured), the procedure seems perfectly analogous to procedures used in drilling exploration in situations where the drill hole sites are selected at random or by some procedure that leads to an approximate random selection. If the entire drill core is assayed by dividing it into equal increments, the drill hole may be regarded as a cluster, and the estimation procedures given by Cochran (1963) and by Mendenhall, Ott, and Scheaffer seem perfectly applicable. The consequences of spacing the drill hole sites at equal or otherwise regular intervals, as is commonly done, are not known.

A great deal of geologic and geochemical sampling is not designed at all, especially in geochemical exploration. No reason can be given for this, but it is unfortunate because geochemical exploration is an expensive endeavor and every precaution should be taken to acquire data that are subject to rigorous evaluation.

The sampling associated with drilling exploration is always carefully planned, but the designs used are generally not thought of as belonging to one or more of the conventional types referred to above. If the drill holes are unequally spaced over the ore body, various schemes are used to derive weighting factors for estimating weighted means, as pointed out in an early section of the syllabus. If the holes are equally spaced, as on a rectangular-grid pattern, such weighting is unnecessary.

V. The problem of independence of samples.

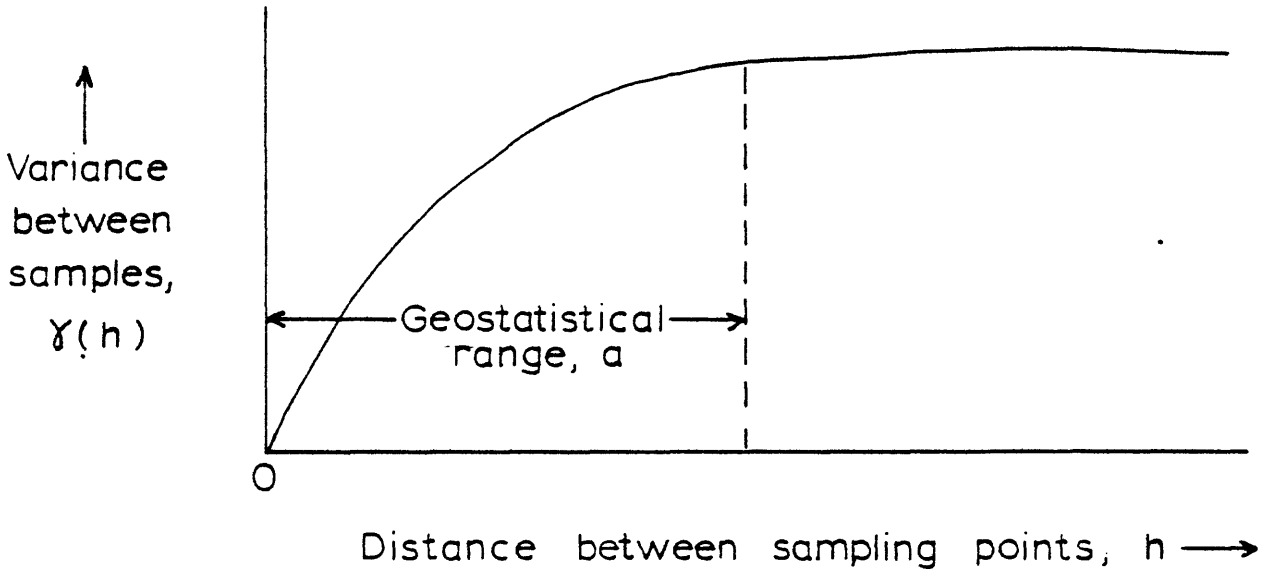
The importance of using randomization procedures in field sampling and laboratory analysis has been stressed in earlier parts of the syllabus. Unless randomization procedures are used, the geochemical values cannot be regarded as independent with any comfortable assurance, and one can never be certain of the degrees of freedom available for the estimation of variances and confidence intervals about the estimated means. In fact, there are some researchers who maintain that geologic samples are not independent even where randomization procedures were used (See remarks of David and Dagbert, 1974, p. 167, concerning the work of Krumbein and Slack, 1956). This attitude is based on the fact that geochemical or assay values are <u>continuous variables</u> and are

spatially ordered. That is, values from samples that were separated in space by less than a certain distance are correlated rather than independent, and only samples from points beyond this range of distance provide independent geochemical values. These researchers seem unaware of the fact that randomly selected samples from an ordered population are, nevertheless, random, just as systematic samples from a randomly ordered population are also random (Cochran, 1963, p. 214). It is possible that the source of confusion is in the interpretation of what is meant by independent samples (observations). Observations are independent when they are unrelated. Correlation among observations arises when the samples are collected in groups or clusters, as when samples are taken close together from more widely separated drill holes. If the samples from one drill hole are high in the measured attribute (e.g., assay value) and those from the other drill hole are comparatively low, the observations will be correlated according to the definition given by Cochran (1963, p. 242). In this situation, the variance estimated from equation (13) will be biased, as will the variance of the mean estimated from equation (19). The reason is that although we may have n_α drill holes with n_β samples from each, we do not have $n = n_\alpha n_\beta$ independent observations even though the drill hole sites and the samples were selected by randomization procedures. If these selections were random, however, analysis of variance procedures could be used to obtain unbiased estimates of the variances within and between drill holes, and equation (23) would give an unbiased estimate of the variance of the grand mean (i.e., the

average grade of the deposit). Ordered populations, analogous to the continuous variables of "geostatistics," are nothing new to conventional statistics.

The question of independence of observations arises in problems other than those of estimating the variance or the confidence intervals about a mean. For example, the probability tests associated with multivariate procedures, for example, require knowledge of the number of degrees of freedom and, consequently, of the number of independent observations. If the samples had been collected according to any procedure other than simple random sampling, especially if they had been collected in any kind of clusters (including drill holes), and if the population is ordered in any way, not all of the samples will be independent. According to the principles of geostatistics (Matheron, 1963), two or more samples are independent only if they were taken from points separated by distances equal to or greater than the geostatistical range. Estimation of this range is based on the fact that samples from a rock body tend to be increasingly different with increasing distance between the points from which they were collected--up to some limiting distance beyond which the relationship disappears. This limiting distance is the geostatistical range and is estimated from a variogram such as that shown in figure 17. Variograms can be estimated from the variance components estimated on the basis of hierarchical sampling designs as well as from the techniques of geostatistics (Miesch, 1975). Knowledge of this range can be useful in determining how sample values can be averaged in

Fig. 17


Variance between samples, γ(h)
Geostatistical range, a
O
Distance between sampling points, h

order to arrive at independent observations. For example, if a multi-level sampling design had been used, wherein each level was determined by the spacing between sampling units, the range would indicate the minimum dimensions of the units that could be averaged in order to obtain average values that could be regarded as independent.

VI. Fundamental properties of geochemical maps.

Two properties of geochemical maps that must be considered before deciding on the sampling design and procedures are resolution and stability. The resolution of the map pertains to the amount of detail--specifically, the distances between adjacent sampling localities. If the localities are closely spaced, small-scale features of the geochemical pattern over a region can be identified, but if they are widely spaced, only the gross features of the pattern may be described. The stability of the geochemical map pertains to its reproducibility--that is, the similarity that would exist among maps derived from subsequent repetitions of the entire experiment, including both sampling and laboratory analysis.

The stability of a geochemical map will depend on the geochemical variation among sampling localities compared with the variance of their means. That is, if the localities are vastly different, the locality means used to construct the geochemical maps need be known only approximately. On the other hand, if the differences among localities are subtle, the locality means must be estimated more precisely. As is shown by equation (19) and by equations of the

type in (23), the variance of a locality mean is determined by the variance within localities and the number of independent samples from each.

Prior to the initial geochemical program, the variances between and within sampling localities can only be guessed at from previous experience, and the results from the initial program may show that the expected stability of the geochemical map was not achieved. In this situation, the initial results can serve to estimate the amount of additional sampling required to obtain any desired degree of map stability. Procedures for doing this are described in the following sections of the syllabus.

VII. Sampling designs for geologic and environmental studies.

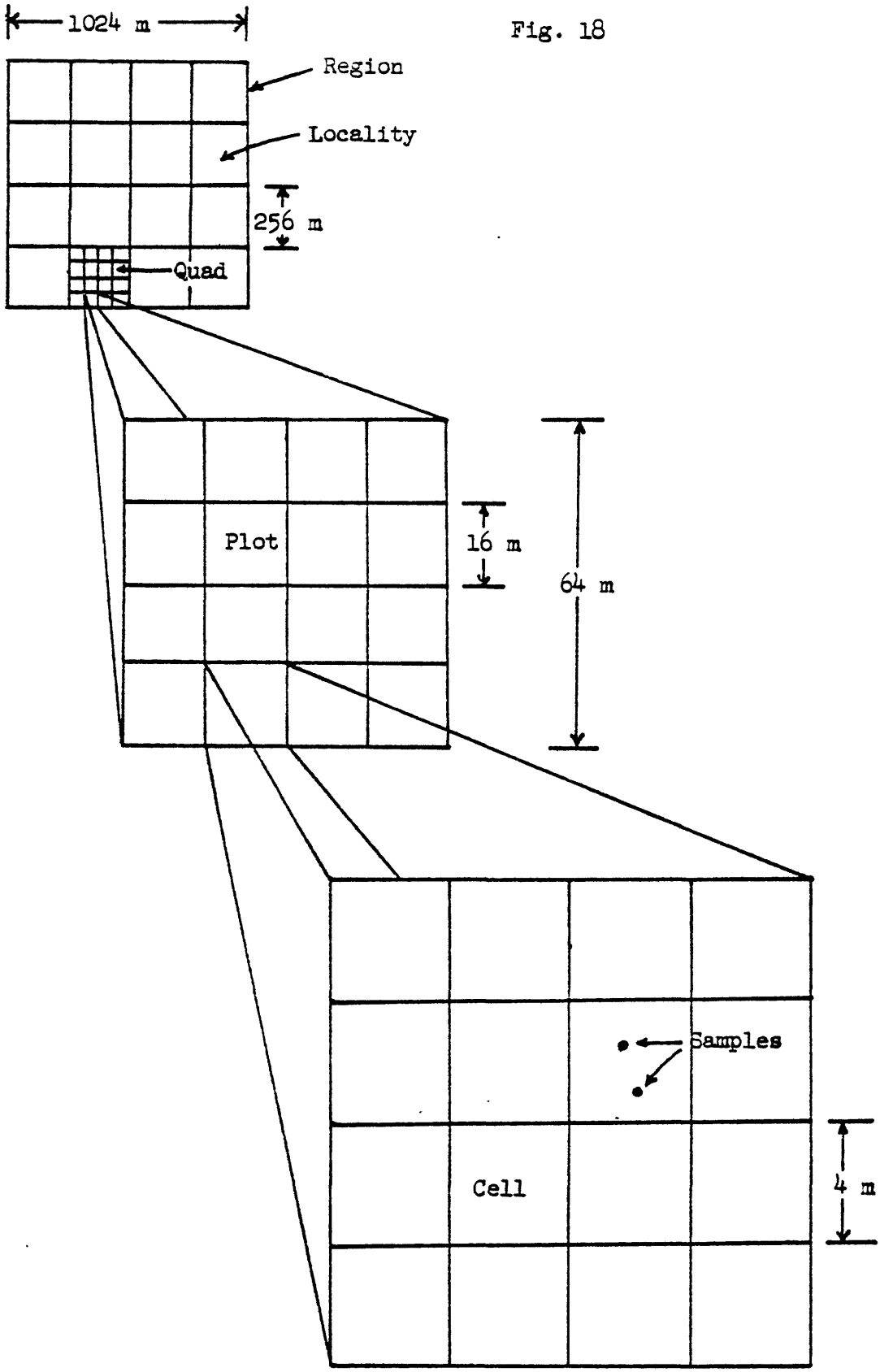
Most of the many varied objectives in geochemical sampling fall into either one of two principal categories: (1) to describe the regional chemical variability throughout a region or within some selected geologic unit, or (2) to detect anomalous features of the geochemical pattern--that is, geochemical "anomalies." The first of these objectives is generally present in geochemical sampling that is part of some broader geologic investigation, such as sampling to describe facies changes in sediments or to define compositional gradients in plutonic rocks or regionally metamorphosed complexes. The first of these objectives is also appropriate in sampling programs designed to describe geochemical environments in support of research in epidemiology or to provide the geochemical baselines necessary in the detection and measurement of chemical

pollution. The second type of objective is characteristic of sampling programs in geochemical exploration and will be treated in the following section of the syllabus.

The principal problems in the design of sampling programs to describe the regional chemical variability are in the selection of spacings between sampling localities (resolution) and in determining the number of samples to collect from each. Actually, neither of these questions can be answered until something is known about the type of regional and local variabilities that are present. A qualitative impression might be obtained by field examination, but it would be difficult to express these observations in terms of variances. The only good way to obtain the information necessary to design an efficient sampling program is to conduct a preliminary geochemical survey.

Suppose that we are sampling soils and that the region was square with 1,024 meters on a side. A regularly shaped region (square or rectangle) is not required, but may be better for purposes of illustration. The region might be divided into 16 square localities, each 256 meters on a side, as shown in figure 18. Then the localities could be subdivided into quadrangles (64 m across), quadrangles divided into plots (16 m across), and plots divided into cells (4 m across). Suppose now that from each of the 16 localities we select two quadrangles at random, from each quadrangle we select two plots at random, from each plot we select two cells at random, and from and from each cell we select two samples at random. With this design

Fig. 18


1024 m
Region
Locality
256 m
Quad
Plot
16 m
64 m
Samples
Cell
4 m

the localities are our master sampling localities, the quadrangles are sublocalities within localities, and so forth. The sampling procedure would lead to 256 samples; we shall assume that quadrangles, plots, cells, and the samples from within cells were selected according to some appropriate randomization procedure so that no bias was introduced in the selection of any of these items. Above all, it will be necessary to assume that if bias was introduced, the bias tended to be constant across all elements of the sampling design--variable bias was absent in the sampling.

Let us suppose that the laboratory procedures to be used in analyzing the samples are new to us and that we have reason to be concerned about the adequacy of the laboratory precision for the task at hand. If we regard the sampling experiment as a pilot one that will probably be followed by a much more detailed, and so much more costly, geochemical survey, it may be advisable to crush each of the 256 samples, homogenize the crushed aggregate, and then split each into two equal parts. We now have 512 samples for laboratory analysis; one-half of the laboratory expenditure will be in the interest of establishing the precision of the laboratory procedures.

Using a table of permuted random numbers similar to that in figure 12, we will place the 512 samples in a randomized sequence and number them from 1 to 512. (Extreme care should be taken to preserve a record of the field sample number corresponding to each random number; the field number should identify the exact position of the sample with respect to the sampling design.) The 512 samples

are then submitted to the laboratories. There will be no way for the laboratory to know which samples are splits of which others, and we are assured of obtaining an unbiased fair estimate of the laboratory precision.

The statistical model on which the sampling is based is:

$$x_{ijklm} = \mu + L_i + Q_{ij} + P_{ijk} + C_{ijkl} + S_{ijklm} + A_{ijklmp} \qquad (40)$$

where L represents the localities, Q represents the quadrangles, P represents the plots, C represents the cells, S represents the samples, and A represents the analyses. Thus, each chemical value, x_{ijklmp}, is regarded as being determined by the grand mean, μ, for the entire region plus a deviation related to the particular locality from which the sample came, the quadrangle, the plot, and so forth. The final term is the deviation of the particular analytical value from the true value for the sample; that is, the analytical error. The total variance among the 512 analytical values will be partitioned as follows:

$$\sigma_x^2 = \sigma_L^2 + \sigma_Q^2 + \sigma_P^2 + \sigma_C^2 + \sigma_S^2 + \sigma_A^2 \qquad (41)$$

On receipt of the analytical data for the 512 samples from the laboratory, the data will be derandomized--that is, ordered by analysis within sample, sample within cell, cell within plot, and so forth. Analysis of variance procedures will then be used to estimate the variance components in equation (41) as:

$$s_x^2 = s_L^2 + s_Q^2 + s_P^2 + s_C^2 + s_S^2 + s_A^2 \qquad (42)$$

Examination of the frequency distribution of the data may suggest that the population frequency distribution is asymmetrical. If so, inasmuch as the terms in the model (equation 40) have zero means and are uncorrelated, it will be reasoned that one or more of the population frequency distributions of the components are asymmetrical also. Some transformation of the data, x_{ijklmp}, will then be sought that yields a frequency distribution that is at least approximately symmetrical. More often than not, for data pertaining to minor chemical constituents, a logarithmic transformation will appear satisfactory. This will be especially true if we find that the means and variances for localities, quadrangles, etc., are correlated. Also, if the population frequency distributions of the components of the model are symmetrical on a log scale, we will have reduced the likelihood of variable bias caused by computational procedures (section II-4 of the syllabus) and will have provided a better basis for tests of the statistical significance of the six sources of variation in the data. In addition, we will be examining proportional rather than absolute variation in the data, as has long been the practice of experienced field geochemists (section I-5 of the syllabus). If a log transformation is used, the left side of the model in equation (40) changes to "$\log x_{ijklmp}$" and all the terms on the right side are in units of logarithms rather than in units of percent or parts per million. Similarly, all of the variances in equations (41) and (42) are variances of log concentration. The antilogarithms of their square roots are geometric deviations.

The estimates of the variance components will provide information regarding the nature of the geochemical profile across the region, and this in turn can be used to design an efficient final sampling plan that will lead to a geochemical map with any desired degree of stability. Our sampling localities for the final sampling program will be defined and spaced according to the type of geochemical profile present--so that the major features of the profile can be described by the map. That is, the resolution required for the map will depend on the type of geochemical profile that is present. If the samples are analyzed for more than one chemical constitutent, we may find that a different final sampling plan is required for each of them. However, this is generally not the case. If the mineralogical constituents of rocks and soils vary highly in concentration on local scales, for example, so will all of the chemical constituents they contain, and if the mineral constituents and their relative proportions tend to be uniform on local scales, their chemical constituents, more often than not, will tend to be uniform also.

The two samples collected within each cell of the sampling design were each collected at randomly selected points that could be as much as 5.6 meters apart ($\sqrt{4^2+4^2}$). According to an equation by Ghosh (1951, p. 24), however, the average distance between sampling points would be $D = 0.521a = 2.08$ meters, where "a" is equal to 4, the dimension of the square cell. The variance component, A_S^2, therefore, will be taken as a measure of all variation in the region that is on scales of two meters or less. That is, any differences between soil

samples from within two meters of each other will contribute to the component, s_S^2. If this component were found to be high in relation to all of the other components, we would conclude that nearly all of the variance in the region is local in character and that the spacings among the sampling localities in the final sampling program would have to be less than two meters (how much less is unknown) in order to describe the major part of the geochemical variation. Any wider sampling interval would result in failure to describe the major variation in the region. A geochemical map of extremely high resolution would be required.

Ghosh's equation applies to points within a square area, and in a strict sense, not for squares within squares. However, as an approximation, we shall assume that the average distance between cells within plots is one-half the dimension of the plot, that the average distance between plots within quadrangles is one-half the dimension of the quadrangle, and so forth. The scales of variation associated with the first five variance components, then, are:

Component	Range of scales (meters)
s_S^2	Less than 2
s_C^2	2 - 8
s_P^2	8 - 32
s_Q^2	32 - 128
s_L^2	greater than 128

The variance component, s_L^2, is affected by variations on scales of 128 to 256 meters, but also by variations among localities; the centers

of localities are separated by distances that range from 256 meters to the total distance across the region of investigation.

If the estimated variance component, s_c^2, is high and all the others are low, we will know that most of the geochemical variability is on scales between two and eight meters. That is, samples from within a cell (two meters apart on the average) tend to be similar as do the average values for plots, quadrangles, and localities. The component, s_c^2, will be large because cell averages tend to differ more than could be expected to have resulted from variability within cells. Also, the observed differences among the averages for plots, quadrangles, and localities will be no greater than might be expected from their uncertainties due to within cell variation. In this situation, the geochemical profile would have a somewhat broader "wave length" than would have been the case if most of the variation had been between samples from within cells. The "wave lengths" would tend to be in the range from two to eight meters. The major part of the geochemical variability could be mapped only by spacing sampling localities at a distance of two meters.

If the variance component, s_L^2, were found to be large in comparison with all of the others, the geochemical profile would be known to be smooth and undulating with "wave lengths" of 128 meters or more. The major part of the geochemical variability could be mapped by spacing sampling localities 128 meters apart.

It should be noted that each quadruple increase in the distance between sampling localities results in a reduction of the number of localities to one-sixteenth. An increase of the sample locality spacing from two meters to eight meters, for example, reduces the number of sampling localities from 262,144 to 16,384, and an increase from eight to 32 meters reduces this to number to 1,024. It is apparent that if most of the geochemical variability is on small scales, the number of samples required may be astronomical, and the intended geochemical survey of the region may be impractical. In this situation, it should be concluded only that no large scale geochemical variation is present and that the geochemical character of the region can be described in terms of statistics that specify a frequency distribution that is applicable for any sub-region within the region as a whole.

On the other hand, if the sampling locality spacing that was found to be appropriate is a practical one in terms of the resources that will be available for the final sampling program, an efficient sampling design can be devised. The first concern will be to construct a geochemical map that is stable--one that shows few, if any, major features that may have resulted from accidents of sampling. The degree of difficulty in the sampling problem will depend on the variation among the sampling localities relative to the variation within them, or by a variance ratio, V:

$$V = \frac{N_V}{D_V} \tag{43}$$

where the numerator, N_V, is the variance among the localities and the denominator, D_V, is the variance within them.

Suppose that it was decided to construct a map that described the geochemical variations on scales of 128 meters or more, so that the centers of the sampling localities were placed 128 meters apart. The number of sampling localities will be 64. The intention would be to describe only the variance that contributed to the component, $\hat{\sigma}^2_L$. The appropriate dimensions of the sampling localities would depend on the scales of the variation within them. For example, if all of the variation within the master sampling localities were on scales of two meters or less, the localities for the final sampling program need be no more than two meters across. However, if variation within the original master sampling localities existed on scales up to 128 meters, the localities for the final sampling program should be of this dimension. In either event, the numerator of the variance ratio is taken as $\hat{\sigma}^2_L$, and the denominator is the sum of the other five components.

The values to be plotted on the geochemical map will be sampling locality means. This is done in the interest of map stability, although at the cost of map resolution. The question is, how many geochemical values are required for each mean in order to achieve the map stability that is desired. The answer can be given in only an approximate way. It is necessary first to define what has been called the variance mean ratio, v_m, as:

$$v_m = \frac{N_v}{D_m} \qquad (44)$$

where N_V is as previously defined and D_m is the variance of the sampling locality means. For this example, D_m is estimated by an obvious extension of equation (23):

$$D_m = \frac{s_Q^2}{n_Q} + \frac{s_P^2}{n_Q n_P} + \frac{s_C^2}{n_Q n_P n_C} + \frac{s_S^2}{n_Q n_P n_C n_S} + \frac{s_A^2}{n_Q n_P n_C n_S n_A} \qquad (45)$$

where n_Q is the number of quadrangles within each locality, n_P is the

number of plots within each quadrangle, and so forth. It will be noted that the variance ratio, V, is determined by the nature of the compositional variability in the rock or soil unit, whereas the variance mean ratio, V_m, can be increased or decreased by changing the number of quadrangles within localities, the number of plots within quadrangles, and so forth. With a sampling plan such as the one used in this example, a more correct estimate of D_m would be obtained by introducing the finite population correction terms as used in equation (23a). It will be seen that if this were done, the five correction terms for the five terms on the right side of equation (45) are, respectively, 0.875, 0.984, 0.998, 1.0, and 1.0, and obviously not highly important. However, as one increased the subscripted values of n in equation (45), the correction terms would become increasingly important. If, for example, n_q were increased from two to eight, by sampling one-half of the 16 quadrangles within each locality, the correction term would become 0.5, and if all 16 quadrangles were sampled, the correction term would be zero, causing the first term on the right side of equation (45) to vanish completely. This is reasonable inasmuch as the selection of quadrangles cannot be a source of error if all quadrangles are sampled.

At this stage in the development of geochemical sampling theory, we are not highly certain just how large V_m should be in order to construct a geochemical map of satisfactory stability; we know for certain only that the larger V_m is, the better. Computer simulation experiments by R. R. Tidball (see Connor and others, 1972, p. 9), and confirmed by the writer, suggest that if V_m is less than about 1.0,

the true geochemical pattern for the region will not be clear from the geochemical map. As v_m is increased above 1.0, the true geochemical pattern becomes increasingly clear from the map of the sample data. Maps based on values of v_m equal to about 3.0 or more appear to reflect the true geochemical patterns very well.

Equations (44) and (45)--with or without the correction terms--provide the basis for determination of the number of samples required per sampling locality and the way they should be spaced. The objective is to increase v_m as much as possible with the smallest number of samples and lowest possible cost in the field. If, for example, the largest term in D_m were s_c^2, an increase in v_m would require an increase in n_q, n_p, or n_c, or increases in all of them. At the extreme, if all were increased to their limit, 16, the first three finite population correction terms would be zero, and the value of D_m would depend only on the final two terms of equation (45). Ordinarily, in a situation like this, such extensive sampling is not required, and D_m can be reduced sufficiently just by increasing n_c. This would result in lower field costs because it would require less time and travel than increasing either n_q or n_p.

If the largest source of error in the sampling locality means were found to be in the selection of quadrangles (i.e., s_q^2 in equation 45 is large), the only way that D_m could be effectively reduced would be to increase n_q. Thus, the field costs for the final sampling program would be larger than otherwise necessary, and there would not be much that could be done about it. If, on the other hand, the major source of

error was found to be in the laboratory procedures (i.e., s_A^2 is large), no additional sampling may be required--a more precise analytical method would have to be found, or the number of analyses per sample, n_A, would have to be increased.

If it was found from the variance components that the resolution of the geochemical map would have to be greater than would be obtained by spacing the sampling localities at intervals of 128 meters, the variance mean ratio, v_m, will have to be different. Say, for example, that s_L^2 is small, containing only a few percent of the total variance, s_χ^2, but that the next component, s_Q^2, were large. The numerator of v_m, then, would be set equal to $s_L^2 + s_Q^2$, or to s_Q^2 only, and the denominator would be:

$$D_m = \frac{s_P^2}{n_P} + \frac{s_C^2}{n_P n_C} + \frac{s_S^2}{n_P n_C n_S} + \frac{s_A^2}{n_P n_C n_S n_A} \qquad (46)$$

The sampling localities would be spaced at intervals of 32 meters, and 1,024 localities would be required. The geochemical map would describe components of the variability on scales greater than 32 meters. Components on smaller scales would be averaged over in computation of the sampling locality means. If the two components, s_L^2 and s_Q^2, comprised, say, 40 percent of the total variance, s_χ^2, the final geochemical map would describe about 40 percent of the total geochemical variation in the region (assuming that the analytical variance is relatively small). The v_m ratio would be increased to any desired value by adjusting the subscripted n's in equation (46).

If the first two components, s_L^2 and s_Q^2, were both of appreciable magnitude and the sampling localities were spaced at 32 meters, a choice must be made as to whether the numerator of v_m should be set

equal to $s_L^2 + s_Q^2$ or simply s_Q^2. If set equal to s_Q^2, the sampling requirements for achieving a v_m ratio of 1.0 or greater will be larger, but the final geochemical map would be more stable and, perhaps, sufficient to describe geochemical variations on all scales greater than 32 meters. If both s_L^2 and s_Q^2 were included in the numerator of v_m, increasing v_m to 1.0 or more would only insure that the gross features of the true geochemical pattern would appear on the final geochemical map.

Knowledge of the variance components may also serve to develop other final sampling designs that are structured differently from the design used in the preliminary survey. For example, if the components s_L^2 and s_C^2 were large with respect to s_Q^2 and s_P^2, the final design might consist of 64 sampling localities spaced 128 meters across, each locality consisting of an area eight meters across. The number of samples required from each locality would be determined from equations (44) and (45); the numerator of v_m would be set to s_L^2 and the denominator would be:

$$D_m = \frac{s_C^2 + s_S^2 + s_A^2}{n} \qquad (47)$$

where n is the number of samples per locality that are to be collected and analyzed. Because s_Q^2 and s_P^2 are small, little of the variability will be lost by confining the dimensions of the sampling localities to eight meters and spacing them 128 meters apart.

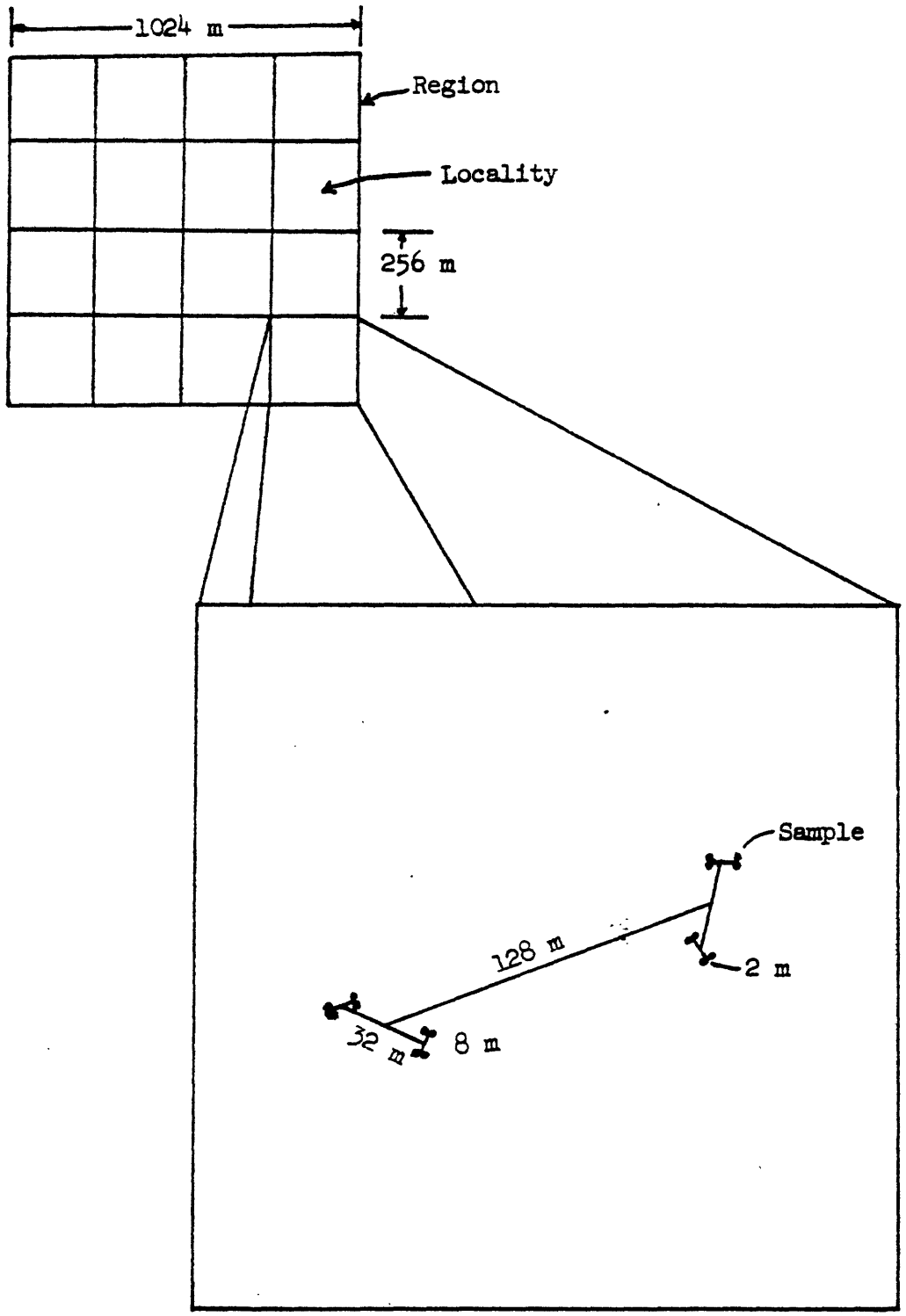
The sampling design outlined in figure 18 is modeled after one used by Krumbein and Slack (1956) in a study of radioactivity in a shale bed that occurs throughout much of the Illinois Basin. A

different design that might accomplish the same objective was used by Youden and Melich (1937) in a study of the variation in soil pH in two small areas of the eastern U.S. A design of the type described by Youden and Melich, could be adapted to the example discussed in this section of the syllabus by using the multiple "barbell" arrangement illustrated in figure 19. All bars of the barbells are oriented in directions selected by randomization procedures, but the lengths of the bars are fixed at 128, 32, eight, and two meters. Samples are collected at the ends of the two-meter bars. All subsequent procedures for analysis of the data and interpretation of the variance components are as previously described, except that finite population correction factors seem inappropriate.

VIII. Sampling designs for geochemical exploration.

The sampling designs previously discussed were established partly to determine an appropriate resolution for the final geochemical map. In most geochemical exploration programs, the resolution is established beforehand and depends on the expected size of the exploration target as well as on field and laboratory resources available. The objective is not to describe as much of the geochemical variability as possible, but to examine differences among the sampling localities and to identify any localities that appear to be geochemically anomalous with respect to the others. The sampling localities may consist of grid cells or small areas centered on grid intersections, especially in geochemical surveys of soils, but more commonly consist of stream segments, stream intersections, streams of a given order, or some other

Fig. 19



division of streams for stream-sediment sampling. If the exploration survey is being directed at rocks, the sampling localities could consist of drill holes, stratigraphic sections, or rock outcrops. In geochemical exploration by sampling of lake sediments, each lake may form a convenient and natural sampling locality. Sampling localities in geochemical exploration can be defined in any manner that is consistent and meaningful for the purpose at hand.

One practical sampling model that can be used in geochemical exploration programs is as given in equation (5):

$$x_{ij} = \mu + \alpha_i + \beta_{ij} \quad . \tag{48}$$

As before, μ represents the grand mean concentration of the constituent in all potential samples from the target population, α_i is the difference between μ and the mean for the i^{th} sampling locality, and β_{ij} is the difference between the mean for the i^{th} locality and the geochemical value for the j^{th} sample from the i^{th} locality. Thus, the term β_{ij} is determined by both natural variation within the i^{th} locality and variance arising from analytical imprecision. The variance of β_{ij}, therefore, is a measure of both sampling and analytical imprecision. It would be possible to partition this variance, s^2_β, into sampling and analytical components by adding another level to the model (see equation 4), but, if the design is to remain balanced, this would mean at least doubling the number of samples to be analyzed. A more practical approach might be to select r of the samples at random, crush each of them and then split each into two parts, and randomly intersperse the

splits in with the other samples for submittal to the laboratory. The variance arising from laboratory procedures could then be estimated from the duplicate analyses, x_1 and x_2, by:

$$s_e^2 = \frac{\Sigma(x_1 - x_2)^2}{2n} \qquad (40)$$

and the variance arising from sampling errors by:

$$s_a^2 = s_\beta^2 - s_e^2 \quad . \qquad (50)$$

On occasion, due to the accidents of randomization in selecting the samples for duplicate analysis, the estimate s_e^2 will be larger than s_β^2. In this situation, s_a^2 is taken as zero. One highly practical reason for estimating s_a^2 and s_e^2 separately is to determine whether the total experimental error, s_β^2, might be reduced by employing a more precise analytical method, without further field work. On the other hand, if s_e^2 is small in comparison with s_a^2, the only way to reduce the total experimental error would be to collect more samples from each of the sampling localities.

The initial goal of a geochemical exploration program is to identify parts of the region being explored that may be geochemically anomalous with respect to the rest of the region. A geochemical anomaly is defined by Hawkes and Webb (1962) as an abnormal geochemical pattern, and by Levinson (1974) as a geochemical measurement that deviates from the norm. For the purpose here, we shall define a geochemical anomaly as a sampling locality mean that differs from other locality means by an increment that is sufficiently large to suggest that processes of mineralization may have contributed to its magnitude. The criterion

used to establish the critical magnitude of this increment will be an empirically derived approximation of the "shortest significant range." The anomalies identified in this manner will not all be related to mineralization. Like anomalies identified in any other manner, some will be false anomalies in the terminology of Levinson (1974).

The procedures recommended for geochemical exploration are not vastly different from those presently in wide use, although some different approaches to the design of the sampling program and evaulation of the data are suggested. The procedures consist of the following steps:

1) The target population is identified and the distribution of its available portion is determined. If the available portion is of very limited distribution, this distribution is shown in some manner on a suitable base map. The population can be defined in any manner whatsoever--such as B-horizon soils, stream sediments (-100 mesh fraction, for example) occurring on the downstream side of boulders, limonite coatings on fractures, and so forth--depending on the judgement of the geologist.
2) A total of n_d sampling localities are defined on the base map and spaced according to the discretion of the geologist and the distribution of the available population. The number of localities, n_d, will depend on the size of the region to be explored, on the expected size of the exploration targets, and on the field and laboratory resources that will be available for the exploration program. The

dimensions of the sampling localities will also depend on the judgement of the geologist. As the dimensions increase (limited only by the spacing between localities), the proportion of the total region actually to be explored increases also. However, at the same time, the tendency to mask anomalies of restricted extent, by inclusion of background material, increases also.

3) A number of samples, n_β, are selected from the available population within each sampling locality according to formal randomization procedures. Unless extreme local variation in the nature of the population is noted, $n_\beta = 2$ will probably be sufficient for this first stage of sampling.

4) After completion of step 3, a total of $n = n_\alpha n_\beta$ samples will be on hand. The number of degrees of freedom available for estimating the variance within localities, and the variance of sampling locality means, will be $df = n_\alpha(n_\beta - 1)$ or $df = n_\alpha$ if $n_\beta = 2$ as suggested in step 3. In order to estimate the contribution of analytical imprecision to the error variance of the means, without adding another level to the design (and thereby at least doubling the required number of analyses), 10 to 20 percent of the n samples are selected at random for duplicate analysis. These selected samples are split, and all samples (originals plus splits) are placed in a randomized sequence using a table of permuted random numbers. Precautions should be taken to be sure that the laboratory will have no way to identify the duplicate splits. The samples are then submitted to the laboratory in the randomized sequence.

5) On receipt of the analytical data from the laboratory, the data are examined to determine whether the differences between the two values for the same locality vary with the locality means. Or, especially where n_β is greater than two, an examination is made to determine whether the locality means and variances are related. If the means and variances are related, log transformation of the analytical data is probably called for. Other transformations that yield independent means and variances for the sampling localities may also be used.

6) Using procedures already described, estimates are made of the variance components and denoted by s_α^2, s_a^2, and s_e^2; these are, respectively, estimates of the variances due to differences among sampling localities, the natural variation within sampling localities, and the laboratory imprecision. The sum of s_a^2 and s_e^2, estimated from analysis of variance procedures, is denoted by s_β^2.

7) An F-test is made to determine whether or not the variance component s_α^2 is significantly different from zero. If s_α^2 is found to be non-significant, then so are the differences among the sampling locality means, and construction of a geochemical map at this point is not advised. The map would not be sufficiently stable to serve any useful purpose. On the other hand, geochemical anomalies might still be recognized by the appearance of no more than a few extreme locality means.

8) If the F-test shows that the variance among sampling localities is statistically significant, however, we know that at least one locality (that with the highest mean) is different from one of the others (that with the lowest mean). We could then eliminate the data from these two localities and perform the analysis of variance and the F-test again, possibly eliminating two additional localities. At some point in this procedure, we would find that the F-test showed no significant differences among the remaining localities. In order to proceed, we would have to return to the region and collect more samples from each locality, thereby increasing the F-test's power--that is, its ability to identify differences among means. We can be reasonably confident that no two localities are precisely the same, and so it is evident that continuation of this procedure would eventually show that all the locality means are different--at least in theory. Realization of this points out the fallacy of trying to develop a purely statistical test that will serve to identify geochemical anomalies. The results of the test will always depend partly on the power of the experiment--specifically the number of independent geochemical values from each sampling locality. A more realistic approach might be to develop some operational procedure that will serve to identify sampling localities that are sufficiently distinct from the others to be referred to as anomalous--localities worthy of further field investigation. One such operational procedure will be described in step (9).

9) When the means of the sampling localities are ordered by magnitude, they either form a rather continuous series with no apparent gaps (i.e., notable differences between adjacent means), or such gaps may appear and suggest that the populations for some of the localities are truly different from the others--and, perhaps, geochemically anomalous. This will be especially suggestive if the apparently anomalous localities are closely spaced or are in parts of the region that appear geologically favorable (as indicated, perhaps, by the presence of fractures, alteration, intrusives, etc.).

The statistical question here is "how large must these 'gaps' be to attract our interest?" If all of the locality population means were actually the same, our estimates of the means would differ due to sampling and analytical errors, and gaps of some magnitude would appear as a result. The expected maximum gap would be some multiple of the standard error of the locality means--that is, of the degree to which the population means were known. If the standard error of the locality means is denoted by $s_{\bar{x}}$ and the multiple by the coefficient, C_p, the maximum gap to expect $100p$ percent of the time if all population means are exactly the same is:

$$SSR = C_p s_{\bar{x}} \qquad (51)$$

If the means are in terms of logarithms (base 10), the gaps between adjacent means in the ordered series would be measured by GM_{i+1}/GM_i where GM_i is the i^{th} geometric mean in the ordered series. In this case, the maximum expectable gap at the p probability level is given by:

$$SSF = 10^{SSR} \qquad (52)$$

We shall refer to SSR as the "shortest significant range" and to SSF as the "smallest significant factor."

The remaining question here is in regard to the appropriate values for C_p. The answer is clear if there are only two locality means--an uncommon situation in exploration geochemistry. With only two means, C_p is equal to $t_p\sqrt{2}$, where t_p is Student's t--that is, SSR is equal to the well-known "least significant difference." Duncan (1955) gives tables, for $p = 0.05$ and $p = 0.01$, for "significant Studentized ranges" that are analogous to C_p, but computer-simulation experiments indicate that they are too large if the number of means, n_α, is greater than two, and that they should not be used in the context of this discussion--or in equation (51) as applied here. The reason is that Duncan's test and a similar test by Tukey are designed to compare each computed mean with every other mean in the set rather than to compare adjacent means in an ordered set. The simulation experiments led to some other values of C_p that can be regarded as empirical approximations of the true unknown values of C_p. The approximations of C_p vary with the number of localities, n_α, the number of samples from each locality, n_β, and the selected probability, p. They are summarized in figure 20 for the case where $n_\beta = 2$ and p is chosen as 0.10, 0.05, and 0.01. A dashed curve gives the values of C_p for $p = 0.05$ and $n_\beta = \infty$ and shows that C_p does not change greatly when n_β is increased. If $n_\beta = 3$, the appropriate curve is about midway between the curves for $n_\beta = 2$

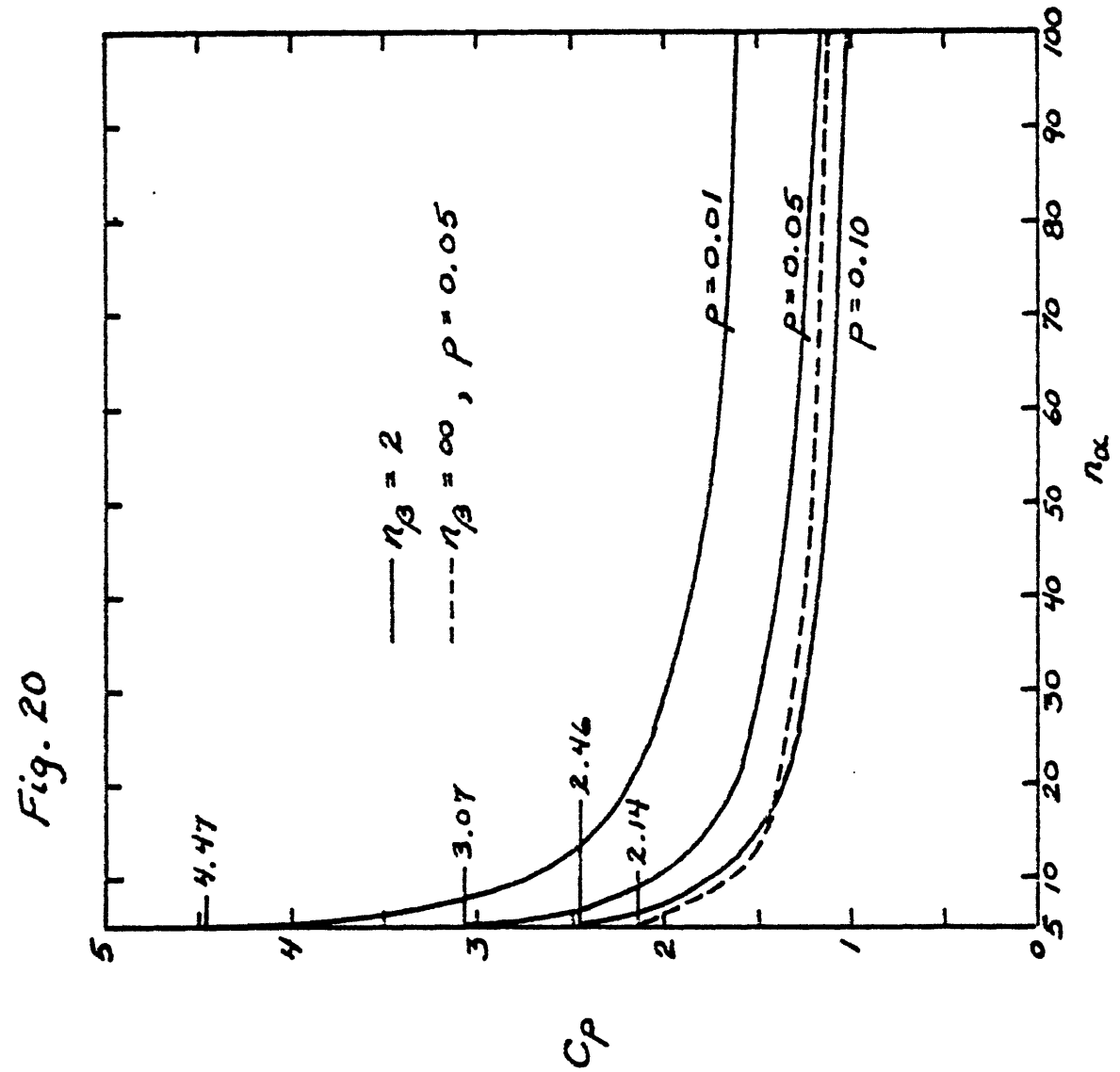


Fig. 20
C_p
n_α
$n_\beta = 2$
$n_\beta = \infty$, $\rho = 0.05$
$\rho = 0.01$
$\rho = 0.05$
$\rho = 0.10$
4.47
3.07
2.46
2.14
0
1
2
3
4
5
10
20
30
40
50
60
70
80
90
100

and $n_\beta = \infty$. If $n_\beta = 10$, the appropriate curve nearly coincides with the curve for $n_\beta = \infty$.

As an illustration of the use of the empirical approximation in the search for a significant gap in a series of ordered means, consider the following seven means--each the average of two:

(1)	(2)	(3)	(4)	(5) ↓	(6)	(7)
0.29	0.30	0.31	0.54	0.67	1.05	1.16

The value of C_p, for $p = 0.05$, from figure 20 is about 2.4. If the standard error of the means was estimated as 0.1, the shortest significant range is:

$$SSR = C_p s_{\bar{x}} = 2.4 \times 0.1 = 0.24 \qquad (53)$$

The vertical arrow in the ordered array of means points to the only gap in the array that is significant at the 0.05 probability level--the sampling localities represented by the 6th and 7th means might warrant further investigation. At least some reason should be sought as to why these localities appear different from the others.

It should be remembered that the test described above suffers from the same shortcoming as the F-test described previously. If the power of the test is repeatedly increased by collecting and analyzing more samples from each locality, eventually it would be possible to place a vertical arrow between each pair of adjacent means, indicating that they are all significantly different from their neighbors in the ordered list. The data available, however, are sufficient to indicate that localities (6) and (7) are different from localities (1) to (5); the difference is sufficiently pronounced to become apparent by collecting

and analyzing only two samples per locality. The available data, on the other hand, provide no justification for referring to any of the localities (1) to (5) as geochemically anomalous.

If the F-test indicated that the variance among sampling locality means, $s_{\bar{x}}^2$, was not significantly different from zero, or if the test, as applied above, failed to identify gaps in the ordered array of locality means, one would have to conclude that none of the sampling localities have been shown to be geochemically anomalous with respect to the others. However, if one increased the power of both tests by collecting additional samples from each locality, so-called anomalous localities might be identified. All of this points to the fact that all degrees of anomalies exist and that each geochemical exploration effort must involve some working definition for an "anomaly." It is suggested that the approximate empirical test is more reasonable than the commonly used criterion that is based on some cut-off or threshold that is equal to some multiple of the average background concentration. The advantage of the empirical approximation is that it varies with the compositional variability within sampling localities and the analytical imprecision. Thus, it considers the errors caused by both sampling and laboratory analysis.

It is also suggested here that the quantity SSR ($C_p s_{\bar{x}}$) or its antilog, SSF, could be used as an appropriate standardized contour interval for geochemical maps--if contours are to be used. Use of this interval would serve to avoid many of the closely-spaced contours that represent non-significant geochemical differences. Alternatively,

each geochemical map should contain the quantity SSR or SSF in the legend or title as an indicator of the stability of the plotted geochemical values.

IX. Suggestions for further reading.

Most textbooks on geostatistical methods give little more than passing reference to the matter of sampling design even though almost all of the statistical methods discussed can be invalidated by improper sampling. It is particularly important that the calculated "degrees of freedom" employed statistical testing be based on the number of independent observations of the variable rather than on merely the number of values that went into the calculated estimate of the variance. The textbooks by Krumbein and Graybill (1965), Griffiths (1967), and Koch and Link (1970) all include chapters on "sampling" and each cites a number of useful references, but the treatments of the subject are far less than is warranted by its importance. A number of papers by Krumbein and his colleagues, however, comprise a highly significant contribution to the matter of sampling in geology and geochemistry; see, in particular, Krumbein (1953, 1955, 1960), Krumbein and Miller (1953, 1954), and Krumbein and Slack (1956). The latter paper has provided a basis for the sampling designs employed in a large number of geochemical surveys conducted by the U.S. Geological Survey. The paper by Krumbein (1960) gives some fundamentally important concepts regarding geological sampling--especially the concepts of target and sampled populations. Some of these concepts, as they apply in geochemical sampling, have been discussed further by Miesch (1967a).

One of the most thorough and statistically rigorous geochemical sampling programs was conducted by A. K. Baird and his colleagues at Pamona College, California. The purpose of the program was to study chemical variations in the Lakeview Mountains pluton of the southern California batholith; the work is described by Baird, McIntyre, Welday, and Morton (1967) and by Morton, Baird, and Baird (1969). A similar study was described by Baird, McIntyre, Welday, and Madlem (1964). These papers are highly recommended.

One of the most extensive applications of statistical methods to the design of geochemical sampling programs has been in a geochemical survey of the State of Missouri. The sampling programs were directed at bedrock, unconsolidated surficial deposits, soils, native vegetation and farm crops, and both surface and ground waters. General descriptions of the sampling designs were given by Connor, Feder, Erdman, and Tidball (1972) and by Miesch (1976). The methods and results for selected soils were given by Tidball (1976); those for selected farm crops and associated soils by Erdman, Shacklette, and Keith (1967a); and those for native vegetation and associated soils by Erdman, Shacklette, and Keith (1976b). Reports from other phases of the survey are in preparation.

A popularized explanation of the application of formal sampling design procedures to geochemical sampling was given by Connor and Myers (1973). Tourtelot and Miesch (1975) gave a general description of formal sampling design procedures for use in environmental geochemistry.

The serious student of sampling will want to consult other references that are not especially geologically oriented. The apparent definitive work on sampling theory is a book by Cochran (1963), although another book by Mendenhall, Ott, and Scheaffer (1971) offers a more elementary treatment and should be extremely useful to beginning students. An excellent introduction to the basic concepts of analysis of variance was given by Tippett (1952), and Bennett and Franklin (1954) gave useful discussions of the application of analysis of variance methods to problems in chemistry. Classical papers on analysis of variance methods are those by Eisenhart (1947) on the assumptions underlying the methods, by Cochran (1947) on the consequences of failure to meet these assumptions, and by Bartlett (1947) on the use of data transformations.

X. Literature cited

Agterberg, F. P., 1974, Geomathematics: Elsevier Scientific Publishing Co., Amsterdam, 596 p.

Aitchison, J., and Brown, J. A. C., 1957, The lognormal distribution: Cambridge at the University Press, 176 p.

Anderson, R. L., and Bancroft, T. A., 1952, Statistical theory in Research: New York, McGraw-Hill Book Co., 399 p.

Baird, A. K., McIntyre, D. B., Welday, E. E., and Madlem, K. W., 1964, Chemical variations in a granitic pluton and its surrounding rocks: Science, v. 146, p. 258-259.

Baird, A. K., McIntyre, D. B., Welday, E. E., and Morton, D. M., 1967, A test of chemical variability and field sampling methods, Lakeview Mountain Tonalite, Lakeview Mountains, southern California batholith: California Div. Mines and Geology Spec. Rept. 92, p. 11-19.

Bartlett, M. S., 1947, The use of transformations; Biometrics, v. 3, no. 1, p. 39-52.

Bennett, C. A., and Franklin, N. L., 1954, Statistical analysis in chemistry and the chemical industry: New York, John Wiley & Sons, Inc., 724 p.

Blais, R. A., and Carlier, P. A., 1968, Applications of geostatistics in ore evaluation: Canadian Inst. Mining and Metallurgy, Spec. Vol. 9, p. 41-68.

Cochran, W. G., 1947, Some consequences when the assumptions for the analysis of variance are not satisfied: Biometrics, v. 3, no. 1, p. 22-38.

Cochran, W. G., 1963, Sampling Techniques: New York, John Wiley & Sons, Inc., 413 p.

Cohen, A. C., Jr., 1959, Simplified estimators for the normal distribution when samples are singly censored or truncated: Technometrics, v. 1, no. 3, p. 217-237.

_____, 1961, Tables for maximum likelihood estimates; singly truncated and singly censored samples: Technometrics, v. 3, no. 4, p. 535-541.

Connor, J. J., Feder, G. L., Erdman, J. A., and Tidball, R. R., 1972, Environmental geochemistry in Missouri - a multidisciplinary study, in Earth sciences and the quality of life - Symposium 1: Internat. Geol. Congress, 24th, Proc., p. 7-14.

Connor, J. J., and Myers, A. T., 1973, How to sample a mountain: Am. Soc. for Testing and Materials, Spec. Tech. Pub. 540, p. 30-36.

Chayes, Felix, 1960, On correlation between variables of constant sum: Jour. Geophysical Res., v. 65, no. 12, p. 4185-4193.

_____, 1962, Numerical correlation and petrographic variation: Jour. Geology, v. 70, no. 4, p. 440-452.

Chayes, Felix, and Kruskal, W., 1966, An approximate statistical test for correlations between proportions: Jour. Geology, v. 74, no. 5, pt. 2, p. 692-702.

David, M., 1969, The notion of "extension variance" and its application to the grade estimation of stratiform deposits, in Weiss, Alfred, ed., A decade of digital computing in the mineral industry: New York, Am. Inst. Mining, Metallurgy, and Petroleum Engineers, p. 63-81.

_____, 1970, Geostatistical ore estimation - A step-by-step case study: Canadian Inst. Mining and Metallurgy, Spec. Vol. 12, p. 185-191.

David, M., and Dagbert, M., 1975, Lakeview revisited - Variograms and correspondence analysis - New tools for the understanding of geochemical data, in Ellicott, I. L., and Fletcher, W. K., eds., Geochemical Exploration 1974, Developments in economic geology [v.] 1: Internat. Geochem. Explor. Symposium, 5th, Vancouver, British Columbia, Canada, 1974, Proc., p. 163-181.

Davis, J. C., 1973, Statistics and data analysis in geology: New York, John Wiley & Sons, Inc., 550 p.

Dixon, W. J., and Massey, F. J., Jr., 1957, Introduction to statistical analysis: New York, McGraw-Hill Book Co., Inc., 488 p.

Duncan, D. B., 1955, Multiple range and multiple F tests: Biometrics, v. 11, no. 1, p. 1-42.

Ebens, R. J., Erdman, J. A., Feder, G. L., Case, A. A., and Selby, L. A., 1973, Geochemical anomalies of a claypit area, Callaway County, Missouri, and related metabolic imbalance in beef cattle: U. S. Geol. Survey Prof. Paper 807, 24 p.

Eisenhart, Churchill, 1947, The assumptions underlying the analysis of variance: Biometrics, v. 3, no. 1, p. 1-21.

Erdman, J. A., Shacklette, H. T., and Keith, J. R., 1976a, Elemental composition of selected native plants and associated soils from major vegetation-type areas in Missouri - Geochemical survey of Missouri: U. S. Geol. Survey Prof. Paper 954-C, 87 p.

_____, 1976b, Elemental composition of corn grains, soybean seeds, pasture grasses, and associated soils from selected areas in Missouri - Geochemical survey of Missouri: U. S. Geol. Survey Prof. Paper 954-D, 23 p.

Ghosh, Birendranath, 1951, Random distances within a rectangle and between two rectangles: Calcutta Mathematical Soc. Bull., v. 43, p. 17-24.

Govett, G. J. S., Goodfellow, W. D., Chapman, R. P., and Chork, C. Y., 1975, Exploration geochemistry - distribution of elements and recognition of anomalies: Jour. Internat. Assoc. for Mathematical Geology, v. 7, nos. 5/6, p. 415-446.

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

Hawkes, H. E., and Webb, J. S., 1962, Geochemistry in mineral exploration: New York, Harper & Row, Publishers, 415 p.

Hubaux, A., and Smiriga-Snoeck, N., 1964, On the limit of sensitivity and the analytical error: Geochemica et Cosmochemica Acta, v. 28, no. 7, p. 1199-1216.

Kendall, M. G., and Stuart, Alan, 1961, The advanced theory of statistics, v. 2, Inference and relationship: New York, Hafner Publishing Co., 676 p.

Koch, G. S., and Link, R. F., 1970, Statistical analysis of geologic data, v. 1: New York, John Wiley & Sons, Inc., 375 p.

Krumbein, W. C., 1953, Statistical designs for sampling beach sand: Trans. Am. Geophys. Union, v. 34, p. 857-868.

_____ , 1955, Experimental design in the earth sciences: Trans. Am. Geophys. Union, v. 36, p. 1-11.

_____ , 1960, The "geological population" as a framework for analyzing numerical data in geology: Liverpool and Manchester Geol. Jour., v. 2, pt. 3, p. 341-368.

Krumbein, W. C., and Miller, R. L., 1953, Design of experiments for statistical analysis of geological data: Jour. Geology, v. 61, no. 6, p.510-532.

_____ , 1954, A note on transformation of data for analysis of variance Jour. Geology, v. 62, no. 2, p. 192-193.

Krumbein, W. C., and Slack, H. A., 1956, Statistical analysis of low-level radioactivity of Pennsylvanian black fissile shale in Illinois: Geol. Soc. America Bull., v. 67, no. 6, p. 739-761.

Krumbein, W. C., and Graybill, F. A., 1965, An introduction to statistical models in geology: New York, McGraw-Hill Book Co., 475 p.

Levinson, A. A., 1974, Introduction to exploration geochemistry: Calgary, Applied Publishing Ltd., 612 p.

Link, R. F., and Koch, G. S., 1975, Some consequences of applying lognormal theory to pseudolognormal distributions: Jour. Internat. Assoc. for Mathematical Geology, v. 7, no. 2, p. 117-128.

Matheron, G., 1963, Principles of geostatistics: Econ. Geology, v. 58, p. 1246-1266.

Mendenhall, William; Ott, Lyman, and Scheaffer, R. L., 1971, Elementary survey sampling: Belmont, Calif., Wadsworth Publishing Co., Inc., 247 p.

Miesch, A. T., 1967a, Theory of error in geochemical data: U. S. Geol. Survey Prof. Paper 574-A, 17 p.

_____ , 1967b, Methods of computation for estimating geochemical abundance: U. S. Geol. Survey Prof. Paper 574-B, 15 p.

_____ , 1969, The constant sum problem in geochemistry, in Merriam, D. F., ed., Computer applications in the earth sciences: New York, Plenum Press, p. 161-176.

_____ , 1975, Variograms and variance components in geochemistry and ore evaluation, in Whitten, E. H. T., ed., Quantitative studies in the geological sciences: Geol. Soc. America, Memoir 142, p.333-340.

_____ , 1976, Methods of sampling, laboratory analysis, and statistical reduction of data - Geochemical survey of Missouri: U. S. Geol. Survey Prof. Paper 954-A, 39 p.

Miesch, A. T., Chao, E. C. T., and Cuttitta, Frank, 1966, Multivariate analysis of geochemical data on tektites: Jour. Geology, v. 74, no. 5, pt. 2, p.673-691.

Morton, D. M., Baird, A. K., and Baird, K. W., 1969, The Lakeview Mountains pluton, southern California batholith, Part II: Chemical composition and variation: Geol. Soc. America Bull., v. 80, no.8, p. 1553-1563.

Olea, R. A., 1972, Application of regionalized variable theory to automatic contouring: Univ. Kansas Center for Research, Inc., Spec. Rept. to Am. Petroleum Inst., Research Proj. 131, 191 p.

Rankama, Kalervo, and Sahama, Th. G., 1950, Geochemistry: Chicago, The Univ. of Chicago Press, 912 p.

Shaw, D. M., 1961, Manipulation errors in geochemistry: A preliminary study: Royal Soc. Canada Trans., v. 55, ser. 3, sec. 4, p. 41-55.

Sichel, H. S., 1952, New methods in the statistical evaluation of mine sampling data: London, Inst. Mining Metallurgy Trans., v. 61, p. 261-288.

_____ , 1966, The estimation of means and associated confidence limits for small samples from lognormal populations: So. African Inst. Mining and Metall. Jour., preprint no. 4, 17 p.

Tidball, R. R., 1976, Chemical variation of soils in Missouri associated with selected levels of the soil classification system - Geochemical survey of Missouri: U. S. Geol. Survey Prof. Paper 954-B, 16 p.

Tippett, L. H. C., 1952, The methods of statistics: New York, Dover Publications, Inc., 395 p.

Tourtelot, H. A., and Miesch, A. T., 1975, Sampling designs in environmental geochemistry, in Freedman, Jacob, ed., Trace element geochemistry in health and disease: Geol. Soc. America, Spec. Paper 155, p. 107-118.

U. S. Geological Survey, 1972, Computer program documentation, Analysis of variance, STATPAC program no. D0038: Unpublished, 7 p. (Available from Computer Center Division, U. S. Geological Survey, Reston, Va., 22092).

Youden, W. J., and Melich, A., 1937, Selection of efficient methods for soil sampling: Boyce Thompson Inst. Contr., v. 9, p. 59-70.

APPENDIX

TABLE A. RANDOM NORMAL NUMBERS, $\mu = 0$, $\sigma = 1$*

01	02	03	04	05	06	07	08	09	10
0.464	0.137	2.455	−0.323	−0.068	0.296	−0.288	1.298	0.241	−0.957
0.060	−2.526	−0.531	−0.194	0.543	−1.558	0.187	−1.190	0.022	0.525
1.486	−0.354	−0.634	0.697	0.926	1.375	0.785	−0.963	−0.853	−1.865
1.022	−0.472	1.279	3.521	0.571	−1.851	0.194	1.192	−0.501	−0.273
1.394	−0.555	0.046	0.321	2.945	1.974	−0.258	0.412	0.439	−0.035
0.906	−0.513	−0.525	0 595	0.881	−0.934	1.579	0.161	−1.885	0.371
1.179	−1.055	0.007	0.769	0.971	0.712	1.090	−0.631	−0.255	−0.702
−1.501	−0.488	−0.162	−0.136	1.033	0.203	0.448	0.748	−0.423	−0.432
−0.690	0.756	−1.618	−0.345	−0.511	−2.051	−0.457	−0.218	0.857	−0.465
1.372	0.225	0.378	0.761	0.181	−0.736	0.960	−1.530	−0.260	0.120
−0.482	1.678	−0.057	−1.229	−0.486	0.856	−0.491	−1.983	−2.830	−0.238
−1.376	−0.150	1.356	−0.561	−0.256	−0.212	0.219	0.779	0.953	−0.869
−1.010	0.598	−0.918	1.598	0.065	0.415	−0.169	0.313	−0.973	−1.016
−0.005	−0.899	0.012	−0.725	1.147	−0.121	1.096	0.481	−1.691	0.417
1.393	−1.163	−0.911	1.231	−0.199	−0.246	1.239	−2.574	−0.558	0.056
−1.787	−0.261	1.237	1.046	−0.508	−1.630	−0.146	−0.392	−0.627	0.561
−0.105	−0.357	−1.384	0.360	−0.992	−0.116	−1.698	−2.832	−1.108	−2.357
−1.339	1.827	−0.959	0.424	0.969	−1.141	−1.041	0.362	−1.726	1.956
1.041	0.535	0.731	1.377	0.983	−1.330	1.620	−1.040	0.524	−0.281
0.279	−2.056	0.717	−0.873	−1.096	−1.396	1.047	0.089	−0.573	0.932
−1.805	−2.008	−1.633	0.542	0.250	−0.166	0.032	0.079	0.471	−1.029
−1.186	1.180	1.114	0.882	1.265	−0.202	0.151	−0.376	−0.310	0.479
0.658	−1.141	1.151	−1.210	−0.927	0.425	0.290	−0.902	0.610	2.709
−0.439	0.358	−1.939	0.891	−0.227	0.602	0.873	−0.437	−0.220	−0.057
−1.399	−0.230	0.385	−0.649	−0.577	0.237	−0.289	0.513	0.738	−0.300
0.199	0.208	−1.083	−0.219	−0.291	1.221	1.119	0.004	−2.015	−0.594
0.159	0.272	−0.313	0.084	−2.828	−0.439	−0.792	−1.275	−0.623	−1.047
2.273	0.606	0.606	−0.747	0.247	1.291	0.063	−1.793	−0.699	−1.347
0.041	−0.307	0.121	0.790	−0.584	0.541	0.484	−0.986	0.481	0.996
−1.132	−2.098	0.921	0.145	0.446	−1.661	1.045	−1.363	−0.586	−1.023
0.768	0.079	−1.473	0.034	−2.127	0.665	0.084	−0.880	−0.579	0.551
0.375	−1.658	−0.851	0.234	−0.656	0.340	−0.086	−0.158	−0.120	0.418
−0.513	−0.344	0.210	−0.736	1.041	0.008	0.427	−0.831	0.191	0.074
0.292	−0.521	1.266	−1.206	−0.899	0.110	−0.528	−0.813	0.071	0.524
1.026	2.990	−0.574	−0.491	−1.114	1.297	−1.433	−1.345	−3.001	0.479
−1.334	1.278	−0.568	−0.109	−0.515	−0.566	2.923	0.500	0.359	0.326
−0.287	−0.144	−0.254	0.574	−0.451	−1.181	−1.190	−0.318	−0.094	1.114
0.161	−0.886	−0.921	−0.509	1.410	−0.518	0.192	−0.432	1.501	1.068
−1.346	0.193	−1.202	0.394	−1.045	0.843	0.942	1.045	0.031	0.772
1.250	−0.199	−0.288	1.810	1.378	0.584	1.216	0.733	0.402	0.226
0.630	−0.537	0.782	0.060	0.499	−0.431	1.705	1.164	0.884	−0.298
0.375	−1.941	0.247	−0.491	0.665	−0.135	−0.145	−0.498	0.457	1.064
−1.420	0.489	−1.711	−1.186	0.754	−0.732	−0.066	1.006	−0.798	0.162
−0.151	−0.243	−0.430	−0.762	0.298	1.049	1.810	2.885	−0.768	−0.129
−0.309	0.531	0.416	−1.541	1.456	2.040	−0.124	0.196	0.023	−1.204
0.424	−0.444	0.593	0.993	−0.106	0.116	0.484	−1.272	1.066	1.097
0.593	0.658	−1.127	−1.407	−1.579	−1.616	1.458	1.262	0.736	−0.916
0.862	−0.885	−0.142	−0.504	0.532	1.381	0.022	−0.281	−0.342	1.222
0.235	−0.628	−0.023	−0.463	−0.899	−0.394	−0.538	1,707	−0.188	−1.153
−0.853	0.402	0.777	0.833	0.410	−0.349	−1.094	0.580	1.395	1.298

*From the RAND Corporation, as reproduced in Dixon and Massey (1957).

Exercise 1.

a) Draw a sample, x_i ($i = 1,10$), from N(100,400).

b) Estimate the population mean, μ, and variance, σ^2.

c) Estimate the standard error of the mean from

$$s_{\bar{x}} = \sqrt{s^2/n} \qquad n = 10$$

d) Given the values of $\bar{x}$ from all members of the class, compute the standard deviation of $\bar{x}$. Compare with $s_{\bar{x}}$ from (c).

Note: The equation $s^2 = \dfrac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}{n-1}$ is algebraically equal to equation (13) in the syllabus and is easier for purposes of computation.

Exercise 2.

a) Draw a sample, x_i (i = 1, 10), from $\Lambda(2.1250, 0.3025)$. Note: Parameters of Λ are logs - base 10.

b) Estimate the population mean and variance of y, where $y = \log_{10}$.

c) Estimate the standard error of the mean log.

d) Given the values of $\bar{y}$ from all members of the class, compute the standard deviation of $\bar{y}$. Compare with $s_{\bar{y}}$ from (c).

e) Compute the ranges $\bar{y} - s_y$ to $\bar{y} + s_y$ and $\bar{y} - 2s_y$ to $\bar{y} + 2s_y$. These limits are in logs; obtain the antilogs (ppm values).

f) Using the answers from (b), compute the geometric mean (GM) and geometric deviation (GD). Now compute the ranges GM/GD to GMxGD and GM/GD^2 to $GMxGD^2$. Compare with limits from (e).

Exercise 3.

a) Using the formulae* given in table 1 and the parameters for the lognormal distribution used in exercise 2 ($\mu = 2.1250$, $\sigma^2 = 0.3025$), compute the true arithmetic mean and standard deviation for the distribution.

b) Now take the antilogs of the 10 log values obtained in exercise (2a) and, from these, estimate the arithmetic mean and standard deviation in the usual manner. Compute the limits for the range from $\bar{x} - a$ to $\bar{x} + a$.

c) Now derive Sichel's t-estimator of the arithmetic mean. Use your results from erercise (2f).

*Recall that the formulae in table 1 use the mean and variance of logs to the base e. The conversion factors for the mean and variance are, respectively, 2.30259 and 5.30190.

$2.1250 \times 2.30259 = 4.8930$
$0.3025 \times 5.30190 = 1.6038$

logs base 10 — logs base e

Exercise 4.

Using the 10 antilog (ppm) values from exercise (3), assume that the lower limit of analytical determination was 100 ppm and exclude all values lower than this amount. What is the detection ratio ? Now estimate the population geometric mean and geometric deviation by the methods of Cohen. Compare these estimates with those obtained in exercise (2f) where all 10 values were used.

Exercise 5.

a) Draw two samples, X_i and Y_i (i=1,10), from N(2,0.25) and N(1,0.49), respectively. Form 10 values of Z_i from $Z_i = X_i + Y_i$.

b) Estimate the variances of X_i, Y_i, and Z_i. Note that the sum of the first two variances is not precisely equal to the third. Why ?

c) Estimate the covariance for X_i and Y_i. Double this and add it to the sum of the variances of X_i and Y_i. The result should be equal to the variance of Z_i.

d) Estimate the correlation between X_i and Y_i.

Exercise 6.

a) Suppose that we have selected 5 samples from a rock unit using randomization procedures and that the "true" values for the 5 samples are as listed under T_i ($i = 1,5$) below. Compute the variance of the 5 values.

b) Now suppose that each sample was analyzed twice and that the analytical errors were as given under E_{ij} below. Note that the variances of the two errors differ from one sample to the next. Compute the error variance for each sample and then the average error variance.

c) The values under X_{ij} below simulate the analytical data that results from the "true" values plus the analytical errors. Perform an analysis of variance to estimate the variance among samples and the analytical error variance. Note that the error variance is the same as the average obtained in part (b) above.

d) Note that the variance among the samples estimated by analysis of variance is less than that obtained in part (a) above. Subtract one-half of the analytical error variance from the latter value. Now the two estimates of variance among the 5 samples should agree. Why does this work ?

i	j	T_i	E_{ij}	X_{ij}
1	1	12	+3	15
	2		-1	11
2	1	15	+5	20
	2		-3	12
3	1	28	-4	24
	2		+6	34
4	1	20	+2	22
	2		0	20
5	1	42	+7	49
	2		-5	37

Note that the mean error is equal for all samples. This insures that the covariance for T_i and E_{ij} is exactly zero.

Exercise 7.

a) Draw two samples, F_{i1} and F_{i2} ($i = 1,10$), from N(0,1). Now given the equations:

$$X_i' = 0.707F_{i1} + 0.707F_{i2} \quad \text{and}$$

$$Y_i' = 0.949F_{i1} + 0.316F_{i2}$$

compute 10 values each of X_i' and Y_i'.

b) Generate 10 values of X_i from

$$X_i = 3X_i' + 2$$

and 10 values of Y_i from

$$Y_i = 2Y_i' + 4$$

The 10 pairs of X_i and Y_i are from populations $N(2,3^2)$ and $N(4,2^2)$, respectively. The bivariate population correlation coefficient is:

$$(0.707 \times 0.949) + (0.707 \times 0.316) = +0.894$$

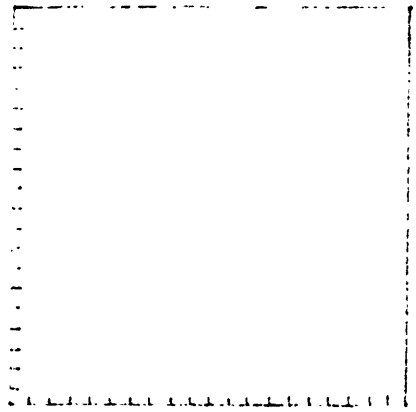
c) Estimate the means and variances of X_i and Y_i from the 10 pairs of values, and the correlation coefficient.

d) About two-thirds of the estimates of the correlation coefficient, from all members of the class, should be in the range from 0.786 to 0.949. Are they ?

Exercise 8.

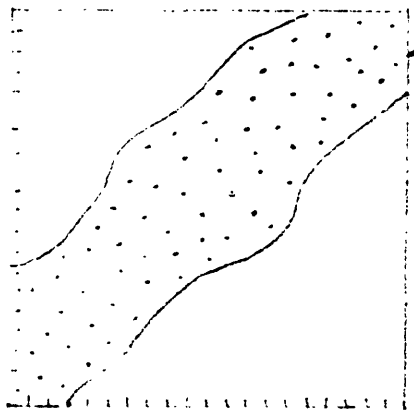
a) Select 10 randomly located points within area "A".

Area "A"



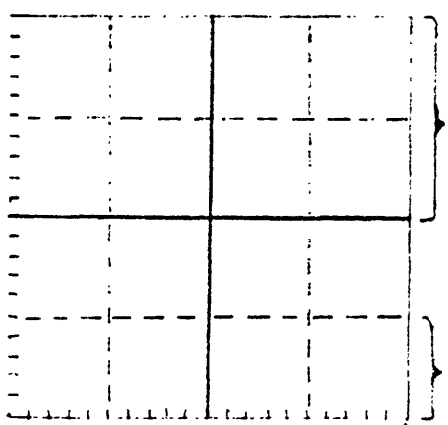
b) Select 10 randomly located points within the sandstone unit of area "B".

Area "B"



c) Sample area "C" according to the model: $x_{ijk} = \mu + a_i + b_{ij} + e_{ijk}$ where x_{ijk} is a sample value, μ is the grand mean, a_i designates a "quad", b_{ij} designates a "subquad" within a quad, and e_{ijk} designates a sampling "point" within a subquad. Let i range from 1 to 4, j range from 1 to 2, and k range from 1 to 3. Taking the three estimated variances as s_a^2, s_b^2, and s_e^2, set up the equation for estimating the variance of the grand mean estimate, using the finite population correction factors (see equation 23a of the syllabus).

Area "C"



Exercise 9.

a) Assume that five localities have been sampled - two samples from each - but that the "true" average value for each locality is 100 ppm. That is, if the truth were known, all localities are exactly the same. Now assume that the sum of the sampling and analytical errors is distributed as $\Lambda(0,0.09)$. Draw two values from this population for each of the five localities.

b) For each of the five localities, add each of the two error values to the log of 100. The 10 resulting values may be taken as the logs of the geochemical data - two data values for each of five localities.

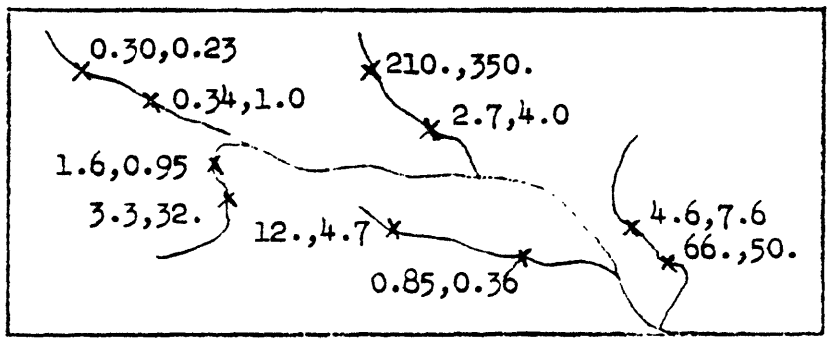
c) Compute the geometric mean for each locality.

d) Using the log data, estimate the variance of the errors due to sampling plus analysis. Compute the empirical approximation of the "shortest significant range". Take the antilog and call this the "smallest significant factor".

e) Order the five geometric means by increasing magnitude. Do any two adjacent means differ by more than the "shortest significant factor". Some should. With $\underline{X}$ members in the class, $0.05\underline{X}$ members should uncover a "false" geochemical anomaly - working at the 0.05 probability level.

Exercise 10.

a) Suppose that a preliminary geochemical survey of an area has been made in preparation for a final geochemical program. Five localities were sampled, two samples were taken at random points from each locality, and each sample was analyzed twice in a randomized sequence. The resulting data are plotted on the map below:



Each first-order stream is a locality. (Data as ppm)

Transform the data to logs (base 10) and estimate the log variance among localities, among sampling points within localities, and between duplicate analyses.

b) Compute the standard error of a locality mean, the empirical approximation of the "shortest significant range" and the "smallest significant factor (SSF)" at the 0.05 probability level.

c) Assume that the estimated variance components, from part (a) are stable and that the cost of collecting a sample is about equal to the cost of analyzing one. Now use the same equations as used in part (b) to develop an optimum sampling plan for the final program. That is, how many samples per locality should be taken in the final program, and how many analyses should be made of each sample?
Hint: "SSF" is a measure of the power of the experiment; a relative measure of the cost per sampling locality is given by $(n_\beta + n_\beta n_\gamma)$, where n_β is the number of samples per locality and n_γ is the number of analyses per sample. First, assume $n_\gamma = 1$ and plot "SSF" against relative cost, then assume $n_\gamma = 2$, etc.

d) What would the relative cost per sampling locality be if we collected 20 samples at each and combined them into a single composite sample which was analyzed only once? What "SSF" should we expect?

Exercise 11.

a) Assume that a region has been sampled with a "barbell-type" design according to the model:

$$x_{ijklm} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} + \delta_{ijkl} + \epsilon_{ijklm}$$

where the subscripted Greek letters represent the following:

α areas 250 Km apart within the region

β plots 50 Km apart within areas

γ sites 10 Km apart within plots

δ samples 2 Km apart within sites

ϵ replicate analyses of samples

The purpose of this initial sampling was to devise a final sampling program that would allow us to construct a geochemical map that describes the geochemical pattern of variation over the region.

The variance components estimated from the initial data are as follows:

$\sigma_\alpha^2 = 0.01$

$\sigma_\beta^2 = 0.36$

$\sigma_\gamma^2 = 0.0004$

$\sigma_\delta^2 = 0.13$

$\sigma_\epsilon^2 = 0.49$

List the ranges of scales of variation represented by each of the 1st four variance components. Draw a rough sketch showing the general nature of the geochemical profile.

b) If you were to conduct the final geochemical program and wanted to describe as much of the geochemical variation as possible, at some reasonable cost, how would you proceed? Would you use the same analytical method? How would you space the sampling localities?

c) Use the variance components to construct a "cumulative variance curve" for the "natural" variation. What percentage of the total natural variance might be mapped by spacing the sampling localities at 2 Km? At 10 Km? At 50 Km? How many more sampling localities would be required to sample at 2 Km rather than 10? What would be gained?

d) Assume that the sampling localities are to be spaced at 10 Km and that we wish to achieve a v_m ratio of at least 3.0. What diameter should each locality have? How many random samples should be taken in each? How many times should each sample be analyzed if the same analytical method is to be used?