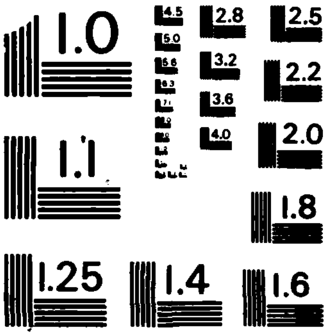
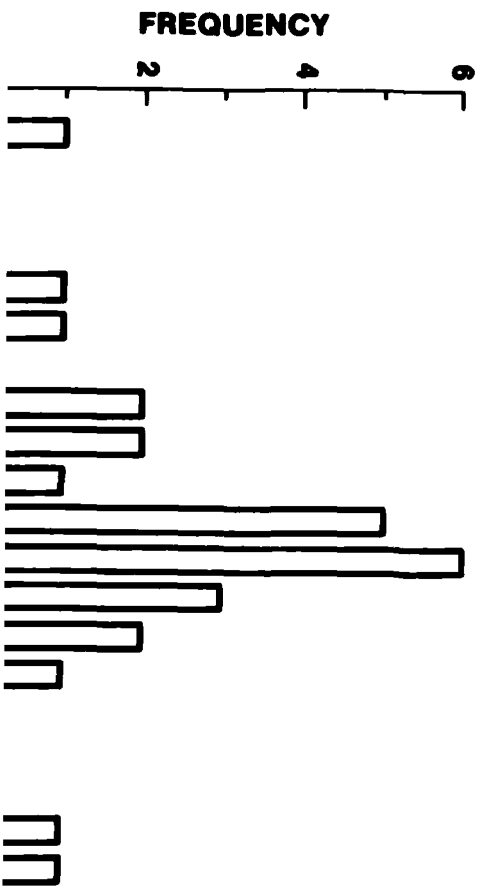
AD-A171 046   ENVIRONMENTAL AND WATER QUALITY OPERATIONAL STUDIES   1/3
STATISTICAL METHODS F..(U) ARMY ENGINEER WATERWAYS
EXPERIMENT STATION VICKSBURG MS ENVIR..   R F GAUGUSH
UNCLASSIFIED   JUN 86 WES/IR/E-86-2   F/G 8/8   NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A


FREQUENCY
2
4
6

12

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
Instruction Report E 86-2

2. GOVT ACCESSION NO.
AD-A171046

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
STATISTICAL METHODS FOR RESERVOIR WATER QUALITY INVESTIGATIONS

5. TYPE OF REPORT & PERIOD COVERED
Final report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Robert F. Gaugush, Technical Editor

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
US Army Engineer Waterways Experiment Station
Environmental Laboratory
PO Box 631, Vicksburg, Mississippi 39180-0631

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
EWQOS Work Unit VIIA

11. CONTROLLING OFFICE NAME AND ADDRESS
DEPARTMENT OF THE ARMY
US Army Corps of Engineers
Washington, DC 20314-1000

12. REPORT DATE
June 1986

13. NUMBER OF PAGES
216

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

E

18. SUPPLEMENTARY NOTES

Available from National Technical Information Service, 5285 Port Royal Road,
Springfield, Virginia 22161.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Reservoir
Statistics
Water quality monitoring
Multivariate statistics
Data analysis
Descriptive statistics
Analysis of variance
Nonparametric statistics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report presents an introduction to the statistical analysis of water quality data. Techniques from the use of simple data displays to the application of nonparametric and multivariate statistics are presented. Methods of data display are presented as a means of becoming familiar with the data and presenting those data. Basic descriptive statistics are discussed as a means

(Continued)

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (Continued).

to summarize the typically large data sets that result from a water quality monitoring program. The application of inferential statistics to make sound conclusions about differences, relationships, or trends within the data is also discussed. All of these methods are presented with concise and clear examples using actual water quality data. The report also provides an introduction to the statistical concerns involved in sample design that are necessary for the proper execution of a water quality monitoring program.

PREFACE

The material presented in this report is part of the Environmental and Water Quality Operational Studies (EWQOS) Program, Work Unit VIIA, Reservoir Field Studies, conducted by the US Army Engineer Waterways Experiment Station (WES) for the Office, Chief of Engineers (OCE), US Army. The OCE Technical Monitors were Dr. John Bushman, Mr. Earl Eiker, and Mr. James L. Gottesman.

The report was completed by the Aquatic Processes and Effects Group (APEG), Ecosystem Research and Simulation Division (ERSD), Environmental Laboratory (EL), WES. The report was written by Drs. Robert F. Gaugush, David C. Blouin, James P. Geaghan, Kenneth H. Reckhow, and William G. Warren, APEG. Dr. Gaugush, who served as technical editor, was associated with APEG through an Interagency Personnel Agreement (IPA) with the Department of Biological Sciences, Kent State University, Kent, Ohio. Drs. Blouin, Geaghan, and Warren were associated with APEG through IPA's with the Department of Experimental Statistics, Louisiana State University, Baton Rouge, La. Dr. Reckhow was associated with APEG through a purchase order with the School of Forestry and Environmental Studies, Duke University, Durham, N. C. This report was prepared under the supervision of Dr. Thomas Hart, Chief, APEG; Mr. Donald L. Robey, Chief, ERSD; and Dr. John Harrison, Chief, EL. Dr. Jerome L. Mahloch was Program Manager of EWQOS. The report was prepared for publication by Ms. Jessica S. Ruff of the WES Publications and Graphic Arts Division.

Director of WES was COL Allen F. Grum, USA. Technical Director was Dr. Robert W. Whalin.

This report should be cited as follows:

Gaugush, R. F., tech. ed. 1986. "Statistical Methods for Reservoir Water Quality Investigations," Instruction Report E-86-2, US Army Engineer Waterways Experiment Station, Vicksburg, Miss.

CONTENTS

STATISTICAL METHODS FOR RESERVOIR WATER QUALITY INVESTIGATIONS

PART I: INTRODUCTION

Background

1. Through its Civil Works Program, the US Army Corps of Engineers (CE) is responsible for the operation of a large number of water resources projects. Over 400 of these projects are reservoirs which are operated for a numb^er of purposes, including flood control, hydropower, recreation, navigation, and water supply. In addition, the CE is required by Federal legislation in the operation of its reservoirs to comply with Federal and state water quality requirements.

2. To provide policy and guidance in addressing Federal and state water quality legislative requirements, the CE has issued Engineer Regulations (ER) on the collection and interpretation of water quality data. Specifically:

- a. ER 1110-2-334, "Reporting Water Quality Management Activities," established the consideration of water quality as an integral feature of CE responsiblities and set out the requirements for monitoring programs and the reporting of water quality data collected at CE reservoirs.
- b. ER 1110-2-415, "Water Quality Data Collection, Interpretation, and Application," established guidelines for water quality monitoring programs and data interpretation at CE projects.

While these ERs establish the general guidelines and requirements for assessing water quality, they do not provide specific assistance to the CE Division and District offices in developing water quality programs, including data analysis and interpretation.

3. To address the national environmental water quality objectives delineated in Federal legislation, in 1978 the Office, Chief of Engineers, instituted a major research effort, the Environmental and Water Quality Operational Studies (EWQOS). The EWQOS Program has addressed a number of environmental quality issues and provided guidance for the design

and operation of CE projects with respect to maintaining or enhancing environmental quality in a manner that is compatible with project purpose.

4. One of the major efforts of the EWQOS Program, Reservoir Field Studies (Work Unit VIIA), was initiated to develop various operational, control, and management procedures to address environmental and water quality problems at CE reservoirs. One specific objective of this research program was to provide guidance on the design of District and Division reservoir water quality sampling programs. This report is intended to provide guidance to field personnel in the data analysis and interpretation phase of a water quality monitoring program.

Purpose and Scope

5. The purpose of this report is to provide to CE Division and District personnel a general introduction to the statistical analysis of water quality data. The major and most common concepts and techniques involved in the statistical interpretation of water quality data will be discussed. It will not be possible to provide a comprehensive or thorough treatment of all of the statistical methods that can be applied to water quality data. This report is not intended to replace statistical textbooks, rather it provides the necessary background for more effective and efficient use of those reference works. Also, this report does not provide specific guidance on the statistical techniques to be used for data interpretation. The application of specific statistical methods will be dependent on and constrained by site-specific features and the data collected in a water quality monitoring program.

6. This report is intended for use by all personnel involved in the design, implementation, and data interpretation of water quality monitoring programs in CE reservoirs. Most of the information contained in this report is discussed in a more detailed manner in other sources. A number of introductory statistics textbooks are identified in the bibliography, and it is suggested that at least one of these be on hand as this report is read. These textbooks also contain mathematical and

statistical tables that will be required for the implementation of the techniques presented in this report.

7. The presentation of the material in this report is divided into several parts. Parts II and III discuss the use of data displays and descriptive statistics, both of which are effective means of summarizing water quality data. The application of inferential statistics is presented in Part IV. Inferential statistics are required to make sound conclusions about differences, relationships, or trends within the data. Part V presents a brief introduction to the statistical concerns involved in sampling program design that are necessary for the proper execution of a water quality monitoring program. A glossary of statistical terms is provided as Appendix A.

PART II: DATA DISPLAYS

Introduction

8. It is good practice in statistical analysis to begin with a study involving graphical displays of the data. That is, before descriptive statistics are calculated from a data set, and before analyses such as correlation, analysis of variance, or linear regression are performed, it is wise to look at various displays of the raw data. The graphs recommended for this task are useful in identifying the need to edit or transform the data prior to conducting the statistical analysis. Most procedures in statistics (e.g., regression analysis, hypothesis testing) derive summary values (e.g., mean and standard deviation) from a data set. Thus, if the inferences drawn from the statistical procedures are to be valid for the entire data set, it is important that the summary statistics represent the entire data set. The graphical displays help guide the choice of any necessary manipulations of the data set and the selection of appropriate statistics (see Part III: Basic Descriptive Statistics) to summarize the data. Examples presented in the following sections should underscore the importance of examining data displays at the beginning of a statistical study.

9. Graphs can also be useful during the course of a statistical study. For example, bivariate scatter plots are helpful in the selection of independent variables for a regression equation. Upon completion of the statistical analysis, the scientist often wisely chooses to present some of the results in graphical form. Not infrequently, conclusions are most effectively conveyed in a graphical display.

Histograms

10. In the most fundamental study, data on a single characteristic are analyzed. For example, the limnologist has data on chlorophyll *a* from a particular reservoir on a particular date and desires to

summarize the information obtained. The limnologist could calculate the mean and standard deviation of the sample data set; alternatively, he could calculate other statistics representing central tendency and dispersion (see Part III). Prior to calculating any statistics for the sample, however, the limnologist should first look at a plot of the data. For data representing a single characteristic (such as chlorophyll *a*), the histogram is often a useful graphical display.

11. As an example, assume that the summer chlorophyll *a* data in Table 1 have just been acquired, and the limnologist would like a "picture" of this sample. As a first cut, the histogram in Figure 1 is plotted. To construct the histogram, the limnologist must first divide the range (highest value to lowest value) into equal-sized intervals. In Figure 1, the range is approximated by 10 to 160 (actually it is 17.7 to 150, but the approximation is good enough) and is divided into intervals of 10 units (micrograms per litre). For each interval, 10 to 20, 30 to 40, and so on, simply count the number of data points that lie in the interval and construct vertical bars with height proportional to that number. So, for example, there are three observations in the 40 to 50 range and six observations in the 60 to 70 range. Thus, the bar for the 60 to 70 interval is twice the height of the 40 to 50 bar.

12. What does the histogram tell us about the sample? Basically, it provides us with a visual image of the distribution of the sample. In specific terms, this means that we are able to quickly see such things as location of the "center" of the sample, amount of "dispersion," extent of "symmetry," and existence of "outliers" in the sample. In Figure 1, the center is clearly identified by the peak in the 50 to 60 interval, and dispersion could perhaps be characterized by stating that about 85 percent of the observations lie between 30 μg/ℓ and 80 μg/ℓ. The histogram is not symmetric, however, and one might want to check on the validity of the two outlying observations at the extreme right.

13. The picture created by the histogram is of considerable value in the selection of descriptive statistics, as is noted in the next section. Some care should be observed in the construction of the

Table 1

Hypothetical Total Phosphorus (TP) and Chlorophyll *a* (CHLa)

Data for Two Sampling Periods

Sample Number	Summer		Fall	
	TP	CHLa	TP	CHLa
1	50	52.1	95	11.3
2	60	55.6	110	20.4
3	60	61.9	140	36.9
4	60	61.2	130	46.6
5	50	46.1	120	12.9
6	58	56.6	110	8.1
7	65	63.8	110	14.1
8	60	25.5	120	23.0
9	110	133.4	90	9.9
10	80	74.8	120	26.2
11	70	59.4	120	26.6
12	70	53.6	120	22.5
13	70	57.6	110	22.1
14	80	76.0	140	21.6
15	90	39.5	110	28.0
16	73	53.6	120	29.0
17	60	37.7	120	34.8
18	70	60.3	140	36.2
19	90	79.5	100	26.8
20	70	63.1	120	27.9
21	120	42.6	120	32.2
22	60	17.7	70	1.2
23	220	150.0	120	35.6
24	65	42.5	150	66.3
25	70	30.0	100	15.6
26	130	67.4	50	4.7
27	90	52.2	80	18.2

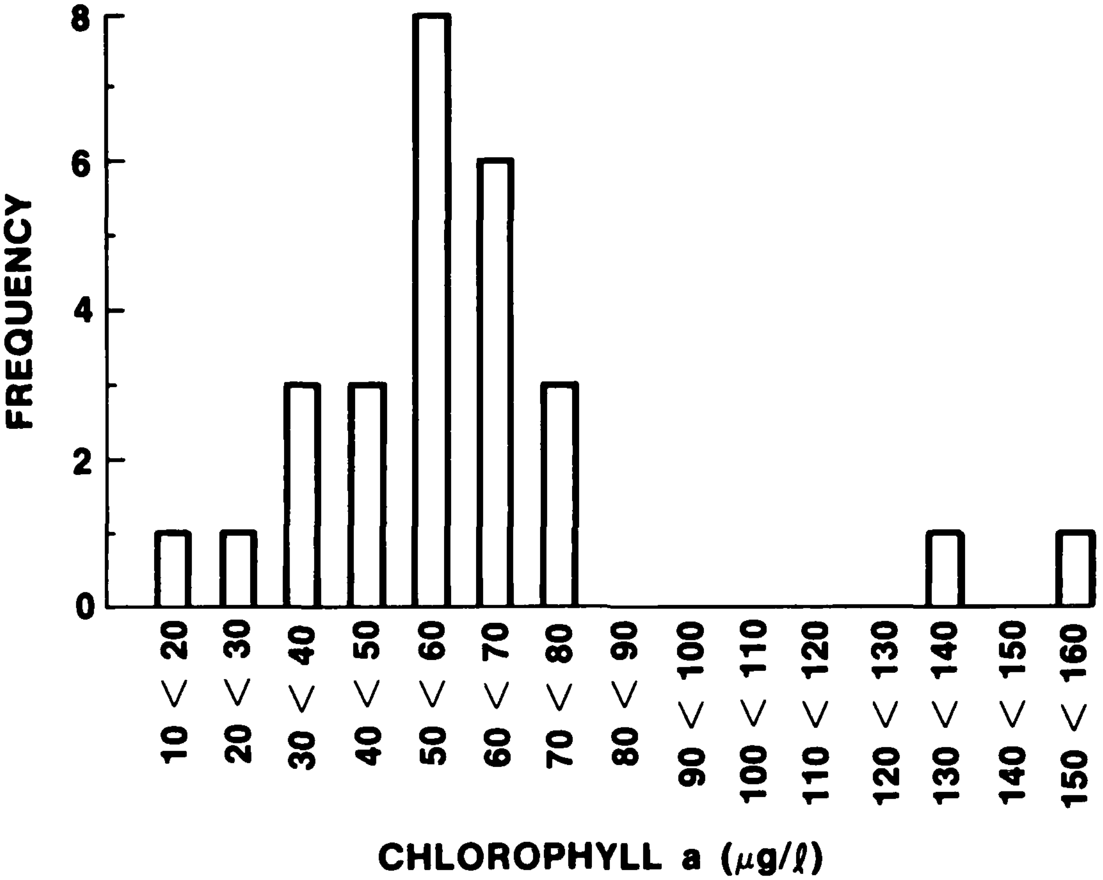
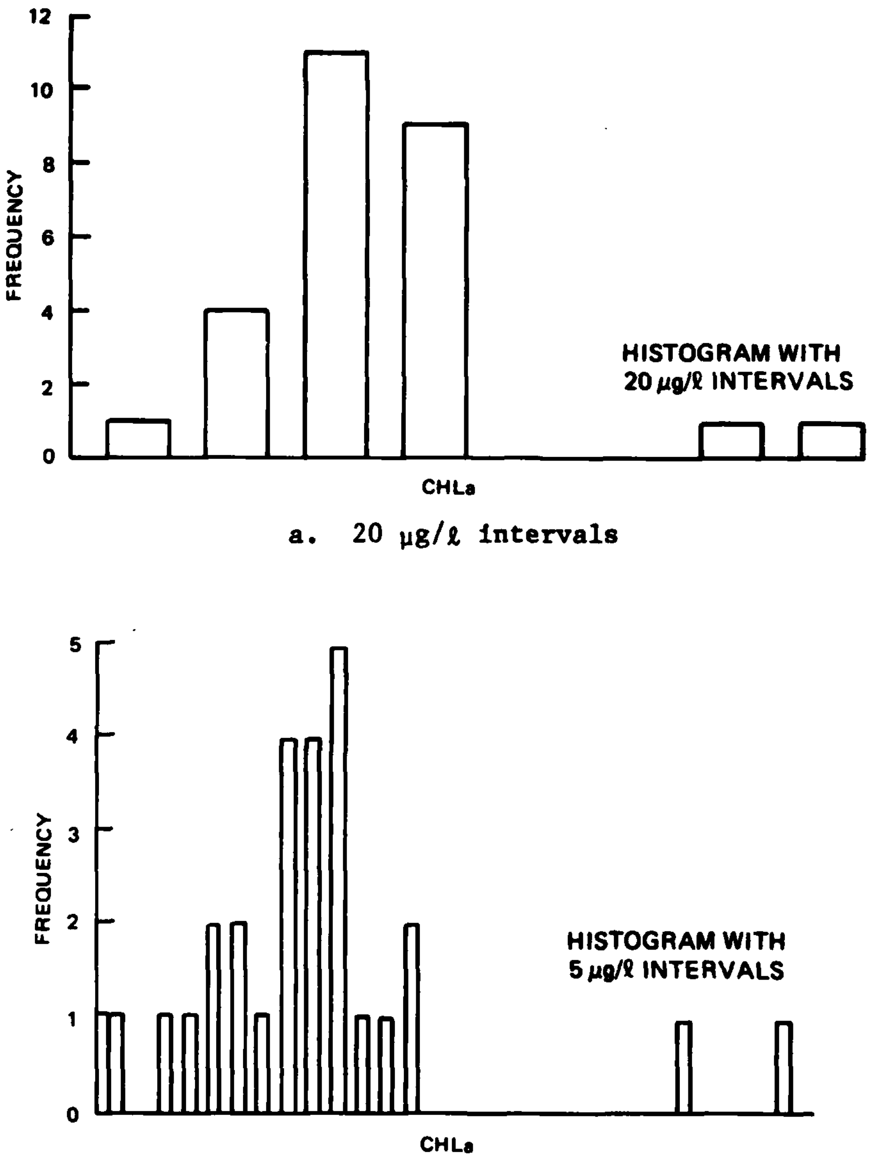


Figure 1. Histogram of summer chlorophyll *a* data from Table 1

histogram, however. With changes in interval size, the histogram may assume different shapes which might affect the inferences drawn. For example, in Figure 2a the chlorophyll *a* data are plotted using an interval size of 20 units. With that scale, the two highest observations are less likely to be considered outliers than they are on the basis of the five-unit interval size histogram in Figure 2b. It is probably good practice to scale the histogram so that the observations are neither too bunched (as in Figure 2a, where 75 percent are concentrated in two intervals) nor too spread out to permit reasonable inferences to be drawn.

14. As noted above, the histogram provides an impression of the extent of symmetry in the sample. Symmetry in a data set is a desirable attribute for two reasons. First, it often means that one can



a. 20 µg/ℓ intervals

b. 5 µg/ℓ intervals

Figure 2. Histograms of summer chlorophyll *a* data

characterize the sample as having a distribution with a shape similar to those symmetric distributions (e.g., the normal and uniform distributions) which are commonly an assumption of statistical analysis. Stating, for example, that a sample approximates the normal distribution conveys useful information to a reader. Beyond that, symmetry implies that the common descriptive statistics such as the mean and standard deviation can be used to provide an adequate summary of the sample (see Part III).

15. The foregoing discussion suggests that it might be useful to apply a transformation (see Part III), if necessary, in order to create symmetry in an asymmetric data set. Fortunately, limnological data are often lognormally distributed, so the choice of transformation is often straightforward. The lognormal distribution is strictly positive (all observations > 0) and it contains skew to the right. As an example, the chlorophyll *a* histogram in Figure 1 approximately fits this description. To check for lognormality, the logarithmic transformation is applied to the data, and a histogram of the transformed data is plotted. Comparison of this histogram with a normal distribution (i.e., a bell-shaped curve) provides a rough test of lognormality; formal tests do exist (e.g., Kolmogorov-Smirnov test or chi-square test) and may be found in many statistics texts (e.g., Wonnacott and Wonnacott 1972, Benjamin and Cornell 1970).

16. To illustrate how a transformation may change the shape of a histogram, the summer chlorophyll *a* data from Table 1 were log-transformed, and the histogram of the logarithms of the chlorophyll *a* observations was constructed in Figure 3. Compare Figure 1 with Figure 3. Note how the logarithmic transformation achieved approximate symmetry. Note also that the observations at the extreme right are less likely to be considered outliers than they were in the original data. In fact, the observation in Figure 3 at the extreme left is further from the mean of the log-transformed sample (the geometric mean) than is either of the points on the right. This is a result of the effect of the logarithmic transformation: to "spread out" low values and "squeeze in" high values. Through the study of histograms of this sample, we are now in a position to determine descriptive statistics and to summarize the data set.

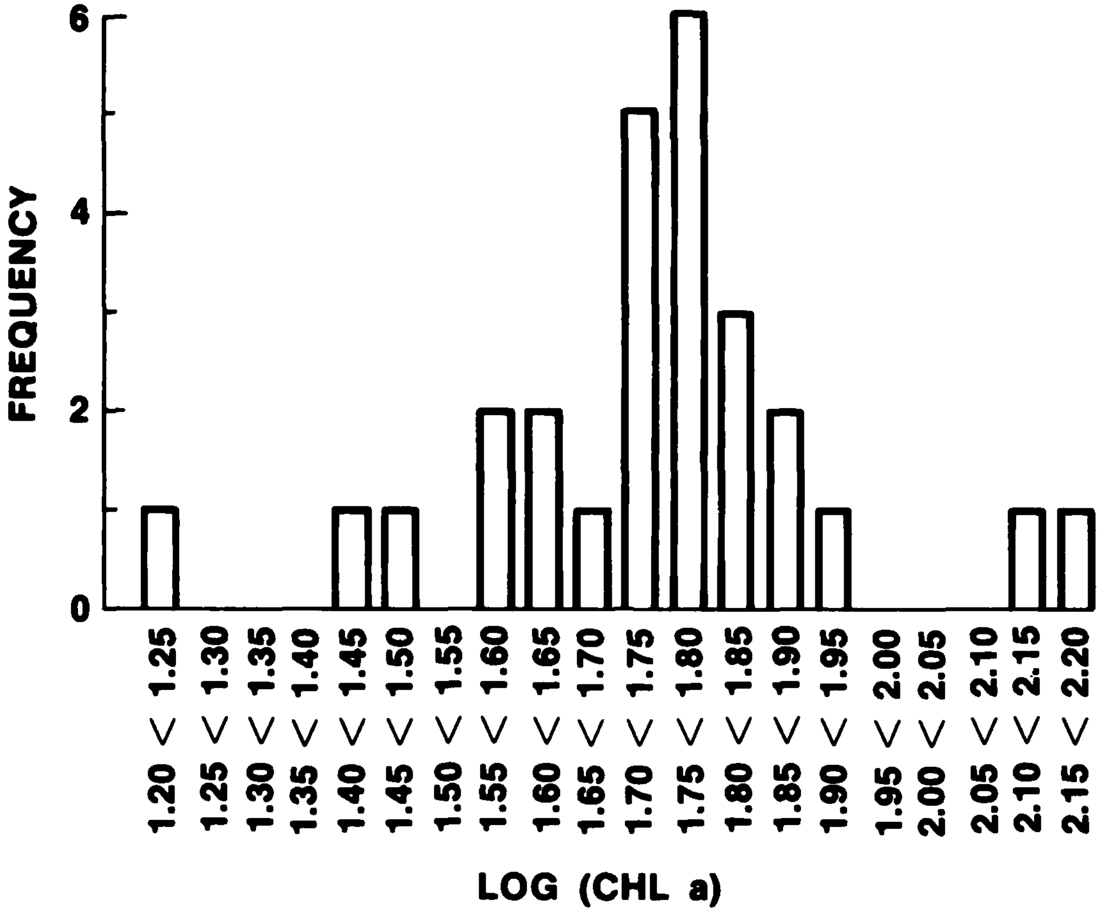


Figure 3. Histogram of log-transformed chlorophyll *a* data

Stem and Leaf Displays

17. An alternative and often informative version of the histogram is the stem and leaf display. Developed by Tukey (1977), the stem and leaf plot provides the shape of a histogram while at the same time presenting the numerical values from the data set. As an example, the stem and leaf display for the summer chlorophyll *a* data in Table 1 is plotted in Figure 4; note that the shape is nearly the same (round-off variations create the slight differences) as the histogram in Figure 1.

18. To construct the stem and leaf diagram, first choose the interval width. The "stem" becomes the digit level corresponding to this interval width (for Figure 4, the stem contains the "tens" digit

0	
1	8
2	6
3	80
4	6033
5	26794842
6	214037
7	56
8	0
9	
10	
11	
12	
13	3
14	
15	0
16	

Figure 4. Stem and leaf display of the summer chlorophyll *a* data

since the interval width is 10 units), and values for the stem are placed to the left of a vertical line. On the right side of this line, the "leaves" are written. For each data point, the leaf is the next digit lower in value than the stems digit. Since the stems in Figure 4 are composed of the tens digit, the leaves are made up of the units digits. Each observation contributes one leaf to the row containing its stem. For the summer chlorophyll *a* data in Table 1, the first observation (52.1 μg/ℓ) results in a 2 (the units digit) placed in the row for the stem value 5 (the tens digit, the second observation) (55.6 μg/ℓ, rounded to 56) results in a 6 placed in the row for stem value 5, and so on.

19. The primary advantage of the stem and leaf display (over the histogram) is that it contains information on the numerical values in the data set (while retaining the ability to provide information on the shape of the sample distribution). There may be advantages to this, particularly when the data are displayed for presentation purposes. Tukey (1977) describes several variations of the stem and leaf display, including an interesting way to look at covariation in bivariate (e.g., chlorophyll and phosphorus) data.

Box Plots

20. Often there is a need to compare two or more samples of the same characteristic (e.g., samples for chlorophyll *a* from two different sampling stations or reservoirs). This comparison may be purely statistical, perhaps using one of the procedures presented in Part IV. Alternatively, a graphical method could be used that provides both a pictorial comparison as well as a statistical comparison. The graphical display that permits this is the box plot.

21. In Figure 5, two box plots are presented, one for the summer chlorophyll *a* data and the other for the fall chlorophyll *a* data in Table 1. (Assume the data were collected at two different times of the year in the same reservoir.) Box plots are based on order statistics (Table 2). These are observations, like the median, that are used to summarize a sample because of their order in a ranking of low value to high value, and convey information on the sample median, dispersion, skew, relative size of the data set, and statistical significance of the median. To construct a box plot for a sample on a single variable, the steps below may be followed (from Reckhow and Chapra 1983):

<u>a</u>. Order the data from lowest to highest.

<u>b</u>. Plot the lowest and highest values on the graph as short horizontal lines. These represent the extreme values for each box plot.

<u>c</u>. Determine the upper and lower quartiles for the data set (see Part III). These values define the positions of the upper and lower edges of the box. Using vertical lines, connect the highest value with the upper quartile and the lowest value with the lower quartile.

<u>d</u>. Plot the median as a dashed horizontal line within the box.

<u>e</u>. Select a scale so that the width of the box represents the sample size (the size of the data set used to construct each box). For example, each centimetre of width could represent 25 observations.

<u>f</u>. Determine the height of the notch (in the box at the median) based on the statistical significance of the median. Based on work by McGill, Tukey, and Larsen

(1978), the height of the notch above and below the median is approximately:

$$\text{Notch limits} = \text{Median} \pm \left[1.7(1.25I/1.35\sqrt{n})\right]$$

where

I = interquartile range
= upper quartile-lower quartile

n = sample size

With this mathematical definition of the notch limits, the notch in the box provides an approximate 95-percent confidence interval for comparison of box medians. Therefore, when the notches for any two boxes overlap in a vertical sense, the medians are not significantly different at about the 95-percent level.

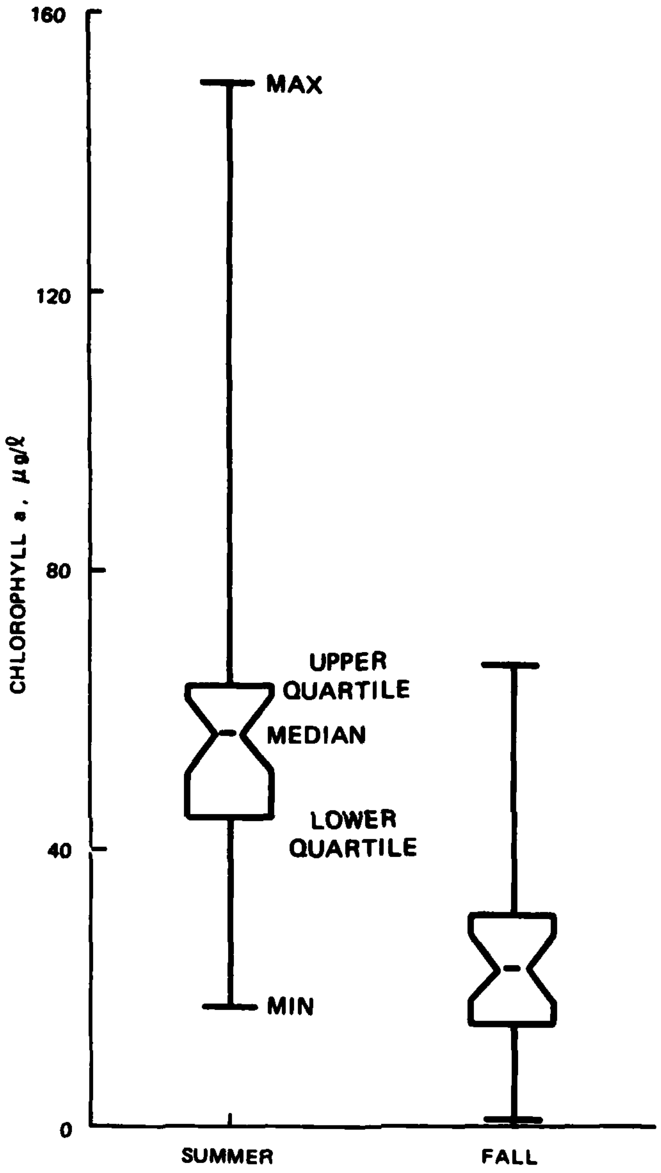


Figure 5. Box plots of the summer and fall chlorophyll *a* data

Table 2

Order Statistics for Chlorophyll *a* Data Presented in Table 1

Order Statistics	Summer	Fall
Minimum	17.7	1.2
Maximum	150.0	66.3
Median	56.6	23.0
Quartiles		
Lower	44.4	14.9
Upper	63.5	30.5
Interquartile range	19.1	15.6
Notch limits	±5.8	±4.7

22. The first of the two box plots in Figure 5 is labeled so that the characteristics mentioned above may be identified. The plot provides information on both a single sample and a comparison among samples. For a single sample we see:

a. An estimate of the center of the sample (the median).

b. measure of dispersion for the sample (the interquartile range).

c. The range (highest value - lowest value) and an impression of skew through a visual comparison of the symmetry above and below the median.

23. For a study involving two or more samples we see:

a. A statistical test of significance in the difference between two medians, based on vertical overlap between notches.

b. A visual comparison of samples, based simply on observing the similarities and differences between features of two box plots.

24. Note that the notches for the two box plots in Figure 5 do not overlap in a vertical sense, indicating that the median chlorophyll *a* observation for the summer sampling date is significantly different from the median chlorophyll *a* observation for the fall sampling date. The skew in the summer sample is evident from the elongated upper

tail, and the summer sample as a whole is largely above the fall sample.

25. Box plots are helpful both in diagnostic work as above or in presenting conclusions about samples following the completion of a statistical study. Reckhow (1979) describes several additions to the basic box plot that might be useful in limnological analyses. Tukey (1977) created the box plot and presents many interesting examples in his book on exploratory methods in statistics.

Scatter Plots

26. Many statistics (e.g., correlation coefficients) and many statistical methods (e.g., regression analysis) are fundamentally concerned with relationships between pairs of variables. Without doubt, the best way to examine a relationship between pairs of variables, a bivariate relationship, is through a scatter plot. The scatter plot is simply a two-variable plot of observations on an x-y coordinate system.

27. In Figure 6, a bivariate scatter plot is presented for the data on summer phosphorus and chlorophyll *a* in Table 1. From the plot, we can examine the distribution of data for each of the variables separately as well as for the two variables together. For example, we can see from Figure 6 that two high observations for chlorophyll *a* tend to stand apart from the rest of the data (which was the same conclusion drawn from the histogram in Figure 1). Likewise, one observation for phosphorus tends to stand apart from the rest of the data.

28. When both variables in Figure 6 are examined together, we see that the point at the upper right of the plot might be considered an outlier. With this point removed, the linear relationship that seems to exist in Figure 6 is much less apparent. Therefore, it might be useful to remove this point from the sample, replot the data, and effectively spread out the remainder of the observations so that we may closely examine the pattern in the cluster of points at the lower left.

29. Basically two characteristics of a bivariate sample are of interest in most statistical studies. First, the analyst often is interested in the linearity or nonlinearity in the relationship. Linear

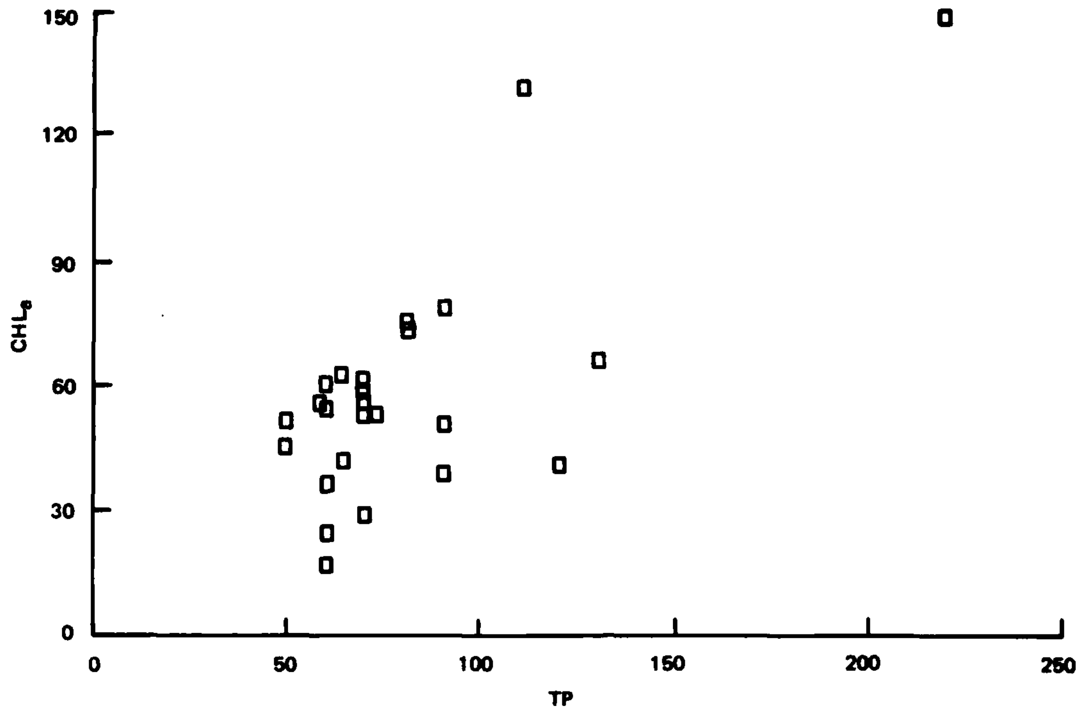


Figure 6. Bivariate scatter plot of the summer total phosphorus and chlorophyll *a* data

relationships are clearly desirable and are necessary for the correct application of correlation analysis and ordinary least squares regression. If the bivariate relationship is nonlinear, it is possible that a transformation (see Part III) can be applied to make it linear. Without question, the scatter plot is the most important diagnostic device for evaluating linearity, and it is often quite helpful in selecting a transformation.

30. The second characteristic of a bivariate sample of particular concern is the presence or absence of outliers. Outliers have no universally accepted objective definition; rather the term is used here to identify observations that stand apart from a cluster of points. We are concerned about outliers because they exert more than their fair share of influence on the value of statistics (such as correlation coefficients and regression coefficients). Statistics and statistical inferences are preferred when they are robust, or in other words, when they change little if any particular observation is deleted from the

sample. Outliers can have a substantial influence on certain statistics; therefore, it is good practice either to transform the data to reduce the influence of the outlier or to carefully examine the outlier to determine if it is a correctly measured, legitimate member of the population sampled. A study of scatter plots is the best way to check for the presence of outliers.

31. The bivariate scatter plot is an extremely useful diagnostic tool. It should always be examined near the beginning of any work involving the study of covariation in pairs of variables. Beyond that, it is the single most effective way to convey information on bivariate relationships in a set of data. Examples illustrating the use of scatter plots are found throughout this manual.

PART III: BASIC DESCRIPTIVE STATISTICS

Introduction

32. When a set of data is quite smal1, one may choose to present the entire data set in a report. For large data sets, the scientist recognizes that to most effectively transfer information he must summarize the data set with a few well-chosen statistics. A choice is made to trade some of the information contained in the entire data set for the convenience of a few descriptive statistics. This choice is usually a good one, provided the descriptive statistics that are selected correctly represent the original data.

33. Some descriptive statistics are so commonly used we forget that they actually represent only one option among many candidate statistics. For example, the mean and the standard deviation (or variance) are statistics used to estimate the center of a data set and the spread on those data. When these statistics are to be used, the scientist should decide beforehand that they are the best choices to describe the aforementioned characteristics of the data set. Often they are (notably for symmetrically distributed data following an approximate normal distribution), so their use is frequently justified. However, as we see below, there are many situations with reservoir water quality data where alternative descriptive statistics are preferred.

34. In the selection of descriptive statistics, it is important that the scientist have a clear understanding of the purpose that the statistic serves. Descriptive statistics are selected because the convenience of a few summary numbers outweighs the loss of information that results when the entire data set is described by the statistics. It is therefore essential that as much information as possible be summarized in the descriptive statistics because the alternative may be a misrepresentation of the original data.

35. Certain specific features of the data set are characterized by using descriptive statistics. For example, the center, or central tendency of a set of data, is probably the most important measure.

Among the candidate statistics are the mean, median, mode, and geometric mean. Once the center of a data set is described, the next important feature requires a statistic estimating the spread, scale, or dispersion. Among the candidates for this task are the range, standard deviation, and interquartile range. These two characteristics of a data set, central tendency and dispersion, are the most common descriptive statistics. Other characteristics, such as skewness and kurtosis, are occasionally important as well. It is useful now to look at some examples that illustrate the choice of descriptive statistics.

Measures of Central Tendency

36. Probably the single most useful statistic summarizing a data set is an indication of the center of the sample. By center we imply the vague notion of the middle of the cluster of points or perhaps the region of greatest concentration of points. Since samples exhibit a variety of distributions when plotted as histograms, it is not possible to unambiguously define the center, and as a result there are several statistical estimators that serve as candidates for determining central tendency or location. Each candidate, as noted below, may be considered to have its own advantages and disadvantages for the task at hand.

Mean (arithmetic)

37. The arithmetic mean, or simply, the mean, is the most frequently used of the central tendency estimators. It is so commonly used that the investigator often loses sight of the true reason for calculating descriptive statistics. The result is that the mean is sometimes calculated as the central tendency statistic in situations where another estimator would be better.

38. The arithmetic mean $\left(\bar{x}\right)$ is the sum of the observations (x_i) divided by the number of observations (n):

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Each observation contributes its magnitude to the sum of the observations and hence to the mean. For symmetric distributions (like the normal or Gaussian distribution), this is desirable and leads to an efficient (minimum variance) estimator. However, as noted in Part II, limnological data are often not symmetrically distributed. Skewed data "pull" the mean in the direction of the skew; this means that a few extremely high observations can pull the mean away from the bulk of the observations and toward the few high data points. In those situations, a robust estimator, like the median or the mode, might be preferred.

Median

39. When a set of data is ordered from lowest to highest value, the median is identified as the middle value. The median is therefore known as an "order statistic" since it is based on an ordering or ranking of observations. When the total number of observations is an even number, leading to two middle values, the median is then the average of the two middle values.

40. The "order" of the median observation is:

$$\text{Median observation} = (n + 1)/2$$

Since the effect on the median of all but the middle-ranking observations is simply to hold a place in the ranking, outlying observations do not pull the median toward the extremes. The median is robust to the influence of any single observation, and thus it is a good statistic to use when the histogram is skewed or unusually shaped.

Mode

41. The mode is the value in the sample that is most frequently observed. In terms of a histogram, the mode is represented by the bar of greatest height. The mode is considered a good estimator for central tendency because the most frequently observed value is usually near the middle of a distribution. An examination of a histogram of the sample will indicate whether, in fact, the mode does correspond with the center.

Geometric mean

42. The geometric mean is the antilog of the mean of logarithmically transformed observations. It is, therefore, a reasonable measure of central tendency for a set of data that exhibit a lognormal distribution. It may also be determined by calculating the n^{th} root of a product of n observations:

$$\text{Geometric mean} = (\Pi x)^{1/n}$$

where $\Pi x = x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n$

$$\text{Geometric mean} = \text{antilog} \; \frac{\Sigma \log (x_i)}{n}$$

43. The data presented in Table 1 for chlorophyll *a*, and reproduced in a histogram in Figure 1, are used to calculate statistics for central tendency; these values are listed in Table 1. Note that the mode is given as the range of values corresponding to the highest bar on the histogram; it is not meaningful to identify a particular chlorophyll *a* value (in units of 0.1 μg/ℓ) as the mode in this example because few values are duplicated. Some skewness is apparent in the histogram in Figure 1, and these data appear to have an approximate lognormal distribution. With skewed data, the mean is "pulled" to the right relative to the median and the geometric mean. Thus, the mean in Table 3 is higher than the median and the geometric mean.

Measures of Dispersion

44. Other than central tendency, measures of dispersion or spread are the most commonly cited statistics used to summarize a data set. Dispersion in a data set refers to the variability in the observations about the center of the distribution. Good measures of dispersion will be obtained from symmetric distributions. Asymmetry, or skewness, will affect the estimate of dispersion so that it overestimates spread in the

shorter tail of the data distribution (while underestimating the spread in the longer tail). A transformation may be used to create a symmetric distribution.

Table 3

Measures of Central Tendency for the Summer Chlorophyll *a* Data in Table 1

Measure	Value
Mean	59.8
Median	56.6
Mode	50-60
Geometric mean	54.6

Standard deviation

45. The most commonly used statistic for dispersion is the standard deviation. Like the mean, the standard deviation has been used so often that it sometimes is thought to be equivalent in definition to dispersion. In fact, like the mean, the standard deviation is strongly affected by extreme values. Thus, the standard deviation for a distribution of data with a long tail to the right (e.g., the histogram in Figure 1) is inflated by the values at the extreme right. It may be preferable to apply a transformation to create a symmetric distribution before calculating the standard deviation.

46. For a sample, the sample variance (s^2) is:

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

and the sample standard deviation (s) is the square root of the variance $\sqrt{s^2}$.

Interquartile range

47. Since the standard deviation is unduly influenced by extreme

observations in asymmetric distributions of data, we would like a robust alternative to the standard deviation (like the median is to the mean) for situations in which the data are skewed but a transformation is undesirable. Fortunately a good alternative exists: the interquartile range. Like the median, the interquartile range is based on order statistics, and thus is unaffected by the magnitude of the extreme observations in either tail. It is calculated as the difference between the observation at the 75-percent level (upper quartile) and the observation at the 25-percent level (lower quartile):

Lower quartile rank order = 1/2 (1 + median rank order)

Upper quartile rank order = 1/2 (1 + n - low quartile rank)

Interquartile range = I = lower quartile value - upper quartile value

The interquartile range is used as the measure of dispersion in the box plot presented in Part II.

Range

48. An easily determined and therefore frequently cited measure of dispersion is the range. The range is simply the maximum value minus the minimum value. Since it is clearly affected by the magnitude of the observations at either extreme, the range should not be relied upon as the sole indicator of variability. Nonetheless, it is often informative to list the range along with one of the other two dispersion statistics mentioned above.

49. In Table 4, measures of dispersion have been calculated for the summer chlorophyll *a* data presented in Table 1. The range, of course, is largest in magnitude. The skewness in the data results in a standard deviation that is next largest in magnitude of those statistics presented. If the two largest chlorophyll *a* observations are removed from the data set, the standard deviation drops from 27.9 μg/ℓ to

Table 4

Measures of Dispersion for the Summer Chlorophyll *a* Data in Table 1

Measure	Value
Standard deviation	27.9
Interquartile range	19.1
Range	132.7
Antilog SD (log CHL*a*)*	24.5

* 1/2 [antilog (mean + std dev) - antilog (mean - std dev)] for log-transformed chlorophyll *a* data.

15.3 µg/ℓ. This is a substantial effect due to only 2 of 27 observations, and it underscores the impact that extreme observations have on the standard deviation.

50. Since two observations greatly affect the value of the standard deviation for the data in Table 1, and if there is no basis for removal of these observations from the data set, then it may be wise to state that the data are skewed right and use one of the other measures of dispersion. In Table 4, both the interquartile range and the antilog SD lie between the two standard deviations previously cited (n = 27 and n = 25), and thus may represent good compromise choices. The antilog SD is a reasonable expression of the standard deviation (in antilog units) for a data set that has a lognormal distribution. Given the familiarity of applied scientists with various measures of dispersion, a good rule of thumb might be to cite the standard deviation and the range for symmetric data sets, and the interquartile range and the range for asymmetric data sets.

Ranks

51. On occasion it is preferable to examine and test a data set on the basis of the rank order of observations. In those situations,

the observations in a data set are simply ordered from low value to high value according to one particular variable. As an example, the data set presented in Table 1 has been ordered according to the magnitude of the chlorophyll *a* observations and is presented in Table 5.

Table 5

Summer Chlorophyll *a* Data of Table 1

Ordered by Magnitude

Order	CHL*a* (μg/ℓ)	Order	CHL*a* (μg/ℓ)
1	17.70	15	57.60
2	25.50	16	59.40
3	30.00	17	60.30
4	37.70	18	61.20
5	39.50	19	61.90
6	42.50	20	63.10
7	42.60	21	63.80
8	46.10	22	67.40
9	52.10	23	74.80
10	52.20	24	76.00
11	53.60	25	79.50
12	53.60	26	133.40
13	55.60	27	150.00
14	56.60		

52. Ranks or ordered data are useful in nonparametric analyses (see section, Nonparametric Analyses) and in exploratory data analysis (see Part II). In particular, when the assumption of normality is not reasonable, or when the underlying probability distribution (generating a set of data) is unknown, rank-based statistics and statistical tests are often appropriate.

Frequencies

53. Once a data set is rank ordered, it is often useful to group the data and present the frequency of observations found within a subsection of the entire range. This is done graphically in Part II (Data Displays); both the stem and leaf diagram and the histogram are graphical displays of the frequency of an observation for equally spaced intervals of the range. For example, the bars in the histogram in Figure 1 have a relative height proportional to the relative frequency of observations within each class.

54. In Table 6, the absolute and relative frequencies are presented for each of the cells (or classes) in Figure 1. The absolute frequency is expressed in terms of number of observations within each class, and the relative frequency is expressed as the percent (of the total number of observations) contained within each class. Cumulative frequencies are also presented in Table 6; these indicate the relative or absolute number of observations less than or equal to a particular class level. Thus, in Table 6, it is indicated that 92.59 percent of the chlorophyll *a* observations are less than 80 μg/ℓ.

Transformations

55. It is often necessary to apply a transformation to reservoir water quality data in order to meet the assumptions of the procedures. For example, methods of estimation (e.g., regression analysis) and hypothesis testing are based on certain assumptions about the observations. In some cases, it is permissible to violate the assumptions without greatly affecting the analysis; alternatively, it is sometimes possible to apply a method (e.g., distribution-free procedures) with less restrictive assumptions. However, there are still likely to be situations in which the assumptions must be approximately met and in which the best approach is to apply a transformation to the data.

56. Transformations are most commonly used in data analysis to reexpress a data set so that it is more consistent with the important

Table 6

Absolute and Relative Frequencies for the Summer Chlorophyll *a* Data in Table 1

			Cumulative	
Class Limits	Frequency	Percent	Frequency	Percent
10 < 20	1	3.70	1	3.70
20 < 30	1	3.70	2	7.41
30 < 40	3	11.11	5	18.52
40 < 50	3	11.11	8	29.63
50 < 60	8	29.63	16	59.26
60 < 70	6	22.22	22	81.48
70 < 80	3	11.11	25	92.59
80 < 90	0	0.00	25	92.59
90 < 100	0	0.00	25	92.59
100 < 110	0	0.00	25	92.59
110 < 120	0	0.00	25	92.59
120 < 130	0	0.00	25	92.59
130 < 140	1	3.70	26	96.30
140 < 150	0	0.00	26	96.30
150 < 160	1	3.70	27	100.00
Total	27	100.00		

assumptions (e.g., normality) of a statistical analysis, and/or to diminish the impact of outlying observations (see section Scatter Plots, Part II). To achieve these objectives, transformations may:

a. Straighten (linearize) a nonlinear relationship between two variables,

b. Reduce skew (achieve symmetry) in a data set for a single variable, and/or

c. Stabilize variance (create constant variance) for a particular variance across two or more data sets.

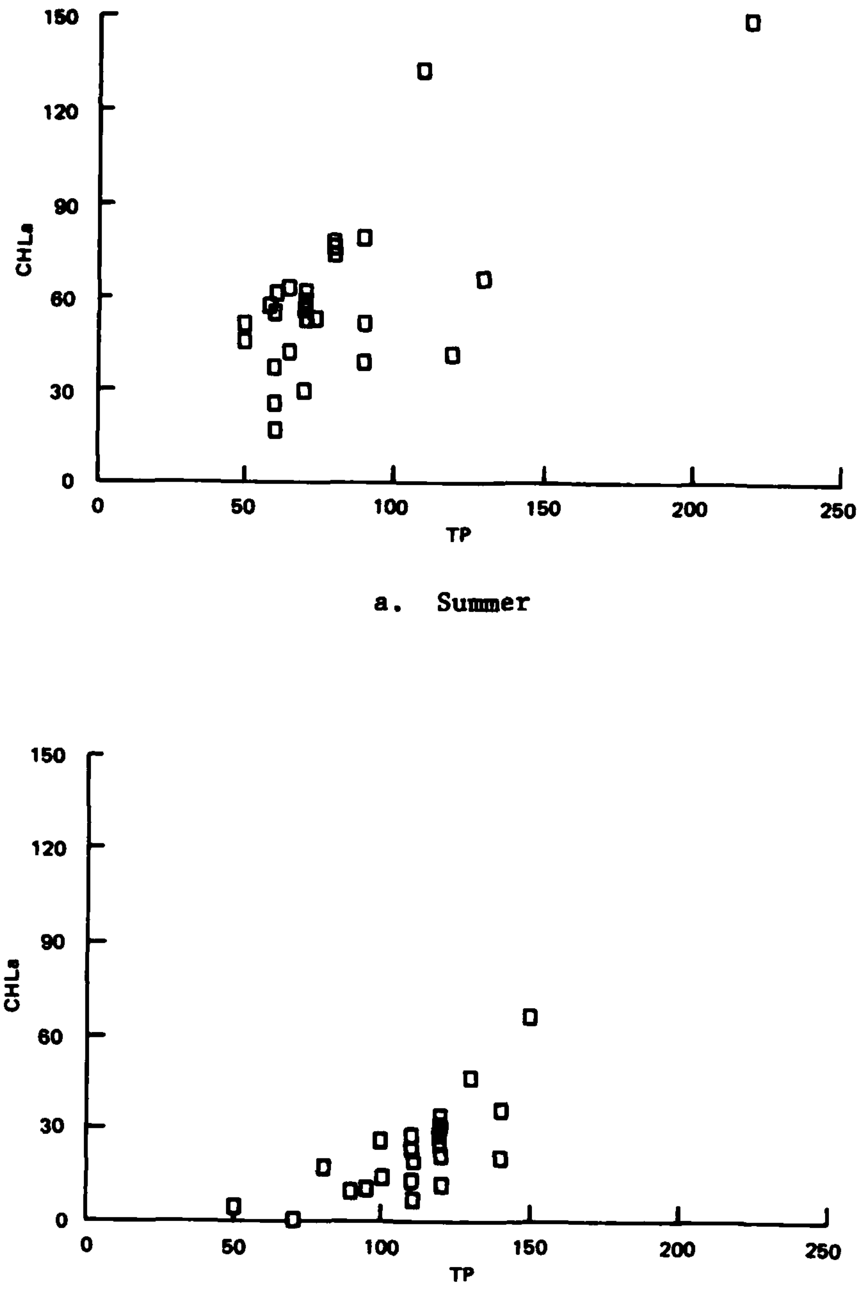
57. Selection of a transformation for these three functions is

beyond the scope of this manual, but fortunately it is quite clearly and simply presented by Velleman and Hoaglin (1981) using transformations of selected order statistics. Reservoir water quality data are often skewed right, exhibiting an approximate lognormal distribution. When this occurs, the logarithmic transformation is generally appropriate if the objective is to obtain a symmetric, approximately normal distribution of data. Reckhow and Chapra (1983, Chap. 6) show how the application of the log transformation to a data set simultaneously addressed problems of nonlinearity and skewness. Achievement of more than one objective with a transformation is actually not uncommon; the investigator should therefore be encouraged to consider data transformations whenever it appears that they may improve an analysis.

58. To illustrate the effect of a transformation (note that the discussion of data displays (Part II) also contains an examination of the transformations), the chlorophyll *a* and total phosphorus (TP) data from Table 1 were plotted, transformed, and then plotted again. The logarithmic transformation was applied, since it is likely to be the most frequently used transformation in limnological studies. Figure 7 presents the untransformed bivariate plots of chlorophyll *a* versus total phosphorus for the two samples; Figure 8 is a log-transformed plot of the data from Table 1.

59. The effect of a log transformation is to "stretch out" data on the left side of a plot and to "pull in" data on the right side of a plot. This is the reason that the log transform tends to create a symmetric distribution from data that are skewed right. Note that the stretching-out and pulling-in effects are observed when comparing Figure 8a with Figure 7a in either a horizontal or vertical direction. The effect of the transformation in Figure 8a is desirable since the "bunched" data near the origin in Figure 7a are spread out in Figure 8a. Correspondingly, the two highest chlorophyll *a* observations are closer to the center of the data in Figure 8a. As a result of these effects, the investigator is likely to obtain more representative statistical summaries of the data under the logarithmic transformation.

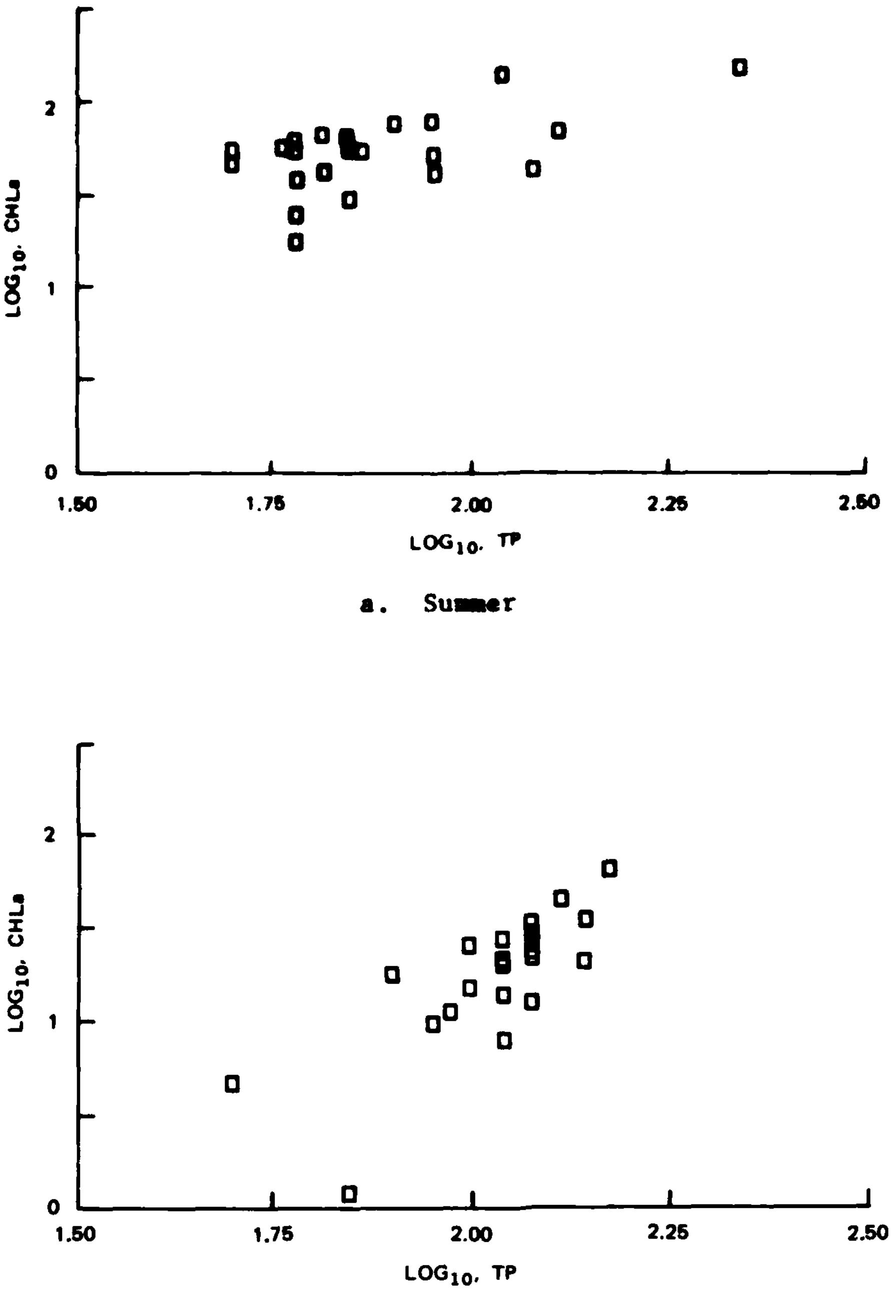
60. In contrast, the stretching and pulling effects that favored



a. Summer

b. Fall

Figure 7. Bivariate scatter plots of total phosphorus and chlorophyll a data



a. Summer

b. Fall

Figure 8. Bivariate scatter plots of the log-transformed total phosphorus and chlorophyll a data

the log transformation above have an undesirable effect on the data in Figures 7b and 8b. Since the untransformed data in Figure 7b are not bunched near the origin (nor are they particularly nonlinear), the log transform spreads the data near the origin and bunches them at the upper right in Figure 8b. In this case, the untransformed data should lead to better statistical summaries. This latter situation is somewhat unusual for limnological data; nonetheless, the investigator should always examine plots of the data to check on the effectiveness of a transformation.

PART IV: ANALYSIS OF WATER QUALITY DATA

Introduction

61. Data analysis will generally fit into one of two categories, estimation of parameters or tests of hypotheses. Since hypothesis testing is basic to much statistical analysis, at least an elementary understanding of the concept is necessary.

62. Commonly one wishes to test hypotheses about the parameters that have been estimated. In particular, in considering the relationship between one variable and another, one will often wish to test hypotheses concerning the slope and/or intercept of the regression. Also, experimentation is frequently directed toward the testing of hypotheses rather than parameter estimation. Both topics, regression and experimental design, will be discussed more fully in subsequent sections.

Population parameters

63. One must first have an understanding of what is meant by the term population parameter. The most common parameters are the mean, represented by μ , and the variance, represented by σ^2 . The mean, or arithmetic average, is the most familiar and commonly estimated parameter. It is calculated as the sum of all the individual observations (Y_i) in the population divided by the total number (N) of observations in the population.

$$\mu = Y_i + Y_2 + \ldots + Y_N = \left(\sum_{i=1}^{N} Y_i \right) \bigg/ N$$

where

N = the size of the population

Y = an observation from the population

i = a subscript value from 1 to N which identifies the individual observation being summed and Σ indicates that the observations are to be summed from observation $i = 1$ to observation $i = N$.

64. Variance is given by the formula

$$\sigma^2 = \sum_{i=1}^{N} (Y_i - \mu)^2/N$$

where previous definitions apply.

65. Variance is obtained from the sum (across all observations in the population from $i = 1$ to N) of the squared deviations of the individual observations (Y_i) from the population mean (μ). This sum of squared deviations is then divided by N , giving an "average squared deviation." If an individual observation deviates greatly from the mean, the variance will be high. If the observations deviate only slightly from the mean, the variance will be low.

66. It will be noted that, in determining the population variance, the deviations were squared. Clearly some of the individual values will fall below the mean, giving a negative deviation, and the remainder will fall above the mean, giving a positive deviation. With regard to sign, the deviations will thus always sum to zero. The sum of the actual deviations is, therefore, not useful as a measure of variability, and some method of considering only the size of the deviations, without regard to their sign, is needed. There are two obvious alternatives, the squares of the deviations and the absolute values, both of which are always positive. The square of the value has some theoretical advantages over the absolute value, so statistical calculations of variability usually employ the variance as defined above.

67. The variance is the average <u>squared</u> deviation of the individual observations from the mean of the population. It is reasonable to convert this value to the same scale as the observations by taking the square root. This is, indeed, what is usually done. The resulting value $\sigma = \sqrt{\sigma^2}$ is called the <u>standard deviation</u>.

68. The calculation described earlier for the variance was the

sum of the N squared deviations in the whole population, which was then divided by N . When dealing with a sample from a population, one usually defines the sample variance as the sum of the squared deviations from the sample mean divided by one less than the sample size. Thus, if the sample size is n , the sum of the squared deviations is divided by n-1 (called the degrees of freedom).

69. The population deviation was calculated as $Y_i - \mu$, where μ is known for the population and does not have to be estimated. For a sample, the deviation is calculated as $Y_i - \bar{Y}$, which uses the sample mean $\bar{Y}$. Since $\bar{Y}$ is calculated from the same sample used to calculate the sample variance (s^2), one degree of freedom is lost, and the denominator is reduced by one. The resulting calculation is

$$s^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2/(n - 1)$$

which gives an unbiased estimate of σ^2 . Using n - 1 rather than n corrects for the bias introduced by using x to estimate μ . The use of n - 1 instead of n is particularly important when n is small. Although it becomes less important as n gets larger, it appears to be good policy always to use n - 1 as the divisor.

Hypothesis testing

70. Repeated mention will be made of assumptions for the various tests and analyses discussed. All analyses assume that the data used are drawn at random from the target population, i.e., the population about which inferences are to be made.

71. We also need to consider the distribution of the sample mean, which exists conceptually rather than in reality. Specifically, a random sample from a population provides a single sample mean. A second independent sample from the same population provides a second sample mean. We could in theory, if not in practice, take an indefinitely large number of random samples of the population and thus obtain an indefinitely large number of sample means. These sample means then have

their own frequency distribution (the distribution of the sample means) which can be determined from knowledge of the population distribution and the sample size. The sample variance, likewise, has a distribution, as does any statistic or function of the sample observations.

72. A further assumption common to the early discussion of hypothesis testing is that the data are normally distributed. Tests of hypothesis have been developed based on the normal distribution because it commonly arises, if not exactly, as an extremely good approximation of the population distribution. Indeed, even if the distribution from which the samples are drawn is not normal, the distribution of the means of various samples will approximate the normal distribution for a sufficiently large sample size. The larger the sample size, the more nearly normal the distribution of the means. Although the rate at which the distribution of a sample mean approaches normality depends on the nature of the population distribution, it is quite rapid for most practical situations, and the normality assumption can be quite viable for fairly small samples.

73. Since the normal distribution plays a central role, it is pertinent that it be examined more closely. A normal distribution can be represented by a bell-shaped curve, three examples of which are given in Fig·re 9. Each of the curves in this figure has the same mean (μ = 20), but different variances (σ^2 = 1, 2, and 3). An objective of a test of hypothesis might be to detect values that are unusually large and therefore probably do not belong to a population with the hypothesized distribution. The taller, narrow curve in Figure 1 has the smallest variance, and the density, or occurrence of observations, is almost zero below 17 or above 23. Therefore, if a value of 15 were observed, it is highly unlikely that the value belongs in the narrowest distribution. However, it may well belong to one of the wider distributions.

74. Unfortunately, there is an infinite number of possible normal curves with different combinations of mean and variance. A method of standardizing the distribution is therefore necessary. The standardized normal distribution is a bell-shaped curve with a mean of zero and a variance of one as shown in Figure 10.

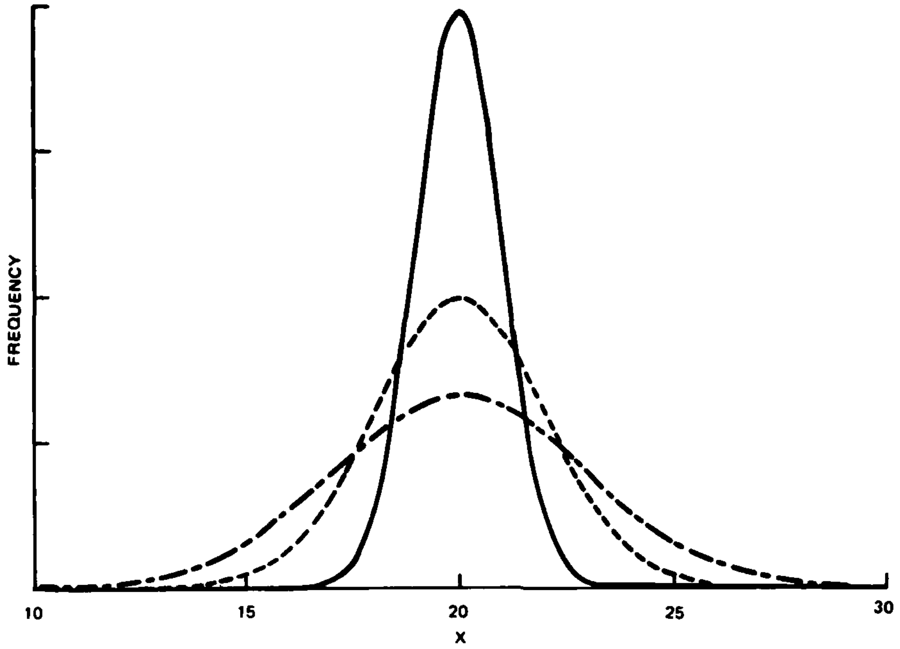


Figure 9. Three normal distributions with a common mean (μ = 20) but different variances

75. Examination of this curve will show that most of the distribution (indeed about 95 percent) is contained in the interval from -2 to +2, although the ends taper off to infinity. Standardization is accomplished by applying the formula

$$\text{Standard normal deviate} = z = \frac{Y - \mu}{\sigma}$$

76. If, for example, a normal distribution is given with a mean of 20 and a variance of 25, then would an observation with a value of 35 be a likely value to draw from the distribution? Standardization produces a value of [(35 - 20)/5] = 3. Ninety-five percent of the standard normal curve falls between -1.96 and +1.96, and 99 percent between -2.576 and +2.576. With an observed value of 35 either (a) the assumed normal distribution (μ = 20, σ^2 = 25) is applicable and a very rare event has occurred, or (b) the distribution from which the observation was drawn is other than that assumed. In this example the event would appear to be so rare that one would be prepared to believe (b) rather

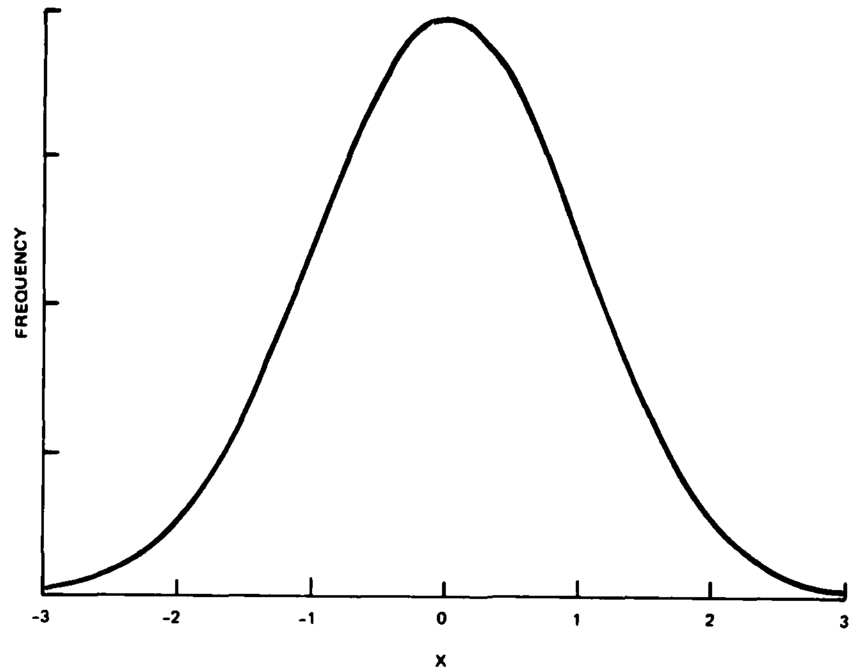


Figure 10. The standardized normal distribution

than (a). It is this type of reasoning that lies behind all statistical hypothesis testing.

One-Sample Hypotheses

77. The simplest type of test involves the comparison of an observed parameter estimate against some hypothesized value. As an example, suppose the objective is to determine if a reservoir chlorophyll concentration exceeds some subjective estimate of trophic state. Chlorophyll concentrations in excess of 10 mg/ℓ are generally considered to be indicative of eutrophy. To test reservoir chlorophyll concentrations against this level, the investigator might obtain samples at randomly selected times during the growing season.

78. The hypothesis to be tested should be stated formally in advance of the study. Hypotheses may be "one-sided" or "two-sided." In the case of the chlorophyll concentrations presented above, the two-sided test would have the null hypothesis stated as "the chlorophyll

concentration is 10 mg/ℓ," and the alternate hypothesis would state that "the chlorophyll concentration is not equal to 10 mg/ℓ."

79. These hypotheses can be stated mathematically as

NULL HYPOTHESIS $H_0: \mu = 10$

ALTERNATE HYPOTHESIS $H_A: \mu \neq 10$

80. However, these hypotheses are of little interest and are not appropriate for use in this situation. What is of interest is the magnitude of the chlorophyll concentration with respect to the value of 10 mg/ℓ. Therefore, the appropriate hypotheses are

NULL HYPOTHESIS $H_0: \mu > 10$

ALTERNATE HYPOTHESIS $H_A: \mu < 10$

81. This null hypothesis states that the parameter (μ) is greater than or equal to 10 mg/ℓ, which is indicative of eutrophy. The alternative hypothesis states that the parameter is less than 10 mg/ℓ, or chlorophyll concentrations are below the eutrophic level. These, then, are the correct hypotheses for the stated objective. The investigator is now faced with the decisionmaking process, the testing of the hypotheses. This will first require an estimate of μ, the target population mean to be tested. Suppose that a sample of 10 observations has been taken during the growing season and that the mean is 11.09 mg/ℓ. Can the reservoir be considered eutrophic? This value is above the boundary of 10 mg/ℓ, but first examine the raw data values below.

RAW DATA VALUES: 5.2, 6.3, 4.1, 13.2, 35.7,
3.5, 3.4, 6.0, 8.8, 24.7

82. Only three of the values exceeded the boundary condition of 10 mg/ℓ; the remaining values were considerably lower. However, we are

not interested in individual observations; we wish to test the estimate of the population parameter (μ) against 10 mg/ℓ. Even with most of the individual values below 10 mg/ℓ, it is possible that the actual population value is greater than 10 mg/ℓ.

83. An appropriate method for the evaluation of the observed mean is by way of the Student t-statistic, the value of which is given by

$$t = \frac{\bar{Y} - \mu_o}{s_{\bar{Y}}}$$

where

$\bar{Y} = \hat{\mu}$, the estimated (indicated by the "hat," ∧) value of the parameter (i.e., the sample mean)

μ_o = hypothesized value (10 mg/ℓ in the example)

$s_{\bar{Y}}$ = square root of the estimated variance of the mean ($s^2_{\bar{Y}}$) (see below for the calculation of $s^2_{\bar{Y}}$)

Note: the sample mean follows a distribution which has its own mean and variance; $s^2_{\bar{Y}}$ is a sample-based estimate of the latter.

84. The statistic provides a measure of the size of the difference between the measured and hypothesized value relative to the variability of the mean. If the t value is large enough, the null hypothesis may be rejected; in this case, the difference is said to be "significant."

85. The problem with simply looking at the mean is that it is impossible to know the characteristics of the distribution unless previous work has been done on the parent population. How large a difference is large enough to indicate that the difference is significant? The advantage with the t statistic is that the characteristics of its distribution are known, provided that the population distribution is normal. Before continuing, the t distribution should be examined.

86. The t distribution is described by a bell-shaped curve, similar to the normal curve discussed earlier. The curve is symmetrical, highest in the center, and tapers on either end. The curve is centered on zero, and about 0.66 of the total area under the curve (66 percent)

is contained in an interval one standard deviation unit above and below the center (the standard deviation is a measure of the variability). An interval from two standard deviation units below the center to two standard deviation units above the center again contains about 0.95 (or 95 percent) of the total area. The difference between the t distribution and the standard normal distribution is that there is but one standard normal distribution. There are many t distributions, one for each possible sample size. For very large samples, the t and standard normal distributions are virtually identical, but the t distribution is wider for smaller samples.

87. Just as important as the area in the center of the distribution is the area in the tail (or tails) of the distribution. Values that fall outside the two standard deviation units from the mean only occur about 5 percent of the time, so they are relatively rare events, not expected to occur very often by random chance. The t table, Table 7, gives the values of t that will be exceeded with specified probability p .

88. If a t value were calculated for a sample of size 10, the sample would have 9 degrees of freedom (d.f.), and the appropriate tabular values would be obtained from the line corresponding to d.f. = 9 . For a two-sided (two-tailed) hypothesis test, a t value of 1.833 or larger would be found by random chance 1/10 of the time, a value of 2.262 or larger would be found 1/20 of the time, and a value of 3.250 or more only 1/100 of the time. If we were to calculate a t value and find it was 2.3, then either the sample is an unusual one (occurring only about 5 percent of the time by random chance) or the true mean of the population is other than hypothesized. We may thus be prepared to conclude that the null hypothesis is false. In so doing, we would err about once in 20 tests, if we always employed a probability level of 0.05.

89. The above demonstrates the central theme in understanding tests of hypotheses. Even if the values come from an underlying population in which the actual mean is equal to or greater than 10, it is still possible to observe individual values that are less than 10. It

Table 7

Representative Portion of the t Table

	Probability Level					
	Two Tails			One Tail		
d.f.	0.10	0.05	0.01	0.10	0.05	0.01
1	6.314	12.706	63.657	3.078	6.314	31.821
2	2.920	4.303	9.925	1.886	2.920	6.965
3	.	.	.	.	.	.
4	.	.	.	.	.	.
5	.	.	.	.	.	.
6	.	.	.	.	.	.
7	.	.	.	.	.	.
8	.	.	.	.	.	.
9	1.833	2.262	3.250	1.383	1.833	2.821
10	1.812	2.228	3.169	1.372	1.812	2.764
20	.	.	.	.	.	.
30	.	.	.	.	.	.
∞	.	.	.	.	.	.

is even possible that the sample mean will be less than 10, although this would be less common.

90. For example, we now evaluate the t value for the data set given above on chlorophyll values, for testing against the standard value of 10.0.

91. The sample mean and variance are calculated first, thus

$$\bar{Y} = \Sigma Y_i/n = 110.9/10 = 11.09$$

$$\Sigma Y_i^2 = 2,279.61$$

92. The sum of squared deviations is given by

$$\Sigma y_i^2 = \Sigma(Y_i - \bar{Y})^2 = \Sigma Y_i^2 - (\Sigma Y_i)^2/n = 2,279.61 - (110.9)^2/10 = 1,049.73$$

and the variance by

$$s^2 = \Sigma y_i^2/(n-1) = 1,049.73/9 = 116.64$$

The variance of the means is not the same as the population variance. Means are expected to be much less variable than the individual observations in the population. Indeed, the variance of the means of samples of size n equals the population variance divided by n and may be estimated as the sample variance divided by n ; thus,

$$s_{\bar{Y}}^2 = s^2/n = 116.64/10 = 11.664$$

The standard error, used for the t test, is the square root of the variance of the means. The t statistic is then calculated as

$$t = \frac{\bar{Y} - \mu_o}{s_{\bar{Y}}} = \frac{11.09 - 10}{3.415} = 0.319$$

93. This value is then compared to the value in the table of t values. The critical value selected depends on the degrees of freedom and the probability of error selected. If a probability of error of 0.05 were selected for this one-tailed example and since there are 9 d.f., the critical value would be 1.833 (one-tailed); so, the test statistic above would not fall outside the range of normal variation. The hypothesis that these values could have come from a population with a mean greater than or equal to 10.0 could not be rejected on the basis of the available data.

94. There is one other point that should be made about hypothesis testing. In doing the test above, the probability of error that was set was for one type of error, called Type I error, or α (alpha) error. This error probability is the "probability of rejecting a null

hypothesis which is true." In the example above, the null hypothesis was $\mu \geq 10.0$, which indicates compliance. Therefore, the probability of erroneously rejecting this hypothesis and concluding that the samples indicated noncompliance was required to be at most 0.05 or 5 percent.

95. There is, however, another type of error, called a Type II or β (beta) error. This is the probability of accepting a null hypothesis which is false. In the above example, since the null hypothesis was not rejected, there is a possibility of a β error. Unfortunately, the probability of a β error depends on the magnitude of the deviation from the null hypothesis, which is unknown. Nevertheless, the t test will, if the assumptions are met, minimize the probability of a β-type error. Therefore, for the data available and under the assumptions previously stated, the t test is considered a "powerful" test because it does have the smallest probability of β error.

Confidence Intervals

96. It has been pointed out that the sample mean can be regarded as having a distribution although usually we will have only one observation, i.e., one sample mean, from that distribution. Further, the standard error plays the same role with respect to the distribution of the sample mean as the standard deviation does to the individual population values. Recall that in the case of a normal distribution, 95 percent of the population values are contained within an interval approximately two standard deviations above and below the population mean--the actual value is 1.96 standard deviation units. We might, therefore, expect to be able to determine an interval about the sample mean, as a multiple of the standard error, which would contain 95 percent of the possible sample means, which are distributed about the population mean. Looked at in another way, this is equivalent to determining an interval that, with high probability, say 0.95, would contain the population mean. We refer to this as a 95-percent confidence interval for the mean; it is a measure of the precision of the sample mean.

97. Because we are dealing with the sample, we cannot determine

such an interval as simply as with the population, given normality of distribution. Recall, however, that the frequency distribution of $t = \frac{\bar{Y} - \mu}{s_{\bar{Y}}}$ is known, where μ is here the actual, but unknown, population mean. We can thus readily determine, usually by reference to tables, a value t_o such that

$$P(t \geq t_o) = \alpha/2$$

where α is a specified, usually small value, say 0.05.

98. That is

$$P\left(\frac{Y - \mu}{s_{\bar{Y}}} \geq t_o\right) = \alpha/2$$

99. On rearranging, this is equivalent to

$$P(Y - t_o s_{\bar{Y}} \geq \mu) = \alpha/2$$

100. Likewise we can determine t_1 such that

$$P\left(\frac{\bar{Y} - \mu}{s_{\bar{Y}}} \leq t_1\right) = \alpha/2$$

101. Indeed $t_1 = -t_o$, i.e.,

$$P\left(\frac{\bar{Y} - \mu}{s_{\bar{Y}}} \leq -t_o\right) = \alpha/2$$

which on rearrangement gives

$$P(Y + t_o s_{\bar{Y}} \leq \mu) = \alpha/2$$

That is,

$$P\ (\mu \leq Y - t_o s_{\bar{Y}}) = \alpha/2$$

$$P\ (\mu \geq Y + t_o s_{\bar{Y}}) = \alpha/2$$

so that

$$P\ (Y - t_o s_{\bar{Y}} \leq \mu \leq Y + t_o s_{\bar{Y}}) = 1 - (\alpha/2 - \alpha/2)$$

$$= 1 - \alpha = 0.95$$

with $\alpha = 0.05$.

102. Thus, $Y \pm t_o s_{\bar{Y}}$ provides a confidence interval for the mean. Loosely, we say that the probability that the population mean is contained within this interval is $1 - \alpha$ (= 95 percent if $\alpha = 0.05$). What we are really saying, however, is that intervals computed this way will, on the average, cover the population mean 95 times out of 100. The probability statement is about the interval and not the population mean which is a fixed, but unknown, quantity. Note that $t = \frac{\bar{Y} - \mu}{s_{\bar{Y}}}$ depends only on the sample and the unknown parameter; such a statistic is called a "pivotal" quantity.

103. The calculations will be demonstrated using the previous values for the chlorophyll concentration example. Degrees of freedom are the same as previously, i.e., d.f. = 9 . However, since confidence intervals are usually placed symmetrically above and below the mean value, a two-sided t value is needed for the 9 d.f. (t = 2.262). The calculations proceed as follows:

$$\bar{Y} = 11.09$$

$$s_{\bar{Y}}^2 = s^2/n = 11.664$$

$$s_{\bar{Y}} = 3.415$$

Lower 95-percent confidence limit (L_1)

$$L_1 = \bar{Y} = ts_{\bar{Y}} = 11.09 - 2.262(3.415) = 3.37$$

Upper 95-percent confidence limit (L_2)

$$L_2 = \bar{Y} + ts_{\bar{Y}} = 11.09 + 2.262(3.415) = 18.81$$

104. In the sense indicated above, we may feel 95 percent sure that the true value of μ, the population mean, lies within the interval 3.37 to 18.81. There is, of course, on the average, a 5-percent chance that the true value of μ falls outside the interval.

Two-Sample Hypotheses

105. In many situations it is necessary to compare two means rather than simply test a mean against a hypothesized value. Examples include the comparison of two reservoirs, two areas in the same reservoir, or the same area at two different times of the year. Obviously, such comparisons require two samples and lead to two sample hypotheses. As with one-sample hypotheses, two-sample hypotheses may be two-sided or one-sided.

Two-sided:

$$H_o:\mu_1 = \mu_2 \quad \text{or} \quad H_o:\mu_1 - \mu_2 = 0$$

$$H_A:\mu_1 \neq \mu_2 \quad \text{or} \quad H_A:\mu_1 - \mu_2 \neq 0$$

One-sided:

$$H_o:\mu_1 \leq \mu_2 \quad \text{or} \quad H_o:\mu_1 - \mu_2 \leq 0$$

$$H_A:\mu_1 > \mu_2 \quad \text{or} \quad H_A:\mu_1 - \mu_2 > 0$$

These hypotheses can be tested using the t test. The t test, as before, requires an assumption of randomly selected samples from normal populations and, strictly speaking, also requires an assumption of homogeneity of variance (i.e., variances of the two populations are equal).

106. The F test is used to test the equality of two population variances. It also arises as a test for differences between two or more means in an analysis of variance (see below). If both populations are normally distributed with equal variances, the distribution of the ratio of the sample variances is known; it is called the F distribution.

107. Unlike the normal and t distributions, the F distribution is asymmetric. Since two variances are estimated, there will be degrees of freedom for the numerator and denominator of the ratio, which will not necessarily be the same.

108. Specifically,

$$F = s_1^2/s_2^2$$

where

s_1^2 = one variance estimate with DF = $n_1 - 1$

s_2^2 = second variance estimate with DF = $n_2 - 1$

109. The F test most commonly tests the hypotheses

$$H_o:\sigma_1^2 = \sigma_2^2$$

$$H_A:\sigma_1^2 \neq \sigma_2^2$$

The two-tailed F test is commonly calculated with the larger variance in the numerator. This is because tables of the F statistic are often used in conjunction with the analysis of variance, in which, perforce, the test is one-tailed. Given this convention, the F statistic becomes larger and larger as the two variances become more different. Therefore, a significantly large F value indicates the inequality of variances, whereas a small F value indicates the equality of variances.

110. Calculated values of the F statistic must be compared with the proper critical value of the F distribution from an F table. The critical value of F is specified by the numerator and denominator degrees of freedom. If the calculated F statistic exceeds the critical value, the null hypothesis $(H_o:\sigma_1^2 = \sigma_2^2)$ can be rejected; whereas, if the F statistic is less than the critical value, the assumption of homogeneity of variance can be accepted.

111. If the assumption of equal variances can be made on the basis of the result of the F test, one can proceed with the t test for the comparison of sample means. The two-sample t statistic is

$$t = (\bar{Y}_1 - \bar{Y}_2)/s_{\bar{Y}_1-\bar{Y}_2}$$

where $\bar{Y}_1 - \bar{Y}_2$ is simply the difference between the two means and $s_{\bar{Y}_1-\bar{Y}_2}$ is the standard error of the difference between the means.

112. The quantities $s_{\bar{Y}_1-\bar{Y}_2}$ and its square, $s^2_{\bar{Y}_1-\bar{Y}_2}$ are statistics that can be calculated from the sample and are estimates of the parameters $\sigma_{\bar{Y}_1-\bar{Y}_2}$ and $\sigma^2_{\bar{Y}_1-\bar{Y}_1}$, respectively. The variance of the difference (or sum) of two variables is equal to the sum of their variances, so

$$\sigma^2_{\bar{Y}_1-\bar{Y}_2} = \sigma^2_{\bar{Y}_1} + \sigma^2_{\bar{Y}_2}$$

Remember that the variance of the mean (i.e., the standard error of the mean) is the variance divided by sample size

$$\sigma^2_{\bar{Y}} = \sigma^2/n$$

Therefore, the variance of the difference between the means can be written as

$$\sigma^2_{\bar{Y}_1-\bar{Y}_2} = \sigma^2_1/n_1 + \sigma^2_2/n_2$$

Given the assumption of equal variances ($\sigma^2_1 = \sigma^2_2$) we can write

$$\sigma^2_{\bar{Y}_1-\bar{Y}_2} = \sigma^2/n_1 + \sigma^2/n_2$$

113. In order to estimate $\sigma^2_{\bar{Y}_1-\bar{Y}_2}$ it is necessary to estimate σ^2 . Given the assumption of equal variances, both samples can be used to calculate a pooled variance $\left(s^2_p\right)$ to estimate σ^2 .

$$s^2_p = \left(DF_1 s^2_1 + DF_2 s^2_2\right) \Big/ (DF_1 + DF_2)$$

where

DF_1 = degrees of freedom, sample 1

s^2_1 = variance of sample 1

DF_2 = degrees of freedom, sample 2

s^2_2 = variance of sample 2

114. The pooled variance is then used to calculate the variance of the difference between the means

$$s^2_{\bar{Y}_1-\bar{Y}_2} = \left(s^2_p/n_1\right) + \left(s^2_p/n_2\right)$$

The standard error of the difference between the means is the square root of the variance estimate

$$s^2_{\bar{Y}_1-\bar{Y}_2} = \sqrt{\left(s^2_p/n_1 \ + \ s_p/n_2\right)}$$

and has degrees of freedom equal to the sum of the degrees of freedom of the two samples

$$DF_p = DF_1 + DF_2$$

115. The calculated t statistic

$$t = (\bar{Y}_1 - \bar{Y}_2)/s_{\bar{Y}_1 - \bar{Y}_2}$$

is then compared to the critical value of t selected from the t table. A sample t statistic larger than the critical value would lead to the rejection of the null hypothesis of equality of means.

116. As an example of the use of the two-sample F and t tests, suppose that phosphorus concentrations have been measured at random from two areas in the same reservoir, and the sample means and variances calculated. These sample estimates are given below.

Area	Number of samples (n)	Mean, μg/ml (Y)	Variance (s^2)
A	7	14	45
B	11	7	20

117. Prior to sampling we have no basis for a hypothesis about which area has higher concentrations or more variation. Therefore, a two-tailed test will be employed for both the F test and the t test. The hypotheses are

$$H_o:\sigma_A^2 = \sigma_B^2 \qquad H_A:\sigma_A^2 \neq \sigma_B^2$$

and

$$H_o:\mu_A = \mu_B \qquad H_A:\mu_A \neq \mu_B$$

118. The F statistic is

$$F = s_A^2/s_B^2 = 45/20 = 2.25$$

For a 5-percent level test, the critical value from the table with degrees of freedom 6 (numerator) and 10 (denominator) is 4.07. (One must check whether the table is designed for one- or two-tailed tests.) This value was not exceeded, so we conclude that 2.25 is not an unusually large F value and that the assumption of equal variances can be accepted.

119. Since the assumption of equal variances is acceptable, the two estimates may be combined into a single estimate of the pooled variance, given by

$$s_p^2 = \left(DF_A s_A^2 + DF_B s_B^2\right) \Big/ (DF_A + DF_B)$$

$$= \frac{[6(45) + 10(20)]}{16} = \frac{470}{16} = 29.375$$

This estimate of the combined variance also has degrees of freedom equal to the sum of the degrees of freedom for the two components (6 + 10 = 16). It can then be used to calculate the standard error for the t test of the two means.

$$s^2_{\bar{Y}_A - \bar{Y}_B} = \left(s_p^2/n_A\right) + \left(s_p^2/n_B\right)$$

$$= (29.375/7) + (29.375/11)$$

$$= 6.687$$

$$s_{\bar{Y}_A - \bar{Y}_B} = 2.620$$

120. The two-sample t statistic is then

$$t = (\bar{Y}_A - \bar{Y}_B)/s_{\bar{Y}_A - \bar{Y}_B}$$

$$= (14 - 7)/2.620$$

$$= 2.672$$

This value would be compared to the tabular t value for 16 d.f. which is 2.120 for a probability of type I error of $\alpha = 0.05$. The assumption of equal phosphorus concentrations would thus be rejected by a 5-percent level test.

121. One problem with the F test is that it is much more dependent on the normality assumption that the t test. Rejection of the null hypothesis (equal variances) may be due to a violation of this assumption rather than a difference in the population variances.

Regression

122. Regression is a type of statistical analysis that is used to express and quantify the relationship between two variables. Its application usually requires the estimation of two or more parameters of a target population. Regression analysis also involves hypothesis testing. Any time a regression is performed, there is an implicit hypothesis that the slope is not zero. A zero value for the slope implies that there is no relationship between the variables.

Linear regression

123. In a quantitative analysis of data, relationships are often observed between two variables. One objective of statistical analysis is to express these relationships quantitatively and, often, to test the magnitude of the relationship, if any, between the two variables.

124. The population regression line follows the average of one variable (Y) at unique values of the other (X). In some cases the relationship is obvious and direct. For example, the salinity of water is often measured as conductivity because increased salinity results in increased conductivity. In other cases the relationship is not as obvious. There are many relationships between water quality variables that can be described quantitatively. The statistical method used to quantify these relationships is called regression analysis.

125. It should be noted that the demonstration of a relationship does not in any way imply "cause and effect." A strong relationship may be demonstrated simply because both variables relate to a third variable

which may not have been included in the analysis.

126. Figure 11 demonstrates the relationship measured between two variables. Pairs of values (X and Y) are measured as sample values. There is an obvious tendency for Y to increase as X increases, indicating that there is some kind of quantitative relationship between the two variables.

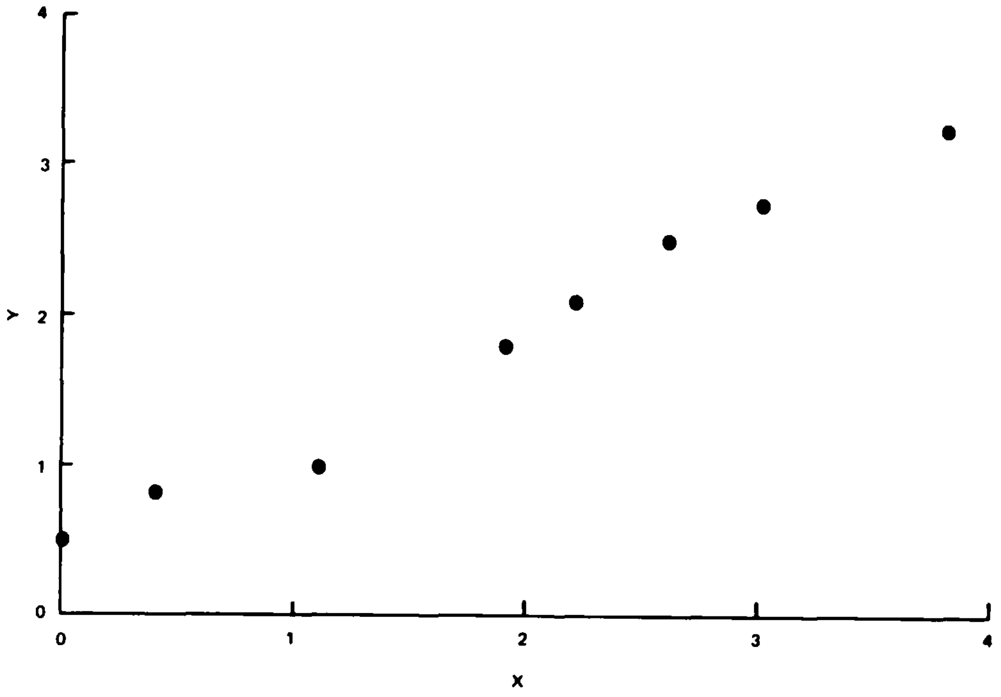


Figure 11. Bivariate scatter plot of X and Y

127. There are two aspects of this line expressing a relationship between the variables X and Y which can be used to describe the relationship quantitatively. The first is the angle that the line makes with the X axis, that is, the slope of the line. As one passes from X = 0 to X = 1 , say, there is an increase in Y . Since the line is straight, for any increase in X of one unit, there is a constant increase in Y . This "increase in Y per unit increase in X " provides a measure of the slope.

128. The second aspect of the line needed to quantify it is some measure of its level. Knowing only the slope, there are an infinite number of possible lines which could be drawn, each parallel to the

other. If we can define one point that the line must pass through, there is then only one line that will satisfy the conditions. This point is usually defined as the value of Y when $X = 0$, or the point at which the line crosses the Y axis. This point is known as the intercept.

129. The simplest linear regression model is represented by the equation

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where

Y_i = an individual observation of variable Y

X_i = a value of variable X

α = a population parameter for the intercept of the regression line between X and Y

β = a population parameter for the slope of the regression line

ε_i = a random variate describing the deviations of the observed points from the line

130. The method usually used to find estimates for the population parameters for a linear relationship is least squares regression. It consists of finding a line that minimizes the sum of the squares of the vertical distances between the points and the line. The bars connecting the points to the line in Figure 12 indicate the distances the sum of squares of which would be minimized in a least squares regression estimate.

131. Some assumptions are made whenever a least squares regression line is fit to a data set. First, we must assume that the relationship that is used is appropriate. If a straight line is used, there is an assumption that the relationship is linear. Other options exist and are discussed later. Several additional assumptions are given below.

<u>a</u>. The units in the sample with any one value of X are randomly chosen from all units in the target population with that value of X . This can be achieved by selecting the units at random from the entire target population.

b. The values of Y at any particular value of X are normally distributed about the regression line.

c. The variance of the Y values is homogeneous, that is, the variance of Y is the same at each value of X .

d. The differences between the points and the line, i.e., the ε_i , are independent; this can be achieved by independent, random selection of the units. Multiple observations in a single unit will not, in general, be independent.

e. All of the variation, due to sampling or other causes, occurs in the Y variable; the X variable is measured without error.

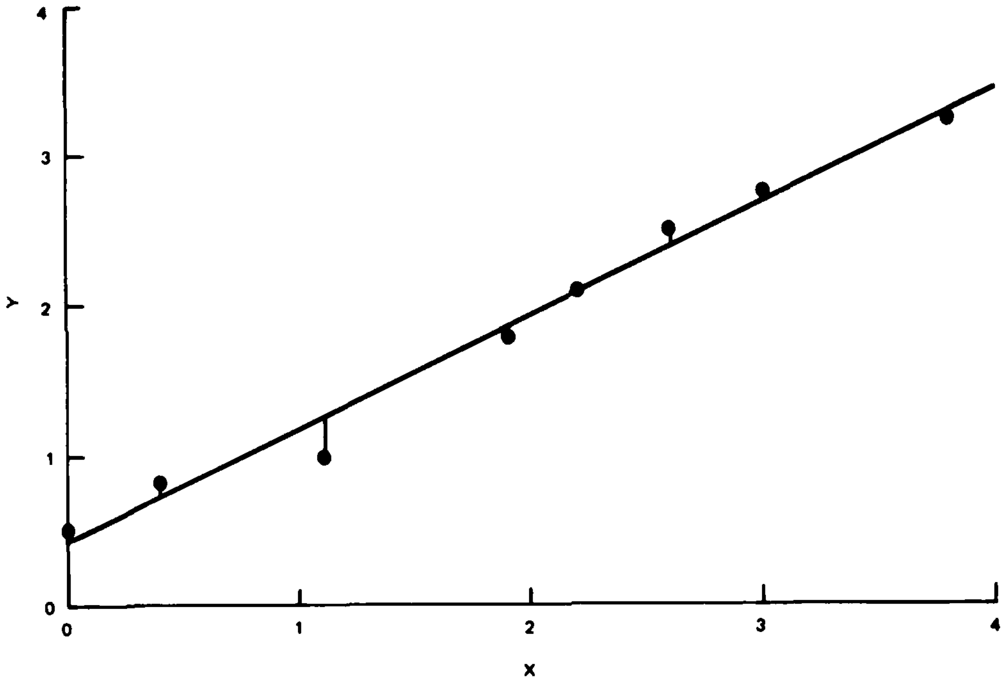


Figure 12. Representation of the distances to be minimized by least squares regression

132. The first assumption is made to ensure that the values chosen are representative of the target population. Such an assumption is appropriate for all statistical analyses. The second assumption is necessary only if hypothesis tests or confidence intervals are required. The statistics employed will then have t and F distributions. The third and fourth assumptions are necessary for the parameter estimates to be

"best," i.e., to be the most precise possible. Sometimes, it is observed that the Y values are more widely scattered as the X value increases. In this case the assumption of homogeneous variance is violated. The difficulty may be overcome by using one of the transformations discussed later.

133. It should be noted that the least squares fit will always yield an unbiased regression line which minimizes the vertical distances, even if the values are not normally distributed, the variances heterogeneous, and the ε_i not independent. Therefore, the estimated values may be useful. The hypothesis tests and confidence intervals would not be correct, however, if the normality assumption is not met.

134. The fifth assumption is necessary because the linear regression techniques used minimize the vertical distance. No consideration is made for variation in X . If the assumption is not met, the estimates will be biased although the bias may be negligible if the measurement errors are small in relation to the standard deviation of the X values employed. Statistical techniques do exist which can address the problem, but are not as easy to use and are not usually applied.

135. As one might gather from the above, regression analysis is generally considered to be robust against violation of several assumptions. This means that adequate results can often be obtained even when minor violations of the assumptions are made.

Fitting simple linear relationships

136. Fitting the regression line requires that estimates of population parameters be obtained from a sample. The values obtained, of course, are not α and β (the population parameters), but rather a and b , the sample estimates of the population parameters. The resulting equation is then

$$Y_i = a + bX_i + e_i$$

where a is an estimate of α , b an estimate of β , and each e_i an estimate of the corresponding ε_i .

137. Six intermediate values must be obtained from the sample data to fit a regression relationship. These are:

n--the sample size
ΣX_i--the sum of the values of the X variable
ΣY_i--the sum of the Y values
ΣY_i^2--the sum of the squares of the Y variables
ΣX_i^2--the sum of squares of the X variable
$\Sigma X_i Y_i$--the sum of the products of the X and Y variables

The slope of the fitted line is calculated as:

$$b = \Sigma xy/\Sigma x^2$$

The lowercase representations used for Σx and Σxy signify that these values have been adjusted about the means, i.e., $x = X-\bar{X}$, $y = Y-\bar{Y}$.

Thus,

$$\Sigma x^2 = \Sigma(X - \bar{X})^2 = \Sigma X - (\Sigma X)^2/n$$

$$\Sigma xy = \Sigma(X - \bar{X})(Y - \bar{Y}) = \Sigma XY - \Sigma X \Sigma Y/n$$

We shall later require

$$\Sigma y^2 = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/n$$

138. The estimated line will pass through the sample mean $(\bar{X},\bar{Y})$. Failure to make the adjustment will force the regression line through the origin.

139. Once the slope is calculated, the intercept can be estimated. This is obtained as

$$a = \bar{Y} - b\bar{X}$$

where

$\bar{Y}$ = mean of the Y variable

$\bar{\bar{X}}$ = mean of the X variable

b = previously calculated slope

140. Once the estimates of the slope and intercept have been calculated, it is necessary to determine how well the regression line "fits" the data. In other words, how much of the variation in the dependent variable is explained or accounted for by the regression with the dependent variable. The total variability in the dependent variable is calculated by computing the sum of the squared deviations $(Y_i - \bar{Y})$, called the corrected total sum of squares

$$SS_{\text{corr. total}} = \Sigma(Y_i - \bar{Y})^2 = \Sigma y^2$$

141. The amount of variability among the Y_i values that is accounted for by the regression is the sum of the squared deviations $(\hat{Y}_i - \bar{Y})$, called the model or regression sum of squares

$$SS_{\text{model}} = \Sigma(\hat{Y}_i - \bar{Y})^2 = (\Sigma xy)^2/\Sigma x^2$$

where the $\hat{Y}_i$ (reads as "Y hat") values are the Y values generated from the regression equation

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

The model sum of squares can also be calculated by

$$SS_{\text{model}} = b\Sigma xy$$

The model sum of squares expresses the amount of variability of the $\hat{Y}_i$ values about the mean of the Y values.

142. The amount of variability not accounted for by the regression is the sum of the squared deviations $(Y_i - \hat{Y}_i)$, called the error or residual sum of squares

$$SS_{error} = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma y^2 - b\Sigma xy$$

The total variability in Y is the sum of the variability accounted for by the regression and the error variability, or

$$SS_{corr.\ total} = SS_{model} + SS_{error}$$

Therefore, given the corrected total and the model sum of squares, the error sum of squares can be calculated by difference,

$$SS_{error} = SS_{corr.\ total} - SS_{model}$$

The SS_{error} expresses the variability due to differences between the sampled values of Y (Y_i) and the $\hat{Y}_i$ values generated from the regression equation.

143. The proportion of the total variation in the dependent variable that is accounted for by the regression is called the coefficient of determination (r^2), where

$$r^2 = SS_{model}/SS_{corr.\ total}$$

Values of r^2 can range from 0 (no relationship between Y and X) to 1 (every value of Y_i lies on the regression line, $Y_i = \hat{Y}_i$).

144. The model and error sum of squares also provide the basis for a test of the significance of a regression. The test is based on the null hypothesis that the slope is zero, or that no relationship exists between Y and X . The hypotheses are

$$H_o:\beta = 0 \quad H_A:\beta \neq 0$$

145. An F test is used to determine if the calculated slope is significantly different from zero. In order to perform this test it is necessary to calculate the model and error mean squares. Mean squares (short for mean squared deviations) are calculated by dividing the model

and error sum of squares by their respective degrees of freedom

$$MS_{model} = SS_{model}/DF_{model}$$

$$MS_{error} = SS_{error}/DF_{error}$$

where

$$DF_{model} = 1$$

$$\begin{aligned} DF_{error} &= DF_{total} - DF_{model} \\ &= (n - 1) - 1 \\ &= n - 2 \end{aligned}$$

Note: DF_{model} for a linear regression will always be 1.

146. The F statistic used to test the null hypothesis is

$$F = MS_{model}/MS_{error}$$

and is then compared to the critical F value with numerator $DF = DF_{model}$ and denominator $DF = DF_{error}$.

147. A t test can also be used to test the null hypothesis of a zero slope, as well as testing the significance of the intercept. The two-tailed hypotheses are

$$H_o{:}\beta = 0 \qquad H_A{:}\beta \neq 0$$

and

$$H_o{:}\alpha = 0 \qquad H_A{:}\alpha \neq 0$$

148. The t statistic used to test the slope is

$$t = (b - 0)/s_b = b/s_b$$

where

b = calculated slope

s_b = standard error of b

149. Similarly, the t statistic used to test the intercept is

$$t = (a - 0)/s_a = a/s_a$$

where

a = calculated intercept

s_a = standard error of a

The standard error of b is

$$s_b = \sqrt{\left(MS_{error}/\Sigma x^2\right)}$$

and the standard error of a is

$$s_a = \sqrt{MS_{error}\left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma x^2}\right)}$$

These t tests are evaluated in the same manner as for those t tests previously discussed.

150. Use of the t test also allows for the testing of one-sided hypotheses about the slope. Either

$$H_o{:}\beta \le c \quad H_A{:}\beta > c$$

or

$$H_o{:}\beta \ge c \quad H_A{:}\beta < c$$

can be tested using a t statistic. Remember that the F test only allowed for the two-sided hypothesis. The t statistic used for these tests is calculated as

$$t = (b - c)/s_b$$

151. As an example of a regression analysis, suppose that an investigator wishes to quantify the relationship between precipitation in a subwatershed of the reservoir and the mean daily discharge from this area into the reservoir. The precipitation may be measured at one site in the subwatershed and the discharge (in 1,000 m/day) could be determined from gage data. Sample data are given below.

Precipitation (X)	Discharge (Y)
0.0	19.3
0.2	20.5
1.3	27.4
1.7	25.7
2.5	34.1
3.2	50.4
6.1	68.4

152. Note that a point has been included which has zero precipitation. This is to illustrate the fact that some discharge is expected even when there is no precipitation.

153. It is obvious from the data that there is some relationship between the precipitation in the basin and discharge into the reservoir. The calculations now proceed as described. The six intermediate values required are

$$\Sigma X = 15$$

$$\Sigma Y = 245.8$$

$$\Sigma X^2 = 58.32$$

$$\Sigma Y^2 = 10{,}585.52$$

$$\Sigma XY = 747.18$$

$$n = 7.$$

154. The next step is to calculate the adjusted sums of squares

and products needed for the calculation of the slope, and the mean values needed to calculate the intercept.

$$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 58.32 - (15)^2/7 = 26.177$$

$$\Sigma xy = \Sigma XY - (\Sigma X \Sigma Y)/n = 747.18 - 15(245.8)/7 = 220.466$$

$$Y = \Sigma Y/n = 245.8/7 = 35.114$$

$$X = \Sigma X/n = 15/7 = 2.143$$

We shall later require

$$\Sigma y^2 = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/n = 10{,}585.52 - (245.8)^2/7 = 1{,}954.429$$

155. The final calculations are then made to obtain estimates for the slope and intercept.

$$b = \Sigma xy/\Sigma x^2 = 220.466/26.177 = 8.422$$

$$a = \bar{Y} - b\bar{X} = 35.114 - 8.422(2.143) = 17.067$$

156. The resulting regression line is shown in Figure 13. The slope is positive, indicating that, as would be expected, Y (the discharge) increases as X (the precipitation) increases.

157. The sums of squares are

$$SS_{\text{corr. total}} = \Sigma y^2 = 1{,}954.429$$

$$SS_{\text{model}} = b\Sigma xy = (8.422)(220.466) = 1{,}856.765$$

$$SS_{\text{error}} = SS_{\text{corr. total}} - SS_{\text{model}}$$

$$= 1{,}954.429 - 1{,}856.765 = 97.664$$

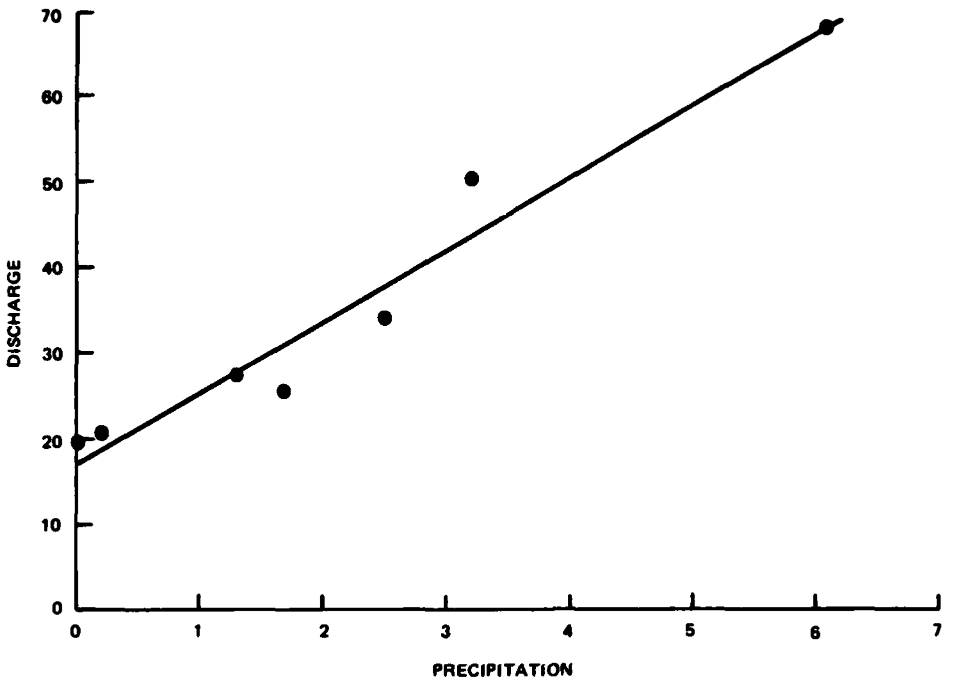


Figure 13. Linear regression of discharge and precipitation

Therefore, the coefficient of determination is

$$r^2 = SS_{model}/SS_{corr.\ total}$$

$$= 1,856.765/1,954.429$$

$$= 0.95$$

The value of r^2 indicates that 0.95 or 95 percent of the variation in the dependent variable is accounted for by the regression.

158. The F statistic used to test the null hypothesis that the slope is zero is

$$F = MS_{model}/MS_{error}$$

where

$$MS_{model} = SS_{model}/DF_{model} = 1{,}856.765/1 = 1{,}856.765$$
$$MS_{error} = SS_{error}/DF_{error} = 97.664/5 = 19.533$$

$$F = 1{,}856.765/19.533$$
$$= 95.059$$

The critical value of F, with $\alpha = 0.05$, numerator $DF = 1$, and denominator $DF = 5$, is 10.0. The calculated F statistic is much larger than the critical value and, as a result, the null hypothesis can be rejected.

159. The t statistic used to test the same hypothesis ($H_o{:}\beta = 0$) is

$$t = (b - o)/s_b$$
$$= 8.422 \bigg/ \sqrt{MS_{error}/\Sigma x^2}$$
$$= 8.422 \big/ \sqrt{19.533/26.177}$$
$$= 9.750$$

The critical value of t, with $\alpha = 0.05$ and $DF = 5$, is 2.571. As with the F test, the null hypothesis can be rejected.

160. Regression lines not only quantify a relationship, but also allow for the estimation of the average value of Y at any specified value of X, whether or not the X value was observed. For example, no precipitations of 5 cm were observed, but the regression relationship can estimate the mean discharge for a precipitation of 5 cm. This would be calculated as

$$\hat{Y} = a + bX = 17.067 + 8.422(5) = 59.177$$

161. The regression line can also be used to estimate values

outside the range of observed data. In the example, the greatest observed precipitation was 6.1 cm. However, the discharge can be calculated for a greater value, 7 cm, 10 cm, or any other number. For example, the estimated discharge resulting from 10 cm of rainfall is

$$\hat{Y} = a + bX = 17.067 + 8.422(10) = 101.287$$

Caution must be employed, however. There is a tacit assumption that the same linear relationship applies outside the range of the data. This may or may not be the case and, unfortunately, cannot be tested unless the range of the data is extended.

162. It is also important to understand that the closer one moves to the extremes of the data, the less precise are the estimates of the mean discharge. The most precise estimate is the one that occurs at the mean of the X values, namely $\bar{Y}$ (since a simple linear regression line always passes through the sample means). The precision is even less outside the range of the data, even if extrapolation of the line is valid. Thus, the precision of the estimate, or the confidence in the predictive ability, decreases as the distance from the mean of the X values increases. This can be graphically illustrated by a confidence interval about the regression line.

Confidence intervals for regression

163. The regression line is an estimate of the true situation for a population, and it can be given a confidence interval. Since the estimate is a line, the confidence interval is a band about the line. It has been pointed out that estimation becomes less precise the farther one moves away from $\bar{X}$. This is reflected into the confidence interval bands which are narrowest (closest to the regression line) at the point where $X = \bar{X}$ and become wider in either direction moving away from $\bar{X}$ (Figure 14).

164. Confidence intervals for the regression line can be calculated as an interval at any point on the regression line and are then connected to give the confidence bands. The equation for the confidence interval at any value of X , say X_o , is

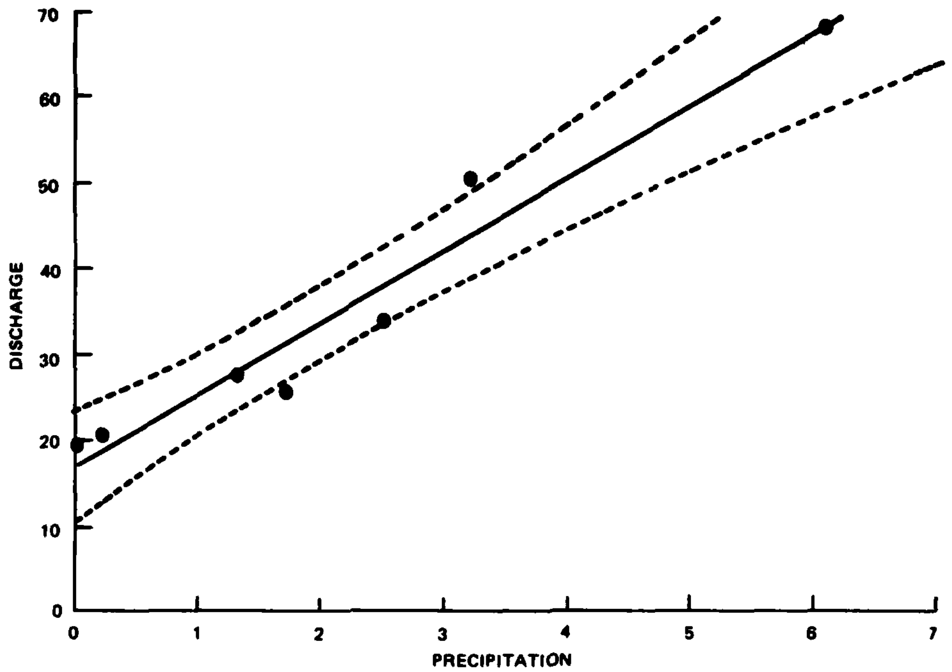


Figure 14. Confidence intervals for the regression of discharge and precipitation

$$\text{Upper limit} = \hat{Y}_o + t_{s\hat{Y}}$$

$$\text{Lower limit} = \hat{Y}_o - t_{s\hat{Y}}$$

where

$\hat{Y}_o = a + bX_o$ (predicted Y value)

$s_{\hat{Y}}$ = standard error of the predicted Y values

t = critical value of t with degrees of freedom equal to those of $s_{\hat{Y}}$.

The standard error of $\hat{Y}$ at X_o is

$$S_{\hat{Y}} = \sqrt{MS_{error}\left[\frac{1}{n} + (X_o - \bar{X})^2/\Sigma x^2\right]}$$

and has n - 2 degrees of freedom. It should be obvious that the standard error is at a minimum for $X_o = \bar{X}$ and will increase as estimates are made at values of X_o farther from the mean.

165. Confidence limits are often expressed as

$$\hat{Y}_o \pm t \frac{1}{n} + \frac{(X_o - \bar{X})^2}{x^2} s^2$$

166. The confidence limit at the mean is

$$Y \pm t \frac{s^2}{n}$$

167. In the example, for a mean discharge at $X_o = 5$, the 95-percent limits can be calculated, given

$$MS_{error} = 19.533$$

$$\bar{X} = 2.143$$

$$n = 7$$

$$t = 2.157 \ (\alpha = 0.05, \ DF = 5)$$

The standard error of $\hat{Y}$ at $X_o = 5$ is

$$S_{\hat{Y}} = \sqrt{19.533 \left[\frac{1}{7} + (5 - 2.143)^2/26.177\right]}$$

$$= 2.980$$

and Y at $X_o = 5$ is

$$\hat{Y} = 17.067 + 8.422(5)$$

$$= 59.177$$

The upper and lower confidence limits are

$$59.177 \pm ts_{\hat{Y}}$$

$$= 59.177 \pm 2.571(2.980)$$

$$= 59.177 \pm 7.662$$

Example with computer output

168. Many applications of statistics employ a computer to obtain the results of the calculations. It will be useful to examine typical output in order to see how the values are likely to be presented. The computer program used is the General Linear Models (GLM) procedure from the Statistical Analysis System (SAS). The example is taken from data on the percent organic content and percent clay of sediment samples from Red Rock Reservoir (Gunkel et al. 1984). The data used in the study are given in Table 8, and the computer output of a simple linear regression is given in Table 9.

169. The computer summary in Table 9 is subdivided into five sections. The first section gives values for the MODEL, ERROR, and the CORRECTED TOTAL sums of squares. The MODEL is simply the sum of squares attributable to the simple linear regression line ($= b\Sigma xy$) and has a single degree of freedom. In multiple regressions and other types of analysis, the MODEL may have many more degrees of freedom. The ERROR is the sum of squared deviations from the regression line ($= \Sigma y^2 - b\Sigma xy$) and provides a measure of random error (i.e., Σe_i^2). The CORRECTED TOTAL sum of squares is the total sum of squares adjusted for the mean ($= \Sigma y^2$). These statistics provide a basis for a test of the simple linear regression model. If there were no linear relationship, i.e., $\beta = 0$, the MEAN SQUARE for the MODEL would also estimate the variance of the ε_i. Thus, under the null hypothesis, the MEAN SQUARE for the MODEL should equal the MEAN SQUARE for ERROR. The ratio of this MEAN SQUARE for the MODEL to the MEAN SQUARE for ERROR would then follow an F distribution, if the ε_i were normally distributed. The F value is here

Table 8

Data Used for Example of Computer Output of Simple Linear Regression

Observation	Transect	Station	Sample No.	Percent Clay	Percent Organic Matter
1	1	A	1	60.0	7.7
2	1	A	2	62.5	8.3
3	1	A	3	40.0	4.9
4	1	B	2	62.5	9.6
5	1	C	1	55.0	7.8
6	1	D	1	57.5	9.0
7	1	E	1	67.5	9.0
8	1	F	1	65.0	9.4
9	2	A	1	30.0	7.1
10	2	B	1	17.5	3.6
11	2	B	2	20.0	3.5
12	2	B	3	17.5	3.5
13	2	C	1	30.0	4.8
14	2	D	1	65.0	9.6
15	2	E	1	47.5	7.3
16	3	A	1	57.5	9.4
17	3	B	1	67.5	9.3
18	3	C	1	52.5	8.9
19	3	C	2	52.5	8.6
20	3	C	3	60.0	8.5
21	3	D	1	32.5	3.9
22	3	E	1	62.5	7.8
23	4	A	1	47.5	7.9
24	4	A	2	50.0	8.1
25	4	A	3	45.0	8.3
26	4	B	1	27.5	6.2
27	4	C	1	42.5	7.2

(Continued)

Table 8 (Concluded)

Observation	Transect	Station	Sample No.	Percent Clay	Percent Organic Matter
28	4	D	1	40.0	6.9
29	4	E	1	60.0	9.6
30	5	B	1	42.5	7.7
31	5	B	1	40.0	6.5
32	5	C	1	40.0	7.6
33	5	C	2	60.0	8.6
34	5	D	1	42.5	5.7
35	5	E	1	52.5	8.5
36	6	A	1	42.5	8.2
37	6	B	1	45.0	7.8
38	6	C	1	37.5	8.1
39	6	D	1	30.0	7.5
40	6	E	1	40.0	8.2
41	7	A	1	42.5	8.2
42	7	B	1	40.0	7.5
43	8	A	1	27.5	6.4
44	8	B	1	32.5	6.9
45	8	D	1	32.5	5.3
46	9	A	1	22.5	4.4
47	9	B	1	20.0	3.9
48	9	C	1	25.0	4.4

Table 9

Example of Computer Output for Simple Linear Regression

(Red Rock Data - Organic Content Regressed on Clay,

Percent - General Linear Models Procedure)

```
                                   Section 1
DEPENDENT VARIABLE:  ORGANIC
SOURCE               DF          SUM OF SQUARES  MEAN SQUARE   F VALUE
MODEL                1           113.80279461    113.80279461  128.68
ERROR                46           40.68033039      0.88435501  PR > F
CORRECTED TOTAL      47          154.48312500                  0.0001

                                   Section 2
R-SQUARE             C.V.        ROOT MSE        ORGANIC MEAN
0.736668             13.0047       0.94040151      7.23125000

                                   Section 3
SOURCE               DF          TYPE I SS       F VALUE       PR > F
PCTCLAY              1           113.80279461    128.68        0.0001

                                   Section 4
SOURCE               DF          TYPE III SS     F VALUE       PR > F
PCTCLAY              1           113.80279461    128.68        0.0001

                                   Section 5
                                 T FOR HO:       PR > T        STD ERROR OF
PARAMETER            ESTIMATE    PARAMETER=0                     ESTIMATE
INTERCEPT            2.50877086    5.73            0.0001      0.43787001
PCTCLAY              0.10743080   11.34            0.0001      0.00947034
```

128.68. As described in the section on hypothesis testing, one must assess whether this is an unlikely value to have arisen by chance alone. Instead of referring to statistical tables to see if the value would occur less than once in 20 times ($\alpha = 0.05$) or less than once in 100 times ($\alpha = 0.01$) by chance, the computer program provides an exact solution ($Pr > F$) and here indicates that there is approximately only one chance in 10,000 that the value would have occurred by chance alone. Note that here the test is, perforce, two-tailed since under the alternative $\beta \neq 0$, the expected value of the MEAN SQUARE for the MODEL must exceed the MEAN SQUARE for ERROR.

170. The second section of the computer output presents some simple summary statistics. The r^2 value is presented and is obtained by dividing the MODEL sum of squares by the CORRECTED TOTAL sum of squares ($b\Sigma xy/\Sigma y^2$). It indicates that the model accounts for approximately 74 percent of the total variation. The ROOT MSE is the square root of the MEAN SQUARE ERROR and is a measure of the variation about the regression line, analogous to a standard deviation. The mean of the dependent variable (ORGANIC material) is given, as is the coefficient of variation (C.V.). The C.V. gives an indication of the amount of error which exists relative to the mean (C.V. = ROOT MSE/MEAN) × 100 percent.

171. The next two segments of the output of this particular computer procedure, viz, GLM, gives two types of sums of squares (Type I SS and Type III SS). For simple linear regression, these sums of squares are equal and provide no new information. They will, however, be discussed in a subsequent section.

172. The last component of the output gives the parameter estimates, both the intercept (b_o) and the coefficient (b_1) of the independent variable (PCTCLAY). These can then be used to form the estimated regression equation.

$$Y_i = 2.5088 + 0.1074 \text{ (PCTCLAY)}$$

This equation provides a method of estimating mean organic content from percent clay content of the substrate. The output also provides the standard error for each of the parameter estimates, t statistics for testing the hypothesis that each parameter equals zero, and the probability that the values would be equaled or exceeded by chance. Observe that the value of t for testing the slope (11.34) is the square root of the F statistic (128.68), i.e., the tests are equivalent.

Multiple regression

173. The concepts applicable to simple linear regression extend to multiple regression. The purpose is still to quantify a relationship between a dependent response variable and some independent variable. However, in multiple regression there is more than one independent variable. Since there may exist interrelationships between the so-called independent variables, the regression coefficients for the dependent variables from a multiple regression are not the same as would be obtained if the dependent variable was regressed on each independent variable separately.

174. The formulas for a multiple regression will not be derived in this manual. For simple two-factor (i.e., two independent variables) multiple regression, formulas are available in some textbooks, but broad application of multiple regressions requires matrix algebra. We will presume that the calculations will be done on a computer and shall therefore concentrate on the interpretation of computer output. Computer table and graphic output presented in this chapter were obtained from SAS procedure GLM (SAS (1981) GRAPH User's Guide; SAS (1982) User's Guide).

175. An example of multiple regression data is given in Table 10. These data represent levels of a pollutant measured at various distances downstream from a source. The pollutant decreases with distance from the source of pollution, but additional factors are expected to influence the results of measurements taken on different occasions. Therefore, two other independent variables have been included in the multiple regression--temperature and discharge. The model is then

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{2i} + \varepsilon_i$$

Table 10

Hypothetical Data for Multiple Regression Example

Observation	Pollutant	Distance, m	Temperature, °C	Discharge m^3/sec
1	15.5	1,000	24	0.8
2	12.9	1,000	20	1.0
3	14.8	1,000	19	1.3
4	10.3	1,000	25	1.8
5	10.7	1,000	15	2.0
6	14.9	2,000	17	0.5
7	6.6	2,000	20	1.0
8	9.5	3,000	21	1.0
9	5.1	3,000	15	2.0
10	7.4	4,000	15	0.5
11	11.9	4,000	24	0.8
12	5.4	4,000	21	1.0

where X_1 , X_2 , and X_3 refer to distance, temperature, and discharge, respectively.

176. Examination of the computer output (Table 11) will illustrate several concepts. The output first indicates that the analysis was performed on a dependent variable called POLLUTNT. The sources of variation are listed as MODEL and ERROR, which will sum to the CORRECTED TOTAL (= Σy^2) as before. Note that the model now has 3 d.f., one for each of the independent variables. Since there were 12 observations, the CORRECTED TOTAL carries 11 d.f., one being lost when the variables are adjusted about their means. Three degrees of freedom were required for the three variables in the model, leaving eight as a measure of random error. The computer provides calculations of the sum of squares for each source, and the corresponding mean square. An F test is calculated for the mean square of the model, using the mean square error. For the example in Table 11, the F test indicates that the regression does account for a significant portion of the total variation. This F test is a joint test of all three of the independent variables and does not indicate which one(s) contribute significantly to the description of POLLUTNT.

Table 11

Computer Output for Multiple Regression Example

(General Linear Model Procedure)

```
DEPENDENT VARIABLE:  POLLUTANT
SOURCE               DF         SUM OF SQUARES    MEAN SQUARE     F VALUE
MODEL                 3         97.73161461       32.57720487     4.84
ERROR                 8         53.82505206        6.72813151     PR > F
CORRECTED TOTAL      11          151.55666667                     0.0331

R-SQUARE              C.V.          ROOT MSE               POLLUTANT MEAN
0.644852             24.9011    2.59386420                    10.41666667

SOURCE               DF         TYPE I SS         F VALUE         PR > F

DISTANCE              1         54.19739726       8.06            0.0219
TEMP                  1          7.19969359       1.07            0.3312
DISCHARG              1         36.33452376       5.40            0.0486

SOURCE               DF         TYPE III SS       F VALUE         PR > F

DISTANCE              1         77.32993941       11.9            0.0095
TEMP                  1          1.69303058        0.25           0.6294
DISCHARG              1         36.33452376        5.40           0.0486

                                  R FOR HO:       PR > T          STD ERROR OF
PARAMETER        ESTIMATE       PARAMETER=0                        ESTIMATE
INTERCEPT        17.59409833        3.02          0.0166          5.82622109
DISTANCE         -0.00225161       -3.39          0.0095          0.00066415
TEMP              0.11241376        0.50          0.6294          0.22409610
DISCHARG         -3.78579922       -2.32          0.0486          1.62909000
```

177. That portion accounted for by the model is given by the r^2 value, which is the ratio of the MODEL sum of squares (= $\Sigma_j b_j \Sigma_i (X_j)^2$, the b_j being the estimate of the β_j) to the CORRECTED TOTAL sum of squares (Σy_i). Thus 65.49 percent of the variation was accounted for by the model. There are several important points to be made about these calculations. Usual practice is to adjust about the mean (i.e., to fit an intercept not necessarily equal to zero), but there are some exceptions. An option in the computer algorithm will provide results not adjusted about the means, but the r^2 value then does not have the same interpretation.

178. Following the synopsis of the model are the results of the individual independent variables for the regression. As before, two types of sums of squares are provided, but in the case of multiple regression they differ. They are the "sequential sums of squares" or, in the jargon of SAS GLM, Type I SS, and partial sums of squares or Type III SS.

179. The sequential sums of squares are commonly, but not necessarily, larger than the corresponding partial sums of squares; indeed, Table 11 provides an example of a case where one partial sum of squares is larger. The sequential sums of squares depend on the order in which they were entered into the computer program. Here distance was entered first, and on its own accounted for 54.1974 of the sum of squares carried by the model. This is exactly the same as would be obtained if a simple linear regression was done for POLLUTNT on DISTANCE, with no other variables in the model. TEMP was the second variable entered into the model, so the variable POLLUTNT was already adjusted for DISTANCE when TEMP was entered. A model containing DISTANCE and TEMP carries a sum of squares of 61.3971 (not given explicitly) so that contribution of TEMP, given that DISTANCE is already in the model, is 61.3971 - 54.1974 = 7.1997. DISCHARG was last to enter the model, so the effect of DISTANCE and TEMP jointly was already included. Accordingly, the sum of squares carried by DISCHARG, given that DISTANCE and TEMP are already in the model, is 97.7316 - 61.3971 = 36.3345.

180. The partial sums of squares for each variable are obtained as if all other variables were already included in the model; i.e., they give the contribution to the total sum of squares carried by each variable adjusted for the previous inclusion of all other variables jointly. Thus, the partial sum of squares for DISCHARG is here equal to the sequential sum of squares since this was the last variable entered. Clearly, the sequential sums of squares will differ according to the order that the variables are entered. The partial sums of squares are independent of this ordering.

181. The appropriate sums of squares for testing hypothesis depend on the circumstances. These tests are specific to the model so if one of the variables were to be omitted, or another variable added, the values would change.

182. For instance, in the example, the partial sums of squares show that there is virtually no merit in including TEMP in the model in addition to DISTANCE and DISCHARG (i.e., the hypothesis $\beta_2 = 0$ can be accepted). The sequential sums of squares also show that there would be little or no merit in including TEMP in addition to DISTANCE, but neither sequential nor partial sums of squares provide a test of adding DISCHARG to DISTANCE or vice versa. Nor do we see how a model that includes TEMP and DISCHARG would perform. For these we would have to fit two-variable models or enter the variables in a different order.

183. The final section of output from the computer contains the parameter estimates, or regression coefficients. These values, as with simple linear regression, can be used to estimate the mean pollution, given values of distance, temperature, and discharge. The equation for this example would be

POLLUTNT = 17.5941 - 0.00225*DISTANCE + 0.1124*TEMP - 3.7858*DISCHARG

184. Note, however, that the regression coefficients for a variable are always calculated with adjustments for all other variables in the model. The parameter would have to be reestimated if TEMP were excluded.

Curvilinear models

185. All of the regression models demonstrated so far have been linear, the defining feature being that the response can be written as a linear combination of the parameters, that is as the sum of the parameters each multiplied by known quantities. Models with parameters that are multiplied, raised to powers, or are powers of known or unknown quantities are not linear models. Some models with these characteristics can be linearized by transformation, e.g., the taking of logarithms, and thus permit parameter estimation by linear least squares methods. There are also models that involve the inverses of variables, or powers of variables, or even trigonometric or exponential functions of variables, and thus appear curvilinear, but since the parameters satisfy the defining feature of linear models these are still, strictly speaking, linear models. Thus, for example

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 \, 1/X_{2i} + \varepsilon_i$$

and

$$Y_i = \alpha + \beta_1 \sin X_{1i} + \beta_2 e + \varepsilon_i$$

are linear models.

186. These can be treated as cases of multiple linear regression with, e.g., $X^*_{1i} = \sin X^*_{1i}$, $X^*_{2i} = e^{X_{2i}}$. The simplest cases of this type are, perhaps, polynomials, in which the dependent variable Y is modeled as the sum of the independent variable X raised to successively greater powers, e.g.

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$

187. The family of polynomial curves is described graphically below. The simplest is

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

and is just simple linear regression as discussed previously. As shown in Figures 15a and 15b, simple linear regression may have either a positive or negative slope. Next in the series is the quadratic

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

This will fit a parabolic-shaped curve and may be concave (Figure 15c) or convex (Figure 15d). Each additional model in the series incorporates an additional power term one higher than the preceding (cubic, quartic, etc.). Thus

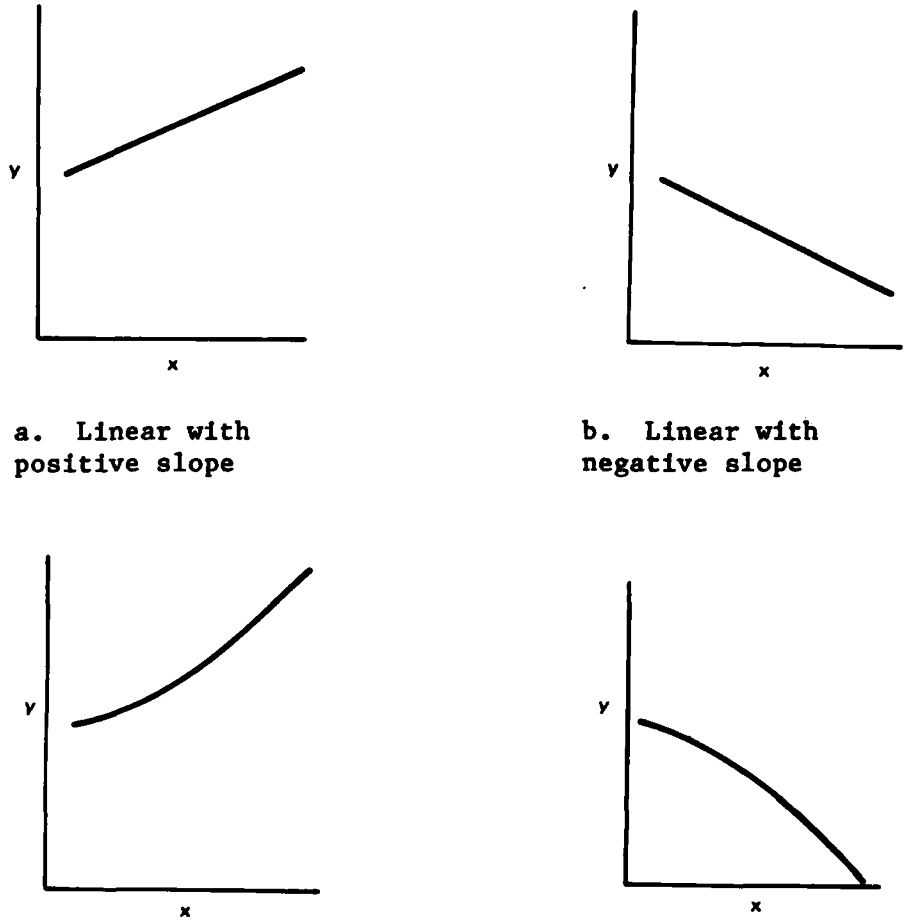
$$\text{cubic} \qquad Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$

$$\text{quartic} \qquad Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \varepsilon_i$$

Each additional term allows for a possible change in direction of the curve. Thus, high-order polynomials can be used to fit rather complicated patterns of lines along some independent variable.

188. Nevertheless, a polynomial equation reveals very little about the underlying nature of the relationship it is attempting to describe. Principally, it provides a very flexible method of fitting complex curvature and does allow for the detection of pattern and the demonstration that portions of the variation can be described by some pattern. An example of the application of a polynomial regression based on the data in Table 12 is given in the computer output in Table 13. A plot of the original data and the resultant polynomial is given in Figure 16.

189. The fitting of a polynomial is one situation in which there is a logical ordering for the entry of the variable into the model (linear, quadratic, cubic, etc.), so for testing purposes sequential sums of squares are usually appropriate.



a. Linear with positive slope

b. Linear with negative slope

c. Concave quadritic

d. Convex quadritic

Figure 15. Family of polynomial curves

Table 12

Hypothetical Data for the Polynomial Regression Example

Observation	Station No.	X	Oxygen	Month
1	118	1	2.2	6
2	118	2	1.3	7
3	118	3	0.1	8
4	118	4	0.0	9
5	118	5	2.1	10
6	118	6	7.3	11
7	118	7	7.0	12
8	118	8	10.1	13
9	118	9	10.4	14
10	118	10	10.5	15
11	118	11	10.9	16
12	118	12	8.4	17
13	1308	1	0.5	6
14	1308	2	0.0	7
15	1308	3	0.1	8
16	1308	4	0.0	9
17	1308	5	2.5	10
18	1308	6	7.9	11
19	1308	7	8.2	12
20	1308	8	11.3	13
21	1308	9	11.4	14
22	1308	10	11.6	15
23	1308	11	11.7	16
24	1308	12	8.9	17

Table 13

Example of Polynomial Regression

General Linear Models Procedure

```
DEPENDENT VARIABLE:  OXYGEN
SOURCE               DF      SUM OF SQUARES   MEAN SQUARE      F VALUE
MODEL                 3       475.46737152    158.48912384   136.48
ERROR                20        23.22596182      1.16129809     PR > F
CORRECTED TOTAL      23       498.69333333                     0.0001

R-SQUARE             C.V.          ROOT MSE   OXYGEN MEAN
0.953426             17.9108     1.07763542     6.01666667

SOURCE               DF            TYPE I SS     F VALUE       PR > F

X                     1       386.56031469    332.87           0.0001
X*X                   1         9.14807859      7.88           0.0109
X*X*X                 1        79.75897824     68.68           0.0001

SOURCE               DF          TYPE III SS     F VALUE       PR > F

X                     1        33.26566326     28.65           0.0001
X*X                   1        70.20390107     60.45           0.0001
X*X*X                 1        79.75897824     68.68           0.0001

                                  T FOR HO:      PR > T      STD ERROR OF
PARAMETER           ESTIMATE     PARAMETER=0                   ESTIMATE

INTERCEPT         4.69343434     3.76           0.0012       1.24670068
X                -4.26674529    -5.35           0.0001       0.79720586
X*X               1.08565046     7.78           0.0001       0.13963081
X*X*X            -0.05867651    -8.29           0.0001       0.00708021
```

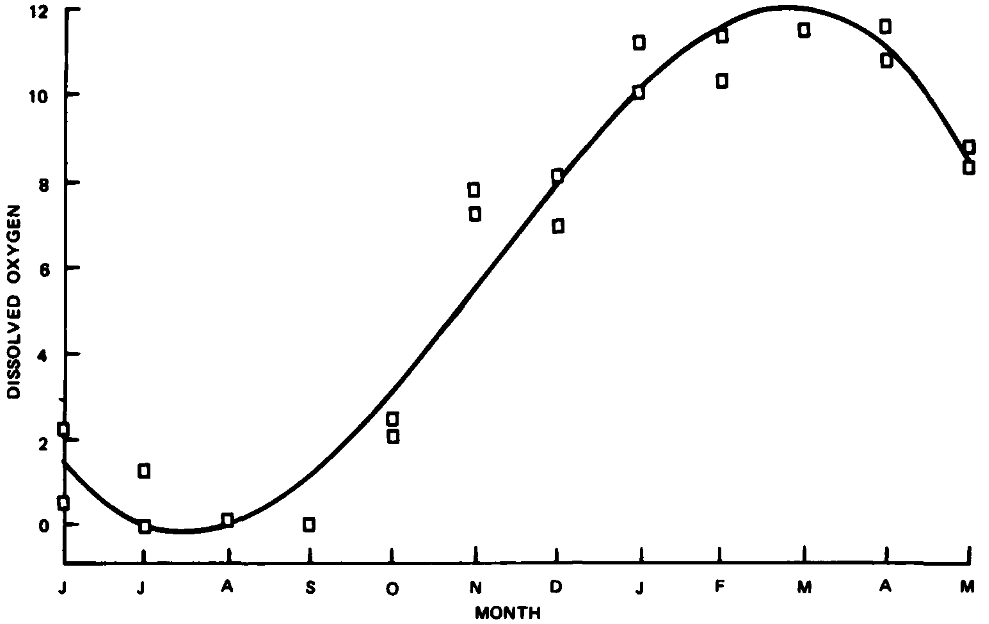


Figure 16. Polynomial regression of dissolved oxygen concentration as a function of time

190. In the real nonlinear models, the parameters are nonadditive. These can sometimes be linearized by transformation of the variables, particularly logarithms. Examples of such models are given below.

Semilog model

Nonlinear equation $Y_i = \beta_o e^{\beta_1 X_i} e^{\varepsilon_i}$

Linearized equation $\log_e (Y_i) = \log_e (\beta_o) + \beta_1 X_i + \varepsilon_i$

or $Y_i^* = \beta_o^* + \beta_1 X_i + \varepsilon_i$

Log-log model

Nonlinear equation $Y_i = \beta_o X_i^{\beta_1} e^{\varepsilon_i}$

Linearized equation $\log_e (Y_i) = \log_e (\beta_o) + \beta_1 \log_e (X_i) + \varepsilon_i$

or $Y_i^* = \beta_o^* + \beta_1 X_i^* + \varepsilon_i$

Note that a multiplicative error is assumed. This has been written as e^{ε_i}. Since $e^o = 1$, $\varepsilon_i > 0$ implies that $e^{\varepsilon_i} > 1$ and $\varepsilon_i < 0$ implies $0 < e^{\varepsilon_i} < 1$.

191. The semilog model is commonly associated with exponential growth or decay; it is useful whenever a dependent variable Y is expected to increase or decrease as a proportion of itself over time or some other independent variable X . The slope of the line β_1 provides a measure of the proportional increase (decrease) per unit of X (e.g., Y may be said to increase by 0.053 or 5.3 percent per unit of time). It can be used, for example, to describe the degradation of chemicals over time in aquatic systems.

192. The log-log model is sometimes used to describe nonlinear relationships where the variables X and Y increase proportionately, and to calibrate instruments. The equation may be expressed as

$$\frac{Y_i}{X_i} \beta_1 = \beta_o e^{\varepsilon_i}$$

which, apart from the random error, indicates that the ratio of Y to X^{β_1} is a constant (β_o). If $\beta_1 = 1$, the model implies that Y is a constant fraction or multiple of X . Note that the taking of logarithms is still appropriate because of the multiplicative error; the situation is not equivalent to

$$Y_i = \beta_o X_i + \varepsilon_i$$

If β_1 is not equal to unity, the relationship is curved. In such relationships, if the error is not multiplicative, e.g., if

$$Y_i = \beta_o e^{\beta_1 X_i} + \varepsilon_i$$

or

$$Y_i = \beta_o X^{\beta_1} + \varepsilon_i$$

the taking of logarithms will no longer linearize the equation and special nonlinear least squares have to be applied. These are outside the scope of this manual. They are also required by the more complex models that cannot be linearized by any transformation.

Multisource regression

193. We are sometimes faced with regression data from more than one source, for example data collections from each of several lakes. It may be of interest to know not simply that the relationship is of the same form for each source but whether the relationships are quantitatively equivalent. We shall here assume that the relationships are of the same form, i.e., involve the same structure of the variables, and examine the second question.

194. Assume that three sets of bivariate (X and Y) data have been collected and each set of data has been fit to a simple linear regression. The three regression equations are

$$Y_{1i} = a_1 + b_1 X_{1i} \qquad i = 1, 2 \ldots n_1$$

$$Y_{2i} = a_2 + b_2 X_{2i} \qquad i = 1, 2 \ldots n_2$$

$$Y_{3i} = a_3 + b_3 X_{3i} \qquad i = 1, 2 \ldots n_3$$

Also assume that all slopes (b_1, b_2, and b_3) are significantly different than zero.

195. The hypotheses to be tested are

$$H_o:\beta_1 = \beta_2 = \beta_3 \qquad H_A:\beta_1 \neq \beta_2 \neq \beta_3$$

The basic calculations necessary to compare the slopes were computed during the linear regressions for each data set. For each data set we need

$$\Sigma x^2 = \Sigma(X_i - \bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/n$$

$$\Sigma y^2 = \Sigma(Y_i - \bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/n$$

$$\Sigma xy = \Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \Sigma XY - \Sigma X\Sigma Y/n$$

and the error sum of squares and error degrees of freedom

$$SS_{error} = \Sigma y^2 - (\Sigma xy)^2/\Sigma x^2$$

$$DF_{error} = n - 2$$

The values necessary for the comparison of slopes are presented in Table 14.

196. The values of the three error sums of squares can be added to provide what may be called a "pooled" error sum of squares

$$SS_p = SS_1 + SS_2 + SS_3$$

with

$$DF_p = DF_1 + DF_2 + DF_3$$

The values of Σx^2 , Σxy , and Σy^2 must also be summed

$$A_c = A_1 + A_2 + A_3$$

$$B_c = B_1 + B_2 + B_3$$

$$C_c = C_1 + C_2 + C_3$$

and from these sums a "common" error sum of squares can be calculated

$$SS_c = C_c - \left(B_c^2/A_c\right)$$

These calculations are also presented in Table 14.

197. An F statistic is used to test the null hypothesis that the three slopes are equal

$$F = \frac{(SS_c - SS_p)/(k - 1)}{(SS_p/DF_p)}$$

where k = number of slopes being compared (in this example $k = 3$). This calculated F statistic would be compared with a critical value of F having $k - 1$ numerator degrees of freedom and denominator degrees of freedom of DF_p .

198. If the calculated F is smaller than the critical value, the null hypothesis ($H_o:\beta_1 = \beta_2 = \beta_3$) can be accepted. If, and only if, this hypothesis can be accepted, it is possible to determine if the intercepts are equal. Given that the slopes are equal, if the intercepts are all equal the regressions for each of the three data sets are identical.

199. The hypotheses to be tested are

$$H_o:\alpha_1 = \alpha_2 = \alpha_3 \qquad H_A:\alpha_1 \neq \alpha_2 \neq \alpha_3$$

Table 14

Calculations Required for a Multisource Regression

Data Set	Σx^2	Σxy	Σy^2	SS_{error}	DF_{error}
1	A_1	B_1	C_1	$SS_1 = C_1 - (B_1^2/A_1)$	$DF_1 = n_1 - 2$
2	A_2	B_2	C_2	$SS_2 = C_2 - (B_2^2/A_2)$	$DF_2 = n_2 - 2$
3	A_3	B_3	C_3	$SS_3 = C_3 - (B_3^2/A_3)$	$DF_3 = n_3 - 2$
"Pooled"	--	--	--	$SS_p = \sum_{i=1}^{3} SS_i$	$DF_p = \sum_{i=1}^{3} DF_i$
"Common"	$A_c = \sum_{i=1}^{3} A_i$	$B_c = \sum_{i=1}^{3} B_i$	$C_c = \sum_{i=1}^{3} C_i$	$SS_c = C_c - (B_c^2/A_c)$	

To perform this test it is necessary to combine the three data sets and calculate Σx^2 , Σxy , and Σy^2

$$(\Sigma x^2)_T = \Sigma X_T^2 - (\Sigma X_T)^2/n_T$$

$$(\Sigma xy)_T = \Sigma X_T Y_T - \Sigma X_T \Sigma Y_T/n_T$$

$$(\Sigma y^2)_T = \Sigma Y_T^2 - (\Sigma Y_T)^2/n_T$$

where

X_T = X value from the set of all Xs

Y_T = Y value from the set of all Ys

$n_T = n_1 + n_2 + n_3$

An error sum of squares can be computed from these values as

$$SS_T = (\Sigma y^2)_T - (\Sigma xy)_T^2/(\Sigma x^2)_T$$

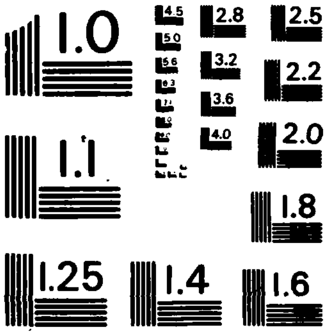
200. An F statistic is used to test the hypothesis that the three intercepts are all equal

$$F = \frac{(SS_T - SS_c)/(k - 1)}{(SS_c/DF_p)}$$

where SS_c , DF_p , and k are as defined for the F test comparing slopes. This calculated F statistic would be compared with a critical value of F having k - 1 numerator degrees of freedom and denominator degrees of freedom of DF_p .

201. These methods extend in a straightforward manner to simple or multiple regressions or more than three sources. Indeed one may conclude that the same parameters satisfy some but not all sources. Likewise some, but not necessarily all, may have the same slope (i.e., be parallel) but be distinct, i.e., have different intercepts. Any combination is possible; usually the objective is to find the most

AD-A171 046 ENVIRONMENTAL AND WATER QUALITY OPERATIONAL STUDIES 2/3
STATISTICAL METHODS F..(U) ARMY ENGINEER WATERWAYS
EXPERIMENT STATION VICKSBURG MS ENVIR.. R F GAUGUSH
UNCLASSIFIED JUN 86 WES/IR/E-86-2 F/G 8/8 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

parsimonious representation that gives a satisfactory fit.

202. The case where one tests for differences between intercepts, under the assumption that all the lines are parallel, is the conventional analysis of covariance. There is a tendency recently, however, to refer to the whole multisource regression situation as the analysis of covariance.

Analysis of Variance

203. As the terminology implies, an analysis of variance (ANOVA) is a procedure for quantifying sources of variability in a set of data. Of prime importance is the realization that an ANOVA does not identify the sources or components of variance per se. The researcher specifies the components and the ANOVA merely assesses the amount of variability accounted for by the factors which have been postulated to explain the variability in the data.

204. The factors or cause-and-effect relationships are specified explicitly in a model. However, model formulation is characteristically different from many modeling endeavors. A large number of modeling enterprises are typified by seeking the one model, from various classes of models, which best describes the data. Classic examples are physical or mathematical models of ecosystems. The ANOVA is based on a single class of models called general linear models. Within this class, model construction is a direct consequence of the sampling or experimental design used to gather the data. Only after model formulation does the ANOVA come into play. The ANOVA ascertains the degree to which model components and their interrelationships account for the variability observed in the data.

205. In summary, model formulation is dependent upon the questions posed by the researcher and the design of the sampling or experimental process followed to obtain the data. The ANOVA operates as the analytical tool for generating the answers. The questions asked may be exploratory or take on the character of a fact-finding mission, but most

often the questions will be posed as "a priori" statements or hypotheses about what one expects to find.

Two hypothetical examples

206. Consider a study dealing with the chlorophyll concentration of surface water samples from two reservoirs. Table 15 presents the data. The sampling design involved the random selection of five stations located in the main pool of each reservoir, followed by the drawing of three samples at each station. The sampling design may have been a consequence of research objectives to determine (a) if reservoirs differ in mean chlorophyll concentration; (b) if a significant degree of heterogeneity exists among stations and if reservoirs differed substantially in this heterogeneity; and (c) if sampling variability within stations was relatively high or low.

207. Based on these objectives and the sampling design, the model would be defined. Three model components are explicitly required: a component reflecting between-reservoir mean differences, a component dealing with variability among stations within each reservoir, and a component expressing sample heterogeneity within stations. No other components are provided for by the objectives and sampling design. If different or additional objectives had to be met, the sampling design and model would have to be modified in order to accommodate the needs of the researcher. An ANOVA performed on these data would quantify the degree to which model components explained the variability inherent in the data.

208. Consider a second example, experimental rather than sampling. Assume that a researcher wanted to identify the limiting nutrient that was controlling algal productivity during summer stratification in the near-dam region of a given reservoir. Identification of the limiting nutrient was carried out using the Algal Assay Procedure (National Eutrophication Research Program 1971). This procedure uses nutrient additions (usually phosphorus, nitrogen, and EDTA (ethylenediaminetetraacetic acid), separately and in combination) to filtered water samples to which a test algae has been added. After 14 to 21 days, the standing crop (usually expressed as dry weight/litre or

Table 15

Chlorophyll Concentration of Surface Water Samples

Reservoir	Station	Sample			Mean		
		1	2	3	Station	Reservoir	Overall
A	A1	11.3	9.9	14.1	11.8	20.9	16.0
	A2	26.8	29.0	26.8	27.5		
	A3	12.9	8.1	12.4	11.1		
	A4	34.7	29.0	32.1	31.9		
	A5	22.1	21.6	22.5	22.1		
B	B1	14.1	15.6	11.3	13.7	11.0	
	B2	8.4	9.9	7.9	8.7		
	B3	4.7	1.2	3.3	3.1		
	B4	12.9	11.3	14.1	12.8		
	B5	15.6	18.2	16.7	16.8		

cells/litre) of the algae is determined. Under controlled laboratory conditions of light and temperature, the maximum standing crop of the algae is related to the amount of limiting nutrient initially available. The results of this experiment are given in Table 16.

209. The experiment involved taking 24 subsamples (eight treatments × three replicates per treatment) from a single epilimnetic sample taken from the near-dam area of the reservoir. There was not sufficient space in any single environmental chamber for all of the subsamples to be processed together, so the investigator randomly assigned the subsamples to three different environmental chambers. The random assignment of the subsamples to the chambers was necessary because identical conditions (i.e., light and temperature) between chambers could not be assured. Within any given chamber, all subsamples are considered to be at identical conditions (except for the experiment treatment applied).

210. Once the experiment is designed and executed, the ANOVA model is predetermined. For the example presented above, the model includes only one component, the effect of the nutrient additions. The ANOVA would provide the information. The null hypothesis of equal standing crop for all treatments can be readily tested. Casual

inspection of the means (Table 16) suggests that phosphorus is the limiting nutrient and that combinations with nitrogen and EDTA have little effect on the standing crop.

Table 16

Dry Weights (mg/litre) of the Test Algae After 14 days of Incubation

Treatment	Block Chamber 1	Chamber 2	Chamber 3	Mean
CONTROL	7.6	8.4	7.8	7.9
+ P	30.5	34.2	32.3	32.3
+ N	9.6	10.1	9.9	9.9
+ EDTA	8.4	9.3	10.4	9.4
+ P, + N	33.7	35.8	34.2	34.6
+ P, + EDTA	32.3	33.1	33.5	33.0
+ N, + EDTA	10.9	11.4	11.2	11.2
+ P, + N, + EDTA	34.3	35.0	33.8	34.4

Fundamentals of design

211. Although a complete discussion of sampling and experimental design is beyond the scope of this text, a discussion of some basic principles and major issues is worthwhile. The major objective of any design, sampling or experimental, is to generate the most powerful, efficient, and accurate results relative to the research questions at hand within the constraints imposed by time, money, and manpower.

212. All designs (sampling or experimental) leading to an ANOVA are characterized by a random sample of units from each of the series of treatment-populations. Henceforth, the terminology "treatment-populations" will be used in a generic sense to refer to the populations under consideration in sampling designs or the controlled manipulations employed by the researcher in experimental designs. Likewise, "sampling units" and "experimental units" will be used to refer to the primary

sample of units from each treatment-population. For instance, our first hypothetical example involved the random selection of five stations from each reservoir and the subsequent selection of three water samples from each station. Accordingly, reservoirs would represent treatment-populations, and stations would be the primary sampling units. Water samples would simply represent the process of secondary sampling or subsampling.

213. Our second example is more complex, but the complexity provides the flexibility to introduce additional concepts and issues. It should be clear that the treatment-populations are the eight experimental manipulations. The manipulations used were under the control of the investigator or, in other words, the investigator chose a specific experimental design. A design is the plan followed in selecting units from populations (i.e., a sampling design) or in applying treatments to units (i.e., an experimental design).

214. In summary, all designs can be characterized by a random sample of units from a set of treatment-populations, but specific designs are differentiated according to the plan followed in selecting sampling units from populations or applying treatments to experimental units. Sampling units could be lakes (units) selected from different regions (treatment-populations) of the country, sediment core samples (units) drawn at different locations (treatment-populations) of a reservoir, coves (units) selected from different reservoirs (treatment-populations), or water samples (units) drawn at different times of the year (treatment-population) and/or at different depths (treatment-population). Two examples of experimental units are limnetic enclosures (units) supplied with different nutrients (treatments) to study nutrient limitation, or littoral stations (units) treated with different herbicides (treatments) to determine effectiveness for macrophyte control.

215. Even though some might claim this characterization of designs to be an oversimplification, the fact remains that the myriad of designs, models, and associated analyses of variance are based on the concept of a random sample of experimental or sampling units from a set of treatment-populations. The complexity arises not from the

fundamental process but from (a) the procedures followed in conducting the entire sampling or experiment plan, (b) the presumed or known underlying structure of the treatment-populations, (c) the number of factors that need to be controlled so as to lead to valid and reliable conclusions, and (d) the procedures followed in performing the analysis. The result is virtually an infinite array of designs and computational algorithms that are situation specific but conceal the underlying simplicity.

216. Selection of design cannot be separated from the research objectives. The design that best meets the needs of the researcher must be found. Sometimes objectives vary over a wide range, and no design may be available to maximize efficiency for all. Prioritization is not negative; certain objectives might be sacrificed, entirely or partially, in order to obtain a design that is best for those objectives of primary concern. Moreover, different designs may be selected to satisfy different sets of objectives. Fortunately, all designs that have been developed and have been documented are extensions of two fundamental building-block designs which are the topic of the next section.

The two basic designs

217. The two studies discussed above exemplify the two designs on which all designs are based. These designs are called "completely randomized design" (CRD) and the "randomized block design" (RBD). A CRD may be defined as follows: Given a set of treatment-populations, a simple random sample of elements is selected from each population or a simple random sample of units is evaluated under each treatment.

218. Consider a study designed to compare three reservoirs in terms of the total phosphorus concentration of the surface water during a particular time of the year. Table 17 presents the data. Sampling merely involved taking one water sample from each station of a randomly selected set of stations. Note that the three reservoirs represent the treatment-populations while stations represent the sampling units from each cove. The situation typifies a CRD, or a simple random sample of units from a series of treatment-populations. It is true that more than one water sample could have been drawn at each station. This added

Table 17

Total Phosphorus Concentration (μg/ℓ) of Surface Water Samples Drawn at Randomly Selected Set of Stations from Three Reservoirs

	Reservoir		
Station	1	2	3
1	31	32	22
2	28	30	21
3	27	30	23
4	30	31	
5		34	

sampling stage does not define the design. As was the case with the comparative study on reservoirs (Table 15), the primary stage of sampling defines the design, and as such the examples of Tables 15 and 17 are cases of a CRD.

219. An RBD, the second basic design, involves a two-stage process. Initially, the sampling or experimental units are grouped into "blocks." The grouping is based on one or more organismic or environmental characteristics which are presumed to affect or influence the response of a unit. The fundamental premise is that variability within a block is quite small relative to the variability between blocks; homogeneity within versus heterogeneity between is the rule. Subsequent to blocking, random sampling occurs as in a CRD, but the sampling process is implemented block by block.

220. The experiment reported in Table 16 followed an RBD. The three environmental chambers constitute the blocks and homogenization prior to subsampling ensured the similarity of initial nutrient concentrations within each block (i.e., low within-block variability). Variability between blocks (due to differences in light and temperature between environmental chambers) is controlled by the design.

221. Randomized block designs have wide applicability. Consider a study with the objective of assessing water quality as a function of

depth (let us say surface, shallow, deep). Suppose further that several stations on the reservoir have been established as monitoring points. Now suppose that during a particular month, several samples are selected at each depth. If this were the case, then an RBD would result where stations would constitute the blocks while depths would represent the treatment-populations.

222. A slight variation on the theme of depth sampling involves time sampling and again would yield an RBD. Restricting assessments to surface water, several samples might be drawn each month for 12 months. Stations would again constitute the blocks, but month (time of sampling) would represent treatment-populations.

223. Some care must be exercised in time and depth sampling. What at first glance appears to be an RBD is oftentimes not. In addition, a treatment-population factor in one instance may not be in another. Regarding the first point, consider time sampling of surface water on a reservoir. If water samples are drawn at distinctly different stations each month, a CRD results, not an RBD. This is so because each month constitutes the treatment-populations and a different set of stations, not a replicated set of stations, is sampled each time.

224. Regarding the second point, consider a comparative study on two reservoirs. Suppose each month a different set of stations is sampled on each reservoir. The result would be an RBD, but reservoirs would represent the treatment-populations and month would constitute replicated time sampling.

225. These were just a few examples of completely randomized designs and randomized block designs. As the basic designs, they can be combined in interesting ways to produce other designs having wide applicability in water quality and field research. Of particular importance is a class of designs that are called "split-plot designs" (SPD). Under certain conditions these designs are called "repeated measures designs."

226. Consider a combination of depth and time sampling. Suppose an established set of monitoring stations is sampled each month for 12 months. The result would be an RBD with stations as blocks and time as treatment-populations. If, in addition, depth sampling were to be

executed, then each station at each time would in effect be split as a function of depth in the water column. In concept, water quality measurement is repeated for each station-month combination, but the repeat measure is equivalent to depth sampling.

227. Other examples are direct. Consider a random set of stations from each of two reservoirs at a particular time of the year. If this were the case, the design would be a CRD by virtue of the fact that a random set of units (stations) is selected from each treatment-population (reservoir). Add depth sampling at each station and the results would be an SPD--each station is split according to depth in the water column. Notice that the only difference between this example and the previous example is the base design. Here the basic design is a CRD, while the base design of the previous example was an RBD.

228. In summary, many designs are available on which to base sampling or experimental studies. The key issue is to select the design which optimizes the quality of information for meeting the objectives of the research. Although only three designs have been discussed (CRD, RBD, SPD) and only a few examples have been given, these three designs and related examples characterize a large body of the studies performed in water quality research.

229. In order to appreciate fully how designs vary in maximizing efficiency in answering research questions, an understanding of methods of data analysis is required. As has been pointed out, the questions of the researcher dictate the design which in turn defines the model with its component sources of variability. The ANOVA quantifies these component sources so as to provide objective criteria on which to base one's inferences and conclusions.

One-way analysis of variance

230. To develop and explain the fundamentals of models and analyses of variance, attention will be restricted to the example of the three reservoirs and a random sample of stations from each reservoir (Table 17). This example provides a minimum level of complexity, but within this level of complexity, all basic concepts and principles can be demonstrated and explained. Formally, this example represents a

CRD with experimental error and leads to a one-way ANOVA which tests the hypothesis that no differences exist between treatment populations. The precise meaning of this terminology will become clear as the discussion proceeds.

The model

231. Based on the design followed in gathering the data, only two sources of variability are explicitly allowed, the variability due to differences between reservoirs and the variability due to differences among stations within each reservoir. The function of the model thereby becomes one of expressing total phosphorus (TP) concentrations (the dependent variable) as a function of treatment-population parameters (the independent variables). The model (typical of all CRDs with experimental error and one-way ANOVA) is given by

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{1}$$

where

X_{ij} = TP concentration for the j^{th} sampling unit (surface water sample at the j^{th} station) from the i^{th} treatment-population (reservoir)

μ = overall mean TP concentration

α_i = effect of i^{th} treatment-population
= $\mu_i - \mu$

μ_i = mean TP concentration for the i^{th} treatment-population

ε_{ij} = residual or random error effect
= $X_{ij} - \mu_i$

232. The model in Equation 1 expresses the dependent variable (X_{ij}) as an additive partition of a sequence of terms (μ, α_i, ε_{ij}). Within the context of an ANOVA, all terms have a unique meaning. The terms are linear functions of treatment-population mean parameters, μ and μ_i (i.e., the parameters are not in any nonlinear form such as logarithmic or exponential). Hence, ANOVA models are called "population linear additive models." The model includes only a simple dependent

variable (X_{ij}) which is functionally related to the sequence of terms (μ, α_i, ε_{ij}). Models with the feature of one dependent variable and multiple independent variables form the basis of "univariate analyses of variance." Models involving multiple dependent variables and multiple independent variables lead to "multivariate analyses of variance" (MANOVAA) and will not be considered in this work.

233. The ANOVA is designed to partition the total variability into its component parts. The terms in the model define the partitions. Expressing the model in Equation 1 in a slightly different fashion (Equation 2), the partitioning accomplished by an ANOVA and its interpretation are fairly direct and easy to comprehend.

$$X_{ij} - \mu = \alpha_i + \varepsilon_{ij} \qquad (2)$$

$$= (\mu_i - \mu) + (X_{ij} - \mu_i)$$

According to the models in Equations 1 and 2, the total deviation (X_{ij} - μ) is decomposed into two additive partitions, namely $\alpha_i = \mu_i - \mu$ and $\varepsilon_{ij} = X_{ij} - \mu_i$. These two partitions give rise to the decomposition of the total variability. The two partitions are appropriately called (a) the variability due to or explained by between-treatment-population effects and (b) the variability due to the heterogeneity which exists withi each treatment-population. Alternative and more commonly used terminology is "the treatment-population source of variance" and "the residual or error source of variance," respectively.

234. The ANOVA is based on the component sources as identified in the model, and the essence of an ANOVA is simple. As the treatment-populations become more distinct, the treatment-population source of variance accounts for a larger portion of the total variance relative to that which is accounted for by the residual, error on within-treatment-population variance. In Figure 17 the treatment-population becomes more distinct as a function of increased mean (μ_i) separation. In Figure 18 the treatment-populations become more distinct as a function of reduced error variance.

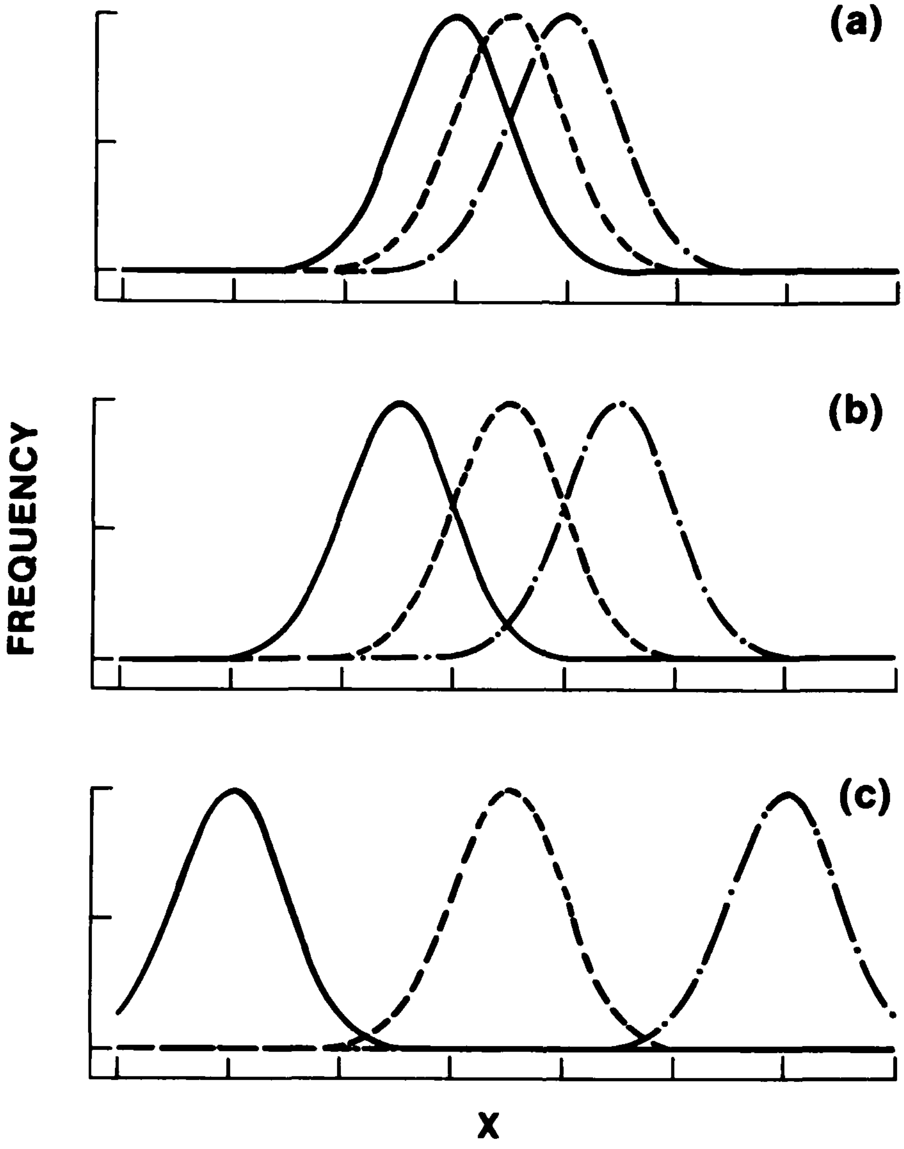


Figure 17. Distinction between treatment populations as a function of increased separation of the means

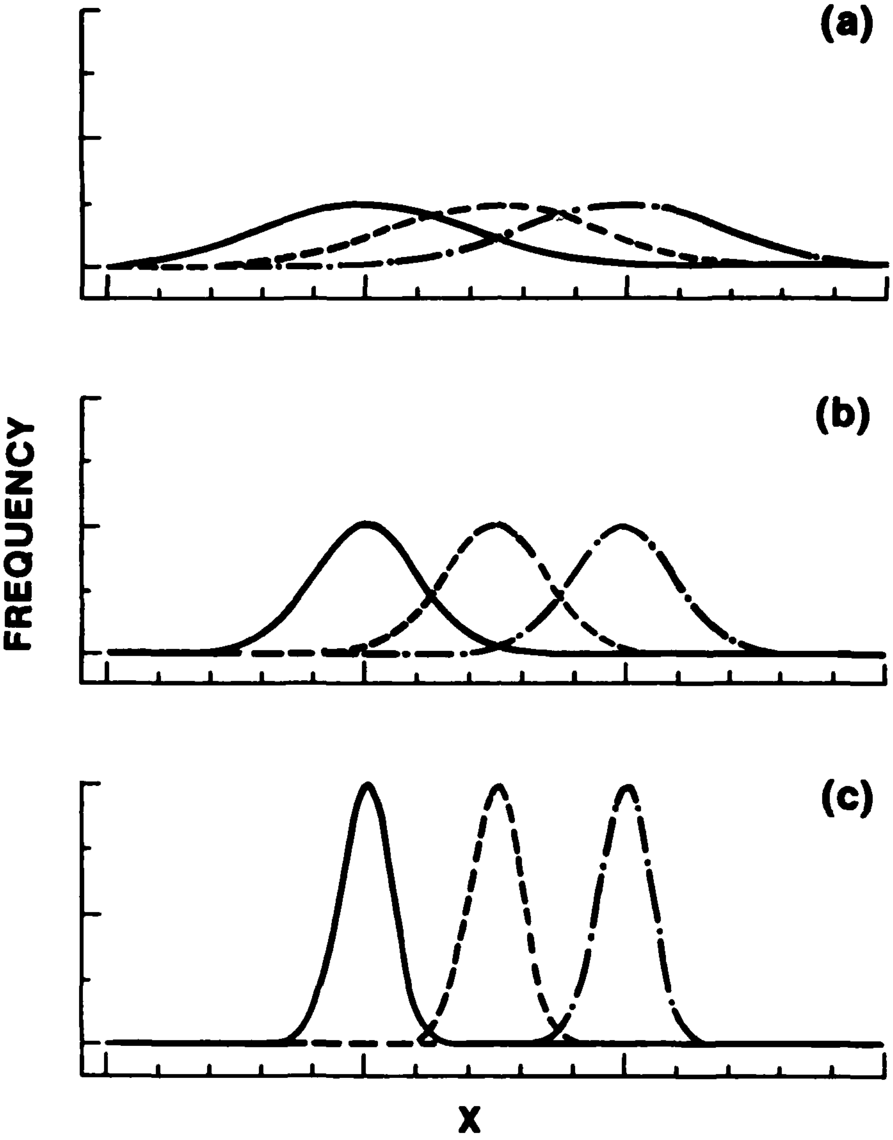


Figure 18. Distinction between treatment populations as a function of reduced error variance

235. The treatment-population distributions in Figures 17 and 18 conceptually represent all possible TP concentrations (X_{ij}) for surface water samples from the i^{th} reservoir. Consequently, μ refers to the overall mean TP concentration across all three reservoirs. Similarly, the total variance across all three reservoirs represents the failure of all observations to be the same and equal to the overall mean (μ). Therefore, it is the totality of all deviations ($X_{ij} - \mu$) in the model in Equation 2 which gives rise to the total variance.

236. In the two-sample t test, where it is assumed that the variances of the two sampled populations are equal, the common population variance σ^2 was estimated by the pooled variance

$$s_p^2 = (SS_1 + SS_2)/(DF_1 + DF_2)$$

where

SS_i = sum of squares, sample i

DF_i = degrees of freedom, sample i

The ANOVA also assumes the equality of variance

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

and the estimated population variance common to all k groups is given by

$$s_p^2 = SS_{error}/DF_{error}$$

where

$$SS_{error} = \sum_{i=1}^{k} \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \right]$$

$$DF_{error} = N - k$$

The quantities SS_{error} and DF_{error} are often referred to as the error sum of squares and the error degrees of freedom, respectively. The pooled variance s_p^2 is the best estimate of the variance common to all k groups.

237. To test the null hypothesis (no differences exist between the k groups), it is necessary to determine the amount of variability between the k groups. This variability is given by the group sum of squares

$$SS_{groups} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

and is associated with

$$DF_{groups} = k - 1$$

degrees of freedom.

238. It is also necessary to consider the total variability in the data collected, which is given by

$$SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

with

$$DF_{total} = N - 1$$

degrees of freedom.

239. The sums of squares given above can be more easily calculated from the following "machine formulae":

$$SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2 - \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}\right)^2 \Bigg/ N$$

$$SS_{groups} = \sum_{i=1}^{k} \left[\frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i}\right] - \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}\right)^2 \Bigg/ N$$

The error sum of squares is calculated by difference

$$SS_{error} = SS_{total} - SS_{groups}$$

240. To complete the calculations required for the one-way ANOVA it is necessary to divide the groups and error sums of squares by their respective degrees of freedom:

$$MS_{groups} = SS_{groups}/DF_{groups}$$

$$MS_{error} = SS_{error}/DF_{error}$$

Dividing a sum of squares by the degrees of freedom results in a variance which is called a mean square (MS) in the ANOVA. Mean square is short for "mean squared deviations from the mean." Table 18 presents a summary of the calculations required for a single-factor ANOVA.

241. Hypothesis testing with an ANOVA is based on the ratio of the two sources of variance

$$F = MS_{groups}/MS_{error}$$

The interpretation is direct. The larger the F value, the more distinct the treatment-population distributions become and a greater amount of the total variability is accounted for or explained by

Table 18

Summary of the Calculations for a One-Way ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Total $(X_{ij} - \bar{X})$	$SS_t = \Sigma\Sigma X_{ij}^2 - (\Sigma\Sigma X_{ij})^2/N$	$DF_t = N - 1$	
Treatment-population $(\bar{X}_i - \bar{X})$	$SS_\alpha = \Sigma[(\Sigma X_{ij})^2/n_i] - (\Sigma\Sigma X_{ij})^2/N$	$DF_\alpha = k - 1$	$MS_\alpha = \frac{SS_\alpha}{DF_\alpha}$
Residual or error $(X_{ij} - \bar{X}_i)$	$SS_e = SS_t - SS_\alpha$	$DF_e = DF_t - DF_\alpha$	$MS_e = \frac{SS_e}{DF_e}$

Notes: k = number of treatment populations, n_i = number, and $N = \sum_{i=1}^{k} n_i$.

treatment-population (group) differences. If the calculated F value is at least as large as the critical F value (with numerator DF = DF_{groups} and denominator DF = DF_{error}), then the conclusion that the treatment-populations are not equal can be made. If the calculated F is less than the critical value, the conclusion is not warranted.

242. Tables 19 and 20 present the computational steps involved in the one-way ANOVA for the TP data from the three reservoirs. Based on the observed F value (31.24), the null hypothesis of no differences can be rejected. However, the exact meaning of an F test in an ANOVA is inherently tied to what the mean squares (MS_{groups}, MS_{error}) estimate.

Residual (error) effects and the mean-square error

243. All analyses of variance reduce to situations involving a random sample of units from a series of treatment-populations. The dependent variable represents the property of the units which is purportedly influenced by the treatment-population to which the unit belongs. As such, the dependent variable is expressed as a function of the treatment-population parameters. But in drawing inferences about these population parameters based on sample data and the results of an ANOVA, certain assumptions are made about the dependent variable. The assumptions that typify virtually all analyses of variance are given in Equation 3, which states that the X_{ij} values are distributed normally and independently about a mean μ_i with variance σ_i^2 .

$$X_{ij} \sim NID (\mu_i, \sigma_i^2) \qquad (3)$$

244. These assumptions say that each observation comes from a normal population with a particular unknown mean and variance. Naturally, the parameters are unknown, but sample estimates are available. However, when inferences are based on the sample data (Table 17) and the ANOVA (Table 20), an additional assumption is made. The within-treatment-population variances are assumed to be equal to a common value, denoted $\sigma_i^2 = \sigma^2$ for all i . The assumptions of Equation 3 thereby become those of Equation 4.

Table 19

Calculations for a One-Way ANOVA Using the Data from Table 17

	Reservoir	Reservoir	Reservoir	
	31 28 27 30	32 30 30 31 34	22 21 23	
n_i	4	5	3	$N = 12$
$\sum$ $\sum_{j=1}^{n_i} X_{ij}$	116	157	66	
$\bar{X}_i$	29.0	31.4	22.0	
$\frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i}$	3,364.0	4,928.8	1,452.0	
$A = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = 339.0$				$DF_{total} = N - 1 = 11$
$B = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2 = 9{,}769.0$				$DF_{groups} = k - 1 = 2$
$C = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i} = 9{,}744.8$				$DF_{error} = N - k = 9$
$SS_{total} = B - (A^2/N) = 192.25$				
$SS_{groups} = C - (A^2/N) = 168.05$				$MS_{groups} = \frac{SS_{groups}}{DF_{groups}} = 84.03$
$SS_{error} = SS_{total} - SS_{groups} = 24.20$				$MS_{error} = \frac{SS_{error}}{DF_{error}} = 2.69$

Table 20

Analysis of Variance on the Data of Table 19

Source of Variation	SS	DF	MS
Total	192.25	11	
Reservoirs (i.e., groups)	168.05	2	84.03
Error	24.20	9	2.69

$$F = \frac{MS_{groups}}{MS_{error}} = \frac{84.03}{2.69} = 31.24$$

$\alpha = 0.05 \qquad F_{2,9} = 4.26$

$$X_{ij} \sim \text{NID}\ (\mu_i,\ \sigma^2) \qquad (4)$$

245. The consequences of this last constraint are immediate. Since each treatment-population has the same variance, each sample variance (see Table 18) estimates the common value. In turn, since all sample variances estimate the common variance, the sample variances can be pooled to provide a combined sample estimate of the within-treatment-population variance. It is precisely this combined or pooled sample estimate of the within-treatment-population variance which is called the mean-square error (MS_e) in the ANOVA. The idea of pooling is simple and can be shown using the data of Table 19. After pooling (i.e., adding) the individual sum-of-squares (i.e., $SS_e = 0.1875 + 0.272 + 0.14 = 0.5995$), divide the pooled sum-of-squares by the pooled degrees of freedom (i.e., $df_e = 3 + 4 + 2 = 9$) to generate the mean-square error ($MS_e = 0.5995/9 = 0.0666$).

246. Since the assumption of equality of variances (Equation 4) refers to within-treatment-population variability and since the deviations ε_{ij} of the models in Equations 1 and 2 yield the within-treatment-population variance, the assumptions of Equation 4 are typically stated in terms of the residual (error) component of the model as given in Equation 5, which states that residual (error) effects are

distributed normally and independently about a mean 0 and common error variance. The moral is that the MS_e estimates the common within-treatment-population variance (σ_ε^2). Hence, Equation 6 says that the expected value of the MS_e is σ_ε^2.

$$\varepsilon_{ij} \sim \text{NID}\ (o,\ \sigma_\varepsilon^2) \qquad (5)$$

$$E\ (MS_e) = \sigma_\varepsilon^2 \qquad (6)$$

Treatment-population effects: fixed or random?

247. Often a researcher faces a dilemma in deciding which of several or many treatment-populations he should study intensively. The researcher's dilemma may be phrased as follows: "Does the researcher want to know about differences between the three reservoirs and only the three reservoirs that were sampled?" or "Does the researcher want to draw inferences about all reservoirs in the region based on the subset of three reservoirs?" If the answer to the first question is "yes," the researcher is dealing with fixed effects. If the answer to the second question is "yes," the researcher is dealing with random effects.

Expected mean-squares for fixed effects (Model I ANOVA)

248. Fixed effects refer to a set of treatment-populations which is finite either by nature or by an "a priori" selection process; more importantly, inferences are limited to only those treatment-populations that are sampled. If the researcher samples all or some of the treatment-populations and if the researcher's concern is with the sampled treatment-populations and only those evaluated, the model in Equation 1 is a fixed effects model, and the analysis of variance is called a Model I ANOVA. Consequently, for the data of Table 17 and its analysis (Tables 19 and 20), if the researcher's interest lies solely in the three sampled coves, then cove effects would be considered fixed.

249. Due to the nature of fixed effects, the inferential process conventionally involves either or both estimation of

treatment-population means or mean differences and tests of hypotheses about significant differences among means. Fixed effects mean-squares are consistent with the nature of the inferences. The expected mean-square for the treatment-population source of variance is written in the form of Equation 7, and its meaning and interpretation are linked to the quadratic portion of the expected mean-square (Equation 8) and the hypothesis (H_0) of conventional interest. Thus,

$$E(MS_\alpha) = \sigma_\varepsilon^2 + Q(\alpha) \qquad (7)$$

where $Q(\alpha)$ is a function of the treatment-population effects (α_i = reservoir effects)

$$Q(\alpha) \text{ involves } \sum_{i=1}^{p} \alpha_i^2 \qquad (8)$$

H_0:all treatment-population means are equal or all treatment-population effects are zero (9)

250. Since the F value is the ratio of the MS_α to the MS_e , the interpretation of the F test statistic is direct. If the null hypothesis (H_0:the hypothesis of no differences) is true, then (a) all treatment-population distribution and means would be superimposed and, as such, equal to each other and equal to the overall mean; (b) all effects would be equal to zero and, as such, $Q(\alpha)$ would be equal to zero; (c) the $E(MS_\alpha) = \sigma_\varepsilon^2 = E(MS_e)$, and (d) the researcher would expect the F value to be close to one. Conversely, if the null hypothesis was not true (i.e., all means were not equal), then (a) the treatment-population effects would show larger and larger deviations from zero as the treatment-population distributions become more distinct; (b) $Q(\alpha)$ would show a corresponding increase; and (c) the researcher would expect the F value to get larger and larger since the $E(MS_\alpha)$ would increase relative to the $E(MS_e)$.

251. If the F value reached the selected significance level, the inference would be that significant differences exist among the means. A natural question of the researcher would be, "Which means are different from each other?" The issue is that an ANOVA and its F test do not lead to the inference that all means are different from each other. A significant F test merely claims that differences do exist. Specific questions about particular means and mean differences (i.e., post-ANOVA type questions) require techniques which supplement the ANOVA. These techniques are the subject of a later section.

Expected mean-squares for random effects (Model II ANOVA)

252. Random effects are considered by many to be conceptually more elusive than fixed effects. This should not be so. Random effects refer to a random sample of treatment-populations. Consider the dilemma of the researcher regarding which reservoirs to study intensively. If the researcher wished to infer something about the variability among all reservoirs across an entire region, then he certainly will not be able to study all reservoirs. He could, however, draw a sample of reservoirs at random, compute the variance of the reservoir effects sampled (denoted s_{α}^{2}) and let this sample variance estimate the variance of the entire population reservoir effects (denoted σ_{α}^{2}). In essence, this is precisely what happens in random effects models or Model II ANOVA.

253. For random effects, the nature of the hypothesis changes from that for fixed effects. The hypothesis (Equation 10) is stated in terms of the variance component being estimated. Likewise, the expected mean-square (Equation 11) is a function of the variance component.

$$H_0 : \sigma_{\alpha}^{2} = 0 \quad (10)$$

$$E(MS_p) = \sigma_{\varepsilon}^{2} + k\sigma_{\alpha}^{2} \quad (11)$$

254. The meaning of the F test is clear. If the null hypothesis (Equation 10) is true, the $E(MS_{\alpha}) = \sigma_{\varepsilon}^{2} = E(MS_e)$ and the researcher would expect the F value to draw closer to one as the variance drew

closer to zero. Conversely, if the variance of cove effects (σ_α^2) were large, the F value would increase and the inference would be that a significant degree of variability exists among the reservoirs of the region. Naturally, the researcher would want to estimate this variance. As with post-ANOVA tests for fixed effects, this estimation enterprise will be pursued later.

255. Several issues need to be raised and explained before concluding the discussion of random effects. Random effects, by definition, involve a two-stage random sampling process. The first stage involves the random selection of a set of treatment-populations while the second stage involves the random sampling of units from each treatment-population. There are cases, however, where the first stage will not actually involve a random sampling process. If this occurs, a minimum requirement is that the researcher must be willing to assume that his set of treatment-population represents a random sample.

256. The second issue is tied to the first. In the presence of random effects, special assumptions (Equation 12) are made in addition to these made about error effects (Equation 5). The assumptions of Equation 12 read, "The α values are normally distributed about a mean of zero with variance σ_α^2 . These assumptions are not problematic in concept. They simply mirror the assumptions made about residual or error effects which are, in actuality, random effects themselves.

$$\alpha_i \sim (o, \sigma_\alpha^2) \qquad (12)$$

257. The third issue is simply a point of clarification. The models of an ANOVA need not be simply fixed effects or random effects models. The structure or nature of the treatment-population may be partially random. If so, the models are called "mixed models." Examples of these types of models are discussed in a later section.

258. The final issue is related to the nature of the inferential process in random effects models. Sometimes a completely valid F test is unavailable. This situation occurs in studies involving unbalanced data (i.e., unequal sample sizes from each treatment-population). The

nature of the inferential process thereby becomes one of estimating variance components rather than specific tests of hypotheses. Variance component estimation is discussed in more detail later. As a final note, the problems caused by unbalanced data within the content of random effects do not occur in fixed effects models.

Fixed effects and arrangements of treatment-populations

259. Inferences about fixed effects are normally restricted to (a) tests of hypotheses about significant differences between treatment-population means, and (b) point or interval estimation of treatment-population means and mean differences. Remember, sampling or experimental designs refer to the plan or procedure followed to obtain a random sample of units from a series of treatment-populations. Except for the prior and more detailed discussion of fixed and random effects, little has been said about how the set of treatment-populations to be studied might be defined.

260. Often treatment-populations differ according to more than one dimension or can be distinguished by more than one factor. Consider a study designed to evaluate surface water quality in two coves (c_1, c_2) as a function of two depths (d_1, d_2) across three seasonal months (m_1, m_2, m_3) and suppose further that several water samples are drawn at each depth from each cove during each month. As a result, 12 treatment-population combinations exist, but the treatment-populations are distinguished by three factors or dimensions, cove at two levels, depth at two levels, and month at three levels. The cross-classification of these factors yields the 12 treatment-population combinations. The factors are said to be cross-classified because each level of each factor, or each cove sampling, is conducted during the same months at the same depths. This property of cross-classification yields a set of treatment-populations which are said to follow a "factorial arrangement of factors."

261. Contrast this factorial structure to the following example. Suppose the study involved two different reservoirs (r_1, r_2) but two coves from one reservoir (c_1, c_2 from r_1) and three coves from the

second reservoir (c_1, c_2, c_3 from r_2). Suppose further that the sampling of coves was done during different months. For instance, let the 12 months be denoted by m_1, $m_2 \ldots m_{12}$ with the following sampling scheme r_1c_1 during m_1, m_{10} ; r_1c_2 during m_4, m_2, m_9 ; r_2c_1 during m_3, m_4 ; r_2c_2 during m_5, m_{11}, m_{12} ; r_2c_3 during m_2, m_7, m_8 . Note immediately that 12 treatment-populations exist based on three factors (reservoir, cove, month) which are not cross-classified, different coves within each reservoir, and different sampling dates for each cove. This property of different levels of a factor at different levels of other factors yields what is termed a "nested" or "hierarchical" structure.

Factorial arrangements (main and interaction effects)

262. The function of an ANOVA in fixed effects models is to determine whether significant differences exist among treatment-population means. Table 21 and Figure 19 present the mean water quality (WQ) for each of the 12 treatment-populations which are based on a 2 by 2 by 3 factorial arrangement of three factors (cove, depth, month). The raw data of Table 21 are totally contrived to expedite the subsequent discussion. Results of the ANOVA are given in Table 22.

$$\overline{CDM}_{ijk} = \text{mean water quality for the } i^{th} \text{ cove } (i=1\ldots c=2) \text{ at the } j^{th} \text{ depth } (j=1\ldots d=2) \text{ during the } k^{th} \text{ month } (k=1\ldots m=3) \qquad (13)$$

263. As can be seen easily in Figure 19 and Table 21, the ANOVA would lead to the conclusion that differences between means exist, and water quality varies as a function of cove, depth, and month. The question arises, "How might treatment-population effects be explained?" According to the factorial structure of the treatment-populations, two types of effects can be identified, "main" and "interaction" effects.

264. The distinction between main and interaction effects is quite simple. Questions about main effects can be stated, "Are there significant differences among the means at each level of a single factor?" For instance, questions about whether coves differ in mean

Table 21

Hypothetical Water Quality for a 2 by 2 by 3

Factorial Arrangement of Three Factors--

Cove, Depth, Month

Cove (i)	Depth (j)	Month (k)	Sample (l)	WQ (ijkl)	Mean ($\overline{CDM}_{ijk}$)
1	1	1	1	1.8	2.0
1	1	1	2	2.2	
1	2	1	1	2.0	2.0
1	1	2	1	3.7	
1	1	2	2	4.3	4.0
1	1	2	3	4.0	
1	2	2	1	8.4	
1	2	2	2	7.5	
1	2	2	3	7.6	8.0
1	2	2	4	8.5	
1	1	3	1	5.5	
1	1	3	2	6.5	6.0
1	2	3	1	7.8	
1	2	3	2	8.2	8.0
2	1	1	1	1.0	1.0
2	2	1	1	0.9	
2	2	1	2	1.1	1.0
2	2	1	3	1.0	
2	1	2	1	3.6	
2	1	2	2	4.4	4.0
2	2	2	1	8.1	
2	2	2	2	7.9	8.0
2	1	3	1	0.8	
2	1	3	2	1.2	1.0
2	1	3	3	1.0	
2	2	3	1	3.0	3.0

water quality and whether mean water quality changed as a function of depth are main effect questions, and three main effect questions exist (one for each factor).

265. Questions about interaction effects refer to interrelationships among the levels of the different factors and can be stated, (a) "Is the effect of a factor the same across levels of other factors?" or (b) "Are the mean differences among levels of a factor consistent across levels of other factors?" If the answer is "yes," the factors do

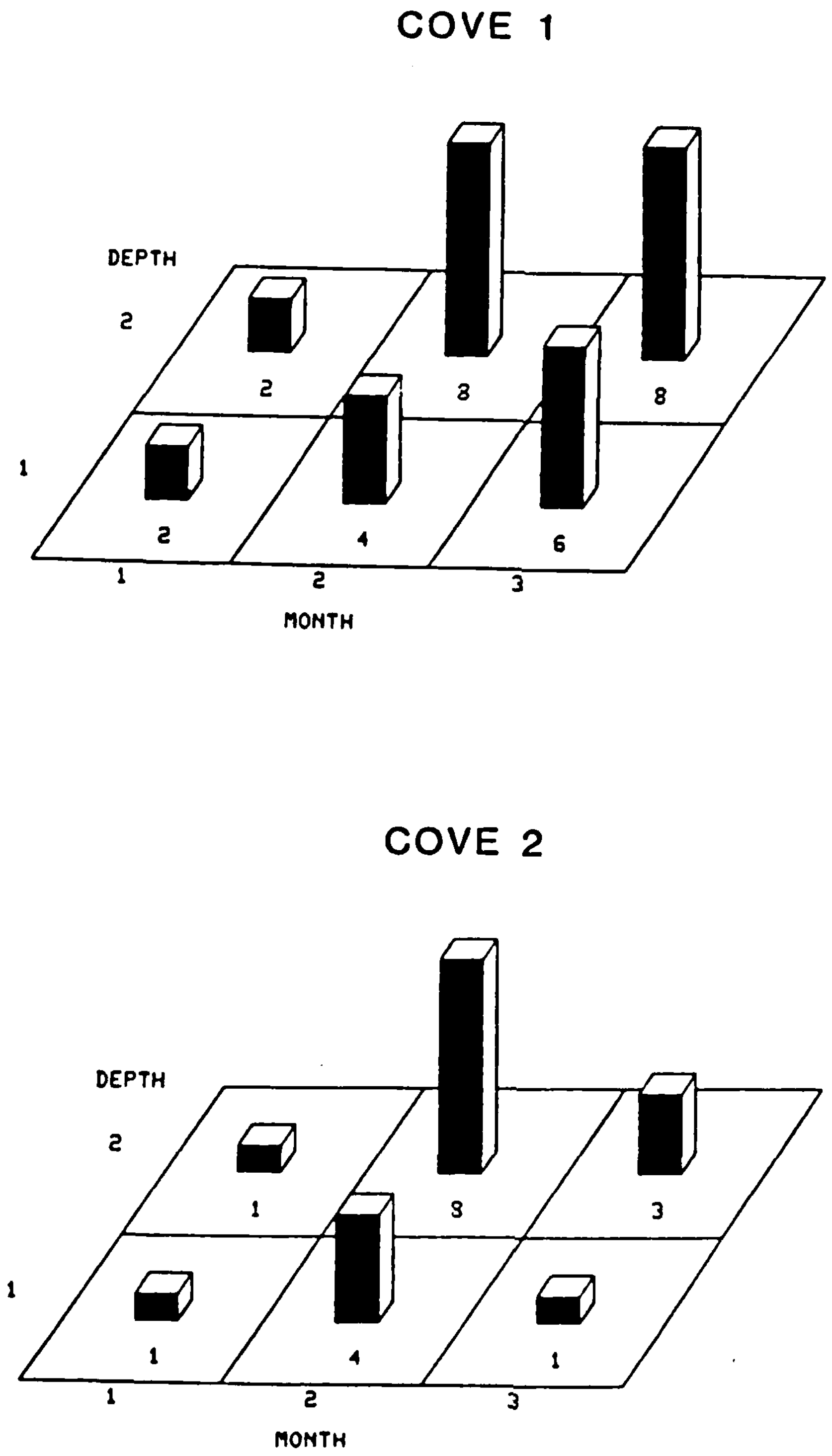


Figure 19. Mean water quality for the 12 treatment-populations

Table 22

Results of ANOVA for the Data of Table 21

Source	df	SS	MS	F	Pr > F
CDM	11	209.538	19.049	126.99	0.0001*
Error	14	2.100	0.150		
Total	25	211.638			

* Significant at the 0.0001 level.

not interact and the individual factor effects are said to be "additive." If the answer is "no," an interaction effect is said to exist.

266. In our example, four interaction effects are possible: interaction effects for each pair of factors (i.e., the three two-factor interactions of C with D, C with M, and D with M) and an interaction effect for all three factors (i.e., one three-factor interaction). For instance, the two-factor interaction (C by D) would concern whether the change in water quality as a function of depth was the same for both coves. The three-factor interaction (C by D by M) would refer to whether or not the interrelationship for C and D was the same across months.

267. Generalization to higher order factorial structures is fairly direct. Consider a four-way cross-classification system (arbitrarily, let the four factors be A, B, C, and D). If this were the situation, four main effects (A, B, C, D), six two-factor interactions (AB, AC, AD, BC, BD, CD), four three-factor interactions (ABC, ABD, ACD, BCD), and one four-factor interaction (ABCD) are possible.

268. To understand the nature of main effects and interaction effects, the means of Table 21 ($\overline{CDM}_{ijk}$) and Figure 19 require additional summarization. Table 23 and Figure 20 present the means as defined in Equations 14-16, to explain main effects, while Table 24 and Figure 21 present the means, as defined Equations 17-19, to explain the two-factor interactions. Table 25 and Figure 22 present the means of Table 21 and Figure 19, but the presentation is in a form that is better suited for

Table 23

Summarization of Means and Mean Differences for Each Main Effect

Factor	Means	Mean Differences
Cove (i)		
C_1	5.6	$C_1 - C_2 = 2.8$
C_2	2.8	
Depth (j)		
$\bar{D}_1$	3.1	$\bar{D}_1 - \bar{D}_2 = -2.4$
$\bar{D}_2$	5.5	
Month (k)		
M_1	1.4	$\bar{M}_1 - \bar{M}_3 = -2.9$
M_2	6.2	$\bar{M}_2 - \bar{M}_3 = 1.7$
M_3	4.3	

interpretation of the three-factor interaction.

$\bar{C}_i$ = mean for the i^{th} cove (14)

$$\left(\sum_{j=1}^{d} \sum_{k=1}^{m} \sum_{l=1}^{s} WQ_{ijkl}\right) \Big/ n_i$$

n_i = number of samples from the i^{th} cove

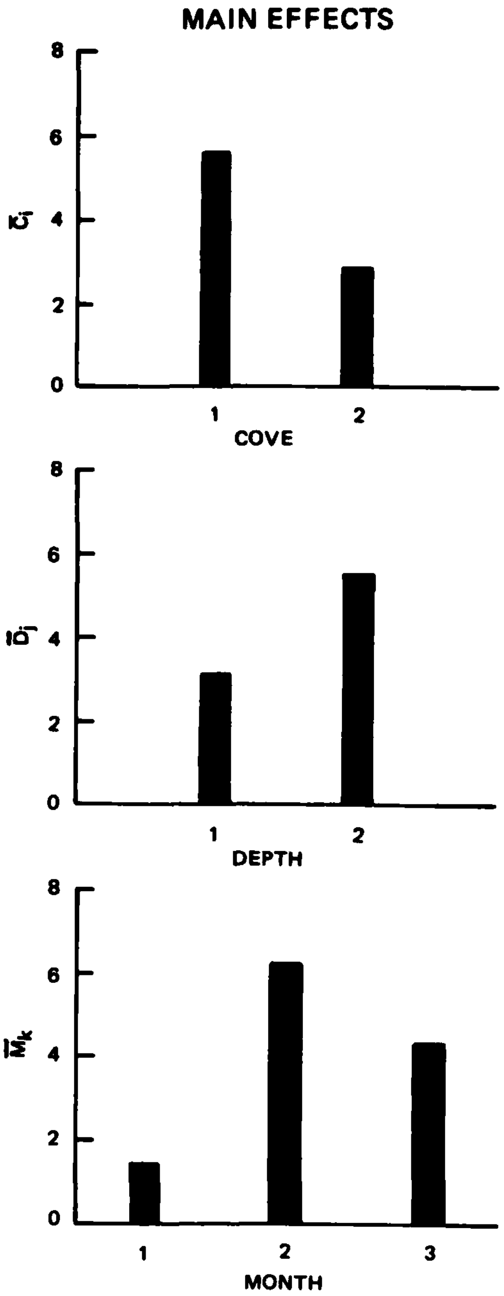


Figure 20. Main effects for the 12 treatment-populations

$\bar{D}_j$ = mean at the j th depth (15)

$$\left(\sum_{i=1}^{c} \sum_{k=1}^{m} \sum_{l=1}^{s} WQ_{ijkl}\right)\Big/ n_j$$

n_j = number of samples from the j th depth

Table 24

Summarization of Means and Mean Differences for Each Two-Factor Interaction

	$\overline{CM}_{ij}$ Cove		Cove Differences by Month	Difference by Month of Cove Differences
Month	C_1	C_2	$(\bar{C}_1 - \bar{C}_2)$ by M	$(M_1 - M_3, M_2 - M_3)$
M_1	2.0	1.0	1.0	-4.5
M_2	6.3	6.0	0.3	-5.2
M_3	7.0	1.5	5.5	

	$\overline{CD}_{ij}$ Cove		Cove Differences by Depth	Difference by Depth of Cove Differences
Depth	C_1	C_2	$(\bar{C}_1 - \bar{C}_2)$ by M	$(D_1 - D_2)$
D_1	4.0	2.0	2.0	-1.4
D_2	7.1	3.7	3.4	

	$\overline{MD}_{jk}$ Depth		Depth Differences by Month	Difference by Month of Depth Differences
Month	D_1	D_2	$(\bar{D}_1 - \bar{D}_2)$ by M	$(M_1 - M_3, M_2 - M_3)$
M_1	1.7	1.25	0.5	3.8
M_2	4.0	8.0	-4.0	-0.7
M_3	3.0	6.3	-3.3	

$\bar{M}_k$ = mean during the k^{th} month (16)

$$\left(\sum_{i=1}^{c} \sum_{j=1}^{d} \sum_{l=1}^{s} WQ_{ijkl} \right) \Big/ n_k$$

n_k = number of samples from the k^{th} month

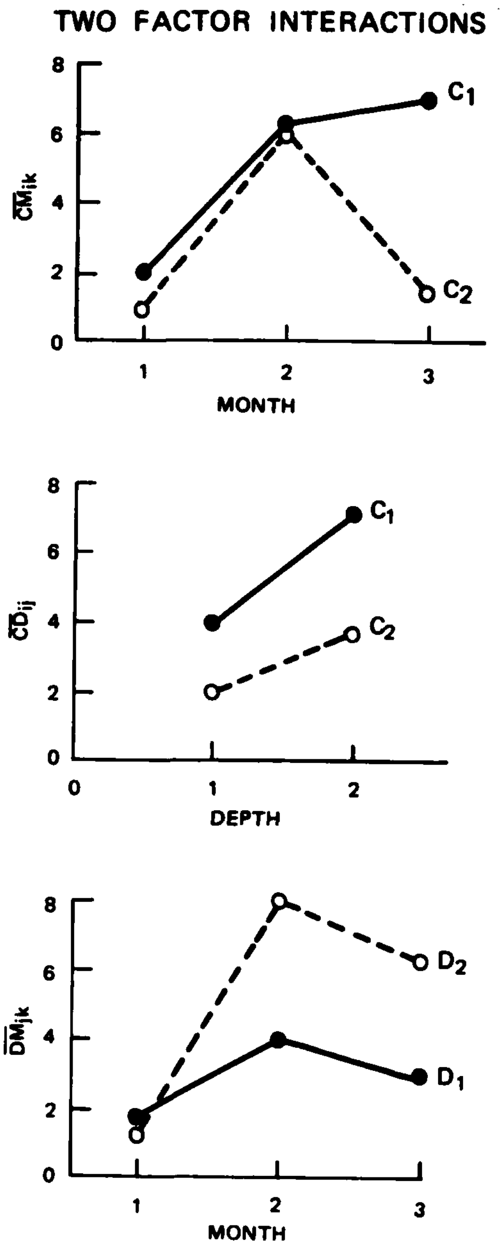


Figure 21. Two-factor interactions for the 12 treatment-populations

$\overline{CD}_{ij}$ = mean at j th depth for i th cove (17)

$$\left(\sum_{k=1}^{m} \sum_{l=1}^{s} WQ_{ijkl}\right) \Big/ n_{ij}$$

N_{ij} = number of samples from the i th cove at the j th depth

Table 25

Summarization of Means and Mean Differences for the Three-Factor Interaction

		$\overline{CDM}_{ijk}$ Cove		$(\bar{C}1 - \bar{C}2)$	CM Interaction	Difference of CM Interactions
Depth	Month	C_1	C_2	by M	at D_1	$(D_1 - D_2)$
1	M_1	2.0	1.0	1.0	-4.0	
	M_2	4.0	4.0	0.0	-5.0	
	M_3	6.0	1.0	5.0		
						0.0
						0.0

		$\overline{CDM}_{ijk}$ Cove		$(\bar{C}1 - \bar{C}2)$	CM Interaction	Difference of CM Interactions
Depth	Month	C_1	C_2	by D	at D_1	$(D_1 - D_2)$
2	M_1	2.0	1.0	1.0	-4.0	
	M_2	8.0	8.0	0.0	-5.0	
	M_3	8.0	3.0	5.0		

$\overline{CM}_{ik}$ = mean during kth month for ith cove (18)

$$\left(\sum_{j=1}^{d} \sum_{l=1}^{s} WQ_{ijkl} \right) \Big/ n_{ik}$$

n_{ik} = number of samples from the ith cove during the kth month

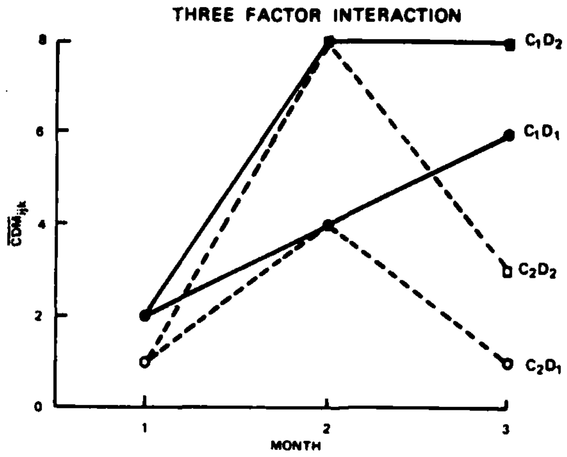


Figure 22. Three-factor interaction for the 12 treatment-populations

$$\overline{DM}_{jk} = \text{mean during } k^{th} \text{ month at } j^{th} \text{ depth} \qquad (19)$$

$$\left(\sum_{i=1}^{c} \sum_{l=1}^{s} WQ_{ijkl} \right) \Big/ n_{jk}$$

n_{jk} = number of samples from the j^{th} depth during the k^{th} month

269. The essential feature of any effect (main, interaction, nested, etc.) is that each is defined in terms of differences among means. Hence, every effect has a set of one or more defining contrasts. The final objective of the present discussion is to develop the contrast matrix of Table 25 and to establish a set of rules for specifying the defining contrasts for any effect. An adequate appreciation of a contrast matrix will not only allow for the presentation of many more examples than would otherwise be possible but will also significantly expedite the discussion of future topics.

270. Only one contrast is needed to establish the presence or absence of a cove main effect. If water quality does not depend on the cove from which the water was drawn, the cove means ($\bar{C}_i$) in Table 23 would be equal and the difference between cove means would be zero

(i.e., $C_1 - C_2 = 0$). As is obvious, such is not the case (i.e., $C_1 - C_2 = 2.0$). Water quality at C_1 was different than at C_2. Only one contrast among means is required because the cove factor has only two levels. A general rule (Rule 1) for main effects is that the required number of defining contrasts is or : less than the total number of levels. By identical arguments, a main effect for depth exists. The magnitude of the water quality variable increased with depth. Naturally, more depths could have been sampled to provide a more thorough picture of change in water quality.

271. A month main effect exists along with the cove and depth main effects. Water quality changed with time. Since the factor of month had three levels, only two defining contrasts are required even though three contrasts are possible (i.e., $M_1 - M_2$, $M_1 - M_3$, $M_2 - M_3$). Table 23 gives the contrasts, $M_1 - M_3$ and $M_2 - M_3$.

272. The means and mean differences show that, with time, water quality rises and then falls, but not to the level observed for the first month. A second general rule (Rule 2) for main effects is that the set of all possible defining contrasts consists of all possible differences among means. A convenient selection consists of those differences involving the mean for the last level of the factor. As a precautionary note, other contrasts are possible (e.g., $M_1 - 2M_2 - M_3$ is a contrast), but Rule 2 provides a convenient and easily generated set.

273. Having chosen the two defining contrasts, a third general rule (Rule 3) for main effects (in fact, for any effect) is that all defining contrasts must be zero in order to establish the absence of a main effect. Rule 3 can be shown easily with some simple algebra. Merely let one contrast be different from zero (e.g., suppose $M_1 - M_3 = 0$ but $M_2 - M_3 = 2$). By the first contrast $M_1 = M_3$, and by the second contrast $M_2 = 2M_3$. Therefore, M_2 is twice as big as M_1 and, as such, not all means are equal (i.e., a main effect exists). Rule 3 also illustrates an extremely important principle--the presence of an effect does not imply that all means are different. If an ANOVA indicates the presence of an effect, then the inference is simply that

differences exist somewhere. Precisely where these differences occur requires supplementing the ANOVA with other techniques. Many specific techniques are available for ascertaining precisely which means are different from each other. These techniques are discussed later, but all are based on the concept of contrasts among means.

274. Two-factor interactions are more complex than main effects, but certain principles are the same. Attention is restricted initially to the C by D interaction (Table 24, Figure 21). A CD interaction exists; the difference between C_1 and C_2 is exactly the same at both depths.

275. To define a two-factor interaction, simply perform a double-contrast operation (Rule 4): (a) choose a factor and evaluate its defining contrast(s) at each level of the second factor, then (b) apply the defining contrast(s) of the second factor to the results of the contrasts from (a). Referring to the C by D interaction in Table 24, the defining contrast for cove differences (i.e., $C_1 - C_2$) is applied separately at each depth. The defining contrast for depth differences (i.e., $D_1 - D_2$) is then applied to the cove differences. The result is not zero and, by Rule 3, an interaction exists. The double-contrast operation yields Rule 5, which specifies the number of defining contrasts required to establish any interaction effect--the number of defining contrasts equals the product of the number of defining contrasts for each factor involved in the interaction (e.g., since both the C main effect and the D main effect had only one defining contrast, only one defining contrast is required for the C by D interaction).

276. Identical arguments are used in Table 24 to establish the presence of a C by M and D by M interaction. Figure 21 shows that the change in water quality as a function of time is different depending upon cove (C by M) and depending upon depth of sampling (D by M). Restricting attention to the C by M interaction (same arguments apply to the M by D interaction), the defining contrast for coves $(C_1 - C_2)$ is applied for each month and then, by Rule 4, the defining contrasts for month $(M_1 - M_3, M_2 - M_3)$ are applied to the cove contrasts. By

Rule 5, the C by M interaction requires two defining contrasts and, by Rule 3, all must be zero to claim no interaction. Obviously, such is not the case for either C by M or M by D interaction.

277. Evaluation of the three-factor interaction (C by D by M) simply requires generalizing Rule 4 to produce Rule 6: to evaluate any higher order interaction: (a) establish the defining contrast(s) for any interaction which is 1 degree lower than the interaction of concern and (b) apply the defining contrasts of the ignored factor to the defining contrasts of the lower order interaction. Rule 6 is followed in Table 25. The defining contrasts for the C by M interaction are established at each depth and then the defining contrast for depths is applied. By Rule 5, the three-factor interaction requires two defining contrasts, and by Rule 3, no three-factor interaction exists because all defining contrasts equal zero.

278. The results of all main effect and interaction effect comparisons may be summarized. By Figure 19, differences existed between means. Not all effects (main and interaction) contributed to the differences witnessed in Figure 19--the differences were not due to the C by M nor the C by D by M interaction, even though a quick glance at all ongoing fluctuations across coves, months, and depths would seem to indicate otherwise (Figure 19 or 22). Consequently, visual inspection is insufficient for establishing higher order interaction effects. Finally, main effects and interaction effects constitute only one set of contrasts among means. Many contrasts are possible; the researcher's task is to select the set of contrasts which has the most meaning for the research questions at hand.

Hierarchical arrangements

279. "Nested" or "hierarchical" arrangements are not that far removed from a factorial arrangement. Structurally, the difference is that there are different levels of one factor at each level of another factor. For instance, a study may involve two reservoirs and three recreational areas on each reservoir, thereby yielding six treatment-populations. The nature of the inferences about the

treatment-population effects is somewhat different from those under a factorial structure. To describe the inferential process, the data of Table 26 can be used. In Case I there is a reservoir main effect ($\mu_{.j}$ are not equal), but a reservoir main effect does not exist in Case II.

280. A word of caution is required regarding main effects in a hierarchical structure. The marginal means are based on different levels of another factor. Hence, main effects comparisons are not made across common or comparable levels of other factors. This word of caution also points to the character of nested effect questions. Rather than asking about comparability of factor levels as a function of other factors, as is done with factorial structures, hierarchical structure inquiries take the form, "Are there differences among the particular levels of a factor within each level of other factors?" Based on the data of Table 26 this question must be answered in the affirmative in both cases.

281. It probably goes without saying that as the number of factors and number of levels of each factor increase, the interpretation of higher order interaction effects and nested effects becomes a more difficult enterprise. In actuality, this is true only for higher order interaction effects. The interpretation of higher order nested effects is incredibly easier. The reason lies in the definition of the effects. In the case of higher order interactions, the effects are stated in terms of differences of differences of differences. For instance, if the situation involved a four-way interaction of A , B , C , and D , then the four-way interaction would ask if a three-way interaction (such as A by B by C) was constant across the fourth factor (D). But evaluating the three-way interaction across the fourth would involve, "Are the differences (among C) of the differences (among B) of the differences (among A) constant over D?" But for nested effects, the question is simply one of whether or not differences do exist among levels of the factor which is nested.

282. Suppose a three-factor hierarchical arrangement of A , B within A (B/A) , and C within B within A (C/B/A) is employed. Letting

Table 26

Two Hypothetical Situations for a 3 by 2 Factorial Arrangement of Area (Factor A at 3 Levels) and Depth (Factor B at 2 Levels)--All Entries Are Means (μ_{ij}, $\mu_{i.}$, $\mu_{.j}$)

	Area (i)			
Reservoir (j)	A_1	A_2	A_3	$\mu_{.j}$
Case I				
B_1	3	5	1	3
B_2	5	7	3	5
$\mu_{i.}$	4	6	2	
Case II				
B_1	3	5	1	3
B_2	1	1	7	3
$\mu_{i.}$	2	3	4	

μ_{ijk} denote the ABC_{ijk} means, the effects may be defined as $\alpha_i = (\mu_{i...} - \mu)$, $\beta_{j/i} = (\mu_{ij} - \mu_{j..})$, and $\gamma_{k/ij} = (\mu_{ijk} - \mu_{ij.})$. Notice that these effects are 0 if and only if the means within a group are equal. For instance $\beta_{j/i} = 0$ if and only if all β means (μ_{ij}) within the i^{th} level of A are equal to the overall i^{th} A mean ($\mu_{i..}$). Thus, nested effects exist if differences among means within a group simply exist.

Post-ANOVA Hypotheses

283. Researchers often need to make specific contrasts or test specific hypotheses about differences among means, hypotheses which are not answered directly by main effects, interaction effects, nested effects on combinations thereof. For instance, the analysis of Table 19 tests the omnibus hypothesis, "Are there differences among coves?" However, the researcher may wish to know whether pairwise differences exist, that is, which coves are different.

284. Define a contrast by

$$L = \sum_{i=1}^{a} c_i \mu_i$$

where

μ_i = i th treatment-population mean

c_i = contrast multiplier for the i th mean

285. Suppose the researcher wants to test whether cove 1 differs significantly from cove 2.

286. Then the hypothesis would be

$$H_o : \mu_1 - \mu_2 = 0$$

where $c_1 = 1$, $c_2 = 0$, and $c_3 = -1$.

287. On the other hand, suppose the researcher wants to know if the average of the means for the first two coves is significantly different from the mean for the third cove.

288. Then the hypothesis would be

$$H_o : \frac{\mu_1 + \mu_2}{2} - \mu_3 = 0$$

where $c_1 = 1/2$, $c_2 = 1/2$, and $c_3 = -1$.

289. To test these hypotheses, an F test statistic may be defined as follows:

$$F = \frac{L^2}{MS_e \sum_{i=1}^{a} \frac{(c_i)^2}{n_i}}$$

where $L = \sum_{i=1}^{a} c_i \bar{x}_i$.

290. Referring to Tables 17 and 19, and the latter hypothesis,

$$F = \frac{[1/2(29.0) + 1/2(31.4) - 22.0]^2}{2.69 \frac{(1/2)^2}{4} + \frac{(1/2)^2}{5} + \frac{(-1)^2}{3}}$$

$$= \frac{67.24}{1.20}$$

$$= 56.03$$

291. The question is whether the F test statistic is significant. It is here that many methods are available which differ according to control of the Type I error (α) rate so that when several contrasts are made, the desired α-level is maintained. Two methods are Bonferroni's and Scheffe's. The Bonferroni method defines the critical F value for α/k with degrees of freedom equal to 1 and df_e . For instance (see Tables 17 and 19), if $\alpha = 0.05$ and five contrasts are made, $F = 10.6$ for $\alpha/k = 0.01$, $df = 1$, and $df_e = 9$. The Scheffe method uses $F^* = (a-1)F$ where F is defined for α and degrees of freedom equal to $a-1$ and df_e . For instance (see Tables 17 and 19), if $\alpha = 0.05$ and five contrasts are made, then $F^* = 2(4.26) = 8.52$ where 4.26 is F for $\alpha = 0.05$ and degrees of freedom equal to 2 and 9.

292. As a final note, the Bonferroni method is the preferred method for a priori planned comparisons and if the number of contrasts is less than or equal to the number of means. As the number of contrasts increases and if the contrasts are mainly a matter of post hoc searches for differences among means, the Scheffe method would be preferable.

293. Before leaving the topic of contrasts, we will address a special set of contrasts, most often called orthogonal polynomials, which are particularly useful for trend analyses. These contrasts apply

when the treatments represent equally spaced quantitative levels such as time or distance. For instance, suppose the levels of the treatment-populations represent 4 months. The research question might be, "across months, what is the nature of the trend in the means?" Because there are 4 months, the trend might be linear, quadratic, cubic, or some combination of linear, quadratic, and cubic. Multipliers for polynomial contrasts for up to four treatment levels are given in Table 27.

Nonparametric Analyses

294. The motivation underlying the use of nonparametric techniques is quite simple and direct. Nonparametric (NP) methods rely generally on a set of assumptions that is less stringent than the set required by the analyses presented in the previous chapters. Basic to all previously discussed ANOVA was a set of assumptions about the underlying distribution of the treatment-populations. The restrictions were that the dependent variable (usually stated in terms of residual effects) comes from underlying continuous distributions which are normal with equal variances. At face value these constraints appear quite restrictive. Many would argue that those conditions are rarely satisfied. For instance, if there are in fact no mean-differences among the treatment-populations, then the significance level (α = Type I error rate) represents the probability of declaring mean-differences based simply on chance or the random sampling process (that is, α - percent of the time significant differences will be detected merely by chance). However, the Type I error rate statement is true if and only if the assumptions of the model are, in fact, satisfied. The question thereby becomes, "If one or more of the conditions of the model are not satisfied, are alternative methods available which maintain 'correctness' in the probability inferences?" The answer is a resounding "yes."

295. Nonparametric analyses of variance yield valid inferences about treatment-population differences yet rely upon satisfying a less stringent set of conditions. However, as is usually the case in the world, one does not get something for nothing. Depending upon the

Table 27

Multipliers for Polynomial Contrasts for up to Four Treatment Levels

Contrast	Treatment 1	2	3	4
Linear	-1	1		
Linear	-1	0	1	
quadratic	1	-2	1	
Linear	-3	-1	1	3
quadratic	1	-1	-1	1
cubic	-1	3	-3	1

degree of mildness in the constraints imposed for NP methods, the loss incurred may lie in (a) less specificity in the precise nature of the differences among the treatment-population that the methods are sensitive to or (b) less power in detecting differences which do in fact exist. Parenthetically, in the latter case, the loss in power is relative to the particular methods used but more importantly is relative to the degree of violation of the assumptions of the parametric analyses. While some NP techniques incur minimal loss in power efficiency, others maintain equal or greater power efficiency depending upon the degree of violation of the parametric assumptions.

296. The fundamental point of departure from parametric methods lies in the assumed form of the underlying treatment-population distributions. Whereas a parametric ANOVA requires normal populations, NP analyses do not. In fact, an NP ANOVA does not even require similarity of distribution form across treatment-populations. Similarly, whereas a parametric ANOVA requires the normal populations to possess equal spread (variance), an NP ANOVA does not. Hence, NP methods are often called "distribution-freer" methods. Some clarification of the term "freer" is warranted. The use of "distribution-freer" does not say that no distributional assumptions are made. Quite the contrary. The terminology simply implies that less restrictive conditions are required. For instance, NP methods share with parametric methods two assumptions--the

residual effects are continuous and distributed independently.

297. Beyond these two basic assumptions of continuity and independence, NP methods vary in their assumptions, and this variation is reflected in the precise meaning of the estimation and hypothesis testing processes. NP analyses of variance techniques are, in a strict sense, sensitive to differences between treatment-populations other than those selected by mean-separation. However, this sensitivity to differences in shape or dispersion (spread) is minimal. However, if assumptions such as equivalence of form (not necessarily normal) and/or spread are valid, NP methods test for and estimate mean-differences. For instance, some NP techniques require in addition to the two basic assumptions of continuity and independence only the assumption that treatment-population distributions are symmetrical regardless of their form or spread.

298. The mechanics of NP ANOVA are conceptually easy to comprehend. The estimation and test processes are based on transforming the original data by ranking the data. Hence, NP analyses are oftentimes referred to as "analyses of variance by ranks." The meaning of the procedures is simple. Suppose we desire to compare three treatment-populations and suppose a random sample of five experimental units is selected from each. Let us rank the observations in ascending order irrespective of the treatment-population yielding the data. As a result and under the presumption of no ties, the ranks will range from 1 to 15 units of 1. Now if no difference between treatment-populations exists, one would expect the average of the ranks from each sample to be equal. Conversely, if substantive differences exist (e.g., Population A > Population C > Population B), the means of the sample ranks should reflect the underlying inequalities. Therein lies the meaning of NP techniques.

299. Three nonparametric analyses will be illustrated: one-way ANOVA by ranks for a completely randomized design (Kruskal-Wallis test), two-way ANOVA by ranks for a randomized block design (Friedman test), and correlation (Spearman's rho).

300. The Kruskal-Wallis test for a one-way ANOVA by ranks is

based on the H statistic, calculated as

$$H = \frac{12}{N(N + 1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N + 1)$$

where

n_i = number of observations in treatment-population i

k = number of treatment-populations

$N = \sum_{i=1}^{k} n_i$ = total number of observations

R_i = sum of the ranks of the n_i observations in treatment population i

301. The calculated H statistic is compared to a chi-square (x^2) value with degrees of freedom equal to the number of treatment populations minus one (DF = k - 1). If the calculated H statistic exceeds the chi-square value, the null hypothesis can be rejected.

302. As an example of a one-way ANOVA by ranks, consider the data presented in Table 28. Twelve samples for chlorophyll *a* were taken from each of three reservoirs, and the data ranked in ascending order. In nonparametric tests, population parameters are not used in statements of hypotheses, so the hypotheses in this example are stated as

H_o:chlorophyll concentration is the same in all three reservoirs

H_A:chlorophyll concentration is not the same in all three reservoirs

303. The statistics necessary to calculate H are

$$n_1 = n_2 = n_3 = 12$$
$$k = 3$$
$$N = 36$$

and

$$R_1 = 206 \quad R_2 = 163 \quad R_3 = 297$$

The H statistic is

$$H = \frac{12}{36(36 + 1)} \left[\frac{(206)^2}{12} + \frac{(163)^2}{12} + \frac{(297)^2}{12} \right] - 3(36 + 1)$$

$$= 0.009(13,101.167) - 111.0$$

$$= 118.029 - 111.0$$

$$= 7.029$$

The critical value of x^2 with DF = 2 and $\alpha = 0.05$ is 5.991, so the null hypothesis of equal chlorophyll concentration can be rejected.

304. Just as with the parametric one-way ANOVA, the nonparametric one-way ANOVA indicates only whether significant differences exist. Rejection of the null hypothesis by the one-way NP ANOVA does not indicate which of the treatment-populations are different. Nonparametric multiple comparisons can be performed in a manner similar to the Student-Newman-Kuels test by using rank sums instead of means.

305. In order to perform the multiple comparisons, the rank sums from the Kruskal-Wallis test are arranged in increasing order of magnitude. Pairwise differences between rank sums are then computed. The standard error is calculated as

$$SE = \sqrt{\frac{n(np)(np + 1)}{12}}$$

where

n = number of observations in each treatment population

p = range of rank sums

Table 28

Data for a One-Way ANOVA by Ranks

Reservoir (i)					
1		2		3	
CHL*a*	Rank	CHL*a*	Rank	CHL*a*	Rank
59.4	16	44.7	6	82.3	29
60.6	18	60.5	17	87.4	31
65.9	22	55.0	13	67.0	23
51.1	11	64.2	21	61.1	19
63.1	20	92.1	33	73.8	26
52.6	12	69.2	24	77.8	28
59.2	15	49.8	10	88.5	32
45.1	8	56.5	14	96.4	34
83.1	30	34.8	4	23.5	1
29.2	2	46.0	9	32.0	3
77.6	27	42.2	5	114.9	36
71.4	25	44.8	7	100.8	35
n_i	12		12		12
ΣR_i	206		163		297
$\bar{R}_i$	17.17		13.58		24.75

Note that this multiple range test requires that there be equal numbers (n) of data in each of the treatment populations.

306. Using the example of chlorophyll *a* concentrations from three reservoirs, the rank sums can be ordered

Rank of rank sums (RR_i)	1	2	3
Rank sum (R_i)	163	206	297
	(Res 2)	(Res 1)	(Res 3)

and the multiple comparisons are given in Table 29. Based on the multiple comparisons it can be concluded that chlorophyll concentration in reservoir 3 is greater than that in reservoirs 1 and 2 and that

Table 29

Data for a Nonparametric Multiple Range Test

Comparison (R_j vs. R_i)	Difference ($R_i - R_j$)	p	SE	q	q_c
3 vs. 1	134	3	36.479	3.672	3.314
3 vs. 2	91	2	24.495	3.715	2.772
2 vs. 1	43	2	24.495	1.755	2.772

chlorophyll concentration is the same in reservoirs 1 and 2. These conclusions can be summarized by

$$\overline{163 \quad 206} \quad 297$$

or

$$\overline{\text{Res 2} \quad \text{Res 1}} \quad \text{Res 3}$$

307. Friedman's test is a nonparametric method that can be applied on a randomized block design. Remember that a randomized block design consists of b blocks and t treatments. To perform Friedman's test, the data within each block are ranked (i.e., values are ranked with respect to members of the given block) and then the ranks are summed for each treatment. The test statistic x_r^2 is calculated as

$$X_r^2 = \frac{12}{btn^2(nt + 1)} \left[\sum_{i=1}^{t} R_i^2 - \frac{\left(\sum_{i=1}^{t} R_i \right)^2}{t} \right]$$

where

b = number of blocks

t = number of treatments

n = number of observations per cell

R_i = rank sum for the i^{th} treatment

The calculated x_r^2 is compared to a critical x^2 value with degrees of freedom equal to the number of treatments minus one (DF = t - 1). If the calculated x_r^2 exceeds the critical x^2, the null hypothesis can be rejected.

Table 30

Data for Two-Way ANOVA by Ranks

	Month (i)					
	June (1)		July (2)		August (3)	
Station	SRP	Rank	SRP	Rank	SRP	Rank
A	7	9.0	3	8.0	1	2.5
	8	10.5	2	6.0	2	6.0
	8	10.5	1	2.5	2	6.0
	10	12.0	1	2.5	1	2.5
B	7	10.0	4	8.0	1	3.0
	7	10.0	3	7.0	1	3.0
	7	10.0	2	6.0	1	3.0
	9	12.0	1	3.0	1	3.0
C	11	11.0	4	5.5	1	2.5
	8	8.0	6	7.0	1	2.5
	9	9.0	4	5.5	1	2.5
	12	12.0	10	10.0	1	2.5
D	17	9.0	5	4.5	3	1.0
	25	12.0	4	2.5	4	2.5
	18	10.0	7	6.5	5	4.5
	21	11.0	9	8.0	7	6.5
R_i		166.0		92.5		53.5

308. An example of a two-way ANOVA by ranks for a randomized block design can be based on the data in Table 30. Soluble reactive phosphorus (SRP) concentrations were measured in four replicate samples from each of four stations sampled in June, July, and August. In this

example, stations are the blocks and months represent the treatment populations.

309. Note that the ranking process is applied within each block (station) and separately for each block. Once the ranks are assigned, the rank sums can be calculated by summing the ranks within each treatment-population (month). The rank sums are

June	$R_1 = 166.0$
July	$R_2 = 92.5$
August	$R_3 = 53.5$

Also, for this example,

$$b = 4$$
$$n = 4$$
$$t = 3$$

310. The null hypothesis to be tested is:

H_o:SRP concentration is the same for June, July and August

with the alternative hypothesis

H_A:SRP concentration is not the same for June, July and August

311. The calculated x_r^2 statistic is

$$x_r^2 = \frac{12}{(4)(3)(4)^2[(4)(3) + 1]} \left[(166.0)^2 + (92.5)^2 + (53.5)^2\right]$$

$$- \frac{(166.0 + 92.5 + 53.5)^2}{3}$$

$$x_r^2 = \frac{12}{2,496.0} \left[38,974.5 - \frac{97,344.0}{3}\right]$$

$$x_r^2 = (0.00481)(6,526.5) = 31.39$$

The critical value of x^2 with DF = 2 and $\alpha = 0.05$ is 5.991 so the null hypothesis can be rejected.

312. Nonparametric or Spearman's correlation is simple and direct. This method of correlation is useful when the bivariate data are not normally distributed. To perform a rank correlation, simply rank separately the x and y data and compute the correlation coefficient as shown for a simple linear correlation using the ranks rather than the raw data.

313. The data in Table 31 can be used for an example of rank correlation. Assume that these flow and concentration data were taken at the major inflow to a reservoir and then ranked in ascending order. To calculate the Spearman rank correlation coefficient r_s, it is first necessary to compute

$$\Sigma x_R^2 = \Sigma X_R^2 - (\Sigma X_R)^2/n$$

$$\Sigma y_R^2 = \Sigma Y_R^2 - (\Sigma Y_R)^2/n$$

$$\Sigma x_R y_R = \Sigma X_R Y_R - (\Sigma X_R)(\Sigma Y_R)/n$$

where

X_R = rank of the X value
Y_R = rank of the Y value
n = number of bivariate pairs

Table 31

Data for Rank Correlation

Flow	Rank (X_R)	Concentration	Rank (Y_R)
77.10	10	99	10
8.00	5	13	1
21.70	7	47	8
37.70	9	41	7
17.70	6	86	0
4.30	4	33	4
2.47	3	35	5
2.44	2	19	2
2.23	1	26	3
26.70	8	40	6

$\Sigma X_R = 55.0$ $\Sigma Y_R = 55.0$

$\Sigma X_R^2 = 385.0$ $\Sigma Y_R^2 = 385.0$

$\Sigma X_R Y_R = 364.0$

NOTE: $\Sigma X_R = \Sigma Y_R$ and $\Sigma X_R^2 = \Sigma Y_R^2$ only because there were no ties in either the ranks of X or the ranks of Y

In the example,

$$\Sigma x_R^2 = 385.0 - (55)^2/10 = 82.5$$

$$\Sigma y_R^2 = 385.0 - (55)^2/10 = 82.5$$

$$\Sigma x_R y_R = 364.0 - (55)(55)/10 = 61.5$$

The rank correlation coefficient is calculated as

$$r_s = \frac{\Sigma x_R y_R}{\sqrt{\Sigma x_R^2 \Sigma y_R^2}}$$

and for the example

$$r_s = \frac{61.5}{\sqrt{(82.5)(82.5)}} = 0.745$$

314. To determine the significance of the rank correlation, a t test is used in the same manner as was used for the simple linear correlation coefficient.

Multivariate Data Analysis

315. Many studies of reservoir water quality involve multiple variables, multiple samples, and/or multiple water bodies. In those situations, statistical studies will be both univariate (and involve methods described elsewhere in this manual) and multivariate. The multivariate methods described below can be used to greatly enhance the limnologist's understanding of water quality relationships in and among reservoirs. Equally important, computer programs (like SAS) with multivariate methods are available that may be used as easily as are their univariate counterparts. As with the application of all statistical methods, however, the use of multivariate methods must occur with consideration of the assumptions behind their inferential use.

Issues that can be addressed with multivariate analysis

316. The types of research questions that can be examined with multivariate statistical methods can be conveniently grouped into a relatively few categories. For the methods discussed herein, these categories are:

a. Characterization of the strength of a relationship between and/or among variables (multiple and canonical correlation).

b. Classification of variables or observations (cluster analysis).

c. Examination of structure within a system (principal component and factor analysis).

d. Development of predictive relationships for assignment to predefined groups (discriminant analysis).

Each of these tasks or functions for multivariate methods is described briefly below.

317. The "strength" of a bivariate relationship is often conveniently expressed in terms of a simple correlation coefficient. Multivariate analogs to the simple correlation coefficient exist for two situations. First, multiple correlation is used to describe the degree of relationship between a single dependent variable and a combination of two or more independent variables. This is the situation that occurs in multiple regression, and therefore the multiple correlation coefficient is a useful indicator of the goodness of fit of the multiple regression model. In truth, the multiple correlation coefficient is just a simple correlation coefficient with a new variable created that is a function of two or more original variables. The second multivariate analog for simple correlation is canonical correlation. In effect, this describes the situation with linear combinations of multiple dependent and multiple independent variables. Thus, the canonical correlation coefficient may be used to describe the strength of a relationship between a linear combination of nutrient variables and a linear combination of biomass-related variables. The linear combinations are defined by the canonical correlation procedure in order to maximize the degree of correlation between these two sets of variables.

318. The multivariate procedure called cluster analysis may be used to take "objects" and group them into categories that are based on the relative similarity of the objects as expressed in a set of prespecified variables. For example, with trophic state data (on several variables) for a number of reservoirs, the limnologist can use cluster analysis to create groups of reservoirs (based on these trophic state variables) that may then be labeled as specific trophic state categories (e.g., oligotrophic, eutrophic, etc.). Alternatively, the objects may be sampling stations within a reservoir, and cluster analysis may be used to group the stations according to similarity in sampling results. In that manner, cluster analysis may be used to identify redundant

stations if sampling effort is to be reduced.

319. It is not uncommon that data acquired on multiple variables actually represent one or only a few fundamental characteristics. For example, multiple nutrient and biomass data can all be considered to represent the single concept trophic state. The companion procedures, principal components and factor analysis, can be used to extract this simple structure (if it exists at all) from a multivariate data set. In other words, these procedures may be used to define a linear function of the variables which represents the "common element" in the data. For this example, the univariate composite that results might be called a trophic state index.

320. Cluster analysis, as noted above, is used to create groupings of observations on the basis of the similarity of observations as represented by a set of multivariate data. No groups were defined a priori. Discriminant analysis, on the other hand, is based on predefined groups of observations. With group membership established beforehand, discriminant analysis can be employed to define a linear function of independent variables that may be used to predict group membership for a new observation. For example, reservoirs could be preassigned to trophic state classes on the basis of existing biomass and water clarity data. Discriminant analysis may then be used to develop a linear model, perhaps based on nutrient loading data from these reservoirs, that can be applied to predict the trophic state of a previously unclassified reservoir (from the nutrient loading estimates).

Important assumptions for multivariate statistical inference

321. There are certain assumptions, and in a more general sense, certain conditions, that should be met or at least considered when applying multivariate methods. This requirement is not different from similar requirements for univariate statistical analysis. In fact, the specific conditions or assumptions that are important are essentially the same as those noted to be of concern in previous sections devoted to nonmultivariate statistical methods. For additional information on this topic, Tabachnick and Fidell (1983) is recommended.

322. A "condition" of a data set (that is not an assumption as such, but can affect most of the assumptions discussed below) concerns outliers or influential data points. As noted in Part III concerning descriptive statistics, observations that are far removed from the bulk of the data points (outliers) can have a major effect on the value of commonly calculated statistics, such as the mean, standard deviation, and variance. Since these statistics are often used in multivariate procedures, outliers can affect the results of multivariate analysis. One approach (Gnanadesikan 1977) is to use robust analogs of the mean and variance; however, this will affect inferential statements (e.g., significance tests) and it is not clear what adjustments need to be made to employ statistical tests with the robust statistics. A better approach may be to carefully screen the data and apply a transformation if necessary. In essence, the methods recommended in Parts II and III may be used to do this screening. This can be undertaken for each variable individually, and in most cases this will serve to identify all multivariate outliers.

323. Collectively, the key assumptions for the multivariate methods concern normality, independence of observations, constant variance (homoscedasticity), and linearity. It is important to realize that the assumptions do not hold for all methods nor do they necessarily hold for all applications of the same method. Further, it is likely that even when the assumptions are necessary, mild violations of an assumption (with the possible exception of the independence assumption) are of little consequence.

324. The assumption of normality is of concern when hypothesis tests, significance levels, or confidence intervals are determined because these procedures require normality. Although the assumption may refer to multivariate normality, it is often adequate to simply assess the normality assumption on each variable individually and apply normalizing transformations where necessary. While univariate normality does not guarantee multivariate normality, it will probably be adequate for most applications.

325. Independence of observations is an important assumption whenever statistical tests that are a function of sample size are applied. The problem occurs because lack of independence means that the effective sample size (based on the amount of nonredundant information in the data set) is less than the actual sample size. Therefore, to properly conduct statistical tests with dependent observations, an effective sample size should be calculated for use in testing procedures. Alternatively, some observations could be eliminated from the data set such that the remaining observations are independent. For example, if the data set consisted of weekly dependent observations, but it was determined that biweekly observations were independent, then one simple (but perhaps inefficient) solution is to eliminate every other observation and conduct statistical tests on biweekly data. Of all the assumptions listed above, the independence assumption is most critical on the basis of the consequences of violation.

326. The assumptions of linearity and homogeneity of variance at times are necessary for statistical tests (statistical inference) and more commonly are important as "conditions" that affect interpretation in descriptive use of multivariate methods. When one or both of these conditions is a problem, the result is that the correlation matrix (or covariance matrix) for the multiple variables does not correctly or adequately represent relationships. For example, if a relationship between two variables is nonlinear, but the multivariate analysis is run for a linear model between the variables (i.e., a linearizing transformation was not applied beforehand), then the result will not reflect the true (nonlinear) relationship. This in turn can affect the conclusions drawn by the investigator. It is a good idea, therefore, to examine the data in univariate and bivariate plots (see Part II) to check for linearity and homoscedasticity. Corrections (e.g., a linearizing or variance stabilizing transformation) made on the basis of univariate and bivariate examination of the data should usually satisfy the multivariate assumptions and conditions.

327. Characterization of relationship strength: canonical correlation. As noted above, the strength of a multivariate relationship

can be assessed using either multiple or canonical correlation. Since multiple correlation is almost always associated with multiple regression, the reader interested in multiple correlation is referred to the section on regression analysis. The discussion in this section focuses on canonical correlation; useful references on this topic include Tabachnick and Fidell (1983) and Green (1978).

328. Canonical correlation is used to identify and estimate a linear function (called a canonical variate) of one set of variables that is maximally correlated with a linear function of a second set of variables. Additional canonical variates, which are uncorrelated with the first set, may also be identified. The procedure results in information that is primarily descriptive in nature, and thus it has been used less frequently than have other multivariate methods that facilitate hypothesis testing and/or prediction.

329. Generally of interest to those who apply canonical correlation is the extent of relationship between two set of descriptors (or variables) for objects (e.g., reservoirs) under study. For example, one may be interested in the relationship (if any exists) between reservoir water chemistry and cell counts for dominant algal species to see if certain conditions (e.g., low inorganic nitrogen concentration) favor (covary with) blue-green dominance. The canonical correlation between a set of water chemistry variables (nutrients, etc.) and a set of variables indicating cell counts for major algal species in a multireservoir study could be quite helpful.

330. Canonical correlation may also be used to see how many "common elements" are contained within two sets of variables. For example, canonical correlation may be applied to a set of reservoir water chemistry variables and a set of reservoir geomorphology and hydrology variables. From this analysis, the first canonical variate might represent trophic state as determined from the canonical weights or coefficients, which in that case might be highest for nutrients in the first set of variables and for depth in the second set of variables. In addition, the second canonical variate could represent reservoirs located primarily in the southern United States with the largest

canonical weights on variables such as conductivity, alkalinity, and reservoir volume.

331. When two or more pairs of canonical variates are identified, the investigator can express the relative importance of each canonical variate pair on the basis of the percent overlapping variance (equal to the squared canonical correlation coefficient) between the two sets of original variables. For the hypothetical example discussed above, it may be found that the first canonical variate pair represents 60 percent overlapping variance and the second canonical variate pair has 15 percent overlap. This information helps the investigator understand the extent of commonality within a set of variables. In addition, one can calculate the percent variance explained by a canonical variate within each of the two sets of original variables. For example, in the previous hypothetical example, the investigator may find that the first canonical variate explains 70 percent of the variance in the nutrient variables and 30 percent of the variance in the hydrology-geomorphology variables.

332. When canonical correlation is used in hypothesis testing or to justify statements of statistical significance, an assumption of multivariate normality is necessary. As noted in the previous section, this is often adequately satisfied by creating data distributions that are approximately univariate normal. Most applications of canonical correlation are descriptive; in that case, multivariate normality is desirable but not necessary. Descriptive applications can yield misleading information, however, if the data distributions are highly skewed or exhibit outliers. It is wise, therefore, to use transformations if necessary to create roughly symmetric univariate data distributions, and to carefully examine the validity of any outlying data points. If outliers cannot be removed from the data set on the basis of substantive reasons, then it is recommended that two analyses be run--one with the outliers and one with the outliers excluded. Both analyses could be reported so the reader could directly relate inclusion/exclusion of the outliers to the canonical correlations.

333. Finally, canonical correlation depends upon linear relationships at two points in the analysis. First, the variables within each of the two groups of data are combined in linear canonical variates. Therefore, the data should be transformed if necessary so that a linear model is appropriate. Second, since the analysis is based on the simple (product-moment) correlation coefficient, bivariate correlation coefficients should reflect the actual linear relationships in the data. Again, transformations may be necessary to achieve this. For example, if the relationship between Secchi disk depth and all other nutrient-related variables exhibits a hyperbolic pattern, and the inverse of Secchi disk depth straightens (linearizes) the relationship, then the inverse of Secchi disk depth should be used in the canonical correlation analysis.

334. Canonical correlation - example. This example and some of the other examples presented in this section on multivariate analysis are based on a data set from Walker (1981). The data, presented in Table 32, concern water chemistry in 43 Corps of Engineers reservoirs.

335. The data set consists of seven variables. Three of the variables (pH, alkalinity, and conductivity) relate to acidity and dissolved salts. Four of the variables (total phosphorus, total nitrogen, Secchi disk depth, and chlorophyll *a*) are related to trophic state. For most of the studies, all variables but pH are log-transformed to create distributions that are closer to univariate normal than are the distributions of the untransformed variables.

336. Given the uses of canonical correlation and the composition of the sample data set, it seems appropriate to use canonical correlation to examine the relationship between a linear function of the acidity-salinity variables and a linear function of the trophic state variables. This was done using the SAS CANCORR procedure. Some of the major features of the SAS output are summarized below.

337. For this example, the PROC CANCORR statement simply identified the data set and labeled one set of variables (the acidity-salinity variables) as "VAR" and the other set (the trophic state variables) as "WITH" as required in the procedure. Table 33 presents a portion of the

Table 32

Multivariate Reservoir Data

Reservoir	pH	Conductivity µmhos/cm	Alkalinity mg/ℓ	Total Phosphorus µg/ℓ	Total Nitrogen µg/ℓ	Secchi Disk Depth m	Chlorophyll α µg/ℓ
1	7.11	74	11	10.7	941	3.49	4.9
2	8.06	273	105	96.5	1,724	1.20	18.0
3	7.48	503	33	50.0	677	1.09	8.1
4	6.74	795	20	24.2	1,152	1.23	2.6
5	6.69	816	25	40.1	1,409	1.20	4.0
6	8.48	652	294	277.0	1,739	0.75	41.3
7	7.92	422	79	58.8	1,428	0.86	14.6
8	8.00	208	53	59.8	1,042	0.93	26.8
9	7.57	138	23	20.9	735	2.18	3.7
10	8.25	294	48	41.3	945	0.96	16.4
11	7.56	559	72	167.2	2,887	0.28	10.9
12	8.25	397	134	127.0	1,765	0.44	67.1
13	8.11	493	158	102.5	3,106	0.40	10.8
14	7.97	429	134	186.6	3,204	0.46	26.1
15	8.24	323	128	40.4	950	1.10	22.8
16	7.40	1,350	42	10.6	523	2.24	5.6
17	7.10	307	10	12.8	858	3.53	6.2
18	8.02	1,849	178	103.9	3,070	0.67	15.8
19	7.37	572	30	31.1	763	1.64	7.2
20	8.07	835	81	44.2	838	1.19	8.4
21	7.67	644	73	124.8	750	0.69	13.3
22	7.56	626	38	15.6	477	1.72	3.8
23	8.04	1,001	119	45.1	569	1.77	10.0
24	7.89	573	52	57.8	629	0.76	8.9
25	7.65	815	62	10.2	443	4.32	3.6
26	8.03	1,512	133	84.4	2,091	0.56	17.4
27	8.04	1,676	164	69.3	4,352	0.98	17.1
28	7.68	1,136	63	71.1	1,211	0.72	23.5
29	7.64	116	47	30.7	547	2.14	3.9
30	8.34	198	127	15.9	527	3.96	4.0
31	8.22	203	98	29.4	636	2.27	9.1
32	8.16	392	176	221.1	1,858	0.19	9.5
33	7.81	399	81	90.0	854	0.45	4.4
34	8.08	1,520	129	211.7	1,287	0.37	27.0
35	8.09	163	75	37.7	837	1.64	6.6
36	6.37	61	22	96.9	686	0.58	5.1
37	8.17	342	109	131.0	854	0.63	17.4
38	8.06	1,345	156	68.1	1,193	0.34	16.0
39	8.36	785	203	94.5	1,169	0.85	18.9
40	8.20	351	144	45.2	1,526	0.47	8.4
41	8.41	528	201	125.3	1,033	0.61	27.8
42	8.17	675	175	72.9	460	2.32	7.4
43	7.21	23	17	13.0	247	2.49	2.4

SAS output: the canonical correlations, the approximate standard error, the F statistic, the Pr > F.

338. Since there are three variables in the smaller of our two groups (the VAR group), three canonical correlations are estimated (rows 1, 2, and 3 in the upper entry of Table 33). Each canonical variable is uncorrelated with all other canonical variables except its corresponding canonical variate from the opposite data set. The first pair of canonical variables (represented by row 1) is constructed so that it maximizes the correlation between a linear combination of the VAR variables with a linear combination of the WITH variables. The second pair of canonical variables is also constructed to maximize this correlation, except that it must also be uncorrelated with the first pair of canonical variates. This continues until all canonical variates are estimated.

339. In Table 33, the canonical correlations are given for the three pairs of canonical variates. Note that the first two correlations are reasonably high. Recall that the interpretation of the canonical correlation coefficient squared is the percent overlapping variance. Thus, for the first pair of canonical variates, there is about 54 percent overlapping variance (0.733^2). This means that 54 percent of the variance in the first VAR canonical variate is explained by the first WITH canonical variate. If these canonical variables are in turn highly correlated with the original variables (this is examined below), then the first canonical variate may describe a high level of commonality between the two sets of variables.

340. The approximate standard errors for the first two canonical correlations are relatively small in comparison to the magnitude of these correlations. This suggests that the correlations may be significantly different from zero. Confirmation of this point is given in the F statistic and the probability level for the F statistic which shows significance at >0.01 level. In these tests, the F statistic is determined for the canonical correlation in its row as well as for all lower canonical correlations simultaneously. Thus, the F statistic in row 2 represents the simultaneous tests of canonical correlation

Table 33

Canonical Correlation Analysis

Canonical Variate	Canonical Correlation	Approximate Standard Error	F Statistic	Degrees of Freedom	Pr > F
1	0.733	0.071	5.336	12	0.000
2	0.618	0.095	4.202	6	0.001
3	0.317	0.139	2.118	2	0.134

Multivariate Test Statistics and F Approximations

Statistic	Value	F	Degrees of Freedom	Pr > F
Wilks' lamda	0.237	5.767	12	< 0.001
Pillai's trace	1.100	5.502	12	< 0.001
Hottelling-Lawley trace	1.978	5.715	12	< 0.001
Roy's greatest root	1.161	11.034	4	< 0.001

coefficients two and three. The interpretation of these tests is that only the first two canonical correlation coefficients are significant at the 0.01 level.

341. The next grouping of the output in Table 33 presents four statistics used to evaluate the significance of the canonical correlations as a set. While each statistic is slightly different from the others, they all test this feature of overall significance. Pillai's trace may be the most robust of the four (Tabachnick and Fidell 1983), but since Wilks' lambda is by far the most commonly reported of the four, it is the one we discuss here. Since tables for the four statistics are uncommon, each statistic is evaluated for significance using an F statistic approximation (F tables are far more common).

342. Wilks' lambda is defined as

$$\Pi\left[1 - (CC)^2\right]$$

where

Π = product operator

CC = canonical correlation

343. Since Wilks' lambda is the product of one minus the canonical correlations squared, it will be low when the canonical correlations are high. Based on the F approximation, Wilks' lambda indicates a set of highly significant canonical correlations for this sample data set.

344. Table 34 presents raw and standardized coefficients for the canonical variables. The raw canonical coefficients are used directly with the original (mostly log-transformed) variables, whereas the standardized canonical coefficients were estimated for variables that were normalized (i.e., subtract the mean and divide by the standard deviation). These coefficients are used to write the linear canonical variates for each set of variables. For example, using the raw canonical coefficients in Table 34, the first canonical variates are:

$$V1 = -1.478(\mathrm{pH}) + 4.402\ [\log(\mathrm{ALK})] - 0.268\ [\log(\mathrm{COND})]$$

$$W1 = 2.822\ [\log(\mathrm{TP})] + 0.546\ [\log(\mathrm{TN})] + 0.706\ [\log(\mathrm{SECCHI})] + 0.030\ [\log(\mathrm{CHL}a)]$$

Table 34

Canonical Coefficients

	Canonical Variable				Canonical Variable		
Variable	V1	V2	V3	Variable	W1	W2	W3
			Raw Canonical Coefficients				
pH	-1.478	3.846	-0.225	Log(TP)	2.822	-2.250	-3.427
Log(ALK)	4.402	-3.947	-1.318	Log(TN)	0.546	-0.986	4.015
Log(COND)	-0.268	1.003	2.769	Log(SECCHI)	0.706	0.180	-1.956
				Log (CHLα)	0.030	4.313	0.239
			Standardized Canonical Coefficients				
pH	-0.719	1.871	-0.110	Log(TP)	1.088	-0.867	-1.321
Log(ALK)	1.599	-1.433	-0.479	Log(TN)	0.146	-0.264	-1.076
Log(COND)	-0.108	0.404	1.117	Log(SECCHI)	0.232	0.059	-0.643
				Log(CHLα)	0.010	1.463	0.081

345. For interpretation purposes, however, the standardized canonical coefficients are often most useful as they provide a relative measure of the importance of each variable in determining the canonical variates. For this example, we can see that alkalinity (log(ALK), with standardized canonical coefficient = 1.599) is the most important determinant of V1, followed by pH (-0.719); conductivity (log(COND), with -0.108) is least important. For W1, phosphorus concentration (log(TP), with 1.088) is by far the most important variable. Thus, the first canonical variable is, to a great extent, describing a relationship between alkalinity and phosphorus that is not shared with the other variables (except perhaps pH).

346. Table 35 presents correlations between the original seven variables and each of the canonical variables. Note that, except for sign differences, the correlations in Table 35 often (but not always) convey the same information as do the standardized canonical coefficients in Table 34. Their relationship is somewhat analogous to the relationship between simple correlation coefficients and multiple regression coefficients.

Table 35

Correlations Between the Original Variables and Their Canonical Variables

	Canonical Variable		
Variable	V1	V2	V3
pH	0.616	0.758	-0.215
Log(ALK)	0.934	0.356	-0.039
Log(COND)	0.461	0.227	0.858
	Canonical Variable		
	W1	W2	W3
Log(TP)	0.990	-0.035	-0.039
Log(TN)	0.685	-0.009	0.711
Log(SECCHI)	-0.793	0.139	-0.270
Log(CHL*a*)	0.742	0.656	0.116

347. Another interesting statistic is the proportion of variance (pv) in the original variables that is explained by the corresponding canonical variate. This is easily calculated by summing the appropriate squared correlation coefficients from Table 35 and dividing by the number of original variables. Thus:

$$pv = \sum_{i-1}^{K} (r_{icv})^2/(K)$$

where

r_{icv} = correlation between original variable i and canonical variable cv

K = number of original variables

348. Using the correlations in Table 35, the proportion of the variance in the acidity-salinity variables that is explained by V1 is:

$$pv = [(0.6155)^2 + (0.9335)^2 + 0.4605)^2]/3$$
$$pv = 0.487$$

349. Thus about 49 percent of the variance in the acidity-salinity variables is explained by the first canonical variate (V1).

350. It is also interesting to determine what proportion of the variance in one set of original variables is explained by the other canonical variate. This is called "redundance" (rd), and we can calculate it from:

$$rd = (pv)(CC)^2$$

where pv is the proportion of variance (determined above) and CC is the canonical correlation coefficient. For example, the redundancy for W1 and the acidity-salinity variables is:

$$rd = (0.4897)(0.733)^2$$
$$rd = 0.262$$

351. This means that about 26 percent of the variance in the acidity-salinity variables is explained by the opposite canonical variable (W1). Thus, W1 is a relatively poor predictor of the acidity-salinity variables. The redundancy for the trophic state variables and canonical variate V1 is 0.353, indicating that V1 is slightly better as a predictor of "opposite" original variables (than is W1). These low redundancies are not surprising, confirming our original belief that there is not a strong relationship between trophic state and acidity-salinity.

352. Cluster analysis. Cluster analysis is a classification method that may be used to group or identify similar objects. These objects may be reservoirs, sampling sites (or dates) within a reservoir, or water quality variables (e.g., nitrogen, chlorophyll, alkalinity, etc.) measured in one or more water bodies. The criterion of similarity may be defined in several ways; in most applications, however, it is based on either the correlation coefficient or the Euclidean distance (which is a function of the sum of squared differences between attributes). Before discussing clustering criteria, though, let us first consider the types of problems that might be fruitfully studied using cluster analysis. For additional information on cluster analysis, Davis (1973) or Green (1978) may be consulted.

353. In summarizing water quality studies among reservoirs, it is often informative to classify the water bodies according to various criteria. Using cluster analysis, the investigator could classify the reservoirs according to their similarities on any group of variables desired. For example, trophic state classification would occur if the analysis is confined to accepted trophic state variables. Or, reservoir similarities in general could be identified when all water quality variables are included in the cluster analysis.

354. Within a single water body, an ongoing sampling program may be designed to gather data at specific points in space and time. Given limited resources for sampling, the scientist may want to know something about the redundancy in the sampling program. In specific terms, he may want to know how much information is lost if one or more sampling

stations or dates are discontinued. When a cluster analysis is performed on the sampling stations and/or dates (in consideration of serial or spatial correlation), the analyst obtains a measure of similarity (or redundancy) among stations/dates. In conjunction with other statistical analyses, the cluster analysis may be used to aid the decision on sampling effort reduction.

355. We might be interested in the covariation or similarity among variables in a data set containing measurements on several variables. These data could be taken within one or more water bodies. For example, we might want to know if quantitative information on cultural activities in the watersheds of reservoirs is related to (covaries with) any of the measured water quality variables. Alternatively, we might want a single statistical "picture" of the similarities among the variables that we have measured. Cluster analysis, with an accompanying dendogram for display, can be used to do these analyses.

356. Three methods of clustering are available for the investigator using the latest version of SAS. Since these algorithms are typical of those that are used in other statistical computing packages as well, we will focus our discussion on these three methods.

357. Clustering of cases using SAS is accomplished in a hierarchical manner. At the beginning, all cases are assumed to belong to separate clusters, and at each step cases (or groups of cases) are clustered together according to one of the three clustering criteria. The three clustering criteria, or methods, are the centroid method, Ward's method, and average linkage according to squared Euclidean distances.

358. Centroid cluster analysis is based on the distance, or similarity, between the centroid (or mean) of each cluster. (Remember that at the start of the cluster analysis, each case is considered a separate cluster. By case, of course, we mean one row in the data matrix, which could be one lake, or one sampling station, if the study involves a cross section of lakes or of sampling stations.) According to this criterion, at each step the cases or clusters separated by the

smallest Euclidean distance are joined together. The Euclidean distance is defined as:

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{0.5}$$

for the distance between points i and j for the k^{th} variable. Euclidean distance may be calculated for variables in the original metric or for standardized variables; standardization is discussed later in this section. Note that correlation between variables is not considered in Euclidean distance.

359. Ward's method is based on the within-clusters sum-of-squared deviations (from the cluster mean). At each step the union of all pairs of clusters is considered. For the candidate cluster being considered, the sum of squared deviations of the cases from the cluster mean is (calculated here for the k^{th} variable):

$$SSD = \sum_{i=1}^{n} (x_i - x)^2$$

360. At each step, the new cluster formed is the candidate cluster that has the smallest sum-of-squared deviations.

361. In the average linkage or group average method, distance is defined as the average squared Euclidean distance between pairs of observations in a cluster. At each step, this distance is determined for a pair of cases consisting of one pair from each cluster. The two clusters that are joined together are those with a minimum value for this distance measure.

362. Among the three hierarchical clustering methods available with SAS, Ward's method and the average linkage method have been found to be two of the best approaches. The centroid method might be favored in certain situations because it is more robust (than the other two methods) to outliers (since it is based on the cluster mean, and not on

a single case). It is also to be noted that Ward's method has a tendency to result in clusters with roughly equal numbers of cases, while the average linkage method tends to yield clusters with approximately the same variance. It is not possible to unambiguously rank the three methods and identify the best clustering criterion for all situations. Therefore, it is recommended that the investigator choose one of the three methods to use on most problems and become familiar (through experience) with that method. For problems where the final clustering of cases has important implications, the investigator would be wise to apply all three methods and select clusters on the basis of the combined results.

363. Prior to beginning a cluster analysis, the investigator must decide if the analysis is to be undertaken with standardized variables. To standardize a data set, each observation is replaced by the deviation of the observation from the mean, divided by the standard deviation. Use of standardized variables has the effect of removing the influence of the units of measurement from the results of the analysis. For example, using unstandardized variables, the variance for a data set measured in micrograms per litre will be much higher (10^6 higher) than the variance for the same data expressed in milligrams per litre. Thus, if one variable in a multivariate data set is expressed in micrograms per litre and all other variables are expressed in milligrams per litre, the former variable is going to excessively dominate the variance-covariance matrix. This in turn will affect the results of all procedures, like cluster analysis, that are based on the variance-covariance matrix. Alternatively, if variables are standardized prior to analysis, the results of multivariate analyses are effectively based on the correlation matrix. The standardized variables are unitless, so any linear change in units will not affect the results.

364. While in most cases standardization is to be recommended, there are situations where standardization will adversely affect an analysis. For example, within-cluster differences can be reduced by standardization. In those situations, if the within-cluster variance were known, standardization should be based on that term. Thus, it is

recommended that unstandardized variables be used if all variances are of approximately the same magnitude; otherwise, the variables should be standardized prior to analysis.

365. Although some investigators have attempted to use formal statistical tests on the results of a cluster analysis, this practice is in general not recommended. The major difficulty is that the data would be asked to both form the clusters and test their significance. Ideally one would want the data groupings specified beforehand; otherwise, the degrees of freedom for statistical testing are affected by using the data to specify the hypothesis (the clusters).

366. If no formal statistical inference is to be undertaken with the results of a cluster analysis, then there are no formal assumptions to be invoked prior to the analysis. However, it is good practice (as noted in the introductory section) to work with data that are symmetrically distributed with no obvious outliers. Transformations (selected from inspection of univariate and bivariate plots) to achieve this symmetry are recommended for use as needed.

367. Cluster analysis - example. This example, like the previous one, is based on the data set from Walker (1981) representing water chemistry in 43 Corps of Engineers reservoirs. Here, using the SAS CLUSTER procedure, we group the lakes in clusters on the basis of similarity in log total phosphorus concentration, log total nitrogen concentration, log Secchi disk depth, and log chlorophyll *a* concentration. Ward's method was used for clustering.

368. Table 36 contains some of the optional output from the SAS procedure. The biomodality statistic indicates the possibility of two or more distinct clusters in the frequency distribution; a value of 0.555 or greater is evidence of this situation. None of the biomodality statistics in Table 36 is that large, suggesting that there is more likely a continuum of change in the variables as opposed to abrupt changes. The eigenvalue statistics, particularly the proportion (of variance "explained" by each eigenvector), indicate the dimensionality of the data. When a high proportion of the variance is represented by one eigenvalue, this means that there is substantial correlation among

Table 36

Cluster Analysis

Variable	Bimodality
Log(TP)	0.421
Log(TN)	0.346
Log(Secchi)	0.382
Log(CHL a)	0.382

Eigenvalues of the Correlation Matrix

Eigenvalue	Proportion of Variance	Cumulative Proportion of Variance
3.004	0.751	0.751
0.456	0.114	0.865
0.422	0.106	0.971
0.118	0.029	1.000

Number of Clusters	Cubic Clustering Criterion
1	0.000
2	-0.468
3	-1.548
4	-1.420
5	-1.710
6	-1.523
7	-1.500
8	-1.457
9	-1.574
10	-1.577

the variables. For this example, the first eigenvalue represents 75 percent of the variance in the data. This is high, and it suggests, not surprisingly, that the four variables are correlated. For our purposes in this cluster analysis, it means that the clusters created will primarily represent the characteristic that all four variables have in common, which might be called "trophic state."

369. The last section of Table 36 presents the cubic clustering criterion (CCC). This is probably the best available indicator of the number of clusters that represent the groupings within the data set. A local peak value of CCC identifies a "number of clusters" that may define an acceptable grouping of the data. For our example, notice that there are local peaks for 2, 4, and 8 clusters. On the basis of CCC, these could be the final candidates for appropriate data groups on the basis of the four variables.

370. To aid in the selection of clusters from among the identified final candidates, a tree diagram should be developed. The tree diagram, or dendrogram, is shown in Figure 23 for our example. Similar observations or clusters (as defined by CCC in this case) are joined first (near the bottom). The higher the point of joining, the less similar are the members of a cluster. SAS output not shown in this manual presents various similarity or distance measures that are recomputed each time a new cluster is formed in this hierarchical procedure.

371. Visual inspection of Figure 23 clearly shows the two- and four-cluster options, but only through scrutiny is the eight-cluster option identified. Further, since the clustering variables all represent trophic state, we might select the four-cluster option as best representing these data (producing four trophic states). To see this, we look at the range of values in the four clusters for some of the trophic state variables. Numbering the clusters 1 through 4 on the basis of left-to-right position on Figure 23:

a. Total phosphorus concentration (mg/m^3)

Cluster 1:	10.2-30.7
Cluster 2:	24.2-72.9
Cluster 3:	40.4-131.0
Cluster 4:	69.3-277.0

```
  NUMBER
    OF
 CLUSTERS                           RESERVOIR NUMBER

               1 2 3   2 1 2 4   2 2 1 3 2 3 4           2 3 4 1 1 2 3 3 3 3 4   3 1 1 1 3 1 1 2 2
             1 7 5 0 9 9 6 2 3 3 0 4 9 5 3 1 2 4 5 2 7 8 8 9 1 0 5 1 7 3 6 8 0 6 4 2 1 3 2 4 8 6 7

    1        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

    2        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

    3        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXX

    4        XXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXX

    5        XXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXX XXXXXXXXXXXXX

    6        XXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXX XXXXXXX XXXXX XXXXXXXXXXXXX

    7        XXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXX XXX XXXXXXXXXXXXXXXXXXX XXXXXXX XXXXX XXXXXXXXXXXXX

    8        XXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXX XXX XXXXXXXXXXXXXXXXXXX XXXXXXX XXXXX XXXXX XXXXXXX

    9        XXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXX XXX XXXXXXXXXXXXXXXXXXX XXX XXX XXXXX XXXXX XXXXXXX

   10        XXXXXXXXXXXXXXX X XXXXXXXXXXXXXXX XXX XXXXXXXXXXXXXXXXXXX XXX XXX XXXXX XXXXX XXXXXXX
```

Figure 23. Dendrogram for the cluster analysis example

b. Chlorophyll *a* concentration (mg/m^3)

Cluster 1:	2.4-6.2
Cluster 2:	2.6-10.0
Cluster 3:	4.4-27.8
Cluster 4:	9.5-67.1

372. While there is some overlap among the clusters, it should be apparent from the phosphorus and chlorophyll levels that we have succeeded in grouping the lakes into four trophic states. It should be realized that the overlap is less serious when all variables are considered simultaneously (e.g., some lakes with low chlorophyll *a* also have low Secchi disk depths due to nonalgal turbidity); this, of course, is a reason for use of a multivariate (as opposed to univariate) analysis.

373. The user of SAS CLUSTER will note that several optional measures of similarity or distance may be computed. These statistics are beyond the scope of this manual; however, information on these measures may be obtained either from the references on multivariate analysis identified above, from Everitt (1974), or from Hand (1981).

374. Examination of structure: principal components and factor analysis. Principal components analysis (PCA) and factor analysis (FA) are used to create a relatively small number of new variables (called "factors") from a larger number of original variables. With PCA, these factors are estimated as simple linear functions of the original variables; each factor is orthogonal (at right angles in a graphical sense) to all other factors. The practical use of these factors may be based on the belief that the observed variables in fact represent only a small number of underlying characteristics, with the relationships or commonality among the observed variables expressed in the covariance or correlation matrix. Thus, by using the observed variable correlations to create a few common factors, the investigator may, for example, increase his understanding of the underlying structural relationships among reservoir water quality variables.

375. One problem that PCA is particularly well suited to address is the estimation of a trophic state index. It is well accepted that

trophic state (a subjective concept) is a function of nutrient concentrations, biomass levels, water clarity, etc. Since there is no universally accepted, objective way to create a trophic state index from these variables, it may be reasonable to use a mathematical procedure (PCA) to extract the common element from these variables and call this common element a measure of trophic state. PCA can thus be used to define a linear function of the trophic state-related original variables. This function will be the linear function that "explains" a maximum (i.e., more than any other linear function) of the variance contained in the original data set. Since this "first principal component" describes (through the linear function) the common element in the trophic state data, it is reasonable to assume that the first principal component is a good trophic state index. Further, since the first principal component maximizes the explained variance, it is the best trophic state index defined in this manner.

376. Another advantage of the use of PCA to define a trophic state index is that the principal component is also the linear function of the original variables that creates a maximum spread among the observations. Thus, if the investigator wants to distinguish reservoirs on a trophic state basis, the first principal component is the linear function that separates the cases (e.g., reservoirs) as much as possible. This facilitates the ranking of reservoirs according to trophic state.

377. PCA and FA can also be used to define the number and type of underlying factors within a data set. For example, a large cross-sectional reservoir water quality data set may contain data on phosphorus, nitrogen, chlorophyll, Secchi depth, alkalinity, pH, conductivity, calcium, magnesium, sodium, aluminum, chloride, sulfate, and silicon. It is likely that this multivariable, or multidimensional, data set actually reflects a much smaller number of true structural factors or dimensions. For example, the trophic state variables may all reflect one underlying dimension: trophic state. Correspondingly, many of the cations and anions may covary (have high bivariate correlation coefficients), and thus represent another underlying dimension. These two

dimensions, or factors, could be estimated as linear functions of the original variables using PCA and FA.

378. It may be of interest to the investigator to compare data sets from two groups of reservoirs (e.g., those in the southeastern United States versus those in the Southwest) to see if the underlying structural relationships among variables are different. This could be done using PCA and FA. For each separate data set, the factors could be estimated. Then the comparison between the two regions would proceed through a comparison of the functional forms of the factors. Specifically, the investigator asks: are the individual factors (from each of the separate data sets) composed of the same combinations of variables with approximately the same weights (coefficients)? This question could be answered in an informal manner through simple inspection of the factors, or it could be answered with a formal hypothesis test.

379. Even though PCA and FA can be used to formally test hypotheses, most applications of these methods are exploratory. In fact, confirmatory FA involves enough additional complexity that it will be ignored in this presentation. Thus, it is assumed here that all applications of PCA and FA are exploratory.

380. While the applications of PCA and FA considered herein are strictly exploratory, it is still a good idea to work with data distributions that are "well behaved." This means that, if possible, data distributions should be approximately symmetric without outlying or influential points. Ideally, data for a single variable should be approximately normal, and data for any pair of variables should be bivariate normal. While this is not strictly required, approximate normality will lead, in the long run, to inferences that correctly represent the data. Of course, transformation should be considered if it is determined that it will result in a more desirable distribution of data.

381. PCA and FA may be conducted from either the correlation matrix or the covariance matrix. As noted above when this option was discussed for other procedures, there are problems of scale with the covariance matrix. Specifically, the magnitude of the covariance will

change as the units of measurement change. Thus, unless the investigator is working with variables that all have approximately the same magnitude, it is recommended that PCA and FA be based on the correlation matrix.

382. The analysis usually begins with PCA. Principal components analysis is used to reexpress the original variables in a set of orthogonal factors. Each factor is a linear function of one or more of the original variables. PCA is a mathematical procedure; it is used to maximize the variance in the original data explained by each factor. Thus, the first principal component is the linear function that explains the maximum variance in the original data. The second principal component is orthogonal to the first component, while explaining the maximum of the remaining variance unexplained by the first component. This continues until all the variance in the original data is explained by the orthogonal components.

383. Principal components analysis is usually conducted to reduce the dimensionality in a data set, or in other words to reexpress the information contained in several variables into a smaller number of factors or components. Thus, it is common to retain only a few factors, as PCA is effective only if much of the original variance is explained in a relatively few factors. In addition, PCA is generally effective only if each component has a substantive (i.e., limnological) interpretation. In one of the examples mentioned above, for example, it was desired that one of the components have a trophic state interpretation and the other component represent cations and anions. Sometimes an obvious substantive interpretation of PCA occurs; when it does not, however, factor analysis can be used to redefine the factors slightly so that interpretation is enhanced.

384. The interpretation of the factors is based on the composition of the factors. The factors are linear functions of one or more of the original variables; thus, the relative contribution of these original variables to each factor is the basis for interpretation. This contribution is measured by the coefficient, or weight, for each variable in the linear function. For example, in the example mentioned

above, if the first principal component is composed of most of the original variables, but only those related to trophic state have high coefficients, it is reasonable to interpret this component as having a trophic state interpretation. A cation-anion interpretation for the second component would be appropriate if the second component was weighted heavily on the original cation-anion variables.

385. It is possible that PCA will not result in easily interpretable components, particularly since PCA is an optimizing routine that is ignorant of any need for interpretation. Factor analysis may then be used as it allows the investigator to reorient the components (now called factors) so that interpretation is facilitated. In a graphical sense, the components form orthogonal axes (axes at right angles to each other). Factor analysis involves rotation of these axes, and this changes the relative weight each of the original variables has for a particular factor. Thus, factor analysis is performed to create a new set of factors (axes) for which a logical grouping of original variables has the highest weights. This will then allow a substantive interpretation for the new factors.

386. The rotation or creation of new axes in factor analysis does not have to involve strictly orthogonal axes. Oblique, or nonorthogonal, rotation can be undertaken; this results in factors that are correlated (orthogonal PCA and FA result in uncorrelated factors). Oblique rotation is considered if it is believed that the underlying structure involves factors that are correlated. The assumption, of course, in orthogonal PCA and FA is that the underlying structure involves strictly uncorrelated factors.

387. If the interpretation resulting from PCA is unsatisfactory, and axis rotation using FA seems necessary, then it is recommended that the investigator try several of the available rotation algorithms available through programs like SAS. Experience with factor analysis rotation for a particular type of data set is clearly an asset. The example presented below may help to explain concepts that are still confusing.

388. Principal components and factor analysis - example. The SAS procedure, FACTOR, was used on the seven-variable Walker data set. One

of the attractive features of this program is that the SAS documentation is relatively good; this is fortunate because the great number of available options means, in effect, that many of them cannot be dealt with in this example. The procedures that are illustrated below, however, are the most commonly used methods.

389. In this example, we begin with PCA and then apply varimax rotation. Principal components analysis is the most commonly used method for creation of the initial orthogonal components or factors from the original variables. On occasion, one of the factor extraction procedures available is SAS, such as principal factors (PF) extraction, might be used. In brief, PCA works with all of the variance in the original data, both common variance and unique variance. PF, on the other hand, works only with common variance. Thus, to obtain a general summary of the data, PCA is the preferred choice. Tabachnick and Fidell (1983) provide a clear explanation of the differences between the initial component and factor extraction options.

390. All seven variables in the Walker data set are log-transformed and used in this analysis. The variables are: pH, conductivity, alkalinity, phosphorus concentration, nitrogen concentration, Secchi disk depth, and chlorophyll *a*.

391. Table 37 contains a summary of the PCA. For PCA, the communality referred to in Table 37 is always one; for PF, the communality will lie between zero and one, and represents the common variance among variables. The eigenvalues in Table 37 indicate the amount of variance in the original seven variables that is explained, or represented, by each of the orthogonal (perpendicular) components. For this example, we see that the first principal component explains 59.82 percent of the variance, and the first three components explain 86.99 percent of the variance. This suggests that the variability in the original seven variables might be reasonably summarized in perhaps three orthogonal components.

Table 37

Principal Components and Factor Analysis: Eigenvalues

	Factor						
	1	2	3	4	5	6	7
Eigenvalue	4.188	1.098	0.804	0.405	0.318	0.136	0.052
Proportion of variance	0.598	0.157	0.115	0.058	0.045	0.019	0.007
Cumulative proportion of variance	0.598	0.755	0.870	0.928	0.973	0.993	1.000

392. The factor pattern matrix in Table 38 is a matrix of correlation coefficients between the original seven variables and the seven factors. For this example, note that the correlations between factor 1 and six of the seven variables are relatively high. This means that the first factor (the first principal component) is a good summary descriptor of essentially all of the variables (particularly phosphorus, alkalinity, and chlorophyll). If it was our objective to create a single (linear) index value for each reservoir summarizing the measurements on these variables within that reservoir, then the first factor in Table 38 is a good choice. We should then produce a matrix of standardized factor scoring coefficients. These are analogous to standardized regression coefficients. For each reservoir, to calculate the score on factor 1, the scoring coefficient for each variable is multiplied times the standardized (subtract the mean and divide by the standard deviation) value of that variable; these terms are then summed. Thus, for a particular reservoir:

$$PCA_1 = \Sigma b_i z_i$$

where

PCA_1 = calculated value for the first principal component

b_i = factor scoring coefficient for variable i

z_i = standardized value for variable i

Table 38

Principal Components and Factor Analysis: Factor Pattern Matrix

	Factor						
Variable	1	2	3	4	5	6	7
pH	0.693	0.675	-0.136	0.048	-0.110	-0.133	0.119
Log(COND)	0.553	0.017	0.815	-0.125	0.111	-0.015	0.033
Log(ALK)	0.859	0.411	0.007	-0.146	-0.195	0.115	-0.144
Log(TP)	0.878	-0.255	-0.242	-0.199	0.084	0.220	0.104
Log(TN)	0.741	-0.430	0.127	0.407	-0.290	0.019	0.015
Log(SECCHI)	-0.790	0.464	0.173	0.273	0.026	0.236	0.036
Log(CHLα)	0.850	0.093	-0.127	0.293	0.404	-0.023	-0.060

393. For this example, however, our objective is to create components or factors that can be given a water quality interpretation. With that concern in mind, factor 1 is less attractive since it seems to represent all of the variables. On the other hand, the factor-variable correlations in Table 38 suggest that factor 2 is primarily an indicator of pH, and factor 3 is primarily an indicator of conductivity. Further, factors 4 through 7 seem relatively unimportant on the basis of both the factor-variable correlations and the eigenvalues. Thus, we request a factor rotation retaining only the first three factors.

394. As noted above, factor rotation is often used to reorient the factors with respect to the original variables so that substantive interpretation is facilitated. Varimax (orthogonal) rotation is the most commonly used procedure, so it is used here. In essence, the varimax procedure increases the effect of a variable on a factor for those variables that are highly correlated with the initial components, and decreases the effect for those variables that are not highly correlated with the initial components.

395. Table 39 contains the results of the varimax rotation. The orthogonal transformation matrix in Table 39 converts the first three principal components into the three new orthogonal factors created using

Table 39

Principal Components and Factor Analysis: Varimax Rotation

Orthogonal Transformation Matrix

Factor	1	2	3
1	0.721	0.617	0.315
2	-0.646	0.763	-0.016
3	-0.250	-0.192	0.949

Rotated Factor Pattern Matrix

	Factor		
Variable	1	2	3
pH	10	97 *	8
Log(COND)	18	20	95 *
Log(ALK)	35	84 *	27
Log(TP)	86 *	39	5
Log(TN)	78 *	10	36
Log(SECCHI)	-90 *	-16	- 8
Log(CHL*a*)	58 *	62 *	15

Communality Estimates

Variable	
pH	0.954
Log(COND)	0.971
Log(ALK)	0.906
Log(TP)	0.894
Log(TN)	0.750
Log(SECCHI)	0.868
Log(CHL*a*)	0.747

varimax. The rotated factor pattern yields the new matrix of factor-variable correlations; the correlations are multiplied by 100, and the highest values are marked with an asterisk for ease of interpretation. The communalities (squared correlations) indicate how much of the variance for each of the seven variables is shared with the other six. Finally, standardized scoring coefficients are presented in Table 40, permitting calculation of factor scores using the standardized variables as shown above.

396. The results in Table 39 indicate that we have achieved our objective of interpretable factors. Based on the rotated factor pattern (the factor-variable correlations) in Table 39, the first factor is effectively a trophic state index, describing primarily the trophic state variables phosphorus, nitrogen, Secchi disk depth and, to a lesser extent, chlorophyll *a*. The second factor is an indicator of acidity, as it is most highly correlated with pH and alkalinity (and to a lesser extent, chlorophyll *a*). The third factor is an indicator of dissolved solids or salinity, as it is largely a function of conductivity. Thus, we have created three orthogonal composites of the original seven variables, and each of the three factors has a clear substantive meaning. The standardized scoring coefficients could then be used in conjunction with standardized variables to calculate values for these trophic state, acidity, and salinity factors for each reservoir.

397. Predicting group membership: discriminant analysis. Discriminant analysis is used to define a linear function of predictor variables that may be employed to predict group membership for a particular case (e.g., lake). The dependent variable in discriminant analysis is categorical, as it identifies group membership. Conceptually, it is helpful to think of discriminant analysis as somewhat analogous to regression analysis; both procedures are often employed to define a linear relationship that may be used to predict the value of a dependent variable. In discriminant analysis, this dependent variable is categorical (e.g., trophic state), while in regression analysis the dependent variable is frequently continuous (e.g., phosphorus concentration).

Table 40

Principal Components and Factor Analysis:

Standardized Scoring Coefficients

Variable	Factor 1	 2	 3
pH	-0.235	0.604	-0.118
Log(COND)	-0.168	-0.101	1.004
Log(ALK)	-0.096	0.411	0.067
Log(TP)	0.377	0.010	-0.216
Log(TN)	0.341	-0.220	0.212
Log(SECCHI)	-0.463	0.165	0.138
Log(CHL*a*)	0.131	0.220	-0.088

398. For example, discriminant analysis can be used to develop a model for the prediction of trophic state. To do this, the investigator can use a cross-sectional data set of lakes and reservoirs, containing data on trophic state predictor variables (nitrogen, phosphorus, water clarity, Secchi disk depth, etc.) along with the trophic state classification for each lake or reservoir. Discriminant analysis is then employed to estimate a linear function of the predictor variables that best classifies the lakes and reservoirs into their preassigned classes (trophic states). Once this model is estimated, it may then be used predict the trophic state for a new lake or reservoir, on the basis of estimates of the predictor variables.

399. Of course, trophic state is not the only limnological classification scheme for which it would be useful to have a predictive model. As noted in Reckhow and Chapra (1983), it might be of value to develop a model for the prediction of oxic versus anoxic status in lakes, or perhaps a model for the prediction of expected dominant algal type on the basis of nutrient and aquatic chemistry.

400. While the development of a predictive model is the primary objective in most applications of discriminant analysis, other useful

information is generated from the application of the procedure. Use of discriminant analysis shows the investigator which variables are most important in the prediction of group membership. It also indicates how effectively one can predict group membership by providing an assessment of the proportion of misclassified cases in a data set. Finally, the discriminant functions (or classification functions) can be used to create a new function that provides an estimate of the probability that a case (e.g., reservoir) belongs in one of the predefined groups. This probability estimate is often a particularly informative way of expressing the confidence one might have in the prediction of group membership. For an excellent treatment of the theory and application of discriminant analysis, see Tabachnick and Fidell (1983).

401. For inferences and group membership predictions to be appropriate, some conditions are recommended and some statistical assumptions are technically required. As with all of the multivariate statistical methods, outliers can adversely affect the results of discriminant analysis. Therefore, it is recommended that the investigator examine histograms and bivariate plots to check for outlying data points. Transformations should be used if necessary to reduce the impact of outliers. (See Parts II and III for additional guidance on the treatment of outliers and influential data points.)

402. The primary statistical assumptions for the application of discriminant analysis are that the predictor variables are distributed according to a multivariate normal distribution within each group and that the variance-covariance matrices are constant across groups. The normality assumption can often be effectively assessed by checking for bivariate normality for any pair of predictor variables, while the variance-covariance matrices can often be compared by eye as these matrices are routinely calculated in most discriminant analysis programs. Informal checks on these assumptions are often adequate, as the results are fairly robust to violations. This is particularly true if the smallest group contains about 20 cases or more and there are only a few predictor variables in the discriminant function (Tabachnick and Fidell 1983). Transformations, of course, can be used if there is

concern over the violation of one of these assumptions.

403. If there are more than two groups and more than two predictor variables, more than one discriminant function may be estimated. Like principal components, each discriminant function is orthogonal to all others, and the first discriminant function is the most effective linear predictor of group membership (with additional functions being successively less effective). Generally, only one or a few of the discriminant functions are retained for the prediction of group membership. A statistical test can be conducted to assess the ability of each of the discriminant functions to determine group membership.

404. Discriminant analysis - example. For this example, we look at a cross-sectional model to be used to predict presence or absence of fish species in a lake. The data set consists of 32 lakes in the Adirondacks of New York State. The dependent, or classification, variable is the observation of presence (1) or absence (0) of brook trout in each lake, assessed during fish surveys. The predictor variables are pH and aluminum concentration (log-transformed).

405. Using SAS, PROC DISCRIM was run using the option POOL=TEST; this tests for equality of the covariance matrices. If the test results are not significant, the pooled covariance matrix is used in the calculations. SAS prints the discriminant function coefficients only when this pooled covariance matrix is used, so this is an important option to consider. POOL=YES causes SAS to bypass the significance test and automatically use the pooled covariance matrix.

406. Several optional statistics may be requested from SAS, including correlation and covariance matrices. For this example, Table 41 presents the within-group correlation matrix. This lists the bivariate correlation coefficients (and their significance levels) for all pairs of variables, calculated separately for the data within each of the predefined groups.

407. Next of interest in the SAS output is the pairwise squared generalized distance (or Mahalanobis distance) between groups. This is the distance between group centroids (based on the group means for each variable) scaled by the within-groups covariance matrix. For this

Table 41

Discriminant Analysis: Within Correlations

	Fish Absent		Fish Present	
	pH	Log(AL)	pH	Log(AL)
pH	1.000	-0.901 (<0.001)	1.000	-0.514 (0.035)
Log(AL)	-0.901 (<0.001)	1.000	-0.514 (0.035)	1.000

Pairwise squared generalized distance = 1.572

example, this distance ($D^2(I/J)$) is estimated as 1.572. Since D^2 is a measure of separation between the two groups, it is of interest to test the significance of this difference. The test is based on conversion of D^2 to an F statistic according to (Green 1978):

$$\frac{m_0 m_1 (m_0 + m_1 - n - 1)}{n(m_0 + m_1)(m_0 + m_1 - 2)} D^2 \sim F_{(n, m_0+m_1-n-1)}$$

where

m_0 = the number of observations in group 0 (absence)

m_1 = the number of observations in group 1 (presence)

n = the number of predictor variables.

For the example (with $m_0 = 15$ and $m_1 = 17$), the F statistic is 6.05. For 2,29 degrees of freedom, the separation of the groups is significant at better than the 0.01 level.

408. Table 42 presents the coefficients for the linear discriminant functions. They may be written as:

$$df_0 = -109.082 + 26.277\text{pH} + 15.880 \log(\text{Al})$$
$$df_1 = -116.772 + 27.885\text{pH} + 15.618 \log(\text{Al})$$

Table 42

Discriminant Analysis: Coefficients for the Linear Discriminant Function

	Fish Absent	Fish Present
Constant	-109.082	-116.772
pH	26.277	27.885
Log(AL)	15.880	15.618

409. These functions are used to classify cases as follows. For each case, the predictor values are substituted into the equations above and observation-specific values for df_0 and df_1 are calculated. A case is then classified into the group yielding the higher value for df . For example, if lake x has pH = 4.8 and Al = 300 μg/ℓ, then

$$df_0 = 56.384$$

$$df_1 = 55.764$$

410. According to this criterion, lake x would be classified in group 0 (fish absence).

411. Table 43 presents a summary of the classification success of the discriminant analysis model. The rows in the 2 × 2 table identify the actual group membership for each case, and the columns identify the predicted group membership (based on the discriminant analysis). A perfect classifier would have nonzero entries along the upper-left-to-lower-right diagonal and zeros in all other cells of the table. For example, in Table 43, 12 and 3 are the entries in the upper row. This means that 12 of 15 observed group 0 lakes are classified correctly as group 0 while 3 are classified incorrectly as group 1. A similar interpretation holds for the second row. Summing along the diagonal, we see

Table 43

Discriminant Analysis: Classification Success

	Fish Absent	Fish Present	Total
Fish absent	12 80.00	3 20.00	15 100.00
Fish present	6 35.29	11 64.71	17 100.00
Total	18 56.25	14 43.75	32 100.00

that 12 + 11 = 23 cases are correctly classified and 32 - 23 = 9 cases are incorrectly classified.

412. In Table 44 all cases are classified according to the discriminant function model using the generalized distance measure in an exponential expression. The exponential equation has range zero to one, so it is given a probabilistic interpretation. Thus, for each case, the probability of membership in each group is calculated, and the case is classified into the group for which the probability is the highest. These probabilities and classifications are presented in Table 44.

413. The probability equation in Table 44 is cumbersome because it is based on the generalized distance of a case from each group mean. Fortunately, probabilities can be calculated more easily from the discriminant functions according to:

$$P(0|df) = \frac{1}{1 + \frac{q_1}{q_o} \exp\ (df_1 - df_0)}$$

where

$P(0|df)$ = posterior probability of group 0 classification

q_0 = prior probability of group 0 classification

q_1 = prior probability of group 1 classification

Table 44

Discriminant Analysis: Classification Probabilities

Lake	Fish*	Classified as: *	Probability	
			Absent	Present
1	1	0 **	0.566	0.432
2	0	1 **	0.063	0.937
3	1	1	0.166	0.834
4	0	0	0.781	0.219
5	0	0	0.823	0.177
6	0	0	0.819	0.181
7	1	1	0.276	0.724
8	1	1	0.198	0.802
9	1	0 **	0.518	0.482
10	1	0 **	0.760	0.240
11	0	0	0.838	0.162
12	0	0	0.600	0.400
13	0	0	0.803	0.197
14	1	1	0.093	0.907
15	1	0 **	0.688	0.312
16	0	0	0.816	0.184
17	1	0 **	0.617	0.383
18	0	1 **	0.269	0.731
19	1	0 **	0.819	0.181
20	0	1 **	0.491	0.509
21	0	0	0.847	0.153
22	0	0	0.847	0.153
23	0	0	0.750	0.250
24	1	1	0.485	0.515
25	0	0	0.730	0.270
26	0	0	0.675	0.325
27	1	1	0.164	0.836
28	1	1	0.128	0.872
29	1	1	0.129	0.871
30	1	1	0.119	0.881
31	1	1	0.106	0.894
32	1	1	0.152	0.848

* Note: 0 = absent; 1 = present.
** Misclassified observation.

414. For now, assume that the prior probabilities are equal; this means that we have no a priori belief that a particular case belongs in one group or the other. In that case, $q_1/q_0 = 1$. For the example presented above (pH = 4.8; Al = 300 μg/ℓ), the probabilities are:

$$P(0|df) = 0.650$$

$$P(1|df) = 1 - P(0|df) = 0.350$$

415. Thus, there is a 0.65 chance that the lake is properly classified in group 0 (fish species absent).

416. Since the equation allows for prior probabilities, we could assign prior probabilities to each group reflecting the relative proportion of the relevant population that belongs to each of the groups. For this example, 15 of the 32 observations in the sample are in group 0 (absence) and 17 of the 32 observations are in group 1 (presence). These proportions can be used as the prior probabilities. Thus,

$$q_0 = 15/32 = 0.469$$

$$q_1 = 17/32 = 0.531$$

417. Using these prior capabilities in the equation above, the new posterior probabilities are:

$$P(0|df,q) = 0.621$$

$$P(1|df,q) = 1 - P(0|df,q) = 0.379$$

418. Notice that, as should be expected, the higher prior probability for group 1 resulted in a drop in the posterior probability for group zero, in comparison to the analysis when the prior probabilities were not explicitly included. Use of a prior probability can be particularly helpful when the groups are quite different in size.

419. In situations where the group sizes are different, the "naive" classification criterion (without consideration of predictor variables) would assign all cases to the group with the largest number cases. Thus, if 80 percent of all cases correctly belong in group 1, the naive classifier would place all cases in group 1 and have an 80-percent classification success rate. This "maximum chance criterion" (Morrison 1969) is appropriate for evaluating the success of discriminant functions (beyond that by chance) if one desires to maximize the proportion of cases correctly classified. Alternatively, if the objective is to correctly classify cases into both groups, the appropriate criterion is based on proportional chance. This is calculated as (Morrison 1969):

$$C_{pro} = (q_0)^2 + (q_1)^2$$

420. For our example, this is

$$C_{pro} = (15/32)^2 + (17/32)^2 = 0.502$$

whereas the maximum chance criterion is

$$C_{max} = 17/32 = 0.531$$

421. Since our objective in this example would probably be to classify cases correctly in both groups, the proportional chance criterion is appropriate. According to that criterion, our model does better than chance if it classifies more than 50 percent of the cases correctly. In fact, for the model development data set, Table 43 indicates that 23/32 = 0.719, or about 72 percent of the cases were correctly classified by the discriminant function model.

422. It must be emphasized that the classification success of the model should actually be evaluated using a data set that is different from the model development data set. It is to be expected that a model developed from a particular data set will yield an overly optimistic classification success rate when evaluated using the model development

data set. Thus, a separate data set should be used to provide an unbiased estimate of the classification error rate. Alternatively, various measures of cross validation (Green 1978) may be employed to estimate classification error.

PART V: SAMPLING PROGRAM DESIGN

Introduction

423. In undertaking a study, an investigator will generally have as his objective either the estimation of some parameter or the comparison of several different populations. Since sampling is the only practical method of carrying out most studies, the researcher is immediately faced with several problems. The eventual discussion of the whole population from a sample involves statistical inference. This means that the true value of the population parameter will never be known, only an approximation or estimate of that parameter. It is necessary to obtain these approximations as accurately and as precisely as possible.

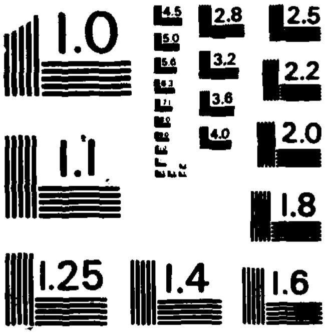
424. Accuracy implies that an estimate of a parameter will, on the average, be centered on the true population parameter and will not be shifted up or down. Estimates that have a consistent tendency to overestimate or underestimate a population parameter are inaccurate and are said to be "biased."

425. Precision is an indication of the reliability of an estimate and refers to the variability between repeated measures of the same quantity. All estimates of parameters will have some variability, but the lower the variability, the higher the precision.

426. The major objective in sampling program design is to obtain as accurate or unbiased an estimate as possible, and at the same time reduce or explain as much of the variability as possible in order to improve the precision of the estimates.

427. A major concern in the design of a sampling program deals with the problem of practicality. Measuring the whole population is impractical. The sampling scheme should provide an estimate that is as accurate and precise as possible, even though the sample may be a very small fraction of the whole. For instance, the objective may be to determine the average phosphorus concentration of a reservoir. The sample may be only a few litres of the millions of cubic metres of water in the reservoir. The number of samples would be small, but with proper

AD-A171 046   ENVIRONMENTAL AND WATER QUALITY OPERATIONAL STUDIES   3/3
STATISTICAL METHODS F..(U) ARMY ENGINEER WATERWAYS
EXPERIMENT STATION VICKSBURG MS ENVIR..   R F GAUGUSH
UNCLASSIFIED   JUN 86 WES/IR/E-86-2   F/G 8/8   NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

placement and an adequate number of samples, a good estimate could still be obtained to meet the study objectives.

428. Another concern in sampling design is cost. More samples are better, but cost increases proportionally with the number of sampling trips made, number of sites sampled, and the number of different analyses performed. The increased number of trips produces a more precise estimate. Unfortunately, the increase in trips or analyses is not directly proportional to the increase in precision, so that doubling the number of sampling trips does not double the precision. The sampling program must be designed to achieve the optimal allocation of the samples. An optimal allocation will be both practical and cost effective by striking a balance between how many samples are needed and how many samples are within budget. If funds are severely limited, the investigator may also be faced with a decision on the feasibility of the study. The question is simply: "Will the results that can be obtained with the available funds produce estimates which are sufficiently precise to meet the study objectives?"

429. The critical element in designing a sampling program is the understanding of variability in both the samples and the target population. If a lake had no variability in phosphorus concentration, one sample from the most convenient site would provide an adequate measure for the whole lake. However, if variability exists (and it always does), some statistical inference procedure is required.

Study Objectives

430. In order to ensure that the sampling scheme is adequate and that it will provide the desired information, it is necessary to state the study objectives clearly. Sampling is facilitated by specifying the narrowest possible set of objectives which will provide the desired information. Several points that should be specified in the study objectives are discussed below.

431. Target population definition is the first step, since the sample must be drawn from the target population. A population, in a

statistical sense, can be defined as the set of all possible values of the variable of interest which might, or do, exist. The target population is a limited subset of the population and is simply the population about which statistical inferences are to be made. The limits of the target population are defined by the objectives of the study. Examples of target populations include dissolved oxygen concentrations in reservoir releases under different operating regimes, phytoplankton abundance before and after some environmental change, or the average phosphorus concentrations in a group of reservoirs. Individual measurements of the population of interest are called observations, and the population parameter being measured is referred to as the variable. This definition will not only describe what is to be sampled, but where and when. All information that limits the population to be sampled should be included.

432. The reason for limiting the target population definition is that much of the variability which exists is not of interest. If dissolved oxygen concentrations in reservoir releases are to be sampled, the investigator need not be concerned with dissolved oxygen concentrations in the rivers and streams entering the reservoir, which would add an additional source of variability. The first step then in addressing the problem of variability is to obtain a clearly defined set of objectives, which in turn includes a definition of the target population.

433. For example, suppose the objective is to sample dissolved oxygen. All oxygen everywhere? No, obviously not. So the definition of the target population is refined, and with each new element of the definition the investigator derives a more homogeneous (less variable) target population--oxygen concentrations in reservoirs, reservoirs in the United States, in the southeastern United States, etc. Eventually the target population may be reduced to a certain size or particular type of reservoir. Sampling may even be restricted to a particular body of water, to particular portions of that body of water, and even to particular depths. The definition depends entirely on the objectives of the study.

434. When the study objective is to test for differences, it is

not uncommon to have some specific reservoirs that are to be contrasted. In this case, the definition of the target population is simplified. However, if the target population calls for the comparison of two or more types of reservoirs in a geographical area, the reservoirs chosen should represent a random sample of all the reservoirs which fall into each of the categories of interest.

435. The definition of the target population above has only considered the spatial limits of the study. Temporal limits must also be defined. Should the oxygen concentration be sampled during the full annual cycle, several annual cycles, or perhaps during only part of the annual cycle? Again this depends on the study objectives. Is the lake to be characterized for one annual cycle (any year), or does the study call for estimates of variability between years? Perhaps the study calls for an evaluation of oxygen conditions in the summer, when hypolimnetic anoxia is expected to occur.

436. The definition of the target population must be broad enough to satisfy the study objectives, but the more narrow the objectives can be made the less variability there is to be taken into consideration. Once the target population and scope of the study have been defined, the investigator should be able to state which variable or variables will be measured. He should also be able to define the spatial and temporal limits of the experiment.

437. Problem identification requires a decision on the nature of the final goal. To begin with, do the researchers wish to estimate a population parameter, such as reservoir phosphorus concentration, or do they wish to test a hypothesis? These two procedures are not mutually exclusive, so the objectives may involve both estimation and hypothesis testing. However, one will usually be chosen as a primary objective, since the sampling allocations may differ for the two goals. For example, a balanced design is a desirable trait in hypothesis testing while an estimation of a parameter may involve sample allocation which is unbalanced but will take into consideration differing variable levels in different areas.

438. The last step in outlining the sampling objectives is to

define exogenous variables to be measured and/or define strata. The purpose of including the measurement of covariables (quantitative variables) and categorical variables or strata (qualitative variables) is to reduce variability and increase precision. For example, primary productivity is obviously dependent on incident light conditions and the attenuation of light with depth. Therefore, in a study of primary productivity, light is an influencing variable which should be taken into consideration even though its measurement may not be part of the study objectives. This exogenous variable can be used to explain part of the variability in primary productivity, and to therefore increase the precision of the estimates.

439. Stratification, or the use of categorical variables, serves a similar purpose (reducing variability in the estimates) in a different manner. In stratification, the area or time frame of sampling is subdivided into smaller more homogeneous units which will have a small amount of variability within the unit. For example, primary productivity varies with season. If primary productivity over an annual cycle is examined, there will be a large degree of variability. However, if each of the seasons is examined individually, there will be much less variability. Therefore, summer productivity may range from 0.3 to 0.5 g C m^{-2} day^{-1}, and relatively precise estimates can be obtained. Relatively precise estimates of the winter productivity, which might range from 0.01 to 0.1 g C m^{-2} day^{-1}, may also be obtained. However, if seasons are ignored, the result is a single less precise estimate ranging from 0.01 to 0.5 g C m^{-2} day^{-1}. Obviously, greater precision results from estimating the mean productivity separately within each of the strata (seasons) than if strata are ignored. The same is true for areas that may have different mean values of the target population. When the areas are combined, the variability will be greater than if the areas are separated as strata.

440. Any consideration in the sampling program which serves to reduce, eliminate, or explain variability is desirable. There will always be some variability which cannot be explained in terms of exogenous variables. This variability is called error. "Error" is not

used to indicate a mistake, but rather it refers to variability in the data that is not accounted for by the sample design. Error can arise from the effect of an unknown (and unmeasured) exogenous variable or from the variability introduced in the measurement of a variable. This source of unexplained variability is a natural property of the target population, and the investigator should not hope to eliminate it, only to reduce the unexplained variability as much as possible.

441. A careful definition of the objectives is a critical step in conducting any study. Objectives that are too broad and poorly defined generally result in an inefficient sampling program which may not adequately meet the objectives of the study. The need for this simple, but often ignored, step of establishing study objectives was described above.

442. At this point the investigator has fully defined his variables of interest, both his target population and exogenous variables, and the scope of the study. This clearly defines the sampling objectives and facilitates the design of the sample placement.

Sample Allocation

443. Once the objectives are clearly stated, sample placement follows fairly easily. Assume for the moment that the investigator must decide how to allocate a given number of samples, the number of which will be referred to as "n." A discussion of how large n must be will follow. The discussions of statistical analyses which follow will make one common assumption, that the sampling has been random. If the objective is to sample the oxygen concentrations in a particular cove of a reservoir, the analysis will assume that every drop of water in that cove had an equal chance of being sampled. This condition is met when the selection of a particular site does not affect the choice of the second and subsequent sites. Other types of sampling will be discussed later, but for the moment, random sampling will be assumed.

444. If sampling is to be random, the individual sites and depths for each sample are selected at random, usually from a random number

table or random number generator. If the sample is to be completely random, sampling dates may even be selected at random across the annual cycle. However, this may not be the best sampling scheme, and it may not address the objectives well.

Identification of strata

445. Strata are subdivisions of a larger population that are used to reduce variability by working with smaller, more homogenous units of the total target population. Strata are generally used where parameters are to be estimated. Each stratum is treated as a separate entity. Estimates of the population of interest are obtained for each stratum separately, and combined for a final estimate. The advantage is that the variances can be estimated separately, and can be added. Each of the variances for the strata is likely to be smaller than if the population had not been stratified.

446. After the strata to be sampled have been identified, such that variability within the strata is more homogeneous than the whole, the next step is to allocate a number of samples to each stratum. This is called stratified random sampling. For instance, suppose a decision has been made to stratify the samples into four seasons and into three depths, with the expectation of relatively little variability within each stratum as compared to the variability which exists between the strata. Then if there were $n = 600$ samples to be allocated, one option is to sample each of the 12 strata (three depths in 4 months) equally. This calls for the placement of 50 samples (600/12) in each of the strata.

447. Sampling schemes other than equal allocation are possible. For instance, if the volume of water at one depth is only 20 percent of the total while the other depths contain 40 percent each, allocation of the number of samples to each depths may be in proportion to the volume of that depth. This is called proportional allocation. This type of allocation can also consider the variability of the strata, in addition to its size. If one stratum is more variable than another, it can receive more sampling effort in proportion to its variability. This yields a greater precision for the more variable strata, and a more

precise estimate overall. One additional consideration is cost. A sample allocation capable of considering cost, in addition to the factors mentioned above, is called optimal allocation. This method considers not only the size and variability of a stratum, but also the cost of sampling each of the strata. In this case, the size of the sample allocation to the strata is decreased in proportion to the inverse of the cost of sampling the strata (i.e., strata that are less costly to sample are allocated a greater number of samples).

448. The mathematical formulation for each of the sample allocations mentioned is discussed below. Allocation of sampling units for the estimation of a population can take three factors into account, the expected size of the strata, the variance of the strata, and the cost of sampling the strata. The three considerations can be expressed in a single f.rmula, which may be simplified if particular factors are not to be considered. The general formula is given by

$$n_i = n\left(\frac{N_i\sigma_i/\sqrt{c_i}}{N_1\sigma_1/\sqrt{c_1} + N_2\sigma_2/\sqrt{c_2} + \ldots + N_s\sigma_s/\sqrt{c_s}}\right)$$

where

s = the number of strata

i = the stratum number (i = 1, 2...s)

N_i = the size of the i^{th} stratum

σ_i = the standard deviation (square root of the variance) of the i^{th} stratum

c_i = the cost of sampling the i^{th} stratum

n = the total number of samples to be allocated

n_i = the number of samples allocated to the i^{th} stratum

449. This formula contains all of the elements in deciding optimal allocation. It allows for consideration of size of the strata, variance, and cost of sampling each stratum. It is not necessary to include every factor. If the cost of sampling each stratum is approximately the same, all c_i factors may be omitted. Also, if the variances are the same or the population sizes are the same, the factor σ_i

or N_i may be omitted, respectively. If the three factors are the same for each stratum, it is interesting to note that the results of the sample allocations are the same for the formula above as for a completely random allocation done without consideration of strata. The three examples below illustrate the use of this equation under differing conditions.

450. Assume that the epilimnion, metalimnion, and hypolimnion of a reservoir are defined as the three strata of interest. Also assume the strata differ in size (volume) such that the epi:meta:hypo ratio is 6:3:1, cost of sampling and variance is constant across strata, and a total of 30 samples are to be allocated. Therefore,

$$\begin{array}{ll} N_1 = 6 & i = 3,\ n = 30 \\ N_2 = 3 & c_1 = c_2 = c_3 = c \\ N_3 = 1 & \sigma_1 = \sigma_2 = \sigma_3 = \sigma \end{array}$$

and the equation becomes

$$n_i = n \left[\frac{N_i(\sigma/\sqrt{c})}{(N_1 + N_2 + N_3)(\sigma/\sqrt{c})} \right]$$

and can be simplified to

$$n_i = n \left(\frac{N_i}{N_1 + N_2 + N_3} \right)$$

The number of samples allocated to each of the strata would be

$$n_1 = 30 \left(\frac{6}{6 + 3 + 1} \right) = 18$$

$$n_2 = 30 \left(\frac{3}{6 + 3 + 1} \right) = 9$$

$$n_3 = 30\left(\frac{1}{6 + 3 + 1}\right) = 3$$

The advantage to this sample allocation is that it emphasizes the size of each stratum. This allocation shifts more samples to strata of greater size.

451. Given the three strata defined above, the variability within strata might not be expected to be constant. Variability might be at a minimum in the well-mixed epilimnion and increase to a maximum in the hypolimnion. Assume that variability is not constant

$$\sigma_1 = 1$$
$$\sigma_2 = 2$$
$$\sigma_3 = 4$$

and that all other terms are as given above. The equation for sample allocation becomes

$$n_i = n\left(\frac{N_i\sigma_i}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3}\right)$$

where $N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3 = 6(1) + 3(2) + 1(4) = 16$. The number of samples allocated to each stratum would be

$$n_1 = 30\left[\frac{6(1)}{16}\right] = 11.25 \simeq 11$$

$$n_2 = 30\left[\frac{3(2)}{16}\right] = 11.25 \simeq 11$$

$$n_3 = 30\left[\frac{1(4)}{16}\right] = 7.5 \simeq 8$$

The result is that more samples are allocated to the more variable metalimnion and hypolimnion. The advantage to this sample allocation is that it provides the lowest possible overall variance in the final estimation of the parameter for all strata combined. The allocation shifts more samples to strata which are more variable. The final variance of each stratum is then reduced proportionally to the sample size allocated to that stratum. The variances of other strata will increase somewhat if the total sample size is kept constant, but the variance of the final combined estimate will be minimized.

452. Cost of sampling might also be expected to differ among the strata. The surface area represented by the epilimnion may be much larger than that of the hypolimnion and, as a result, the time involved in sampling (a cost) the epilimnion could be considerably greater than for sampling the hypolimnion. Assume that cost is not constant

$$c_1 = 2$$
$$c_2 = 1.5$$
$$c_3 = 1$$

and all other terms are as given above. The equation for sample allocation becomes

$$n = n_i \left(\frac{N_i \sigma_i / \sqrt{c_i}}{N_1 \sigma_1 / \sqrt{c_1} + N_2 \sigma_2 / \sqrt{c_2} + N_3 \sigma_3 / \sqrt{c_3}} \right)$$

where

$$N_1 \sigma_1 / \sqrt{c_1} + N_2 \sigma_2 / \sqrt{c_2} + N_3 \sigma_3 / \sqrt{c_3} =$$

$$6(1)/\sqrt{2} + 3(2)/\sqrt{1.5} + 1(4)/\sqrt{1} = 13.14$$

The number of samples allocated to each of the strata would be

$$n_1 = 30 \left[\frac{6(1)/\sqrt{2}}{13.14} \right] = 9.69 \simeq 10$$

$$n_2 = 30 \left[\frac{3(2)/\sqrt{1.5}}{13.14} \right] = 11.18 \simeq 11$$

$$n_3 = 30 \left[\frac{1(4)/\sqrt{1}}{13.14} \right] = 9.13 \simeq 9$$

The advantage to this sample allocation is that it provides the lowest possible overall variance in the final estimation of the parameter for all strata combined for a given investment. Conversely, for a given precision, this method could also define a minimum cost. It incorporates the above considerations but will, in addition, weight the sample allocation to strata which are less costly to sample.

453. The investigator can obtain either a combined mean or a combined total population estimate from a stratified sample. The estimation of a population mean from a stratified random sample is the weighted mean of each of the individual strata, weighted to account for the number of samples in each of the strata. The total population estimate is the sum of the estimates for each of the individual strata. The estimate of the variance for the combined strata is given by the weighted sum of the variances for the individual strata.

Determining the Number of Samples

454. It is possible to estimate how large a sample size is required to achieve a particular level of precision. The objective is to estimate some parameter to within a specified tolerance. The final statement of the results may be expressed as a confidence interval. This statement is expressed in the form

$$P(\text{lower limit} \geq \text{true population parameter} \leq \text{upper limit}) = \text{confidence level}$$

455. There is always a chance that the interval does not contain the true value of the population parameter. For this reason the confidence level, expressed as a probability, will always be less than one, or 100 percent. For instance, the investigator may be 95 percent sure that the interval given contains the true population value. The width

of the interval reflects both the level of confidence and the variability in the population.

456. A more complete discussion of confidence intervals is presented in the section on statistical analysis. For the moment it is only necessary that the reader understand that the objective of calculating a "required minimum sample size" is to obtain a sample size from which a confidence interval can be calculated which will contain the true population parameter within a specified level of tolerance.

457. The formula for calculation of sample size requirements is simple, but obtaining good estimates of the factors contained in the formula is often difficult. The formula is given by

$$n \geq t^2\sigma^2/p^2$$

where n is the sample size needed for a particular probability level of error (or confidence) in the eventual confidence interval calculation. The value of t is given by this probability. The variance for the population is given by σ^2, and a desired level of precision or tolerance is given by p. The variance factor (σ^2) refers to variability in the population. It dictates that for a given level of confidence and precision, more variable populations require larger samples (larger values of n). The factor p is determined by the researcher and gives a level of precision required. For example, if the objective is to estimate phosphorus concentration, the researcher may decide that the estimate should be within 2 μg P/ℓ of the actual concentration. That is, the estimate will be the actual concentration ±2 μg P/ℓ. The values of t, p, and σ^2 used must be provided by the investigator. Many who are not familiar with the process, or with the population to be sampled, may be at a loss to give reasonable estimates to these values.

458. The value t is derived from statistical theory. The researcher must first decide the probability of error in the eventual results of the study. There will always be some probability of error. For example, when the research finally states that the actual concentration is 20 ± 2 μg P/ℓ, there is always a possibility that he will be

wrong. If he sets t to a level which corresponds to a 5-percent error rate, he will be wrong five times in 100 estimates. If he chooses a 1-percent error level, he will be wrong only one time out of 100 estimates, but he will have to take many more samples. Generally, researchers have chosen a 5-percent or 1-percent level of error. The t values corresponding to these levels may be initially approximated as 2 and 2.6, respectively. Actual values of t can be obtained from statistical tables when the expected sample size has been estimated.

459. The estimate of the variance σ^2 should be determined, if possible, from the population to be sampled. This would require a preliminary study of the population and an estimate of the variance. Variance is estimated by the formula

$$\sigma^2 = s^2 = \left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right] \Big/ (n - 1)$$

$$= \left[\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \Big/ n\right] \Big/ (n - 1)$$

where

X_i = value of the i^{th} observation of the target population variable

s^2 = estimate of σ^2, indicated as $\hat{\sigma}2$

n = sample size

If the estimate cannot be obtained with a preliminary sample, an approximation may be obtained from published values from similar populations.

460. The estimate of precision is one of the more difficult to predetermine. If the investigator knows that he would like to determine the final estimate within ±2 μg P/ℓ, this value would be used. The desired precision can also be expressed as a percent of the mean. For example, if the concentration is expected to be about 20 μg P/ℓ, the precision might be requested as 10 percent of this value, or 2 μg P/ℓ. A level of precision of this magnitude is often appropriate for field studies. More refined laboratory studies may require a greater

precision, while preliminary surveys or short studies whose objective is to guide an administrative decision may require less precision.

461. The investigator should understand that the values provided are only rough estimates. It is possible that when the final calculations are made, the tolerance is not as small as originally specified. A similar development of the same minimum sample size estimating formula can be made for tests of hypothesis. Again, the application of the formula does not guarantee that detectable differences can be made at the level originally specified. Still, the formula provides the best estimates available.

Systematic Sampling

462. All of the sampling discussed previously was random or stratified random. Systematic sampling is generally easier to carry out than random sampling. In systematic sampling the first sample placement is generally decided at random within an initial region, and subsequent samples are taken at some constant distance or time from the first. For example, the first sample may be randomly placed near the headwaters of a reservoir, and subsequent samples are taken every 2 miles down the reservoir from the previous sample. More commonly, sequential samples are taken over time. The first sample is taken in January, and additional samples are taken at 4-week intervals after the first. Although this is a popular sampling scheme, there are several subtle problems that can originate from its use. The eventual statistical analysis of data will generally require an assumption of random sampling. A systematic sample may have a variance that is either greater or smaller than a random sample.

463. Systematic sampling in an area, such as over the surface of a reservoir, often entails sampling a grid of stations. This sampling scheme is effective in covering the whole range of variability available in the area, since it will often uncover heterogeneities missed in random sampling. However, the ability of systematic sampling to cover the range of variability also means that the variability can be <u>greater</u> than

if sampling is done randomly. However, additional studies have shown that variances resulting from systematic samples are frequently smaller than those resulting from random sampling. Whether the variance estimate is large or small depends on the distribution of the population being sampled.

464. In either case, the assumption of random sampling required for the eventual statistical analysis will not be met. This is particularly important in hypothesis testing, and it will also affect confidence interval estimates. However, this does not mean that estimates obtained from systematic samples are necessarily biased or lack precision.

465. The detection of the wider range of variability may be desirable in preliminary studies, particularly if the objectives include the definition of strata. Systematic sampling may be better for some applications such as defining strata. However, once the strata are defined, the sampling should be done at random within the strata, if possible, in order to meet the assumptions.

Cluster Sampling

466. Another type of sampling design is cluster sampling, done as either one- or two-stage sampling. This type of sampling is applicable when the sampling unit is an identifiable group of individuals or a cluster of observations. In this case it may be impractical to sample all of the individuals at random, so the clusters are identified and a sample is selected at random. Then the cluster is completely sampled (one-stage) or subsampled (two-stage).

467. As an example, suppose that the nitrogen concentration of first-order streams is to be sampled in a reservoir drainage basin. There may be a hundred small first-order streams in the basin. Compiling a list of all these streams for a random sample would be difficult, and covering the whole drainage basin to provide a random sample may not be practical. An alternative would be to identify individual river basins entering the reservoir, which may only number a dozen, and to

select first-order streams in two stages. First, a random sample of river basins is selected; second, a random sample of first-order streams is selected within each of the river basins selected for sampling. This greatly simplifies the listing of feeder streams for sampling, since maps for only a sample of the river basins must be searched, rather than for all of the river basins. Field sampling is also simplified, since trips need not be made into all of the river basins in the reservoir drainage area.

468. The analysis of cluster samples requires the estimation of variance at two levels, the between-cluster variability and the within-cluster variability. The total variability is a recombination of these two levels. The slightly more complicated calculation of the combined variance may be more than offset by the application of this more practical sampling scheme, and by savings in cost of sampling.

Types of Sampling Programs

469. The basic types of water quality investigations may be placed in several categories depending on the objectives of the sampling program. The objectives may call for parameter estimation, a test of a hypothesis, or the development of a predictive model. The objectives are not mutually exclusive since hypotheses about the estimated parameters may be tested. However, the primary goal should be stated in the objectives.

470. Data collected in any sampling program should ultimately be processed by some statistical technique. Therefore, some important aspects of statistical analysis, such as the hypotheses to be tested and which sources of variability should be included in the analysis, are discussed as part of the considerations of sampling.

Parameter estimation

471. The primary objective in many programs is to document the status of an area. This is often the objective of baseline or pilot studies. These studies are designed to establish the normal levels of parameters prior to impoundment of an area or some other future change.

The eventual analysis resulting from this type of study would generally be some parameter estimation. The variables measured would depend on the objectives of the study. The objective may be to document particular attributes of water quality, such as pH, dissolved oxygen, turbidity, or alkalinity. In this case the analysis may include only the mean values and a measure of variability.

472. Another type of survey is the pilot study of an area, prior to the initiation of a major study. Pilot studies may have as their objective the estimation of parameters and the determination of their range and variance. Another objective of pilot studies may be the identification of strata. Stratification, previously discussed, is simply the subdivision of an area into smaller, more uniform sections. This is an important aspect of reducing the variability of the final estimates produced by a study. Estimates of the costs of sampling may also be obtained from pilot studies.

Tests of hypotheses

473. Another type of study objective may be the testing of some specific hypothesis. There are two basic approaches in this type of study. The objective may be to test the existing water quality against some hypothesized value, or it may be to test the equality of two or more areas, seasons, or reservoirs. Although this type of study may be done as a survey, many aspects of designed experiments may be used to improve the results.

474. A test of hypothesis may be formally represented by the mathematical statement

$$H_o:\mu \geq \theta$$

where

μ = mean of the sampled population

θ = hypothesized value

The alternative hypothesis can be stated as

$$H_A:\mu < \theta$$

475. Note that this set of hypotheses is one-sided, that is, the only alternative hypothesis of interest is one where the parameter estimates are below the hypothesized value. The hypothesis may also be two-sided as in the example below.

476. Tests of hypothesis are not always made against hypothesized values. The objective may be to test for a difference of the means between two areas. Actually there is an hypothesized value here, even though it is not known in advance what mean values the reservoirs will have. The hypothesis in this case is that the <u>difference</u> between the mean values observed for the two reservoirs will be zero. The alternative is that the difference is not zero, indicating that a difference exists.

477. These hypotheses may be represented mathematically as

$$H_o:\mu_1 = \mu_2 \quad \text{or} \quad H_o:\mu_1 - \mu_2 = 0$$

where

μ_1 = mean of the first population sampled

μ_2 = mean of the second population

478. In statistically testing a hypothesis, the primary hypothesis (H_o) is called the "null hypothesis" because it is generally a hypothesis of no difference. The alternative hypotheses are

$$H_A:\mu_1 \neq \mu_2 \quad \text{or} \quad H_A:\mu_1 - \mu_2 \neq 0$$

479. In this case the alternative hypothesis is two-sided since the second area may have either a larger value than the first, or a smaller value than the first. Either case is of interest, and both indicate that the areas are not equal. The hypothesis in this type of example may also be one-sided. For example, the hypothesis may be that one area has a higher dissolved oxygen concentration than some other area.

480. Similar tests of hypothesis may be made to test for effects, that is, changes due to some treatment that has been administered by the

investigator or has occurred naturally. Here again, the hypothesis is one of "no difference exists" against the alternative that "a difference does exist."

Prediction

481. Predicting the value of a variable under certain conditions is done with regression analysis. In regression analysis the objective is generally to demonstrate or document a relationship between one variable called the "dependent" variable and one or more variables called the "independent" variable or variables.

482. A major difference between this type of analysis and those discussed previously is that, in parameter estimation and hypothesis testing, the means being estimated or tested could have been for some particular group, class, or category, such as an area or a particular month. Regression analysis requires the use of a quantitative variable. At least one of the independent variables in a regression analysis will be a quantitative variable, as opposed to a qualitative, categorical, or group variable.

483. Examples of relationships that may be documented by regression analysis are phosphorus-chlorophyll relationships in reservoirs or the relationship between light and primary productivity. Regression analysis can also be used to test for relationships between some variable and a gradient, or to test for trends over time. Predictions done with regression analysis may result either from surveys or from designed experiments.

Sampling a spatial gradient

484. Analysis of a gradient such as a trend with depth or distance calls for the measurement of the variable of interest and for measurement of the gradient variable. In terms of material already discussed, the gradient variable will be treated as a covariable to the variable of interest. As usual, the hypothesis is that the gradient variable will account for some additional variability of the variable of interest.

485. The relationship between a gradient and the response variable may be either a simple linear function or some nonlinear

relationship. If the response of the variable of interest to the covariable is known to be linear, or can be assumed to be linear, the only sampling observations required are from either end of the gradient. If the gradient is not a simple linear function, or if one of the objectives is to test the relationship for linearity, the gradient should be sampled at a number of intermediate points along its range.

486. Two aspects of analyzing a gradient should be taken into account. First, the analysis of a response along a gradient that is not linear requires placing samples along the gradient. The tendency of many investigators is to then spread all of the sample points out along the gradient. The second aspect of analyzing a gradient is the evaluation of the adequacy of the model fitted to the gradient. This will require replication of the samples at numerous points along the gradient.

487. The investigator is then faced with a decision. Is it preferable to sample more points along the gradient with few replicates at each point, or to sample fewer points with more replicates? As stated previously, if the relationship is linear, then only two points need be sampled (sampling the ext emes is preferable). Whenever the relationship is known, even if it is curved, then relatively few sampling points are needed along the gradient. The minimum number of points depends on the relationship that is to be used. In this case, more samples may be used as replicates.

488. If little is known about the relationship, or if the relationship is complex, the necessary number of points along the gradient increases. At the same time, more replicates are required in order to test the adequacy of the proposed model. It is better in this case to spread as many points as possible along the gradient, but to insist on some replication at the sampling points.

489. It is generally good practice for the placement of the points along the gradient to be approximately equidistant, but care should be taken to randomly select points within gradient segments if there is a danger of falling in step with some natural phenomenon. However, equidistant points are not necessarily the optimal distribution

if, for instance, a particular section of the gradient is to be estimated with greater precision.

Sampling over time

490. When samples are taken over time, time can be considered as an additional variable or source of variability. Time can either be viewed as a source of variability that should be blocked into more homogeneous subunits (quarters, seasons, or months) or as a covariable. If the changes in the variable of interest are expected to be small, or stable for long periods with short periods of change, then blocking on time may be desirable. The periods to be blocked would depend on which periods of time are expected to be stable. If treated as a covariable, the preceding discussion on sampling a gradient applies quite well to sampling over time. However, two considerations should be included.

491. First, "time" is not likely to be the actual influencing variable. Changes over the annual cycle, for instance, occur because of changing light intensity, temperature, etc. "Time" can only be used as a surrogate for other variables that are either unknown or whose individual influences on the target variable cannot be discerned.

492. The second consideration is that the relationship with time as a covariable is likely to be complex. Over an annual cycle the relationship with most variables is unlikely to be linear (though relationships may appear linear for short periods of time). Sampling over the annual cycle, with time as a covariable, will therefore call for relatively frequent samples over time, but with some replication within sampling periods. Care should also be taken to sample randomly within each of the sampling periods.

Frequency of sampling

493. The frequency of samples over time depends on the definition of the target population and whether time is to be considered as strata or a covariable. If stratified, the samples within the time strata should be randomly allocated, and their number would reflect the desired precision. If time is to be treated as a covariable, the annual cycle should be subdivided into a larger number of subunits, and sample replicates randomly placed within the subunits. Monthly sampling is usually

adequate to detect the annual pattern of changes with time. If the investigation requires the detection of short-lived phenomena, then more frequent sampling may be required to obtain greater resolution.

Summary

494. There are several aspects of sampling with which the researcher should be familiar before designing a sampling program. The design used will depend on the objectives and on the types of variables that must be taken into consideration. In any design, the primary concern will be variance or variability. Whether the objective is parameter estimation, testing of a hypothesis, or development of a predictive model, the results should have the smallest possible component of unexplained variability. The main concern is to reduce, eliminate, or account for sources of variability.

495. The sampling plan can reduce variability in several ways. One way is to standardize the field sampling techniques as much as possible. Any refinement of technique that will contribute to uniformity will aid in reducing variability.

496. Not all variability can be reduced or eliminated. In some cases the sampling program will have to include the measurement of exogenous variables (variables that are not part of the study but which contribute to the variability) in order to measure and account for additional variability. In other cases, the parameter estimates cannot be done separately for each area. For instance, when used for tests of hypotheses, it is convenient to combine all areas into one analysis in order to obtain a pooled estimate of variability. Blocking provides a measure of variability without "eliminating" it and allows for all data to be combined into a single analysis for testing hypotheses.

497. Previous mention has been made of stratification as a method of eliminating variability. Although there is no real difference in the concept of a "block" or a "stratum," there is a difference in the way in which the two are applied. A block may be a subdivision of a larger population, such that the variability within the individual block is less than for the whole, which is also true of strata. Blocking is a term generally applied to situations in which hypotheses are to be

tested, and it is simply a convenience used to account for a fraction of the variability (the between-block variability). The variance within the blocks is assumed to be the same, and a pooled estimate of the variance is obtained for the best test of the hypothesis. If the variances are not the same, efforts must be made (such as transformation) to ensure their uniformity.

498. Other exogenous variables will also explain or remove a certain part of the variability. For a previous example of primary productivity, it was as important to know the light conditions as it was to know the season. In fact, the two variables, season and light, may be explaining the same variability in productivity. That is, the season variability may be due to changing light, and some of the light variability is due to changing seasons. On the other hand, season may account for some variability which is not explained by light, such as changes in nutrient availability and species composition, so that light may not be able to completely explain the seasonal variability. In this example the light is a quantitative variable called a "covariable" and season is a class variable, where each season is a "block" or category of the class variable.

BIBLIOGRAPHY

* Benjamin, J. R., and Cornell, C. A. 1970. Probability, Statistics, and Decision for Civil Engineers, McGraw-Hill, New York.

* Cochran, W. G., and Cox, G. M. 1957. Experimental Designs, J. Wiley and Sons, New York.

Davis, J. C. 1973. Statistics and Data Analysis in Geology, J. Wiley and Sons, New York.

Everitt, B. 1974. Cluster Analysis, Heinemann Educational Books Ltd., London.

Gnanadesikan, R. 1977. Methods for Statistical Data Analysis of Multivariate Observations, J. Wiley and Sons, New York.

Green, P. E. 1978. Analyzing Multivariate Data, Dryden Press, Hinsdale, Ill.

Gunkel, R. C., Gaugush, R. F., Kennedy, R. H., Saul, G. E., Carroll, J. H., and Gauthey, J. E. 1984. "A Comparative Study of Sediment Quality in Four Reservoirs," Technical Report E-84-2, US Army Engineer Waterways Experiment Station, Vicksburg, Miss.

Hand, D. J. 1981. Discrimination and Classification, J. Wiley and Sons, New York.

McGill, R., Tukey, J. W., and Larsen, W. A. 1978. "Variations of Box Plots," American Statistician, Vol 32, pp 12-16.

Morrison, D. G. 1969. "On the Interpretation of Discriminant Analysis," Journal of Marketing Research, Vol 6, pp 156-163.

National Eutrophication Research Program. 1971. "Algal Assay Procedure: Bottle Test," US Environmental Protection Agency, Pacific Northwest Environmental Research Laboratory, Corvallis, Oreg.

Reckhow, K. H. 1979. "Quantitative Techniques for the Assessment of Lake Quality," EPA-440/5-79-015, US Environmental Protection Agency, Washington, DC.

Reckhow, K. H., and Chapra, S. C. 1983. Engineering Approaches for Lake Management; Vol I: Data Analysis and Empirical Modelling, Boston, Mass.

SAS. 1981. "Statistical Analysis System User's Guide: Statistics," SAS Institute, Cary, N. C.

__________. 1982. "Statistical Analysis System User's Guide," SAS Institute, Cary, N. C.

* Snedecor, G. W., and Cochran, W. G. 1967. Statistical Methods, Iowa State University Press, Ames, Iowa.

* References marked with an asterisk are introductory texts.

* Sokal, R. R., and Rohlf, F. J. 1969. Biometry, W. H. Freeman, San Francisco.

Tabachnick, B. G., and Fidell, L. S. 1983. Using Multivariate Statistics, Harper and Row, New York.

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

Velleman, P. F., and Hoaglin, D. C. 1981. Applications, Basics, and Coupling of Exploratory Data Analysis, Duxbury Press, Boston, Mass.

Walker, W. W., Jr. 1981. "Empirical Methods for Predicting Eutrophication in Impoundments; Report 1: Phase I: Data Base Development," Technical Report E-81-9, US Army Engineer Waterways Experiment Station, Vicksburg, Miss.

* Wonnacott, T. H., and Wonnacott, R. J. 1972. Introductory Statistics, John Wiley and Sons, New York.

* Zar, J. H. 1974. Biostatistical Analysis, Prentice-Hall, Englewood Cliffs, N. J.

* References marked with an asterisk are introductory texts.

APPENDIX A: GLOSSARY

accuracy--the nearness of a measurement to the actual value of the variable measured.

alpha error--see Type I error.

alternative hypothesis--the hypothesis that remains tenable when the null hypothesis is rejected.

beta error--see Type II error.

central tendency--the tendency for a majority of measurements to lie near the middle of the range of the entire set of measurements.

descriptive statistics---a means to organize and summarize data.

frequency distribution--the distribution of the total number of observations among a set of categories.

heterogeneity of variance--variances of different samples are not equal; also referred to as heteroscedasticity.

homogeneity of variance--variances of different samples are equal; also referred to as homoscedasticity.

inferential statistics--a means to make generalized conclusions, based on the inference of characteristics of the population drawn from the characteristics of the sample.

level of significance--the probability level that is considered to be too low to justify support of the hypothesis being tested.

Model I ANOVA--an analysis of variance involving fixed effects, i.e., the levels of a factor are specifically chosen.

Model II ANOVA--an analysis of variance involving random effects, i.e., the levels of a factor are chosen at random.

Model III ANOVA--a factorial analysis of variance that includes both fixed and random effects; also referred to as a mixed model.

nonparametric statistics--statistical methods that draw inferences about populations but not their parameters; since these methods do not make assumptions about the distribution of the sampled populations, they are also referred to as distribution-free statistics.

null hypothesis--statement of "no difference" (see Alternative Hypothesis).

parameter--a quantity characteristic of a population (see Statistic).

parametric statistics--statistical methods that make inferences about a population's parameters; these methods generally assume random sampling, a normal distribution, and homogeneity of variance.

population--the entire collection of measurements about which one wishes to draw conclusions; sometimes referred to as the universe or target population.

power--the probability of rejecting the null hypothesis when it is false and should be rejected.

precision--refers to the closeness of repeated measures of the same quantity.

random sample--a sample in which each member of the population had an equal and independent chance of being sampled.

robust--refers to how sensitive the validity of a given statistical test is to minor deviations from the assumptions of the best.

sample--a subset of all possible measurements of the population.

skewed distribution--a frequency distribution in which the mean and median are not identical.

statistic-- a quantity estimated from sample data; an estimate of a population parameter.

statistical hypothesis--a statement about a statistical population which one seeks to accept or reject on the basis of observed data.

statistical test--a set of rules by which the decision about a statistical hypothesis is made.

symmetrical distribution--a *frequency* distribution in which the mean and median are identical.

target population--the statistical population about which inferences are to be made based on sample data (see Population).

transformation--a mathematical operation applied to sample data to correct for a nonnormal distribution and heterogeneity of variance.

Type I error--the rejection of the null hypothesis when it is in fact true; also called an alpha error.

Type II error--the acceptance of the null hypothesis when it is in fact false; also called a beta error.

variable--a characteristic that varies from one entity to another.

END

DTIC

9-86