GUIDEBOOK FOR BIOMETRIC APPLICATIONS IN OCEAN PULSE ACTIVITIES

Sukwoo Chang and Anthony L. Pacheco

U. S. Department of Commerce
National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Northeast Fisheries Center
Sandy Hook Laboratory
Highlands, New Jersey 07732

SHL Report No. 81-18
(December 1981)

4.2 Non-Parametric Statistics

4.2.1 Wilcoxon Two Sample Rank Sum Test (Mann-Whitney U-test)
4.2.2 Kruskal-Wallis Test
4.2.3 Kolmogorov-Sminov Test
4.2.4 Correlation

4.3 Measures of Association

5. APPLICATIONS

5.1 Community Studies
5.2 Seasonal Abundance
5.3 Succession Studies
5.4 Anaerobe Analysis
5.5 Calorimetry
5.6 Physiological Activities
5.7 Parasite Analysis
5.8 Virology
5.9 Anamalies (morphology, histopathology)
5.10 Nutrient Bioassay
5.11 Pollution Uptake Studies
5.12 Genetic Studies
5.13 Petroleum Bioassay
5.14 Limiting Factors
5.15 Hydrocarbon Exposure Studies
5.16 Benthic Respiration
5.18 Primary Productivity and Phytoplankton Biomass Studies
5.19 Nutrient Studies

6. SYNTHESIS

6.1 Trend Index Interpretation

6.1.1 Determination of Indicator Parameters for Monitoring
6.1.2 Establishing Criteria of the Key Parameters for Monitoring Under Laboratory Conditions
6.1.3 Determination Correlations of the Criteria to Survey Field Data
6.1.4 Interpretation of Natural Fluctuations and Man-Induced Processes
6.1.5 Time Series Interpretation

6.2 Systems Oriented Interpretation

6.2.1 Ecosystem Change Monitoring

6.2.1.1 Food Chain and Energy Dynamics
6.2.1.2 Species Composition and Community Structure

6.2.2 Biomass Change Monitoring

7. FEEDBACK

GUIDEBOOK FOR BIOMETRIC ACTIVITIES IN OCEAN PULSE

1. Introduction:

Ocean Pulse is a program designed for continuous monitoring of and associated research necessary for forecasting the condition of coastal waters in the Northeast region. The goal is to determine the extent to which man's activities, particularly chronic and acute pollution and habitat modification, are affecting the elements of our living resources. The program provides an integrated marine environmental assessment, which an interdisciplinary base incorporating traditional and innovative measurements of resource status.

The Ocean Pulse activity will most likely proceed incorporating procedures analagous in part to that of medical science. The first phase is that of examination. The field and laboratory data, including derived indices, will describe conditions of habitats ranging inshore to offshore, and under various impacts, short term and chronic. Chemical, physiological and population indices will be examined to determine symptoms associated with various impacts. At this point statistical procedures will determine associative linkages and possibly define symptoms heretofore undescribed, (i.e. synergistic effects). The NMFS can be expected to perform in producing data files, analyses and reports including diagnoses and advise on recommendations for treatment. Treatments will come under the domain of government agencies responsible for pollution abatement. The prognosis will rest with public support.

All disciplines in Ocean Pulse collect numeric data from controlled field and laboratory experiments as well as observations and surveys at sea. The data will be accumulated as large sets of physical, chemical and biological variables to include among others physiology, pathobiology, genetics, benthos, oceanography, fisheries. The data sets will require appropriate sampling schemes throughout

the experiment or survey, a requirement particularly crucial at the planning and initial stages. They also demand a proper data mangement system to reach the objectives of monitoring and predicting. Only with an appropriate sampling scheme accommodating for some desired level of precision and analyzed with proper statistical procedures can data lead to meaningful interpretation and conclusions.

This report concentrates on general topics of statistical procedures and their direct applications in the Ocean Pulse program. However, it is worthwhile to review first the requisites for a sound experiment. Next, the selection of a sampling frame and various statistical methods both parametric and non-parametric; analysis and interpretations for the application within tasks of individual disciplines follow. Integration and synthesis for an appropriate monitoring system in terms of an ecological and environmental assessment point of view and a feedback system is described, critical for realignment of initial objectives. Lastly a data flow system is discussed.

2. Prerequisites:

In collecting basic information from samples at sea in the Ocean Pulse Program, an experiment may be defined as a directed course of action aimed at answering through scientific procedures one or more carefully framed questions. In a controlled laboratory experiment the experimenter manipulates at least some of the factors under the study and then observes the effects of his action. Suppose, for example, we have survey and laboratory measurements of biochemical enzyme responses observed under similar environmental stresses on sea scallops. Then, both should be related to each other after establishing monitoring and diagnostic criteria. We can assume the nature of the basic field and controlled laboratory data are linked by similar environmental stress. Without this linkage, there is little to say in interpreting or synthesizing results and only with it, can the experiment succeed in accomplishing the objectives of the proposed project.

There are certain characteristics an experiment must have to succeed. These are requisites of any sound experiment and to achieve these requisites, statistical design of experiments can provide some direction and an appropriate tool for soundness. These are summarized in Table 1 (Natrella, 1963). Recommended references on the general principle of experimentation are Anderson and Bancroft (1952), Cochran and Cox (1957), Cox (1958), Natrella (1963), Wilson (1952) and Yates (1960).

2.1 Establishing Objectives:

The objective is a statement in the form of questions to be answered, the hypothesis to be tested or, of the effects to be estimated. The statement should be lucid and specific. Common faults are vagueness and excessive ambition, i.e., that the program cannot be accomplished within the limitations of time, money, and availability of material, personnel, or other constraints. Establishing an objective is more than writing down a few key words or parameters. A proper setting for objectives depends on purpose, tempered by the physical restrictions on the process of taking measurements and other constraints. An objective should include an account of the range over which generalizations or statistical inferences are to be made. The objective should be described in detail, and an outline of the analysis should be constructed. Then following the details of how the experiment is to be conducted and analyzed.

As examples we should consider the following -- an initial field survey at sea designed to furnish answers to the questions of desired sample size and precision? Are the results of the controlled laboratory exposures with heavy metals to measure stress upon marine organisms similar to levels expected to be found at sea? Can results be used to explain facets of theory not adequately understood before? Are we solely interested in estimates of primary productivity around some particular sites? If not, how will tests of significance be determined for links in determining trophic food chain dynamics?

Once we establish the objectives and decide what we are going to do in the experiment, then observations through some sampling system will provide an estimate of the population being studied. Our ultimate goal is to have small variance (experimental error), bias (systematic error), with mean estimates about the same as the true value. The extent of bias and variance in the experiment are to a large extent independent. We can have estimates having small variance, i.e. differ little among themselves, but with a large bias, so that all the estimates differ greatly from the true value. Bias may arise from a poor method of analysis, but more likely from a poor choice of samples, or from the method from which the measurement or counts are made from the sample.

If the size of replicates or samples increases, then the variance will be reduced, but the bias will remain unchanged. This leads to the discussion below on replication. However, the bias can only be detected and hence eliminated by careful examination of the whole sample procedures from beginning to end and must utilize the concept of randomization.

2.2 Replication:

It is seldom that only one observation in an experiment is regarded as sufficient. Repetitions are considered desirable to confirm results and to form a basis for estimating precision. The precision is concerned merely with repeatability of measurements. This process of replication is especially necessary when the parameter under study is not precisely defined, and is subject to wide variations. When this applies, large numbers for testing may be required, but it is also desirable to make check runs to determine the experimental errors (random errors).

Three main sources of experimental errors may be distinguished. The first is inherent or intrinsic variability in the experimental material to which the treatment are applied. The second is lack of uniformity in physical conduct of the experiment, i.e. failure to standardize the experimental technique. Third is the size of the experiment, in the sense of either providing replicates or including additional treatments.

Whatever the source of the experimental error, replication of an experiment steadily decreases the associated error. But precautions have been taken to ensure that one treatment or factor is no more likely to be favored in any replicate than another, so that the errors affecting any treatment tend to cancel out as the number of replications is increased. The rate at which the experimental error is reduced is predictable from statistical theory. One should avoid two common mistakes: 1) require more precision than the purpose warrants, and 2) obtaining insufficient precision for the purpose. In the first mistake, the experiment will cost more money than is necessary. In the second mistake, the experiment fails to achieve significance.

The basic quantity used to measure experimental errors is the error variance per experimental unit, which is defined as the expected value of the square of the error that affects the observations for a single experiment unit. The square root of this quantity is called the standard error per unit, i.e.

$$\text{standard error} = \frac{\text{error variance per unit}}{\text{no. of replicates}}$$

Hence, to estimate the number of replicates, we need only the error variance per unit (which is usually obtained from the analysis of variance) and the desired or required standard error (precision). Further readings for this are in Cochran and Cox (1957 and Cox (1958).

2.3 Randomization:

One way to eliminate bias is the use of the principle or randomization. The use of a strictly random choice (not some process such as guessing numbers which the experimenter perceives as random), has two aims. The first is the essential one of ensuring that the inevitable prejudices and preferences of the experimenter do not bias the experiment. The second aim is to provide a mathematically sound bias for calculation of approximate probability of error, as well as a statistically meaningful inference for interpretation of the results.

The basic operation of randomization is that of arranging in random order a series of numbered objects. In the more complicated designs this process must be applied several times. An essential feature of randomization is that it be an objective impersonal procedure. Arranging things in random order does not mean just a manipulation into some order that looks haphazard. Methods of randomizing include rolling dice, shuffling numbered cards or drawing numbered balls out of a well-shaken bag. The main method is the use of numerical random tables. It is used as follows: choose a starting point without looking at the tables. For example, write down a number for the page, a number for the row, and a number for the column block. Similarly we can also choose multi-digit random numbers according to the experimental unit or treatment for which we want to establish a random order.

The positive advantages of randomization are assurances that a randomized experiment is more accurate than a corresponding nonrandomized one in which an unskillful assignment to treatments to units leads to systematic bias. Randomization can prevent human bias from entering in the selection of the sample and in making the assignment of treatments or observations. It also assures that the random error of the estimated treatment effects can be measured and their level of statistical significance examined. The concept of randomization was introduced by R. A. Fisher and further readings are in Fisher (1947) and Cox (1958).

Table 1. Some requisites and tools for sound experimentation.

Requisites	Tools
1. The experiment should have carefully defined objectives.	1. The definition of objectives requires all of the specialized subject-matter knowledge of the experimenter, and results in such things as: (a) Choice of factors, including their range; (b) Choice of experimental materials, procedure, and equipment; (c) Knowledge of what the results are applicable to.
2. As far as possible, effects of factor should not be obscured by other variables.	2. The use of an appropriate experimental pattern helps to free the comparisons of interest from the effects of uncontrolled variables, and simplifies the analysis of the results.
3. As far as possible, the experiment should be free from bias (conscious or unconscious).	3. Some variables may be taken into account by planned grouping. For variables not so taken care of use randomization. The use of replication aids randomization to do a better job.
4. Experiment should provide a measure of precision (experimental error).	4. Replication provides the measure of precision; randomization assures validity of the measure of precision.
5. Precision of experiment should be sufficient to meet objectives set forth in requisite 1.	5. Greater precision may be achieved by: Refinements of technique; experimental pattern (including planned grouping); replication.

3. Sampling

Sampling is a method that guides quantitative studies of content, behavior, performance, material and causes of differences. Every sampling system is used to obtain estimates of certain estimates of certain measurements or properties of the population being studied, and the system can be judged by how good the estimates obtained are in the sense of minimizing errors and bias. A good system provides a frequency distribution with a small variance and bias with the estimated mean close to the true value. The requirements for precision and randomization have to be fulfilled.

To extend valid generalizations from samples about characteristics of the population in which we are interested, the samples must have been obtained by a suitable sampling scheme. Such a scheme ensures two basic conditions: 1) all possible samples associated with the sampling scheme must bear a known relation to the characteristics of the population (if the population is small, it is sometimes convenient to obtain the information by collecting the data for the whole of the population); 2) generalizations may be drawn from such samples in accordance with the validity of the mathematical theory of probability. If a sampling scheme is to meet these two requirements, it is necessary that the selection of the individuals to be included in a sample involve some type of random selection, that is, each possible sample must have a fixed and determinate probability of selection.

There are excellent reference books for sampling methods. Yates (1960), is more practical and readable than some of the popular ones, and contains a list of references over all disciplines. For fisheries and marine science, recent publications are available; for instance, Gulland (1966, fisheries biology), Gonor and Kemp (1978, quantitative ecology), Stofan and Grant (1978, phytoplankton), Jacobs and Grant (1978, zooplankton), Swartz (1978, macrobenthos), Mearns and Allen

(1978, small otter trawls), Grosslein (1970, groundfish survey), Saila (1900, sequential sampling for benthos). Excerpts from Grosslein (1970) appear in Appendix I.

3.1 Simple Random Sampling:

The most useful type of selection is simple random sampling. This type of sampling is defined by the requirement that each individual in the population has an equal chance of being the first member of the sample; after the first is selected, each of the remaining individuals in the population has an equal chance of being the second member of the sample; and so forth. For simple random sampling, it is not sufficient that each individual in the population has an equal chance of appearing in the sample, but it is sufficient that each possible sample has an equal chance of being selected.

A useful and widely applicable method of obtaining a truely random sample is by use of random numbers as described earlier. The individuals in the population from which a sample is to be drawn are allotted numbers, and those to be sampled are determined by reference to a table of random numbers. For instance, if a sample of 10 clams or fish has to be taken from a population of 100, and the first 10 random numbers may be, say, 57, 21, 79, 29, 45, 86, 3, 17, 18, and 93, the individuals corresponding to those numbers will be selected. If the number of individuals in the population is not exactly 100, some random numbers occurring will not correspond to numbers to be discarded. For example, if we want to have a sample of 10 from 24 fish, we consider only random numbers ranging from 1 to 24. Two or more digits may be ascribed to each individual, so that the first unit has, for instance, numbers 01 to 04, the second, 05-08 and so forth, the 24th has 93-96, and numbers 97-100 are not used. Or instead of selecting all the units in the sample individually from the random number

table, units may be taken at regular intervals systematically, e.g. every third or seventh of which the first one is chosen by random number. In other words, if the randomly chosen number is three, then we choose for the sample every third individual to reach the required sample size.

If a randomization process is not employed, then it is likely that all individuals in the population will not have equal chances of selection in the sample. If we just "grab a handful" the individuals in the handful almost always resemble one another on the average more than do the members of a sample chosen with randomization process. Cochran, Mosteller and Tukey (1954) pointed out that a "grab" sample tends to underestimate the variability in the population. We should have to overestimate it to obtain valid estimates of variability of "grab" sample means by substituting such an estimate into the formula for variability of means of simple random samples. Thus, using simple random sample formulae for "grab" sample means introduces a double bias, both parts of which lead to an unwarranted appearance of higher stability.

Now suppose that we draw a sample of n units from a population of N units and these units from 1 to n in order of which they are selected. Then a sample of n independent random individuals is taken with values $x_1, x_2, \ldots x_n$, the resulting estimate of the mean value per unit in the population is:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the variance of $\bar{x}$ is expressed by:

$$\text{var}(\bar{x}) = \frac{N-n}{N}\,\frac{s^2}{n}$$

where s^2 = sample variance $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

The factor of $\frac{N-n}{n}$ is derived from the basic sampling scheme without replacement, and for further details and proof, one should consult with Cochran (1977, p. 23).

The above precision of a sample estimate (variance of the mean estimate) or standard error is a measure of absolute error. However, if we deal with precision of a standard error of the estimate over the value being estimated (symbolically expressed by $var(\bar{x})\ /\bar{x}$), then it is expressed in terms of a relative precision of a sample estimate. It is referred to as "the coefficient of variation". Yates (1960) supported the formula for the sample size determination in a random sample as:

$$n = \frac{(\ var(\bar{x})/\bar{x})^2}{(\text{desired sample error in the sample})^2}$$

Calculations of biochemical data using this are given in Appendix II.

3.2 Stratified Random Sampling:

In stratified random sampling, the population is subdivided into groups or "strata" before selection of the sample. These strata may either all contain the same number of units or differing numbers of units. If a uniform sampling fraction is used, the same fraction of the units of each stratum is included in the sample, the units selected being chosen at random from all the units within each stratum. A stratified sample is thus equivalent to a set of random samples on a number of subpopulations, each equivalent to one stratum.

The increase in precision and bias reduction of sample estimates accomplished by stratification depends on the degree of homogeneity that is achieved within strata. In other words, the amount of the variability in the characteristic being estimated is reflected in the differences among the strata. This in turn depends on how effectively strata have been defined.

In establishing a stratum, all information could help classify members of the population into groups which differ from one another with respect to the characteristic being measured or with respect to the cost of collecting data. Each stratum is then sampled independently, and estimates obtained for each stratum.

These can then be combined to give the estimate for the whole population. The variance of this estimate will also be obtained by combining the variances of the estimates within the individual strata. Since the strata are relatively homogeneous, the variance within strata will tend to be small, and possibly the variance of the combined estimate will be smaller than the variance in the population as a whole. This is the rationale for employing stratification procedures in the sampling.

The following steps are required for the stratified random sampling scheme: 1) defining the strata to be utilized; 2) determining the size of sample to be taken from each stratum; 3) selecting the sample from the strata as defined; 4) calculating the estimate from the sample; and 5) evaluating the reliability of the sample estimate with variance estimates.

Suppose the population consists of N individuals, N_i is the i^{th} stratum where $N = \sum_{i=1}^{I} N_i$, and a sample of $N_1, N_2, \ldots, N_I$ units are taken from the I strata respectively. Let X_{ij} be j^{th} values of quantity in i^{th} stratum to be estimated (e.g. length of fish, amount of enzyme, etc.) with $j = 1, 2, \ldots n_i$. The estimated mean value $\bar{x}_i$ in the stratum is:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

and an unbiased estimate of the mean value in whole population is given as the weighted mean of the means of the individual strata (the weighting factor being the total numbers in each stratum)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{I} N_i \bar{x}_i$$

If the variance within the i^{th} strata is an extension of simple ramdom sampling

$$\text{var}(\bar{x}_i) = \frac{1}{n_i} \frac{(N_i - n_i)}{N_i} s_i^2$$

where $s_i^2 = \dfrac{\sum_j (x_{ij}-x_i)^2}{n_i-1}$

then we have an unbiased estimate of the variance of $\bar{x}$ for the overall strata expressed by:

$$\text{var}(\bar{x}) = \frac{1}{N_2} \sum_{i=1}^{I} N_i^2 \, \text{var}(\bar{x}_i)$$

$$= \frac{1}{N_2} \sum_{i=1}^{I} N_i \left[\frac{1}{n_i} \frac{(N_i-n_i)}{N_i} s_i^2\right]$$

$$= \frac{1}{N_2} \sum_{i=1}^{I} N_i \, (N_i-n_i) \frac{1}{n_i} \, s_i^2$$

To determine the sample size in a stratified sampling scheme, the values of the sample size n_i in the respective strata are expressed by Neyman (1934)

$$\frac{n_i}{n} = \frac{N_i \, \text{var}(\bar{x}_i)}{\sum_{i=1}^{I} N_i \, \text{var}(\bar{x}_i)}$$

Although the above equation give the n_i in terms of n, we do not know what n has. The solution depends on whether the sample is chosen to meet a specific or desired variance of the stratified mean (v). If v is fixed, and we substitute the optimum n_i in the formula for var($\bar{x}$), then we have an optimum allocation of n as

$$n = \frac{\sum_i (\frac{N_i}{N})^2 \, [\text{var}(x_i)]^2}{V + \frac{1}{N} \sum_i (\frac{N_i}{N}) \, [\text{var}(\bar{x}_i)]^2}$$

Suppose we minimize the variance of the estimate $\bar{x}$, var($\bar{x}$), for a specified cost of taking the sample or to minimize the cost for a specified value of var($\bar{x}$). The simple cost function is of the form

$$\text{cost} = C = C_0 + \sum_i C_i \, n_i$$

where C_0 represents an overhead or initial cost for a sampling scheme, and C_i is cost per unit varying from stratum to stratum so that the cost is proportional to the size of sample.

Then, the optimum size of sample is:

$$n = \frac{(C-C_o) \sum_i [N_i \ var(\bar{x}_i)/\sqrt{C_i}]}{\sum_i N_i \ var(\bar{x}_i) \sqrt{C_i}}$$

and

$$n = \frac{\sum_i \frac{N_i}{N} var(\bar{x}_i) \sqrt{C_i}] [\sum_i \frac{N_i}{N} var(\bar{x}_i)/\sqrt{C_i}]}{\bar{V} + \frac{1}{N} \sum_i \frac{N_i}{N} var(\bar{x}_i)}$$

Further readings in detail for the optimum allocation problems and the sample size determination in the stratified random sampling scheme are referred to in books by Cochran (1977) and Hansen, Hurwitz and Hadow (1953). The applications for NMFS groundfish survey and its variability estimates with the stratified random sampling method are referred to in Grosslein (1971) and Hennemuth (1976). An interesting application for the structure of New York Bight benthic data using post-collection stratification of samples based on the physical characteristics of each grab sample rather than classical spatial strata classification is given by Walker, Saila and Anderson (1979). Excerpts of this paper are given in Appendix III.

4. Statistical Methodology:

Modern statistics provides research workers with knowledge. However, the extent of statistics makes it difficult to define. It was developed to deal with those problems where, for the individual observations, laws of causes and effect are not apparent to the observer and where an objective approach is needed. In such problems, there must always be some uncertainty about any inference based on a limited number of observations. Hence, statistics is the science, pure and applied, of creating, developing, and applying techniques such that the uncertainty of inductive inference may be evauated.

4.1 Parametric Statistics

A parameter is a measure of some characteristic of a statistical population. For example, the mean and the variance are two such measures which occur in a normal (bell-shaped) distribution. Statistical methods which rest on particular assumptions about the forms of distribution and their parameters are called parametric methods. The most frequently assumed distribution form is normal. For many years the normal distribution has established a pre-eminent position in statistical theory. It deserves its position on two grounds. First, a large number of variables, including sample statistics such as means, appear to be distributed normally or nearly so. Second, non-normal distributions often can be readily transformed to normal form.

4.1.1 Linear Regression Analysis

4.1.1.1 Simple Regression

We consider the problem of statistical inferences which can be made regarding the variability of a dependent variable, y, relative to an independent variable, x. The y's can fluctuate from sample to sample, for example the measurements of fish physiological stress, y (e.g. enzyme level) are affected by the amount of contaminants, x. Furthermore, the x's will also be variable

subject to random fluctuation. As another example, we may wish to examine the rate of primary productivity, y for different environmental variables, x of nutrients observed.

Regression has many uses. Perhaps the objective is only to learn if y depends on x; or prediction of y from x may be the goal. Some wish to determine the shape of the regression curve. Others are concerned with the error in y in an experiment after adjustment has been made for the effect of a related variable x. If you have a theory about cause and effect, employing regression can test this hypothesis.

To satisfy these various needs an extensive account of regression method is required. If a variable y is a linear function of a variable, x, we may have

$$y = \alpha + \beta x + \varepsilon$$

where ε represents some residual or random errors, the amount of y not accounted for in the regression on line of y on x. We postulate that the regression line is selected so that residuals are of a random nature and uncorrelated with each other, with a usual added assumption that the ε are normally distributed with mean 0 and variance ^{2} (Figure 1). Suppose we consider both variables (x and y) are subject to an error measurement which has a joint probability distribution at (x_1, y_1). It is represented by the "mounds" centered over the true point (Figure 2). Similarly, the points (x_2, y_2) and $x_3, y_3)$ are demonstrated.

To estimate the relationship between the y and x variate, n simultaneous observations will be obtained on y and x, i.e.:

$$y_1, y_2, \ldots y_n$$

$$x_1, x_2, \ldots x_n$$

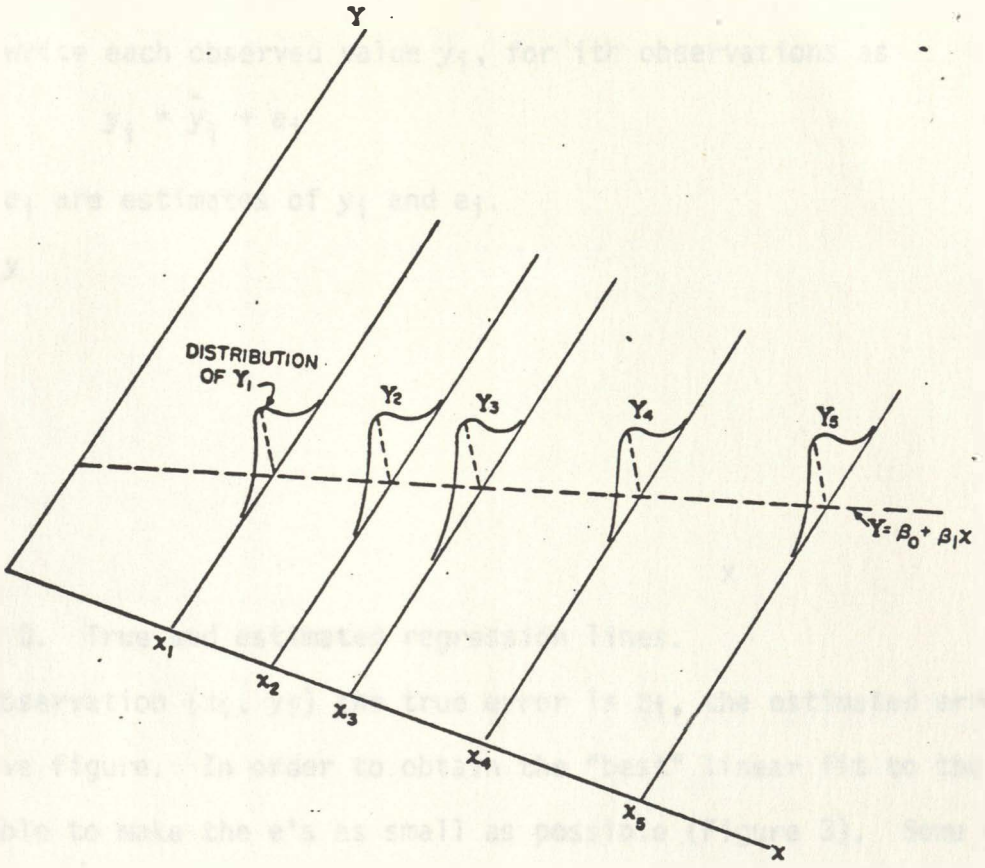

Y
DISTRIBUTION OF Y_1
Y_2
Y_3
Y_4
Y_5
$Y = \beta_0 + \beta_1 X$
X_1
X_2
X_3
X_4
X_5
X

Fig 1

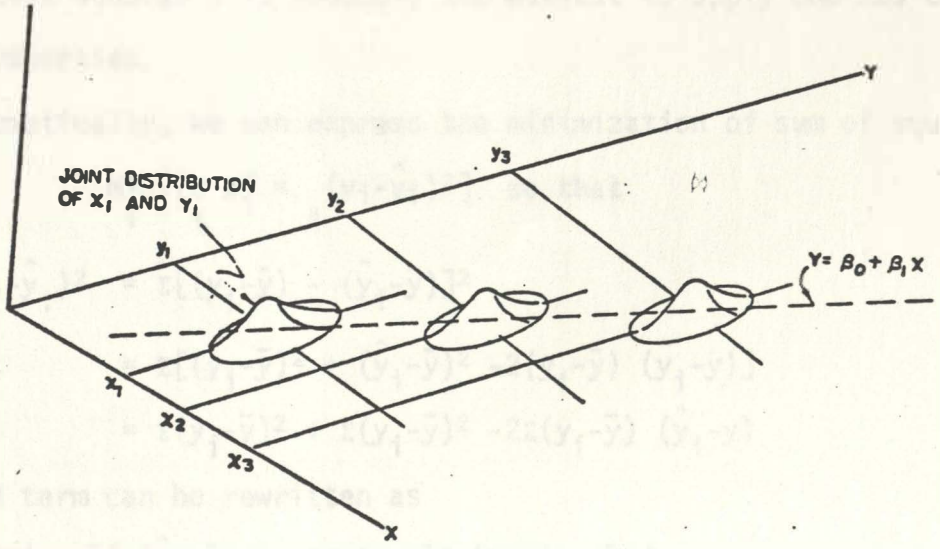

Y
JOINT DISTRIBUTION OF X_1 AND Y_1
y_3
y_2
y_1
$Y = \beta_0 + \beta_1 X$
X_1
X_2
X_3
X

Fig 2

Then, we can write each observed value y_i, for ith observations as

$$y_i = \hat{y}_i + e_i$$

where $\hat{y}_i$ and e_i are estimates of y_i and e_i.

y

x

Figure 3. True and estimated regression lines.

For a given observation (x_i, y_i) the true error is Σ_i, the estimated error by e_i in the above figure. In order to obtain the "best" linear fit to the data, it is reasonable to make the e's as small as possible (Figure 3). Some choices are available to make the e's small;

1) minimize the sums of the absolute values of the e

2) minimize the sum of squares of the e. Method (2) (Called the "method of least squares") is probably the easiest to apply and has certain optimum properties.

Mathematically, we can express the minimization of sum of squares of the e's as $\underset{i}{\text{Min.}}[\Sigma_i e_i^2 = \Sigma_i (y_i-\hat{y}_i)^2]$ so that

$$\Sigma(y_i-\hat{y}_i)^2 = \Sigma[(y_i-\bar{y}) - (\hat{y}_i-\bar{y})]^2$$

$$= \Sigma[(y_i-\bar{y})^2 + (\hat{y}_i-\bar{y})^2 - 2(y_i-\bar{y})(\hat{y}_i-y)]$$

$$= \Sigma(y_i-\bar{y})^2 + \Sigma(y_i-\bar{y})^2 - 2\Sigma(y_i-\bar{y})(\hat{y}_i-y)$$

The third term can be rewritten as

$$-2\Sigma(y_i-\bar{y})(\hat{y}_i-\bar{y}) = -2\Sigma(y_i-\bar{y})[b\,\Sigma(x_i-\bar{x})]$$

It is because, for example, in the case of simple regression (one y and one x are linearly related), i.e.

$$y_i = \alpha + \beta\, x_i + \Sigma_i$$

$$\hat{y}_i = a + b\, x_i \quad \text{and since } a = \bar{y} - b\,\bar{x}$$

$$= (\bar{y} - b\bar{x}) + b\, x_i$$

$$= \bar{y} + b\,(x_i - \bar{x})$$

$$\text{and } \hat{y}_i - \bar{y} = b(x_i - \bar{x}) \text{ ....*}$$

One step further, we know

$$b = \frac{\Sigma(x_i - \bar{x})\ (y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$b\ \ \Sigma(x_i - \bar{x})^2 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) \text{ ....**}$$

So, the third term of the original equation is expressed by:

$$-2\ \Sigma(y_i - \bar{y})(\hat{y}_i - \bar{y})$$

$$= -2\ \Sigma(y_i - \bar{y})[b\Sigma(x_i - \bar{x})] \text{ .....(by *)}$$

$$= -2b\ \Sigma(y_i - \bar{y})(x_i - \bar{x})$$

$$= -2b^2\Sigma(x_i - \bar{x})^2 \text{ ....(by**)}$$

$$= -2\Sigma(y_i - \bar{y})^2 \text{ ....(by*)}$$

Thus, we have

$$\Sigma e_i^2 = \Sigma(y_i - \hat{y}_i)^2$$

$$= \Sigma(y_i - \bar{y})^2 + \Sigma(\hat{y}_i - \bar{y})^2 - 2\Sigma(\hat{y}_i - \bar{y})^2$$

$$= \Sigma(y_i - \bar{y})^2 - \Sigma(y_i - \bar{y})^2$$

Rearrange the terms, we have

$$\Sigma(y_i - y)^2 = \Sigma(y_i - \hat{y})^2 + \Sigma(y_i - \bar{y})^2$$

i.e. (sum of squares about the mean) = (sum of squares about regression) + (sum of squares due to regression)

We can construct an analysis of variance (ANOVA) table in Table 2, which indicates that the mean square error (MSE)

$$MSE = \frac{\text{ss about regression}}{n-2}$$

is an unbiased estimate of σ^2 and relate how to test of regression (test of regression slope; H_o:β=0).

It is also possible to obtain the exact probability distribution on the estimates $\hat{y}_i$, a, b, and thus, the confidence intervals for the predicted line and the individual parameters. However, we will skip a detailed analysis. You should consult one of the following: Anderson and Bancroft (1952), Draper and Smith (1966), Natrella (1963), Snedecor and Cochran (1967), and Steel and Torrie (1960). Ricker (1973) is an excellent reference for details of the functional relation of linear regress-ons in fisheries research problems, particularly in cases where x and y are both subject to random fluctuations.

The comprehensive ANOVA table and other statistics for the simple regression are obtained by an ADP computer program (Dahlberg, 1969). The biomedical computer program, BMDolR (Dixon, 1974) provides similar output for analyses, but has an output difficult to read. Dahlberg's program provides the plots of regression and the analysis of the weighted regression if you have to utilize computational weight factors.

To obtain a degree of confidence that the relationship is indeed linear, a test of deviation from linearity may be derived. We must have more than one value (n_i) of y's for a given value of x_j.

$y_{11}, y_{12}, \ldots, y_{1n_1}$ one n_1 repeat observation of x_1

$y_{21}, y_{22}, \ldots, y_{2n_2}$ are n_2 repeat observation at x_2

$\vdots \quad \vdots \quad \vdots$

$y_{r1}, y_{r2}, \ldots, y_{2,n_r}$ are n_r repeat observation at x_r

Then we can subdivide the quantity of sum of squares about regression (ss error) to two terms of pure error and "lack of fit" sum of squares. The mean square for pure error is expressed by:

$$S_e^2 = \frac{\sum_{i=1}^{y} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{r} n_i - r}$$

Then, the "lack of fit" sum of squares are obtained by subtraction of s_e^2 from ss about regression, i.e.

ss about regression - s_e = lack of fit sum of squares

The test of linearity is:

$$F^* = \frac{\text{ss lack of fit}/r-2}{s_e^2 / \sum_i n_i - r} \sim F_{(r-2,\ \Sigma n_i - r)}.$$

If the test is rejected (i.e. F* statistics is greater than F table value with r-2 and $\Sigma n_i - r$ degree of freedom) the linear regression model appears to be inadequate. If the test is not rejected, the model presents no reason to doubt the adequacy of linearity, and both pure error and lack of fit mean squares can be used as estimates of σ^2. You should consult Draper and Smith (1966) for details.

4.1.1.2 Multiple Linear Regression Analysis

Often it is more realistic and economical to be concerned with the joint effects of a number of independent variables ($x_i, \ldots x_r$) on a single dependent variable y rather than examining each variable separately. As in the simple regression, the simplest and most used functional relationship is a linear one. The multiple linear regression model has the form of:

$$y_i = \alpha + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \ldots + \beta_r x_{r_i} + \varepsilon_i$$

for $i = 1, 2, \ldots n$ and where ε_i follows normal distribution with mean 0 and variance σ^2, and ε_i and ε_j are uncorrelated each other.

The application of the least square technique is the same as described for simple regression. We have to minimize

$$\sum_{i=1}^{r} \sum_{j=1}^{n} (y_{ij} - \hat{y}_{ij})^2$$

The ANOVA table is summarized in Table 3. The formulae for estimating parameters, testing, and confidence limits are omitted, but one should consult Afifi and Azen (1972) or Draper and Smith (1968).

In many regression situations, the experimenter does not have sufficient information about the order of relative importance of the independent variables $x_1, x_2, \ldots x_r$ in predicting the dependent variable y. Testing a hypothesis: $\beta_i = 0$ for each x_i, $i = 1,2,\ldots n$ does not reveal this ordering. Suppose we reject the test on false conclusions that x_1 was the only variable of importance in predicting y.

Then, our question is, which x variables are most important in determining and predicting y. Usually no unique or fully satisfactory answer can be given, but a few approaches have been tried: 1) standard partial regression coefficient (see Snedecor and Cochran, 1967); 2) multiple correlation coefficient (see next section), and 3) stepwise regression procedures (see Afifi and Azen, 1972 and Draper and Smith, 1968).

The solution for the stepwise regression selects a single variable x_i which best predicts y. The second step finds the variable x_j which best predicts y, with the given x_i, the first variable entered. In the steps that follow, either: 1) a variable is entered which best improves the prediction of y given all variables entered from the previous steps; or 2) a variable is

removed from the set of predictors if its predictive ability falls below a given level. The process is terminated when no further variable improves the prediction of y.

The computation of the stepwise regression is obtained by computer program BMD02R (Dixon, 1974).

4.1.2 Correlation Analysis

4.1.2.1 Simple Correlation

In its most general sense correlation denotes the interdependence between quantitative or qualitative data. However, in a more restricted sense we will consider correlation as a measure of the degree of relationship between the variables, independent of the units or terms in which they are originally expressed. A closely related measure may permit you to state the relative amount of variation which is explained by the estimating regression equation. Recalling the expression of sum of squares (SS) in the previous section, the fraction of SS due to regression is expressed by SS about the mean. This is called the coefficient of determinations in the regression analysis, i.e.

$$r^2 = \frac{\text{SS due to regression}}{\text{ss about mean}}$$

$$= \frac{\Sigma\ (\hat{y}_i - \bar{y})^2}{\Sigma\ (y_i - \bar{y})^2}$$

This coefficient is equal to the ratio of the reduction in the sum of squares of deviations obtained by using the linear regression to the total sum of squares of deviations about the sample mean y, which would be the predictor of y if x were ignored. This provides a more meaningful interpretation of the strength of the relation between y and x than the Pearson product moment, the coefficient of correlation is:

$$r = \frac{\Sigma\ (x_i-\bar{x})\ (y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\ \sqrt{\Sigma(y_i-\bar{y})^2}}$$

Squaring both sides,

$$r2 = \frac{[\Sigma(x_i-\bar{x})\ (y_i-\bar{y})]^2}{\Sigma(x_i-\bar{x})^2\ \Sigma(y_i-\bar{y})^2}$$

Dividing by $\Sigma(x_i-x)^2$ or $\Sigma(y_i-y)^2$ we have

$$r = \frac{[\Sigma(x_i-x)\ (y_i-y)]/\cdot\Sigma(x_i-x)^2}{\Sigma(y_i-y)^2}$$

$$= \frac{\text{reduction in SS (y) attributable to x}}{\text{SS (y) about mean = corrected total SS (y)}}$$

$$= \frac{[\Sigma(x_i-x)\ (y_i-y)]^2/\Sigma(y_i-y)^2}{\Sigma(x_i-x)^2}$$

$$= \frac{\text{reduction in SS (x) attributable to y}}{\text{SS (x) about mean = corrected total SS (x)}}$$

In addition,

$$r^2 = \left[\frac{\Sigma(x_i-x)\ (y_i-y)}{\Sigma(x_i-x)^2}\right]\ \left[\frac{(x_i-x)\ (y_i-y)}{\Sigma(y_i-y)^2}\right] = (b_{yx})(b_{xy})$$

where b_{yx} and b_{xy} are the regression coefficient slopes for the regression of y on x and of x on y. Thus, the product of the regression coefficient is the square of the correlation coefficient, inversely the correlation coefficient is the square root of the product of the regression slopes or their geometric mean. Hence, if we are interested in testing whether there is a linear relationship between x and y (Ho: $\alpha = 0$ where α is population correlation coefficient) a statistical test is available (Snedecor and Cochran, 1967 and Steel and Torrie, 1960). In fact, this test is equivalent to testing that the hypothesis $\beta = 0$. While r provides a nice measure of the goodness of fit of the least squares line to the fitted data, its use in making inferences concerning p would seem to be of dubious value in many situations. This is simply because it is unlikely that a phenomenon y observed in natural science, especially marine environmental science, would be a function of a single variable x. The larger reduction in

SS about regression (SS error) could possibly be obtained by constructing a predictor of y based on a set of variables $x_1, x_2 \ldots$. It leads to multiple and partial correlation which will be described below.

A few reminders concerning the interpretation of r are worthwhile.

1) if r = 0.6 as indicative of a linear relation between x and y, this value 0.6 would imply that use of x in predicting y reduces the sum of squares of deviation about the prediction line by only $r^2 = 0.36$ or 36 percent;
2) r is a measure of linear correlation and x and y could be perfectly related in some curvilinear function when the observed value of r is even very low;
3) if the linear correlation coefficients between y and each of two variables x_1 and x_2 were calculated 0.6 and 0.7 respectively, it does not follow that a predictor y using both variables would account for a $(0.6)^2 + (0.7)^2 = 0.85$ or an 85 percent reduction in the sum of squares of deviation. Actually x_1 and x_2 might be highly correlated and therefore contribute the same information for the prediction of y;
4) detecting linear correlation visually from plotted points can be difficult. An unfortunate choice of scale may hide a real correlation or indicate a real one when none is present. A change of scale will also change the slope of regression line. Further with an unfortunate choice of scale, visual detection is further hindered by the fact that the relation between r and r^2 (proportion of the total sum of squares expressed by regression) is not linear;
5) the correlation coefficient is considered only when variables x and y are both subject to random errors.

4.1.2.2 Multiple Correlation

The simple correlation may not be what is desired in situations where the dependent variable is influenced by two or more independent variables. Multiple correlations provides an analysis of the relations among two or more predictor measures. It measures the closeness of representation by the regression plane and may also be regarded as the maximum of the correlation coefficient between the dependent variable and all linear functions of a set of two or more of independent variables. The coefficient is usually denoted by R but is regarded as essentially non-negative; the quantity R^2 being the one which occurs in practice as r^2 in simple correlation.

$$R^2 = \frac{\text{SS due to regression}}{\text{SS about mean}}$$

$$= \frac{\Sigma(y_i-\bar{y})^2}{\Sigma(y_i-\bar{y})^2}$$

Multiple correlation coefficients are strictly applicable only when the total observation, that is $(y_i, x_{1i}, x_{2i} \ldots x_{pi})$ is subject to random error as we have noted in the case of simple correlation. However, regardless of randomness of the observations, these correlation coefficients may be useful for computing and for other reasons. The reminders given for simple correlation coefficients are all valid for multiple correlation coefficients. Recommended readings for multiple correlation are Steel and Torrie (1960) and Kendall (1961).

The computations of multiple correlation coefficients are obtained through computer program BMD02R (Dixon, 1974), stepwise regression analysis.

4.1.2.3 Partial Correlation

The simple correlation and multiple correlation coefficients are measures of the closeness represented by the regression line or plane, i.e. measures between two or more variables. This consideration leads us to examine the correlations between variables when other variables are held as constant, i.e. conditionally upon those other variables taking certain fixed values. These are so-called partial correlations.

Suppose there are three variables. Then we have three simple correlation coefficients among variables: variables 1 and 2, r_{12}; 1 and 3, r_{13} and 2 and 3, r_{23}. The partial correlation is expressed as the correlation between variables 1 and 2 in a cross section of individuals all having the same values of variable 3, $r_{12}(3)$, i.e. the variable is held constant over variables 1 and 2 which are involved in the correlation coefficient computation.

When we come to interpret a measure of interdependence, we often meet difficulties, as when the first variable is correlated with second variables. This may be merely incidental to the fact that both are correlated with another variable or set of variables. This consideration leads to an examination of the partial correlation. If we find that holding the third variable fixed reduced the correlation between two variables, we make the inference that their interdependence arises in part through the agency of a third variable. If the partial correlation coefficient ($r_{12}(3)$) is very small, we infer their interdependence is entirely attributable to that agency, and conversely if the partial correlation is larger than the original simple correlation coefficient (r_{12}) as a measure of dependence between variables, then we make the inference that the third variable was obscuring the stronger correlation or making the correlation.

A useful identity between the partial and multiple correlation coefficients for the set of variables $(y, x_1, x_2 \ldots x_p)$ is

$$1-R^2_{y,x_1,x_p} = (1-r^2_{yx_1})\,(1-r^2_{yx_2(x_1)})\ldots(1-r^2_{yxp(x_1x_2\ldots x_p)})$$

where R^2_{y,x_1,x_p} is multiple correlation coefficient between variable y and $x_1,\ldots,x_p$, $r^2_{yx_2}(x_1)$ and $r^2_{yxp}\,(x_1, x_2,\ldots x_{p-1})$ are the partial correlation coefficients between y and x_2 when x_1 is held as constant, and of y and x_p when other $x_1, x_2\ldots x_{p-1}$ variables are held as constant. For instance, in the above three variable case, we have

$$1 - R^2_{1,2,3} = (1-r^2_{12})\,(1-r^2_{12(3)})$$

where

$$r_{12} = \frac{\Sigma(x_{1i}-\bar{x}_1)\,(x_{2i}-\bar{x}_2)}{\sqrt{\Sigma(x_{1i}-\bar{x}_i)^2}\;\sqrt{\Sigma(x_{2i}-\bar{x})^2}}$$

$$r_{12(3)} = \frac{r_{12}-r_{13}\,r_{23}}{\sqrt{(1-r_{13})^2\,(1-r_{23})^2}}$$

A test of significance of the partial correlation coefficients, e.g. $r_{12(3)}$, is available (Snedecor and Cochran, 1967 and Afifi and Azen, 1972). BMDQ2R (Dixon, 1974) stepwise regression analysis provides computations of the partial correlation coefficient. Utilization of the stepwise regression analysis are referred to in Draper and Smith (1966) and Afifi and Azen (1972).

4.1.3 Multivariate Analysis

As we have seen in the section of multiple regression and correlation, observations on more than one random variable may be made for each individual in the sample. The multivariate analysis is used rather loosely to denote the analysis of data which are multivariate in the sense that each member bears the values of p variables. In regression problems emphasis is placed upon the relationship between the dependent variable on one hand and the set of independent variables on the other hand. In other multivariate analyses, however, all of the random variables are analyzed simultaneously as a random vector having a multivariate distribution. Some multivariate methods are a generalization of the univariate method, while others are unique to multivariate analysis.

Most of the continuous multivariate analyses assume that the underlying distribution of the random vector is a normal multivariate. The justification of this assumption, similar to those in the univariate case, are: 1) many observable phenomena follow an approximate multivariate normal distribution; 2) transformations of some or all of the components of the random vector sometimes induce a multivariate normal distribution; and 3) the central limit theorem for one random variable extends to the multivariate case, that is, summations of many independent and identically distributed random vectors approach multivariate normality.

Anderson (1958) classifies the multivariate analysis into the following categories:

1. correlation (multiple and partial correlation analysis);
2. analogues of univariate statistical analysis (multiple regression, multivariate analysis of variance, generalized T^2-test for discriminant function analysis;

3. problems of coordinate systems (principal components analysis, canonical correlation analysis);
4. more detailed problems (factor analysis);
5. dependent observation (time series problems with serial correlation analysis).

We will cover some of the selected topics including discriminant function analysis, principal component analysis and canonical correlation analysis. Correlation analysis and multiple regression analysis were discussed earlier.

4.1.3.1 Discriminant Function Analysis

Discriminant analysis is a procedure for estimating the position of a measurement on a line that best separates classes or groups. The estimated position is obtained as a linear function of the n measurement values. Since one best line may not exhaust the predictive power of the test battery in distinguishing among the classes, additional discriminant functions, all mutually orthogonal (in the sense that discriminant values are uncorrelated), may be fitted.

The geometric interpretation of discriminant analysis can be seen for the case of two groups and two variables with the assistance of Figure 4, in which the two sets of concentric ellipses represent the bivariate swarms of data for the two groups in idealized form. The variable x, y are slightly positively correlated. Each ellipse is the focus of points of equal density (or frequency) for a group (category). For example, the outer ellipse for group A might define the region within which 90 percent of group A lies, and the inner ellipse concentric with it might define the region within which 75 percent of group A lies. The two points at which corresponding ellipses intersect define a straight line II. If a second line I, is constructed perpendicular to line II, and if the points in the two-dimensional space are

projected onto line I, the overlap between the two groups will be smaller than for any other possible line. The discriminant function therefore transforms the measurement values to a single discriminant value, and that value is the measurement's location along line I. The point b where II intersects I would divide the one-dimensional discriminant space into two regions, one indicating probable observation in group A and the other region for group B. Notice that this figure depends on the equality of the two group variances. If either the variance of x and y or the x,y covariance were different from the two groups, the ellipses for two groups would not have the same shape and orientation, the boundary (line II) would not be a straight line. The size of the two populations do not have to be the same, only teh variance and covariances.

We can consider, similar to the example above, the case of classifying a two-dimensional observation into a one-dimensional normal population, to classify a p-dimensional observation vector $\underline{x}^* = (x^*_1, x^*_2, \ldots x^*_n)$ into one k multivariate normal populations with mean u_i and variance-covariance matric Σ, i - 1,2,...k. Since $\underline{x}^*$ is a realization of a random vector $\underline{x} = (x_1, x_2 \ldots x_p)$, the results presented so far used all p variables $x_1, x_2 \ldots x_p$ to discriminate between k populations. In many applications, however, it is desired to identify a subset of these variables which best discriminates between the k populations. This problem is analogous to that of stepwise regression analysis in an earlier section, in which it was desired to identify a subset of independent variables which best predicts a dependent variable.

This stepwise discriminant procedure is as follows. We first identify the variable for which the mean values in the k populations are most different. For each variable this difference is measured by one-way analysis of variance F statistics and the variable with the largest F is chosen (or entered). On successive steps, we consider the conditional distribution of each variable not entered given the variable entered. Of the variables not entered, we

identify the variable for which the mean values of the conditional distributions in the k populations are most different. This difference is also measured by one-way analysis of variance F statistics. The stepwise process is stopped when no additional variables significantly contribute to the discrimination between the k populations. The computations are obtained by BMD07M (Dixon, 1974), and details are referred to by Afifi and Azen (1972).

4.1.3.2 Principal Component Analysis

The method of principal component analysis is a general technique of displaying interrelations in the data, but it is not a statistical technique which can lead to a decision or a hypothesis. This interrelationship, called the dependence structure, may be measured by the covariances, or equivalently the variances and correlations between variables $x_1 \ldots x_p$. It is possible to find a linear combination $y_1, y_2, \ldots y_q$, $(q<p)$ of $x_1, x_2 \ldots x_p$ which generates the dependence structure between x's. Then, the new variates y's which are independent of each other account in turn for as much of the variation as possible in the sense that the variance of y, is a maximum among all linearly transformed variates, the variance of y_2 is a maximum among all linearly transformed variates orthogonal to y, and so on. Then we have

$$\begin{aligned} y_1 &= \sum_{i=1}^{p} \alpha_{1i} x_i \\ &\vdots \\ y_q &= \sum_{i=1}^{p} \alpha_{qi} x_i \end{aligned}$$

$$\text{with } \sum_{i=1}^{p} \alpha_{1i}^2 = 1, \ldots, \sum_{i=1}^{p} \alpha_{qi}^2 = 1$$

$$\text{and } \sum_{i=1}^{p} \sum_{j=1}^{q} \alpha_{i^1 j^1} = 0 \quad \text{for } i = 1, 2, \ldots, p, \; i \neq i^1$$

$$j = 1, 2, \ldots, q \quad j \neq j^1$$

From these equations, it is seen that new variables y's are uncorrelated and ordered by their variances, viz, cov $(y_j\ y_j) = 0$ for all j, $j \neq J'$, and var $(y_1) \geq$ var $(y_2) \geq \quad \geq$ var (y_q) where cov and var are covariance and variance. Further, the total variance $v = \sum_{j=1}^{q}$ var $(y_j) = \sum_{i=1}^{p}$ var (x_1) are the same after the transformation. In this way, a subset of the first q y's may explain most of the total variance and is therefore a parsimonious description of the dependence structure among the original variable x's. The method of principal components is to determine the coerficients α_{ij}, which are eigenvectors.

Since we assume that $x_1, x_2 \ldots x_p$ have multivariate distribution (not necessarily normal) with mean u and known covariance matrix $\Sigma = (\sigma_{ij})$, we wish to find eigenvector as

$$\text{var } (y_1) = \sum_{i=1}^{p} \sum_{i=1}^{q} \alpha_{1i}\ \alpha_{1j}\ \sigma_{ij}$$

is maximized subject to the condition of $\sum_{j=1}^{q} \alpha^2_{1j} = 1$. Thus, the first principal component explains 100 [var (y_1)]/V percent of the total variance. Likewise we have the first two principal components explain 100[var (y_1) + var (y_2)]/V percent of total variance, and so forth. Hence y_q is the qth principal component, the variables $y_1, y_2 \ldots y_q$ explain $100[\sum_{j=1}^{q}$ var $(y_j)]$/V percent total variance. And we found the set of eigenvectors for each principal component.

To compare the contribution of $x_1, x_2 \ldots x_p$ to y_j we examine the quantities α_{ji}/σ_i, i=1, 2,...p and j=1, 2,...q, and i the standard deviation of x_i, since the correlation between x_i and y_j is given by α_{ji} [var $(y_j)]^{1/2}/\sigma_i$. Furthermore, when the correlation matrix used, then comparison of coefficient α_{ji} is all that is necessary. Hence the larger the coefficient, the larger the contribution of the variables $x_i, x_2 \ldots x_p$ to the principal component, $y_1, y_2, \ldots y_q$.

A geometric interpretation of the principal component with p = 2 is as follows. Each variable x_1, x_2 is represented by a coordinate axis from the origin with mean u, and u_2. Then, as eigenvector specifies its direction and eigenvalue (variance of $y_1$or y_2) specifies the length of an axis of notional ellipse. In principal component analysis, we search for a notation of these axes so that the variable y, represented by the first new principal axes has a maximum variance. The variable y_2 represented by the second of the new axes is uncorrelated with y_1, and has a maximum variance under this restriction. Hence, the first principal component $y_1 = \alpha_{11} x_1 + \alpha_{12} x_2$ is in the direction of the major axis of the ellipse, and second principal component $y_2 = \alpha_{21} x_1 + \alpha_{22} x_2$ is in the direction of the minor axis of the ellipse (Figure 5).

The computations are obtained by BMD01M (Dixon, 1974), and details are referred to by Afifi and Azen (1972).

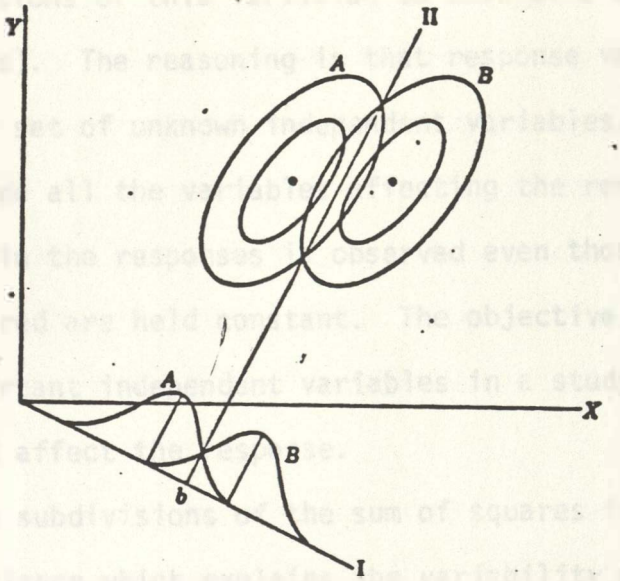
4.1.3.3 Canonical Correlation Analysis

Canonical correlation analysis can be considered a generalization of multiple correlation. In the multiple correlation problem, we have a set of p variables x_1, $x_2 \ldots x_p$ and one variable y; The objective is to find a linear compound of the x-variables that has the maximum correlation with y. In canonical correlation analysis, there is more than one y-variable, and the objective is to find a linear compound of the y-variables. The most suitable class of examples that comes are those where the x-variables are from a different domain than the y-variables. For example, the x-variables could be background variables referring to environmental data, and the y-variables descriptive variables such as the abnormal stages of fish egg embryos. The problem would be to find out whether there is some combination of background variables, that has high correlation with a combination of the y-variables.

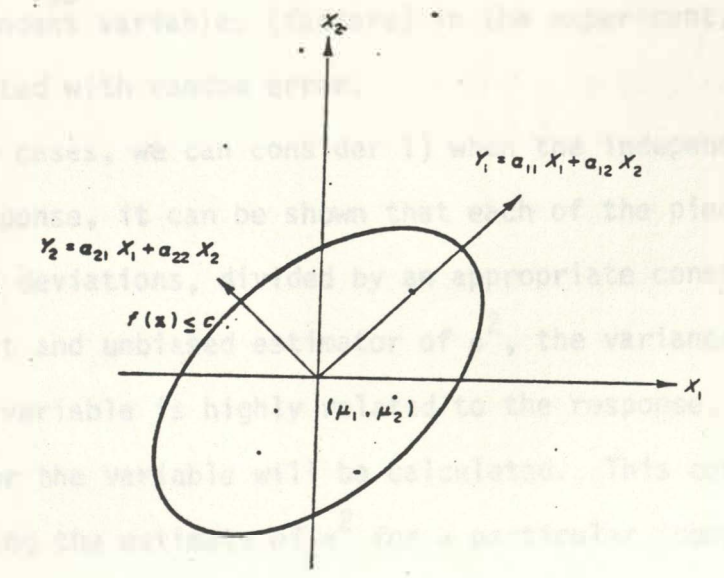
However, after that a pair of linear functions that maximally correlates has been located, there may be an opportunity to locate additional pairs of functions that maximally correlate, subject to the restriction that the functions in each new pair must be uncorrelated with all previously located functions in both domains. That is, each pair of functions is so determined as to maximize the correlation between the new pair of canonical variables, subject to the restriction that they be entirely orthogonal (uncorrelated) to all previously derived linear combinations. The analytical trick is to display the structure of relationships across domains of measurement in the canonical analysis by reducting the dimensionality to a few linear functions of the measures that have maximum covariances between domains subject to restrictions of orthogonality.

The computations are obtained by the BMD09M (Dixon, 1974) and further reading for the canonical correlation analysis see Cooley and Lohnes (1971).

F 4


Y
II
A
B
A
X
B
b
I

F 5


X_2
$Y_1 = a_{11} X_1 + a_{12} X_2$
$Y_2 = a_{21} X_1 + a_{22} X_2$
$f(x) \leq c$
X_1
(μ_1, μ_2)

4.1.4 Variance Analysis

4.1.4.1 Analysis of Variance

The analysis of variance attempts to analyze the variation of a response and to assign portions of this variation to each of a set of independent variables (factors). The reasoning is that response variables vary only because of variation in a set of unknown independent variables. Since the experimenter will rarely include all the variables affecting the response in the experiment, random variation in the responses is observed even though all independent variables considered are held constant. The objective of the analysis of variance is to locate important independent variables in a study and to determine how they interact and affect the response.

Recall the subdivisions of the sum of squares in the regression analysis, or the sample variance which explains the variability of a set of n measurements as the sum of squares of deviations. The analysis of variance partitions the sum of squares of deviation, called the total sum of squares (corrected with mean), say, $\sum_{i,j} (y_{ij}-\bar{y}..)$, into parts, each of which is attributed to one of the independent variables (factors) in the experiment, plus a remainder that is associated with random error.

For cases, we can consider 1) when the independent variables are unrelated to the response, it can be shown that each of the pieces of the total sum of squares of deviations, divided by an appropriate constant, provides an independent and unbiased estimator of σ^2, the variance of experimental error; 2) when a variable is highly related to the response, portion of its sum of squares for the variable will be calculated. This condition can be detected by comparing the estimate of σ^2 for a particular independent variable with that obtained from the sum of squares for error using an F-test. If the estimate for the independent variable is significantly larger, the F-test will reject a hypothesis of "no effect for the independent variable", and produce evidence to indicate a relation to the response.

The basic assumptions for the analysis of variance where tests of significance are attained are 1) independent variable, factor, or treatment effects are additive; 2) experimental errors are random, independent and normally distributed about a zero mean and with a common variance. The assumption of normality is not required for estimating components of variance. In practice we are never certain that all these assumptions hold; often there is good reason to believe some are false. Excellent discussions of these assumptions, the consequences when they are false, and remedial steps are given by Eisenhart (1947), Cochran (1947), and Bartlett (1947). Steel and Torrie (1960), and Cox (1958) summarize this topic in a short but comprehensive discussion.

There are some ways to reduce the effects of uncontrolled variations on the error of treatment (variable or factor) comparison. Error control can be accomplished by the experimental design. The general idea of choosing a suitable design is the common sense one of grouping the units into sets (blocks), all the units into a block being as alike as possible, the assigning the treatments so that each occurs once in each block. All comparisons are then made within blocks of similar units. The variation among units within a block is less than that among units in different blocks, the precision of the experiment is increased as a result of error control. Such blocks of similar outcome are also called replicates. This kind of design is known as a randomized complete block design. Sometimes two or more systems of blocking suggest themselves and it may be desired to use them simultaneously. When the units are simultaneously blocked in two ways this is called the Latin square design. If the units are blocked in three ways simultaneously, the desing is called a Graeco-Latin square design.

As the number of treatments in an experiment increases, the number of experimental units required for a replicate increases. In most cases, this results in an increase in the error, that is, in the variance in the parent population. Designs are available where the complete block is subdivided into a number of incomplete blocks such that each incomplete block contains only a portion of the treatments. The subdivision into incomplete blocks is done according to certain rules, so that the experimental error can be estimated among the units within the incomplete blocks. Precision is increased to extent that the units within an incomplete block are more uniform than the incomplete blocks within a replicate. The split-plot design, balanced incomplete block design, partially balanced lattices and other designs within the incomplete design are discussed fully in Cochran and Cox (1957), Federer (1955), and Kempthorne (1952).

The second approach for an error control mechanism is the utilization of concomitant observation. For example, if you study weight gains, it is useful to consider initial weights. An essential condition has to be satisfied in order that after use of the concomitant observation an estimated treatment of effects for the desired main observation shall still be obtained. This condition is that the concomitant observations should be unaffected by the treatment. In practice concomitant observations should be taken before the assignment of treatments to unit is made or the concomitant observations are made after the assignment of treatment, but before the effect of treatment has had time to develop. The supplementary observation value for any unit must be unaffected by the particular assignment of treatments to units acutally used. The analysis for the concomitant observations is called the analysis of covariance which will be discussed in a later section. The design for the reduction of error is discussed in detail by Cox (1958).

Further details of procedures for analysis of variance are omitted, however, excellent references include Cochran and Cox (1957), Cox (1958), Federer (1955), Kempthorne (1952), Snedecor and Cochran (1967) and Steel and Torrie (1960). Steel and Torrie is the best choice for developing an understanding the analysis of variance.

The biomedical computer program (BMD, Dixon, 1974) package provides the computations of analysis of variance: BMD01V for one way classification, completely randomized block design, BMD02V for factorial design, BMD05V, BMD08V and BMD10V for any hierarchical designs including partially crossed, fully crossed, partially nested, fully nested designs. BMD05V and BMD10V are more flexible for setting up any design problem.

4.1.4.2 Multiple Comparison

In a completely random design (one-way classification model), an analysis is designed to detect a difference in a set of more than two populations means (H_o: $u_1 - u_2.. = u_p$). The hypothesis H_o will be rejected if

$$F = \frac{\text{ss of between treatment/p-1}}{\text{ss of within treatment/}}$$

$$= \frac{MST}{MSE} \quad F\alpha$$

where $MST = \sum_i n_i(\bar{y}_i - \bar{y}..)^2/p-1$

$$MSE = \sum_{i,j} (y_{ij} - \bar{y}..)/\sum_i n_i - p$$

y_{ij} = the jth observation on the ith treatment

$y_{i.}$ = ith treatment mean

$y..$ = mean of all observation

$F\alpha$ = critical value (F table value) based on (p-1) and $\sum_i n_i - p$ degree of freedom for probability of a type I error, α.

If the H_o is not rejected (F statistics for treatment is not significant), the evidence is against rejecting the H_o and specific treatment comparisons

should not usually be necessary. In other words, if F is not significant, the treatment means are regarded as indistinguishable. However, if F is significant (H_o is rejected), the ordinary t-test for the difference between two means is applied to every pair of means. Where the difference of any two means exceeds the critical value they are significantly different, i.e.

$$[\bar{y}_{i.} - \bar{y}_{j.}] \quad t\alpha, (\Sigma n_i - p) \quad \sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}$$

where $t\alpha, (\Sigma n_i - p)$ is tabular value of t for error degrees of freedom. This is called the least significant difference (lsd). The lsd is basically a t-test using a pooled error variance. Since the lsd need be calculated only once and takes advantage of the pooled error variance, its use is seen to be a timesaver as compared with making individual t-tests.

Since the lsd can be and is often misused, some statisticians hesitate to recommend it. The most common misuse is to make comparisons suggested by the data, comparisons not initially planned. For the tabulated confidence levels to be valid, the lsd should be only for independent or nonindependent comparisons planned before data have been examined. A valid test criterion for planned comparisons of paired means, a criterion in considerable vogue both past and present, used the lsd.

The uncritical use of the lsd and the need for other methods of making multiple comparisons among treatment means, especially nonindependent comparison, have led to several other tests, such as Duncan's new multiple range test, Tukey's w-procedure (significantly different, hsd), Student-Newman-Keul's test, Dunnett's test (comparing all means with a control) and Scheffe's multiple contrasts test. Scheffe's method should not be used for paired comparison, but it fits for tests of more complicated contrasts. The testing procedures of other tests are very similar to each other. The references for this topic are by Li (1964) and Steel and Torrie (1960). The computation for Duncan's multiple test is obtained by BMD07V (Dixon, 1974).

4.1.4.3 Analysis of Covariance

It is possible to superimpose upon the simple linear regression model a one-way analysis of variance model. This combination of analysis of variance and regression techniques is called analysis of covariance. Analysis of covariance arises in several situations, but mainly in two: 1) the variable x is introduced to increase experimental precision or is inherent in the problem and must be accounted for in the analysis. One very important assumption in using the covariance method is that variation of the x value is not due to the treatment; 2) the linear relationships are themselves the object of study in several treatment groups.

Let us have available pairs of observations from several samples, which may be arranged in an array as follows:

Sample from Population 1	Sample from Population 2	...	Sample from Population y	Totals x y
x_{11} y_{11}	x_{21} y_{21}	...	x_{r1} y_{r1}	
x_{12} y_{12}	x_{22} y_{22}	...	x_{r2} y_{r2}	
.	.		.	
.	.		.	
.	.		.	
x_{1n_1} y_{1n_1}	x_{2n_2} y_{2n_2}		x_{rn_r} y_{rn_r}	
Total $x_{1.}$ $y_{1.}$	$x_{2.}$ $y_{2.}$	...	$x_{r.}$ $y_{r.}$	$x..$ $y..$
No. of Observ. n_1	n_2	...	n_r	N
Means $\bar{x}_{1.}$ $\bar{y}_{1.}$	$\bar{x}_{2.}$ $\bar{y}_{2.}$	...	$\bar{x}_{r.}$ $\bar{y}_{r.}$	$\bar{x}..$ $\bar{y}..$

Often it is desired to test the null hypothesis of a common line of the form against the alternative hypothesis of the form, i.e.

$$Ho : y_{ij} = \alpha + \beta (\bar{x}_{ij} - \bar{x}..)$$

$$Hi : y_{ij} = \breve{\alpha}_i + \beta_i (x_{ij} - \bar{x}i.)$$

Within each sample α_i and β_i are estimated by $a_i = \bar{y}_{i.}$ and

$$b_i = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})(x_{ij} - \bar{x}_{i.})}{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}$$

The error ss about the ith individual line is

$$(SSE)_i = (SS\ Total)_i - (SS\ Treat)_i = \sum_j (y_{ij} - \bar{y}_{i.})^2 - b_i^2 [\Sigma (x_{ij} - \bar{x}_{i.})^2]$$

Then, total error ss about 5 individual regression lines is

$$SSE = \sum_{i=1}^{r} (SSE)i$$

with N-2r degree of freedom.

Meantime, ss about the single line under the Ho, we have estimate $a=\bar{y}..$ and

$$b = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{..})(x_{ij} - \bar{x}_{..})}{\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2}$$

and the total variability about the single line is

$$SST = \sum_i \sum_j (y_{ij} - y_{..})^2 - b^2 [\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2]$$

With N-2 degree of freedom (d.f.). Consequently, we have the summary table for analysis of covariance as follows:

Source	SS	d.f.	F
Excess explained by r lines	SST-SSE	2(r-1)	$\frac{(SST-SSE)/2(r-1)}{SSE/N-2_r}$
About r lines	SSE	N-2r	
About single line	SST	N-2	

If Ho is accepted, then we may regard the populations as having the same linear relationship of y on x. If Ho is rejected, it may be that the slopes β_i are the same, just the intercepts α_i differed. In other words, the regression lines may be parallel but not coincident. Thus, in some situations, it is also desirable to have a test of the new hypothesis against the same H_a, i.e.

$$H_o = y_{ij} = \alpha_i + \beta(x_{ij} - \bar{x}_{i.})$$

$$H_a = y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.})$$

This is identical to the test of the hypothesis as:

$$H_o : \beta_1 = \beta_2 = \ldots = \beta r = \beta$$

$$H_a : \beta_i \pm \beta_j \text{ for some } i \pm j.$$

Under the H_o, all the samples have a common slope which is estimated by

$$b_c = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i.})(x_{ij} - \bar{x}_{i.})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2}$$

The ss about these parallel lines is

$$SSC = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 - b_c^2 \left[\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 \right]$$

The new analysis of covariance table for one common slope is as follows:

Source	SS	d.f.	F
Excess explained by r lines	SSC-SSE	r-1	$\frac{(SSC-SSE)/r-1}{SSE/N-2r}$
About r lines	SSE	N-2r	
About single slope r different intercepts	SSC	N-r-1	

If Ho is accepted, the lines are parallel. If Ho is rejected, then we may perform another test which is often referred to as the test of adjusted means. The new Ho and Ha is the following: i.e.

$$Ho : y_{ij} - \alpha + \beta (x_{ij} - \bar{x}_{..})$$

$$Ha : y_{ij} = \alpha_{i+}\beta (x_{ij} - \bar{x}_{i.})$$

This is identical to the test of the hypothesis

$$Ho : \alpha_1 = \alpha_2 = \ldots = \alpha_r = \alpha$$

$$Ha : \alpha_i \pm \alpha_j \text{ for some } i \pm j.$$

Under the Ho, we have a new table of analysis of covariance as:

Source	SS	d.f.	F
Explained by r intercepts	SST-SSC	r-1	$\frac{(SST-SSC)/r-1}{SSC/N-r-1}$
About single slope r different intercept	SSC	n-r-1	
About single line	SST	N-2	

When covariance is used in testing adjusted treatment means, it is important to know whether or not the independent variable is influenced by the treatments. If the independent variable is so influenced, the interpretation of the data is changed. This is because the adjusted treatment means estimate the values expected when the treatment means for the independent variable are the same. Adjustment removes part of the treatment effects when means of the independent variable are affected by treatments. This does not mean that covariance should not be used in such cases, but that care must be exercised in the interpretation of the data.

The computations of analysis of covariance are obtained by BMD03V, BMD04V, and BMD09V (Dixon, 1974). General reference books for the analysis of covariance are Li (1964), Snedecor and Cochran (1967), and Steel and Torrie (1960).

4.1.5 Goodness of Fit

The method of measuring the discrepancy between an observed and a theoretical distribution and of deciding when the discrepancy is so large that the theoretical distribution is not a good fit and does not adequately explain

the observed distribution is developed in a simple procedure where all the parameters are known in advance. A not very obvious but perfectly valid relative measure of the discrepancy between an observed (O) and expected frequency (E) is expressed as $(O-E)^2/E$. The sum of these quantities for all classifications (sample events or categories) is an index of discrepancy, which is called the chi-square (x^2) goodness of fit test. The degrees of freedom are the number of categories (k) decreased by one and the number of parameters (m) estimated, i.e.

$$x^2 = \sum_{i-1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \text{with k-1-m degree of freedom.}$$

This test statistic has approximately a x^2 distribution provided expected frequencies are large (five to ten as a minimum). If the expected frequencies are too small in both end categories, they can be pooled into the adjacent categories. However, since the tails of a distribution often offer the best source of evidence for distinguishing among hypothesized distributions, the x^2 approximation is improved at the expense of the power of the test (Steel and Torrie, 1960). Cochran (1942, 1952, and 1954) has shown that there is little disturbance to the 5% x^2 test when a single expected frequency is as low as 0.5. However, in general, the accuracy of the x^2 approximation improves as observed frequencies (O_i) increase. The classification (category) should be chosen so that each observed frequency is not small, that is, it suffices to insure that each $O_i \geq 5$, but the approximation is reasonable even when a few $O_i \geq 2$ and the remaining $O_i \geq 5$.

Suppose that we have a random sample of size n, and selected k class intervals $[x_1, x_2), [x_2, x_3), [x_3, x_4), \ldots [x_k, x_{k+1})$, with say $x_1 = -\infty$ and $x_{k+1} = +\infty$. Let f_i be the observed frequency in the interval $[x_i, x_{i+1})$. To compare an observed distribution with a normal distribution with mean μ and variance 2. i.e. $N(\mu,\sigma\ 2)$, then the expected frequencies are required. To compute expected frequencies, the probabilities associated with each interval are necessary. These probabilities are obtained by $Z_i = (x_i - \mu)/\sigma$, i.e.

$$P_i = P\ (x_i \leq x < x_{i+1}) = P\ \left(\frac{x_i-\mu}{\sigma} \leq Z \leq \frac{x_{i+1}-\mu}{\sigma}\right).$$

So we find the sample mean $\bar{x}$ and variance s^2 and consider them μ and $\sigma 2$. Then each probability on a given interval times the total frequency n, i.e. $E_i = nxp_i$, gives an expected frequency on that interval. We compute now the value of the test statistics defined by the formula as the above

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

with k-1-2 = k-3 degrees of freedom (d.f.), since we lost 2 d.f. for estimating two parameters μ and σ^2. We reject the null hypothesis H_0 (E(x) = O(x)) if $x^2 > x^2\alpha$, k-3 where $x^2\alpha$, k-3 is the critical value from chi-square table with k-3 d.f. and α level of significance, and E(x) in a distribution of expected frequencies (in the above example) $(E(x) = N\ (\mu, \sigma^2)$, and O(x) is a distribution of given observed frequencies. This is a so-called "test for normality".

In a similar way we could test whether a random sample has a Poisson or negative binomial distribution by a goodness of fit test. These are a so-called "test of randomness". Steel and Torrie's book (1960) is a good starting point for reference on this topic. Kendall and Stuart (1961) is an excellent reference for the theoretical structure of a goodness of fit test.

4.1.6 Biological Assay

4.1.6.1 Bioassay

Biological assays are methods for the estimation of natural constitution or potency of a material by means of the reaction that follows its application to living matter. The typical bioassay involves a stimulus (heavy metal, drug, vitamin, fungicide, etc.) applied to a subject (fish, animal, a piece of fish tissue, plant, bacterial culture, etc.). Application of the stimulus is followed by a change in some measurable characteristic of the subject, the magnitude of change being dependent upon the dose. A measurement of this characteristic is the response of the subject. The relationship between dose and response will not be exact, but will be obscured by random variations between replicate subjects.

Typically two preparations are involved, one designated as "standard" and the other as "unknown". Any test preparation of the stimulus, having an unknown potency, is assayed by finding the mean response to a selected dose, and equating this does to that of a standard preparation shown by experiment to produce the same mean response; experimentation with several different doses of one or both preparations is almost always needed in order to accomplish this satisfactorily. The ratio of the two equally effective doses is an estimate of the potency of the test preparation relative to that of the standard.

Bliss (1954) describes three types of bioassay and their underlying assumptions as follows:

1. Comparative assays occur most widely and are of special interest in research. They estimate the relative potency under specified conditions, of two preparations which give a similar response. To determine whether the estimated potency is independent of the level of response requires two or more dosage levels of both the standard and the unknown. To test the assumed

linearity of the dosage response curves requires three or more levels.

2. Analytical assays for biological standardization depend, theoretically upon the following additional assumptions: 1) the standard and the unknown differ only in the concentration of the same active agent, 2) the same relative potency would be obtained with all methods of assay or test organisms, 3) if the stimulus contains two or more active proportions in both the standard and the unknown.

3. Pass or fail assays test whether the unknown preparation meets prescribed standards but do not determine its actual potency. Although comparative or analytical assays are often used instead, they may be relatively less efficient for inspection purposes.

When the response can be plotted linearly against the logarithm of the dose, the relative amounts of the two preparations which produce any given response is estimated by the horizontal distance between two parallel regression lines. Suppose x_S is a dose of a standard stimulus, S, and Y_S is the response measured on a subject receiving this dose under the specified experimental conditions. Let T be a stimulus of the unknown to be compared with assayed against S. We have similarly X_T and Y_T for a dose and response of the unknown preparation. Then, we summarize as two equations:

$$Y_S = a_S + b \log x_S$$

$$Y_T = a_T + b \log x_T$$

There are two parameters (a_S and a_T) for each stimulus and b is identical for S and T. What we want to have is the estimate of potency ($\ln S_T$) which is the difference between equipotent values x, the horizontal distance between the two lines for S and T; i.e.

$$S_T = \exp [(a_S - a_T)/b]$$

$$\ln S_T = (a_S - a_T)/b \ .$$

The detailed treatment of estimation of potency, test hypothesis of potency, test hypothesis on linearity, parallelism and analysis of variance are described well by Finney (1964).

As an example, there may be reason to believe that S_T represents a chemical property of T, the ratio of its content of the active constituent to the corresponding content for S, independent of the particular conditions of experimentation. Provided that measurements of Y_S and Y_T for various doses are made under the same experimental conditions, a requirement usually fulfilled by arranging for simultaneous experimentation with random allocation of subjects to preparations and doses, and an estimate of S_T will then have general validity. Statistical analysis cannot prove that S_T exists and is independent of experimental conditions. The purpose of validity tests, such as the test of parallelism in a parallel line assay, is to examine whether a particular assay experiment shows any indications of departure from the general pattern: Accidental introduction of impurities or other disturbances may be detected by a typical behavior of responses, so enabling a faulty experiment to be discarded and replaced.

Cornfield (1964) has justifiably criticized that certain statistical criteria of validity be met before any assay is regarded as of practical value for relative potency, S_T. Such an idealization may be scarcely relevant to the reality of many assay situations; if the preparations assayed are qualitatively dissimilar, the strict dilution requirements can scarcely be satisfied. The linear regression of response on logarithm dose may not be parallel, yet results of such comparative assays may still seem useful in giving some indication of relative potency. He comments that if the slopes in such as assay do differ

considerably, then there is no alternative other than to treat relative potency as a function of response level. He develops a statistical technique based on representation of relative potency as itself a linear function of the expected response to preparation. Finney (1965) examines the general situation in a broader framework, to see how far Cornfield's proposal conforms to reasonable requirements on the properties of a measure of relative potency. However, Finney stated that Cornfield's considerations deserve further theoretical study as well as experimental approaches and his paper invites discussion rather than acceptance.

The complicated bioassay designs, such as regression analysis with factorial techniques and quantal responses, are referred to in Finney (1964), Bliss (1952) and Bliss (1954).

4.1.6.2 Probit Analysis

In the biological assay data the percentage or proportions of the subject reacting to the doses of stimulus can be converted into probit (probability unit). Bliss (1934) defines the probit as the normal equivalent deviate increased by 5 in order to make negative values very rare. Probit for specific percentage values were tabulated by Bliss (1935), and were reproduced by Fisher and Yates (1964, Table IX) and Finney (1971, Table I). A simplified table, sufficiently detailed for many purposes, is given as Table 4 (Finney, 1971, Table 3.2).

The relation between the probit of the expected response proportion response and the dose is $y = 5 + \frac{1}{\sigma}(x-\mu)$ where μ, σ are mean and standard deviation of the normal distribution estimated from data, and x is the logarithm value of the stimulus (dose) level. Y probit from above tables, is related to p which is the probability derived from the normal distribution as follows:

$$\int_{-\infty}^{y-5} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, dy = p$$

Then, least square procedures are used to estimate the best straight line

passing through the k points (x_i, y_i), i.e. $y_i = (5 - \frac{\mu}{\sigma}) + \frac{1}{\sigma} x_i = \alpha + \beta x_i$. To test whether this probit regression line is well represented with the results of the experiment, the utilization of a chi-square goodness test is appropriate, i.e.

$$\chi^2 = \sum_{i=1}^{I} \frac{(r_i - n_i P_i)^2}{n_i P_i (1-P_i)} \quad \text{with I-2 degree of freedom}$$

where r_i is the observed response out of the n_i samples of i^{th} dose level and p_i is the probability defined as above under the normal curve with y_i, probit of i^{th} dose level. If the test is not rejected, then the probit regression line appears to be a satisfactory representation of the experimental results. Otherwise, we need to find a suitable transformation to analysis and meet the requirement of the experiment. Then $(5-\alpha)/\beta$ is an estimate of the logarithm value of the lethal dose of 50 percent responses (log LD50). The real LD50 value is obtained by taking the value of anti-log LD50. The standard error of log LD50 is approximated by $s = 1/\sqrt{\sum_i n_i w_i}$ where n_i is the sample size of i^{th} dose level and w_i is the weight coefficient for i^{th} dose level (Table 5, Finney, 1971, Table 3.5), if the log LD50 is not very different from the mean value of dosages ($\bar{x}$) in the experiment. This expression makes no allowance for sampling errors in the estimation of β. If log LD50 is far from the mean value of dosages, the standard error is grossly underestimated. It requires adjustment with correction factors, i.e. the variance of log LD50 is expressed as:

$$s^2 = \text{var (log LD50)} = \frac{1}{b^2} \left[\frac{1}{\sum_i n_i w_i} + \frac{(\text{log LD50} - \bar{x})^2}{\sum_i n_i w_i (x-\bar{x})^2} \right]$$

Thus, the confidence limits for log LD50 at the 5 percent level of significance is obtained by log LD50 $\pm$ 1.96 s. If logarithm scale to base 10 is used, we have confidence limits for LD50 expressed as, i.e.:

$$LD50 \pm 1.96 \left[(10^{LD50}) (\log_e 10) (s) \right].$$

For further details, Finney (1971) is appropriate.

In practice, when experimental data on the relation between dose and mortality have been obtained, either a graphical or an exact probit solution (regression as above) can be used to estimate the parameters. The graphical approach is rapid and sufficiently good for many purposes, but for some more complex problems or when an accurate assessment of precision of estimates is required, the exact probit solution is necessary. Both approaches are described with detailed examples in Natrella (1973). The more advanced design problems and foundations of probit analysis are presented by Finney (1971). Computation of the exact probit solution is obtained by the BMD03S (Dixon, 1974) computer program.

Table 4. Conversion table for probits (Y) from response percentage (i.e. r/n where r is number of responses out of n sample size), e.g. Y = 5.39 if (r/n) 100 = 65.

%	0	1	2	3	4	5	6	7	8	9
0	-	2.67	2.95	3.12	3.25	3.36	3.45	3.52	3.59	3.66
10	3.72	3.77	3.82	3.87	3.92	3.96	4.01	4.05	4.08	4.12
20	4.16	4.19	4.23	4.26	4.29	4.33	4.36	4.39	4.42	4.45
30	4.48	4.50	4.53	4.56	4.59	4.61	4.64	4.67	4.69	4.72
40	4.75	4.77	4.80	4.82	4.85	4.87	4.90	4.92	4.95	4.97
50	5.00	5.03	5.05	5.08	5.10	5.13	5.15	5.18	5.20	5.23
60	5.25	5.28	5.31	5.33	5.36	5.39	5.41	5.44	5.47	5.50
70	5.52	5.55	5.58	5.61	5.64	5.67	5.71	5.74	5.77	5.81
80	5.84	5.88	5.92	5.95	5.99	6.04	6.08	6.13	6.18	6.23
90	6.28	6.34	6.41	6.48	6.55	6.64	6.75	6.88	7.05	7.33
-	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
99	7.33	7.37	7.41	7.46	7.51	7.58	7.65	7.75	7.88	8.09

Table 5. The weighting coefficient (w) for the probits, e.g. w = 0.503 if y = 4.2.

y	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.001	0.001	0.001	0.002	0.002	0.003	0.005	0.006	0.008	0.011
2	0.015	0.019	0.025	0.031	0.040	0.050	0.062	0.076	0.092	0.110
3	0.131	0.154	0.180	0.208	0.238	0.269	0.302	0.336	0.370	0.405
4	0.439	0.471	0.503	0.532	0.558	0.581	0.601	0.616	0.627	0.634
5	0.637	0.634	0.627	0.616	0.601	0.581	0.558	0.532	0.503	0.471
6	0.439	0.405	0.370	0.336	0.302	0.269	0.238	0.208	0.180	0.154
7	0.131	0.110	0.092	0.076	0.062	0.050	0.040	0.031	0.025	0.019
8	0.015	0.011	0.008	0.006	0.005	0.003	0.002	0.002	0.001	0.001

4.1.7 Time series analysis

Observations on a phenomenon which is moving through time generates an ordered set known as a time series. The objective of time series analysis, as statistical analysis as a whole, is to arrive at a deeper understanding of the causal mechanisms which generated it, because we wish to extrapolate into the future.

The typical time series may be composed of four parts:

1) Trend or long term movement,
2) oscillations about the trend with a greater or lesser regularity,
3) seasonal effects, and
4) random, unsystematic or irregular components.

We can always represent a series as one of these constituents or sum of several of them. A large part of the traditional theory of time series is devoted to an analysis of the data into such components, so as to isolate them for separate study. However, if we can represent a series as the sum of such experiments, they correspond to independently operating causal systems. The analysis of components of a series is often useful, but it may be misleading. In any case it is not the ultimate object of statistical analysis. The statistical analyses are not detailed here, however, Kendall and Stuart (1966), Davis (1941) and Croxton and Cowden (1947) cover the subject. Croxton and Cowden is the best choice for a starting point to understand and comprehend time series analysis. Kendall and Stuart are theoretical, but cover the subject thoroughly. Davis' book is oriented toward economic time series, not the environmental monitoring aspect, but is still worthwhile. BMD computer programs are available, but these require considerable knowledge for interpretating output.

4.1.7.1 Trend

The concept of trend is difficult to define succinctly. The statistical problem is to decide what type of trend fits the data closely. It must describe data logically. Such a trend is not only an expression of tendencies; but also provides a base from which to measure deviations. Thus, there are two reasons for attempting to describe the trend of a series by some kind of curve fitting. First, it may be desirable to measure the deviations from trend. These deviations consist of cyclical, seasonal and random movements. Second, it may be useful to study the trend itself, in order to note the effect of factors bearing on the trend, to compare one trend with another, to discover what effect trend movements have on cyclical fluctuations or to forecast future trend movements.

The simplest method of describing a trend is a graphical presentation, drawing it free hand or by use of curve-fitting rules. Plots of the data on semi-logarithmic paper tend to straighten out some rate trends. The trend will be a straight line on this type of scale if the series is increasing or decreasing at a constant rate.

If the polynomial is fitted to the whole series by the least squares method, it may produce a linear or curvilinear regression line of Y_t on the time variable t, i.e.

$$Y_t = a+b_1t + b_2t^2 + \ldots + b_pt^p$$

It is clear, however, that to obtain a satisfactory trend curve for marine environmental data, we should have to take a polynomial of rather high order or a somewhat complex general function. This may be not too easy to handle and in any case the coefficients of such a polynomial, being based on higher order term, would tend to be unstable from the sampling viewpoint. A more practical

point is if we add another term to the series, for example if we are keeping an annual series current from year to year, the curve fitting has to be redone each time. Moreover, the trend line may be affected throughout its length. When, therefore, the series has no obvious trend to utilize the polynomial, it is more convenient to use the moving average method.

The moving average method is a simple and flexible mathematical technique of trend fitting. The moving average is to take the first n terms (n being chosen at will), fit a polynomial of degree p, not greater than n-1, to them, and use that polynomial to determine the value in the middle of its range; then to repeat the procedure with next n terms from the second to the $(n+1)^{th}$, from third to $(n+2)^{th}$, and so on, moving on one term at each stage. Unless other considerations require it, we take n to be an odd number, so that the middle point of the range corresponds in time to a value which is actually observed. Otherwise, if we take n to be an even number, the middle point falls halfway between two observed values, or we have to use some value of fitted polynomial other than the middle point which results in a loss of useful symmetry. A simple example of a moving average is illustrated below:

Time Period	Observation	3 year Moving Total	3 year Moving Average
1	22.1	-	
2	23.8	71.6	23.87
3	25.7	74.4	24.80
4	24.9	78.8	26.27
5	28.2	-	

Thus, the moving average is a device for obtaining a series of figures, and the corresponding graph, which represents the general trend because the minor deviations of the series are averaged out in the process of its construction.

4.1.7.2 Seasonality

Perhaps the easiest component to understand and to remove from the time series is the seasonal effect. This is a fluctuation imposed on the series by a cyclic phenomenon external to the main body of causal influences at work upon it. The seasonality, refers to the effects which are annual in period, or applies to any phenomenon generated by strictly periodic natural processes, such as spring and neap variation in tides or daily variation in temperature. We must, however, be careful about extending the notion of seasonality to phenomena which are not demonstrated beyond reasonable doubt to depend on strictly periodic stimuli. For instance, to speak of sunspot variation as a seasonal effect, it may be too extreme to infer seasonality in the climatic and oceanic environment as a function of sunspots, even if the relation between the two were established.

Kendall and Stuart (1966) stated a few approaches to deal with the seasonality factor. A possibility is to use a moving average to eliminate trend before examining the residual values for seasonality. We then, of course, run into the danger of distorting the residuals. However, if we choose the moving average with care, we can minimize this effect so far as seasonal effects are concerned. In fact, if the simple moving average (with equal weights) is equal in extent to the period of a cyclical component, the trend value of the components is zero, so that residual is unimpaired. The effect of trend elimination both on seasonal components and random residuals are treated with spectrum analysis. Readers who are interested in pursuing spectrum analysis should consult the book by Kendall and Stuart (1966).

To treat seasonal effects, we rank the quarters within any one year from 1 to 4 and consider how the ranks vary from year to year. To test

these seasonal indices, we use the model equation $\mu t = y_+ s_q + \Sigma$ for $t = 1, 2..., n$, $q = 1, 2, 3, 4$. The procedure is to assume that each observation is the sum of three effects: a yearly value, y, a seasonal value, s (constant from year to year in proportional effect), and an error term Σ, which is random. If the trend is slow, so that the seasonal effect may be regarded as constant from year to year in absolute (not proportional) magnitude, we have approximately $u = y_t + s_q + \Sigma$, which is an ordinary analysis of variance model. If the trend is not slow, we have to transform the equation as $\log u_t = \log y_t + \log s_q + \log$. Then, the analysis of variance model is also utilized.

4.1.7.3 Oscillation

If we remove the attributable elements to seasonal variation and trend, we shall be left with a series oscillating about some constant value. This movement may be so small as to be virtually negligible. The series, then consists entirely of seasonality and trend. The seasonality and trend may themselves be non-existent, in which case the series is entirely oscillatory. An oscillation in a time series (or more generally, in a series ordered in time and space) is a more or less regular fluctuation about the mean value of the series. In this sense it can be sharply distinguished from a cycle, which is strictly periodic; thus a cyclical series is oscillatory, but an oscillating series is not necessarily cyclical. To fit an oscillatory curve, we can utilize a sine-cosine function curve to adjust the cyclical pattern of observed values, y. A typical curve is expressed as:

$$Y_c = \bar{y} + A \sin \left(\frac{360}{T} X\right)^\circ + B \cos \left(\frac{360}{T} X\right)^\circ$$

where $\bar{y}$ = mean of y

T = the periodicity in time (say month, season, etc.)

$$A = \frac{2}{T} \Sigma\left[Y \sin \left(\frac{360}{T} X\right)^\circ\right]$$

$$B = \frac{2}{T} \Sigma\left[Y \cos \left(\frac{360}{T} X\right)^\circ\right]$$

Y = observed time series variable

x = time period

However, it is found that most environmental data series in practice are not exactly periodic or oscillatory, and that it is difficult to describe them adequately by mathematical curves.

4.1.7.4 Randomness

We have discussed long term trends, seasonal effect and systematic oscillatory behavior. However, some of the time series which we are concerned with in environmental phenomena are clearly expressed by none of the above characteristics. An ordered series of observations could have risen by pure chance. There are many tests for randomness. Kendall and Stuart (1966) suggest a few such as the rank correlation test, difference-sign test, series correlation test and others.

4.2 Non-Parametric Statistics

Non-parametric statistics require no particular assumptions about the form of population distribution. Thereby, a non-parametric statistical test is one whose model does not specify conditions about the parameters of the population from which the sample was drawn. Certain assumptions, however, are associated with most non-parametric statistical tests; i.e. that the observations are independent and that the variable under study has an underlying continuity. These assumptions are much weaker than those associated with parametric statistical tests. Moveover, non-parametric tests do not require the forms of real value that are required for the parametric tests; most non-parametric tests apply to data in an ordinal scale, and some apply also to data in nominal scale.

Non-parametric statistics have a number of advantages: 1) Probability statements obtained from most non-parametric statistical tests are exact probabilities (except in the case of large samples where approximations are available), regardless of the shape of the population distribution from which the random sample was drawn; 2) there are suitable non-parametric statistical tests for treating samples made up of observations from several different populations; 3) since they may use ranks or signs of difference, they are often, though not always, quick and easy to apply and to learn; 4) for the same reasons, they may reduce the work of data collecting.

The non-parametric statistical tests discussed in this manuscript represent only a few of many non-parametric statistical inference methods available. A much larger collection of non-parametric test procedures, along with worked examples, are given in Siegel (1956) and Conover (1971). The popular general statistic books, such as Snedecor and Cochran (1967), Steel and Torrie (1960) and Mendenhall (1975) are good starting points for a better understanding of the topics. Several popular non-parametric statistical tests are computed by DMDP)3S computer program (Dixon 1977).

4.2.1 Wilcoxon Two Sample Rank Sum Test (Mann-Whitney U-Test)

When at least ordinal measurement has been achieved, the Mann-Whitney U-test may be used to test whether two independent groups have been drawn from the same population. This test is one of the most powerful of the non-parametric statistical tests, and it is a most useful alternative to the parametric t-test when the experimenter wishes to avoid the assumptions of the t-test.

Suppose we have samples from two populations, population A and B. The null hypothesis, Ho, is that A and B have the same distribution. The alternative hypothesis, H_A, against which we test Ho, is that A is stochastically larger than B. Let n_1 be the number of cases in the smaller of two independent groups, and n_2 be the number of cases in the larger. Then, the test statistic is computed as follows:

1. Rank all observations in the whole experiment disregarding that the samples are drawn from A and B.
2. Compute the sum of the ranks for each group (T_1 and T_2).
3. Average the ties for rank computation. Each score is given the mean of the ranks for which it is tied.
4. Look at the rank sum from a group which has the smaller sample size. Call this rank sum T.
5. Compute $T' = n_1(n_1+n_2+1) - T$
6. Compute the smaller rank sum with tabulated critical values (Snedecor and Cochran, 1967, and Steel and Torrie, 1960).
7. Reject Ho if the smaller rank sum is less than the critical table value at a given significance level α.
8. If the critical table value is inadequate, we can use the mean and standard deviation of T as

$$\mu_T = \frac{1}{2}\,[n_1(n_1+n_2+1)]$$

$$\sigma_T = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$$

With these and T, we may compute quantity $Z = (T - \mu_t)/\sigma_T$, which is approximately normally distributed with mean 0 and variance 1 as n_1 and n_2 become large. Use the critical values of normal distribution as in the usual (parametric) testing hypothesis procedure.

To use Mann-Whitney U-test procedures, we follow the steps as above then,

$$U = n_1\, n_2 + \frac{n_1(n_1+1)}{2} - T_1$$

or

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2$$

5* Compare the smaller U with tabulated critical values (Siegel, 1956).

6* Reject Ho if smaller U is less than the critical table values at given significance level.

7* Similar way as Wilcoxon's test, a simplified larger sample test can be obtained using the familiar Z statistics. When the population distributions are identical, it can be shown that the Mann-Whitney U-test statistics has the mean and standard deviation of U as

$$\mu_s = n_1 n_2/2$$

$$\sigma_s = \sqrt{\frac{1}{2}\,[n_1 n_2 (n_1+n_2+1)]}$$

Then $Z = (U-\mu_s)/\sigma_s$ tends to distribute normally with mean zero variance 1 as n_1 and n_2 become large. This approximation will be adequate when n_1 and n_2 are both greater than or equal to 10.

The computations are obtained by BMD_)3S (Doxon, 1977) computer program.

4.2.2 Kruskal-Wallis Test

The Kruskal-Wallis one-way analysis of variance by ranks is a useful test for deciding whether k independent samples are from different populations. Sample values almost invariably differ somewhat, and the question is whether the differences among the samples signify genuine population differences or whether they represent merely chance variations such as are to be expected among several random samples from the same population. The null hypothesis for the Kruskal-Wallis test is that the k samples come from the same population or from identical populations with respect to average. The test assumes that the variable under the study has an underlying continuous distribution. It requires at least ordinal measurement of *that* variable.

The procedures for utilizing the Kruskal-Wallis test are the following:

1. Rank all the k sample combined observations in a single series disregarding the samples that are drawn from k samples.
2. Compute the sum of the ranks in each k groups, Ri for i - 1,2,...,k
3. Average the ties which occur between two or more scores, each score is given the mean of the ranks for which it is tied.
4. Compute the test statistics

$$k\text{-}w = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\,[N+1]$$

where $N = \sum_{i=1}^{k} n_i$, the number of all observations in k samples combined, n_1 = number of observations in i^{th} sample, and R_i = sum of ranks in i^{th} sample. This test statistic is distributed approximately as chi-square with degrees of freedom of k-1, for sample size (n_i's) sufficiently large.

5. Reject null analysis Ho if K-W x^2 σ, (k-1) where x^2, (k-1) is the critical value found in the chi-square table with degree of freedom k-1 and α level of significance.

We may recall the concept of multiple comparison technique in the parametric statistical procedures. If the K-W statistic is not significant, the k samples come from the sample population. However, if K-W is significant (Ho is rejected), this suggests that at least two samples are drawn from different populations.

Hence, we want to explore which samples do not satisfy the hypothesis. Where the difference of any two mean rank exceeds the critical value, they are drawn from significantly different populations, i.e.

$$|\bar{R}_i - \bar{R}_j| > \sqrt{\chi^2_{\sigma_1}(k-1)} \quad \sqrt{\frac{N(N+1)}{12}} \quad \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $\bar{R}_i$ = average rank of i^{th} sample

$$= \frac{1}{n_i} \sum_{n=1}^{n_i} R_m$$

We can perform all possible pair-wise testing procedures for better interpretation. Unfortunately, even if we reject the null hypothesis of the Kruskal-Wallis procedure, we cannot detect any difference between i^{th} and j^{th} mean rank difference i.e., we cannot find any $|\bar{R}_i - \bar{R}_j|$ is greater than a given critical value as above. The reader should consult more details of the multiple comparison test and approximation procedure for the Kruskal-Wallis tests in the books by Miller (1966) and Hollander and Wolfe (1973).

The computations are obtained by BMDP)3S computer program (Dixon, 1977).

4.2.3 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov one sample test, is a test of goodness of fit. It is concerned with the degree of agreement between the distribution of a set of sample values (observed scores) and some specified theoretical distribution.

It determines whether the scores in the sample can reasonably be thought to have come from a population having the theoretical distribution. The test involves specifying the cumulative frequency distribution which would occur under the theoretical distribution and comparing that with the observed cumulative frequency distribution.

Let Fo(x) be a completely specified comulative frequency distribution under the null hypothesis Ho. That is, for any values of x, the value of F(x) is the proportion of the case expected to have scores equal to or less than x. S(x) is the observed cumulative frequency distribution (step function) of a random sample of N observations. Where x is any possible score, S(x) = k/N where k is the number of observations equal to or less than x. So under the Ho, it is expected that for every value of x, S(x) should be close to F(x). The Kolmogorov-Smirnov one sample test focuses on the largest of the deviations, i.e.:

$$D = \text{maximum } | F(x) - S(x) |$$

The sampling distribution of D under the Ho is known. We can compare the value of D and critical table value (Siegel, 1956). If $D \geq$ a given critical value, we reject the Ho.

The Kolmogorov-Smirnov two sample test is a test of whether two independent samples have been drawn from the same population or from populations with the same distribution. The two-tailed test is sensitive to any kind of differences in location (central tendency) and dispersion. The one tailed test is used to decide whether the population values from which one of the samples was drawn are stochastically larger than the values from other populations.

Let $S_1(x)$ and $S_2(x)$ be the observed cumulative frequency distribution (step function) of the first and second sample, i.e $S_1(x) = k/n_1$ and $S_2(x) = 1/n_2$. The Kolmogorov-Smirnov two samples test focuses on

$$D = \text{maximum} \quad | \; S_1(x) - S_2(x) \; |$$

The principle of hypothesis testing is the same as the one sample test with different critical table values for the two sample test (Siegel, 1956 and Hollander and Wolfe, 1973). In the case of large sample approximation procedures see Hollander and Wolfe (1973).

4.2.4 Correlation

If, with a given set of experimental data, the requirement is not met or the normality assumption is unrealistic, then use one of the non-parametric correlation coefficients, Spearman rank correlation, Kendall rank correlation, Kendall partial rank correlation and Kendall coefficient of concordance. Non-parametric measures of correlation are available for both nominal and ordinal data. The tests make no assumption about the shape of the population from which the scores were drawn. Some assume that the variables have underlying continuity, while others do not even make this assumption. Moreover, we find that, especially with small samples, the computation of a non-parametric measure of correlation and test of significance is much easier than the computation of the Pearson correlation described earlier.

The detailed procedures of computation and applications are found in Siegel (1956), and the computation for Spearman rank correlation is obtained by BMDP)3S (Dixon, 1977) computer program.

4.3 Measures of Association

Originally we were to attempt summarizing a few indices of measurements useful for marine ecological investigations, i.e. methods to determine measures of similarity, diversity, and clustering. Since there are excellent references for the topics, all well detailed, we did not attempt summarization. Refer to the following:

Similarity - Boesch (1977) and Clifford and Stephenson (1975

Diversity - Clifford and Stephenson (1975), Pielou (1974) and Pielow (1975)

Clustering - Boesch (1977) and Everitt (1974).

5. Applications

None of the quantities involved in Ocean Pulse research can be observed or measured throughout the whole population. Conclusions will be based on the attributes of samples considered representative. If the sampling and analysis are good the interpretation derived may differ little from reality. In order to achieve this the objective will require a thorough grasp of individual subjects and indices developed in allied fields. Some recognition of limitations (probabilities) is necessary for deriving projections of events. The correlations of time series will likely be employed and the topic will be an important part in the synthesis of research findings. At the present time the array of intended test species and enlisted disciplines is noted in Table 4. Each of the activities is considered a promising arbiter of environmental quality. However, the efficiency, reproducibility and other attributes of the studies still remain to be evaluated in many cases. The selection of test species has been derived from the availability of species encountered in sampling gear during early cruises. Nevertheless, these key species must be linked as a part of the tangible ecosystem model which we develop and characterize in our synthesis. The subjects range from phytoplankton, constituent chemicals and chlorophyll through particulate, filter feeding invertebrates and commercially harvestable fish species. The suitability of various statistical tests is discussed below for each of the study disciplines (Table 5). However it must be remembered that after the basic survey series of results there will be material to begin determining time series trends. Only from the integration of individual study results, will there follow an evaluation of ecosystem impacts.

5.1 Community Studies

Descriptions of population makeup by location will require analyses to determine similarities of composition, by species and biomass. These referred to here are of general assemblages in the water column or living on and in the sediments. Special topics treated below are basically of special indicator groups. Succession studies simply measure population changes over time. Correlations and diversity indices are appropriate as are equitability indices. A clustering analysis is available in a computer package which classifies the similarity and dissimilarity hierarchy of organisms. Multiple regression and all multivariate analyses are most appropriate in relating change or differences to background variables.

5.2 Seasonal Abundance of Organisms (Host and Parasitic)

Contaminants, such as heavy metals can be correlated to such variables as substrate, water mass characteristics and climatological phenomena. This presentation is a natural outgrowth of a time series analysis. Regression, correlation, multivariate analysis and analysis of variance are all appropriate. Some non-parametric tests.

5.3 Succession Studies

To measure the natural and unnatural progress of dominant organisms, both water column, substrate biota and fish are considered. This can include an analysis of species interactions (i.e. replacement). Multivariate tests and variance analysis are all appropriate, obviously in a time-series mode.

5.4 Anaerobic Analysis

A special form of population analysis to express the dynamics of the bacterial population. Interest in the enumeration of anaerobic bacteria in sediment, water and animal tissue and the presence or absence of disease producing

organisms. Inshore-offshore interactions will be studied as well as comparative analyses of impacted and control sites. Multiple correlation, analysis of variance, bioassay and clustering techniques are all feasible tests. Changes will be observed seasonally and some non-parametric tests may be found appropriate.

5.5 Calorimetry

Technique is to measure bound carbon in the biota. This may provide an index of condition to measure differences or relate with impacts upon species. This is related to the study for trophic interactions and energy budgets. Regression and correlation techniques, analysis of variance and covariance are principal tests. Some non-parametric tests are appropriate. Correlations will be made to physiological and pathological survey data.

5.6 Physiological Activities

The objective of physiological and biochemical activities is the detection of abnormal variations from baseline norms in a variety of marine animals, including finfish, molluscs and crustaceans. The plan is to sample key species to compare between impacted areas and control stations. Field detected abnormalities will be compared with those noted in laboratory studies. As levels of enzymes and blood are established and compared, many tests are appropriate. These include regression, correlation, multivariate analyses, analysis of variance, bioassay techniques, and profit analysis. Some non-parametric tests will be pertinent. These will be related to temporal and spatial differences. Studies will be coordinated with pollutant uptake studies and pathological findings. There will be an intimate association with the chemistry staff. Tissues used by physiology and biochemistry will be analyzed.

5.7 Parasite Analysis

This is a special form of population study consisting of pathobiological survey and the effects of transmitted parasites and pathogens and their routing levels of selected planktonic and benthic crustaceans. Parasites, gross and histological abnormalities of selected species taken from pristine and contaminated stations will be evaluated.

Blood parasites will be investigated in five finfish species - cod, haddock, yellowtail, herring and silver hake. The object will be to determine the distribution and prevalence. Molluscan pathology will include the target species of sea scallop and tellinoid clams. Pathological observations will include gross and histological examination for abnormalities. Parasite burdens, regression, correlation, multivariate techniques and analysis of variance, are likely techniques. A time series analysis is possible as are also community analyses, such as clustering.

5.8 Virology

Delineation of blood virus characteristics of marine organisms. Five commercially important species have been selected including cod, haddock, yellowtail, herring, and silver hake. Clinical techniques are available for assaying variations from normal. In these species as well as many others, norms have yet to be established on types and incidence. Multivariate, correlation analyses, and analysis of variance are possible choices for analyses, also bioassay and non-parametric techniques. Population measures (diversity, etc.) may also be adaptable as data accure.

5.9 Anomalies

Measures of gross and histopathological effects include type, frequency and distribution. Correlations and both parametric and non-parametric analyses of variance are likely. One target species is Ammodytes. The egg is demersal;

adults spawn along the inside edge of the shelf and are also demersal, spending considerable time burrowed in the sediments associated with degraded habitats.

5.10 Nutrient Bioassay

Initial study is planned as a growth assay employing two phytoplankton species, seasonally dominant in Bight waters. Test data will consist of species growth rates under experimental conditions. Test variables will include nitrogen, phosphorus, metal, vitamins, and chelators. The objective is to assess the influence of key substances known to limit phytoplankton growth. Here all correlation and multivariate analysis techniques are useful. Non-parametric tests are effective tools along with the obvious bioassay and probit analysis applications.

5.11 Pollution Uptake Studies

Levels of metals in sediments and tissue collected from impacted and normal environments will be determined. Subsequent tests can make use of multivariate and correlation analyses, possibly a utilization of bioassay and probit analysis. One aspect will be to compare field data with laboratory exposure. A time series analysis to determine seasonal changes should be considered as well as non-parametric tests.

5.12 Genetic Studies

Studies of miotic figures and embryonic anomalies can utilize both regression and correlation analyses. Correlations will be made with water chemistry. Multivariate analysis should prove particularly useful. A larval development series under laboratory exposures can be analyzed using bioassay and probit techniques. Non-parametric tests are feasible as is the use of a clustering for interpreting field data.

5.13 Petroleum Bioassay

A specific variation of pollution uptake studies and the same statistical techniques pertain.

5.14 Limiting Factors

This study relates to a determination of the sources of mortality of surf clams and an estimate of relative impact. Analyses will include correlation analyses, possibly bioassays if lab exposures are conducted. Probit and non-parametric techniques are appropriate. Clustering could provide a useful analyses of similarities.

5.15 Hydrocarbon Exposure Studies

These will make use of all the correlation analyses. This may be a special variant of the pollution uptake studies. Multivariate analyses and analysis of variance will test effects of various petrochemicals on growth and survival of biota. Non-parametric tests will be useful. Both bioassay and probit analyses are likely choices for obtaining and analyzing data.

5.16 Benthic Respiration

Benthic respiration (seabed oxygen consumption) rates are an indicator of organic loading and other impacts to the benthos. The objective is to detect abnormal variations in organic loading. This requires the establishment of both temporal and spatial baselines of the natural system in both and uncontaminated areas as well as laboratory tests to illustrate the nature of the loading or stress versus the response of the system. Multiple regression, analysis of variance, multivariate analysis, time series analysis and non-parametric methods are most appropriate for achieving the objectives.

5.17 Total Plankton Respiration

Total plankton respiration rates are an index of the rates of decomposition of organic matter (utilization of oxygen) and the concurrent regeneration of nutrients required for phytoplankton growth. The objective is to detect major shifts in the temporal, spatial or size component distribution of plankton respiration. This requires the establishment of temporal, spatial and size component baselines of the natural system in both contaminated and uncontaminated areas as well as laboratory and/or shipboard experiments to illucidate the response of the system to contaminants and/or other stresses. Multiple regression, analysis of variance, multivariate analysis, time series analysis and non-parametric methods are most appropriate for achieving the objectives.

5.18 Phytoplankton Biomass and Primary Productivity

Chlorophyll a pigments are used as an index of phytoplankton biomass. We are particularly concerned with the relationship between eutrophication and shifts in abundance as well as shifts in size classes of phytoplankton (chlorophyll) which may alter marine food chains. Correlation and multivariate analyses will be applied to ascertain relationships between inorganic and organic nutrients, heavy metals, and phytoplankton chlorophyll. Measurements of primary productivity (via ^{14}C methods) will be correlated with phytoplankton biomass measurements, as well as measurements of nutrients, metals, light and other oceanographic data to determine principle forces affecting organic production.

5.19 Nutrient Studies

Inorganic nutrients (nitrates, phosphates, silicates, etc.) and organic nutrients will be related to spatial and temporal distributions of pollutants (metals, hydrocarbons, etc.). Nutrients will also be correlated with physiological assays as well as with primary productivity measurements to determine which nutrients are driving forces behind production. Multivariate analyses and multiple regression tests will be employed.

6. Synthesis

Synthesis of the Ocean Pulse analysis encompasses the effects of natural and man-induced stresses on marine ecosystems and living resources. The program should emphasize not only an integrated trend index analysis for marine pollution problems, but also develop an understanding of an environmental system and living resources as a whole.

The integrated trend analysis is mainly rated on baseline data of the occurrence of marine pollutants, physical and chemical factors and their effects on many species from lower trophic levels to higher levels. The synthesizing trend interpretation also requires basic criteria for monitoring parameters as standard measurements from effects observed under laboratory conditions. These can be extended or extrapolated to the natural marine environment and living resources. The integrated environmental systems-oriented analysis deals with a total environmental system. The natural and man-induced stresses are effects on the food chain dynamics and energy flow system, species composition and community structure, biomass changes and the relationship between living resources and their supporting environment.

6.1 Trend Index Interpretation

6.1.1 Determination of indicator parameters for monitoring

The right selection of ocean monitoring parameters is essential for project success. The parameters are the biological, chemical, and physical factors necessary for a synchronized trend analysis and systems interpretation. They are a crucial linkage of species, nutrients, heavy metals, pollutants, parasites, pathogens and other selected foci. These measured and/or estimated parameters in the water column, sediment and/or organisms determine the trend indices -- their interpretation will provide appropriate monitoring schemes.

6.1.2 Establishing criteria of the key parameters for monitoring under laboratory conditions.

First, pertinent parameters are recognized and determined by their roles within natural and man-induced stress environmental and ecological systems. Then, following the establishment of criteria for describing tolerance limits on biological responses. Directly or indirectly, growth, survival, health, and other attributes of marine organisms influenced by varying environmental quality must be examined. In other words, we have to establish the range of threshold values of parameters which affect survival and influence the process in which key species cope with man-induced stresses (e.g. heavy metal influx) and natural mortality factors (parasites and pathogens). Without these criteria, any monitoring activities are purely exercises of data collection documentation.

An important aspect of establishing criteria is how to consider the problems of multiple exposure of pollutants, heavy metals, or other contaminated matter. Synergistic effects behave in a compounded fashion. These may not be easily interpretable as a single exposure case, or may not be even detectable as the compounded responses. If the measurements of multiple exposure of stimulants are available, the criteria may be obtained by the method of bioassay with factorial designs and may be interpreted by utilization of canonical correlation techniques.

6.1.3 Determining correlations of the criteria to survey field data

The applications of established criteria of the key parameters for marine environmental conditions on various man-induced stresses should be directly utilized from the survey field data. Ideally, onboard inspection and analysis of the samples is desirable to detect abnormalities and for monitoring and diagnosis of marine organism health on a real-time basis.

6.1.4 Interpretation of natural fluctuations and man-induced stress processes, i.e. contrasts of contaminated against pristine areas

This is a logical proposition for monitoring marine environment. However, the interpretation of the results require extreme caution for practical applications in monitoring the marine environment. The spatial and temporal marine environmental conditions from which the samples are obtained are influenced by so many external variables and constraints. These natural variables and constraints make identification of aberrant levels or oscillations extremely difficult and interpretation tenuous.

6.1.5 Time series interpretation

Once a series of observations for many desirable variables is compiled from the field, the examination and interpretation of time series analysis provides the means of monitoring schemes for the environmental fluctuation and changes which are closely related to the abundance of marine organisms and their community structures. As we have described in an earlier section, the analyses of trend, seasonal variation, oscillatory phenomena and random fluctuation processes are required the meaningful interpretation of significant changes in the measured or estimated environmental parameters. Utilizing this basic information will provide timely advice and warning to management so appropriate actions may be taken.

Preferably, the interpretation of marine population cycles or successions and environmental parameters should require extreme caution in environmental assessment. This is simply because many cases of marine population successions and environmental parameters may be essentially natural random fluctuations with serial correlation between the populations and their environment in successive years. We should focus attention upon the processes of marine population dynamics as a whole; upon growth and decline processes, health problems with various environmental limiting factors and carrying capacity of given environments as well as unexplainable environmental changes and their parameters. These lead

to a broadly scoped monitoring scheme for a total ecosystem evaluation for any environmental management.

6.2 Systems Oriented Interpretation

6.2.1 Ecosystem change monitoring

6.2.1.1 Food chain and energy flow dynamics

The study of food chain and energy budget flow dynamics in the marine environment describes the dietary components and interrelation between trophic energy transport. The study also identifies not only the process of competition, predation, interactions and energy flow among organisms, but also estimates the effects of transmitted parasites, pathogens, heavy metals, etc. and their routing from lower to higher trophic levels within the marine environment. Such a continuing monitoring effort will achieve the objectives of the Ocean Pulse.

6.2.1.2 Species composition and community structure

Similarly, analysis of food chain and energy flow dynamics, species composition and community structure changes within a given marine environment will provide a monitoring technique for natural and man-induced stress effects. It requires a standard mechanism or criteria for detecting and distinguishing differences of normal or abnormal conditions. Species composition and community structure changes in a given marine environment, i.e. spatial and temporal variations will be the input for interpretation of marine environmental assessment. The main problem in attaining the stated objectives will be that of establishing an acceptable healthy marine environmental model. Achievement of this model will result from synthesizing the various inputs of individual disciplines. The criteria for defining aberrancy and delineation of causal effects will depend on a long series of insightful analysis.

6.2.2 Biomass change monitoring

The measurement of biomass changes over time is another way to attempt a meaningful monitoring in population changes of marine environment. The measurement of absolute values of total biomass in the marine environment is an ideal, but the actual figure is impossible to obtain. The relative biomass indices are computed on the basis of quantified relative contribution of time periods expressed in terms of an arbitrary standard time period base. The index of the overall species relative biomass throughout the time periods relates to spawning success, survival and growth within a given marine community. However, as an alternative, we can select a few indicator species for monitoring relative biomass changes over time. The choice should be based upon forms in a well delineated and known food web and community structure organization. Both major and minor elements should be included from each trophic level in the subset. It will then be easier to monitor any changes in biomass of the subset. Again, we should emphasize detecting and distinguishing natural fluctuations from those caused by man-induced stresses.

7. FEEDBACK

For our project or analysis to succeed and to minimize the errors between what it is doing and what it intended to do to meet its objective, it must somehow monitor its own activities. It must feed back a portion of its output results for comparison with its input. Finding the cause of defectiveness and the optimum solution for a given problem is usually difficult and requires honest introspection. Thus, trial comparison of several alternatives can determine the best resolution for a given problem. The continuing verification of experimental alternates with realignment of the objectives under given constraints is the feedback process.

The selection of alternatives for the optimum solution should be associated with the prechosen criteria. A criterion is a rule or standard for ranking the alternatives in their order of desirability and indicating the most promising within fixed contingences, i.e. it usually provides a means for weighing cost against performance within fixed contingencies, we must compute for each solution the expected value of effectiveness measured and choose the solution that has the highest expected effectiveness, assuming equal cost. We may also employ the maximum procedure for measure of effectiveness.

For some of these contingencies, there may be available either sufficient data (the constraints imposed on Ocean Pulse are the contents of the data themselves) or sufficient theory so that we know the probability of occurrence of each contingency. At the present time, we do not know how to determine the probability distribution for the system which will deliver the expected measure of effectiveness. Furthermore, if we construct some kind of robustness test for the alternatives and the best solution, then such tests may be used as the

main body of criteria. These robustness tests and expected value criteria should be based upon either some known probability distribution (parametric) or completely distribution-free (non-parametric) method, so that they are mainly dependent on the structure of the system or model, set of alternatives and data themselves.

8. DATA MANAGEMENT*

Ocean Pulse is not a limited study involving but a single discipline. If it were, data management would not need to be formally structured. The testing of the Ocean Pulse project hypotheses will be attainable only through multidisciplinary studies. The goals and objectives are derived from all the disciplines and investigators in any one discipline do not necessarily provide the total input in the resolution of questions. Project activities are interdependent.

The experimental design of each project is essentially determined within the project in consultation with biostatisticians. The project data bank will lie in the NE Regional ADP System at the Sandy Hook Laboratory and its data processing will function in archiving and updating files. Formats used will be amendable to conversion to NODC files. To attain these objectives the following description defines terms and a system to be used in data operations. It is intended to provide a backdrop to researchers in planning their activities.

8.1 Introduction

The goals of Ocean Pulse include:

a. The collection and integration of data sets which assist in understanding the nature and driving forces of complex marine ecosystems.

b. The creation of a data bank for use by a variety of users including the public, scientists, and administrators.

The realization of these goals requires a systematic approach in the organization and storage of data for maximum benefit to users in access and retrieval; we call this approach data administration. The development of this foundation for organizing information is intended to avoid costly duplication of effort wherever possible.

---

* This material adapted from "Data Administration for Marine Ecosystems Analysis", NOAA Tech. Memo. ERL & MESA-36 by P. A. Eisen, A. Sadler, Jr., and M. E. Sheffler.

Information is a decision-making and research tool. Efficient data systems can make a large amount of relevant information readily accessible. The data administrator holds responsibility for convincing scientists that the services provided by data systems can be used in the solution of complex problems. Assuming you have a rudimentary knowledge of computers, we will illustrate some ways to effectively use data administration in marine environmental research.

This report presents a methodology of data administration. This methodology has been adopted in some degree by the other NOAA Programs. We hope its presentation here will encourage a dialogue for scientists and decision makers to the data services they require.

The central aim of our data administration is to make data obtained from research accessible to users. To accomplish this, the responsibility for data archival and retrieval has been transferred from scientists to data centers via the ADP staff. The reason behind this transfer of responsibility is that it both frees investigators from time-consuming tasks and offers several advantages to data users. Direct informal exchanges of data among scientists and others also occurs and can be efficient. The ADP can facilitate informal data exchanges by personal referrals to appropriate sources.

8.1.1 The Freedom of Information Act

In compliance with the Freedom of Information Act, unclassified data and information, whether produced, sponsored, collected, or obtained by the Project, reside within the public domain. It is the policy of NOAA (NOAA Directives Manual: Chapter 21, Section 25) to supply these data and information by load, exchange, or sale (at cost of reproduction) through the ADP Office and the Environmental Data Service (EDS). Requests for data or information are handled expeditiously, usually within ten days when possible.

8.1.2 Data Necessary for Project Success

Two principal tasks of Ocean Pulse are:

1. To identify and describe the major existing ecological systems, processes, stresses, and responses operating in the Middle Atlantic Bight, and define their relationships and rates of change.
2. To determine the types, transport rates, fates and impacts of pollutants, and other people-related stresses on the ecosystem.

The extent to which Ocean Pulse output furthers accurate assessments and predictions of marine cological impacts will be a criterion of its success. Such success is predicted on the type of data acquired and processed, its statistical validity, and the quality of its technical interpretation. Evaluators of the data administration will require user needs to be met properly with sifficiently detailed data.

8.1.3 Initial Project Plans for Data Administration

This framework for data administration is cognizant of the unique nature of study and the need to outline the relationships among participants. Some guidelines for data administration standards and responsibilities follow.

8.2 Analysis of Available Systems

It may be helpful to review the technological perspective on which data administration systems are based. Following that is an analysis of strategies for handling data that are in common use today.

Much of today's computer information technology evolved because of a need for a generalized tool for handling large banks of data reposited on computer storage media (e.g., magnetic and paper tapes, disc packs, punch cards, magnetic core). Out of this need grew Data-Base Management Systems (DBMS), Information Retrieval Systems (IRS), and Management Systems (MIS). Though the differences between the above systems are, in some cases, subtle, we will not concern ourselves

with individual aspects or goals of these systems, but review qualities that are common and fundamental to all three systems.

Data administration technology can be traced back to the late fifties when the success of "generalized" routines were first discussed. These routines can sort the components of any data set (file) regardless of its content. The significance of this work was the proposal that these ideas be extended into other areas, such as data set maintenance and report generation. This generalized processing entails the building of special programs which perform frequently used, common, and repetitive data processing tasks. The benefits of such a generalized approach are the elimination of program duplication, and the amortization of one-time development costs over many applications of the program. Generalized data processing techniques have evolved into a class of sophisticated, generalized systems (DBMS, MIS, IRS) and have helped establish concepts of data administration technology.

The origin of data administration technology also stems from data definition languages development and report generator packages of the fifties. Data definition languages provide a facility for describing data-bases that are accessed by multiple users diverse application programs. Thus, the structure of data can be defined to avoid special programming effort by the user.

The development of report generators stems from the need to produce good reports without large programming efforts. In most cases, report generators can perform complex table transformations and produce sophisticated reports from a data-base. Thus, these allow the user to examine, process, and summarize large volumes of data fairly easily.

The implementation of data administration tools (e.g., DBMS, IRS, MIS) rests on organizational schemes which have been characterized in three commonly used strategies: brute force, piggyback, data-base/key-task. We can also call these strategies: (a) traditional/inflexible, (b) traditional/flexible, and (c) data-base/key-task. The first word of the strategy titles (a), (b), (c), indicates the way data are stored, i.e. using a traditional method or a data-base. A slash separates the strategy titles into a second half which refers to ways that data can be retrieved.

All the strategies use the terms, fields, records, and files. Each data value or piece of raw information a system stores, retrieves, and processes is called an elementary data item. A data item is placed into a named storage location called a field. A collection of data items or fields is called a record. Records are collected into logical units called files. Files are made up of records having an important feature in common (e.g., all from a single cruise).

In the traditional/inflexible and traditional/flexible strategies, data files are the principal structures for organizing data. These data can be distributed into compartmentalized and clearly defined units called files which are loosely linked in some way for retrieval purposes. In this report, a program is a sequence of instructions written in some computer language. The program will always use data, possibly taking the data from files, to perform desired operations.

8.2.1 The Traditional/Inflexible Strategy

This strategy for storing and retrieving data is one of the earliest used techniques and is still common. The word "traditional", describes ways of storing data, means that data are collected into a file, but the data in the file can be read only by a specific program. Each file essentially becomes glued to a specific program, and is not versatile. The retrieval aspect of this

strategy is inflexible because a newly created program cannot simply use data that resides within a given file. If a program is written that needs some data in an existing file, a totally new file must be created, copying the pertinent data from the original file (Fig. 1).

The duplication of effort involved in recopying data into the new file is inefficient and introduces error. If an update of data in one file is made, it must be remembered that values from data are also in other files. The result is that one occurrence of the data is edited, while another is not. The discrepancy may not be noticed until other uses of the file have been made. Tracking the error is time consuming and the original inefficiency is compounded.

This approach to data storage and retrieval also does not take advantage of recent advances in computer hardware. It is now feasible to keep relatively high amounts of data alive in on-line storage systems since computer memory is cheaper today. The development of large capacity disc devices has also greatly reduced the costs of random-access storage. These are invitations to adjust data storage schemes to maximize potential user benefits.

8.2.2 The Traditional/Flexible Strategy

This is the present situation in the Sandy Hook operation. As in the traditional/inflexible strategy, this strategy, of data storage is traditional in that data files are the structures used to organize data, but these data files are constructed to allow data retrieval to become flexible. Figure 2 shows the organization of this strategy. The one-to-one correspondence between data files and recurring programs still holds, but the files are organized so that they are centrally located and available to a team of programmers. When data values from existing files are needed, the values can be pulled from the files and put into a special data pool. Data values not in the files can be added to the special data pool.

The special data pool represents a particular set of data needed to solve a problem. Any number of data sets can be constructed for the special data pool. Data sets in the special data pool can be generated by a looping routine. First, data values are taken from a data file and augmented with additional raw data, thereby forming the special data pool. Then the special data pool is fed into an interface system for special applications (a package combining specialized and commercial software) which produces the desired output. The looping routine can return to a second data file and repeat the process until terminated.

The disadvantage of this strategy lies in the necessity to construct a data pool from the current files. Work has already been done to put the data values into the system, but additional effort must be extended to write a software package that strips the data values from existing files and also inserts new ones into the special data pool. Any advantage that accrues to this flexible data retrieval capability depends on the development of an efficient data-independent interface system for special applications.

8.2.3 The Data Base/Key Task Strategy

This is the system to which we are developing. In the data-base/key task strategy, individual files become an optional means for storing data. Within the data-base storage system, data values are translated into computer readable data which are then merged into a single conceptual storage entity called a data-base. In a rough way, a data-base can be considered a giant file, because the computer readable data are not connected in an arbitrary way. This macrocosm called a data-base is predicated on an underlying logical system devised by defining key-tasks. The definition of key-tasks results from a

comprehensive evaluation as to the types of data that will be collected and the general applications required of the data. The way data are to be used thus plays a role in where a data value is stored within the data-base and how that data value is linked to the rest of the data-base for retrieval purposes.

Figure 3 gives a visual breakdown of the components in the data-base/key task strategy. The cylinder in Figire 3 represents the storage area of data values, i.e., the data-base. Raw data values are coded, inserted into their particular place in the data-base, and exist in that place as computer readable data until it is necessary to examine or update them.

The octagon in Figure 3, the general data-base interface system, contains software that accesses data values and performs operations on data values. If updating data values is desirable for economy or efficiency, the general data-base interface system does to work. This system facilitates the care and grooming of the data-base by the programmer. Since the general data-base interface system accesses data values, it also extracts input data for the running of routine key-task programs.

Bacause specialized sophisticated needs arise and must be accommodated, an additional software system is available. It is called the interface system for special applications and appears in Figure 3 as a six-sided polygon. The interface system for special applications answers ad hoc requests and produces solutions by skillfully utilizing data values made available via the general data-base interface system.

An interface system for special applications is also a feature of the previous traditional/flexible strategy. The data-independent nature of this system is important to both strategies because the versatility of the system

is enhanced. But the data storage differences between the two strategies affect the end results of the interface system for special applications. In the traditional/flexible strategy, the data storage pool must access data values from various files. Each file is built with its unique logical structure. The retrieval of data values from several files requires cognizance of each structure and, therefore, can become unwieldy and inefficient. Given the constraints on the data accessibility, the traditional/flexible strategy yields limited ad hoc reporting programs.

In contrast, the data-base storage system allows the interface system for special applications a greater range. Data values reside in an interlocking structure, the data-base, whereby they can be readily successed. Data retrieval for any needed data values proceeds uniformly by using the general data-base interface system as a tool. As a result, greater responsiveness to ad hoc requests accrues to the interface system for special applications.

One constraint on the use of the data-base/key task strategy for administrating data lies in the definition of key-tasks. If scientists and administrators focus on key-tasks that use much or all of the project's data and require extensive integration of data types, then organizing the data-base becomes complex. In the long run, the data-base/key-task approach is usually the most expedient and cost-efficient approach for data retrieval. However, its successful inplementation depends on the ability to identify key-tasks, and then insure that the data processing personnel, who are responsible for structuring and maintaining the data-base clearly understand them.

8.3 The Design and Rationale Project Data System

For Ocean Pulse, a system that integrates the traditional/flexible strategy and data-base/key-task strategy is planned. A strict application of the traditional/flexible strategy does not respond to the project's needs. Data requests from the public are handled routinely. It is not practical to constantly strip data from existing files to form the special data pool in response to many ad hoc demands. Tagging into files with unique logic structures requires regular modification of programs and subroutines to operate similarly in different data files.

On the other hand, a data-base/key-task approach requires a comprehensive evaluation as to the types of data that will be collected and applications for the data. Definition of key-tasks necessitate that the comprehensive evaluation be an ongoing process, subject to constant revision.

There also is a concern in the scientific community that parallels the invasion of privacy issue raised by the public in regard to large computer systems. Scientists usually have a proprietary attitude about data they have collected and are apprehensive about the possible premature use of the data by others. The data could reside in the data-base after initial reduction but before the scientist has completely edited them (i.e., eliminated all erroneous values). Working via a data-base can raise this concern as well as a concern about data loss and inaccessibility in a big system environment.

In summary, a synthesis of both the traditional/flexible strategy and data-base/key-task strategy can be successful. Most data collectors must organize and specially structure the data values of their own files. The data collectors do this using data formats that are designed by the Data Coordinator.

The prototype of the traditional/flexible strategy has centrally located, individualized and logically unique files from which data values are pulled. The data values are then held inside a special data pool. Within the data system, the use of coordinated data formats, resulting in data files structured for interface, negates the need for the special data pool. When demands are made of the data values in the files, the interface system for special applications, consisting of a high-level programming language, works directly and efficiently with the specially structured data files.

8.4 The Data Catalogue

8.4.1 Background Theory

The data system diagram in Figure 4 shows the data catalogue branching from data collection. The data catalogue is produced through the joint actions of the computer technician and Data Index. The data catalogue is a resource devised to display the current status of data collection efforts. These collection efforts will generate many data files. The data catalogue defines the collected files, what they contain, who has them, and whether they are available for retrieval. The data catalogue can be compared to a card catalogue in a library. The data catalogue is consulted to ascertain the Project's holdings, just as the library card catalogue is consulted to ascertain the library's holdings.

The data catalogue is organized in the following way: Each work unit has one or more cruises undertaken to gather necessary samples. The samples gathered from each cruise are used to measure pertinent parameters (e.g., incident radiation, carbon assimilation rate). Each parameter has common information reported as to its accessibility and sampling frequency (e.g., name of scientist responsible, number of stations). These qualitative details are entered into

the data catalogue. Since the data catalogue is much smaller than the processed data file it describes, it is an efficient tool for locating needed data files.

8.5 Data Archival and Retrieval

Project data services utilize the interface system for special applications (Figure 4). The interface system for special applications is geared to operate through special data storage formats.

8.5.1 Data Formats

The approach used on format development is the specification of a common structure that can be applied to most data sets regardless of content. The result is a set of formats for difference types of data which are linked by a common framework. The consequent degree of standardization has facilitated data retrieval.

Project data sets are put into a structure called network. The theory behind network is as follows. Individual records having the same format are grouped into a record type. A family of record types composes a data file. Each record type must be linked to another record type in some way in order to build the structure of the data file. Linkages of record types are accomplished by connecting each record of one record type (owner records) to any other records of other record types (member records). We say the linkages of all owner records in record type 1 to all member records in record types 2,3...,N define a '1-2-3...-N' set type. The constraints which can be put on set types differentiate networks (e.g., given record types 1,2,3,4, let a set type include owner records in record type 1 and member records belonging only to record type 2).

A form of network is commonly called a tree structure. Here, an owner can have any number of members (a limb can have any number of branches), but the convention used is that no record can act as a member record for more than one set (e.g., no branch is attached to more than one limb). This structure allows us to identify relationships among records in the data file. All direct relationships are inserted onto each record as keys and usurp a certain amount of space in the files.

8.6 Anticipated Requirements

Early in 1979 a questionnaire was circulated to task leaders to determine what information they anticipated gathering, at least for the preliminary phases of OP. The list of questions included type of field and laboratory data, how recorded, sets per station, statistical analyses, format status and objectives.

The response was good but not unanimous. One of our objectives was to determine requirements for building data files, etc. A certain degree of naivete and resistance appeared from some quarters and some guidance is needed to direct investigations in adopting adequate record keeping techniques to inplement computer file record development. Design of formats was requested by investigations at Milford associated with contaminant biochemistry. The genetics group has a format in development. Other format design is needed for pathology, microbiology and some chemistry (unless the present heavy metals format is adopted).

Data volumes have been estimated as follows (Table 8) for the following investigators:

Thomas Robertson Phoel	15,000 cards per year 1,000 cards
Mahoney	approx. 2,000 cards per year
Cohn	approx. 1,000 cards per year
Reid	25,000 cards per year
Radosh	7,500 cards per year
O'Reilly	20,000 cards per year
Evans	50,000 cards per year (2 cruises per year)
Waldhauer	4,000 cards per year (2 cruises per year)
Zdanowicz	7,000 cards per year
MacKenzie	1,000 cards per year
Longwell	?
Murchelano Ziskowski	1,000 cards per year
Calabrese Gould	16,000 cards per year
Thurberg	7,200 cards per year (2 cruises per year)
Graikoski	4,800 cards per year (2 cruises per year

This results in a minimum of 187K cards per year on 2 cruises per year.

The index data file types, and analysis programs are summarized in Table 9.

Future problems are difficult to identify but one procedure should be made eminently clear. The investigators making observations at a given station should all use the same station identification. The integration of data between disciplines will be effectively conducted only if key indices can be identified between files. Contractural arrangements with a systems analyst would be an effective procedure to develop a viable structure for data management.

NODC has developed a number of formats which are on file in the Sandy Hook computer offices. For general information and review the following are available:

Seabed Oxygen Consumption
Water Column Respiration
Index of Relative Importance (stomach analysis)
Zooplankton
Intertidal/Subtidal (sediment, specie, fish, stomach)
Marine Invertebrate Pathology
Trace Elements (heavy metal)
Mutagenesis
Photosynthetically Active Radiation
Primary Productivity 2
Hydrocarbon 2 (sediment, organism, water)
Fish Resource Assessment
Hydrocarbon 1
Primary Productivity
Phytoplankton Specie
Specimen Feeding Studies (food sample content)
Fish Resource Assessment (shellfish)
Water Physic and Chemistry
Marine Fish Pathology
Bacteriology
Fin Rot
Benthic Macrofauna File
Metal in Organisms, Sediment and Water
Sediment Characteristics
Benthic Organisms

	Benthic Resources	Ocean-ography	Prim. Prod.	Chemistry	Micro-biology	Surf clam	Contam-inants	Patho-biology	Genetics
	Reid Radosh	Thomas Robertson Phoel	O'Reilly Evans	Zdanowicz Draxler Waldhauer Matte	Cohn Mahoney	MacKenzie	Greig Graikoski Calabrese Gould Thurberg	Murchelano Ziskowski Sawyer	Longwell
Cruise									
Sta. Grab									
Date									
Time (local GMT)									
Latitude									
Longitude									
°C temperature (Bot. Surf.)							X		
Depth									
Salinity							X		
D.O.							X		
Sediments	X								
% silt/clay	X								
% organic	X								
Sorting index	X								
Macrofauna	X					X			
Metals	X			X			X		
Primary productivity		X	X						
Chlorophyll			X						
Nutrients				X	X				
Hydrocarbons									
Blood chem.							X		
Pathology								X	
Enzymes							X		
Oxygen consumption		X					X		
Bacteria							X	X	
Chemistry			X	X	?		X	X	
Genetics							X		X
Phytoplankton Dist./Abund.			X		X				
Zooplankton									
Applications Programs									
Diversity/equitability	X								
Cluster analysis	X				X				X
Length frequency						X			

Table 6. Ocean Pulse interaction between studies and environmental elements.

	Community Diversity Equitability	Seasonal Abundance	Succession	Anaerobes	Calorimetry	Physiological Activity	Parasites	Virology	Anomalies (gross-histo)	Nutrient Bioassay	Uptake	Genetic	Petroleum Bioassay	Limiting Factors	Hydrocarbon Exposure	Benthic Respiration	Plankton Respiration	Primary Productivity	Nutrient Studies
Water																			
Temperature	X															X	X	X	X
Constituents	X			X												X	X	X	
Chlorophyll		X																X	
Phytoplankton	X	X	X							X						X	X		
Sediments	X	X		X												X			
Benthos	X				X		X°		X°		X					X			
Nekton					X		X°		X°								X		
Mysids						X			X						•				
Isopods						X													
Euphausids						X													
Crangon						X									0				
Rock Crab															0				
Lobster						X									0				
Tellinoid Mussel						X	X		X										
Scallop						X			X						0				
Surf Clam						X								X	0				
Ocean Quahog						X													
Herring							X	X											
Flounders																			
Winter												X*	**						
Yellowtail							X	X				X*	**						
Windowpane												X*	**						
Hake																			
Red													**						
Silver							X	X											
Sand Lance									X										
Cod							X	X				X*	**						
Haddock							X	X				X*	**						

° - Crustacean
* - Finfish eggs available
**- Fish larvae and gonad development
0 - Eggs and larvae (invertebrate)

9. FUTURE PERSPECTIVES

We have passed through a description of several stages in the development of Ocean Pulse activities: the examination or sampling procedures required for collecting laboratory and field survey data, then the treatment phase with applications of statistical methods, next the diagnosis phase synthesizing the various research unit results with feedback refinement procedures. The final step in the series is construction of recommendations to implement management practices for ocean welfare. Recommendations will draw from some predicting capacity to anticipate emergencies. This predictive ability may evolve from the development of ecosystem models. Although Ocean Pulse will be primarily concerned with biological aspects of modeling, the economic models must also be considered.

Future coastal ocean management activities could proceed along these steps:

1. Construct a small-scale model based on well established ecological links. Derivations include food chain and energy budget dynamics.
2. Expand the elementary model to a total ecosystem model. Linkages between the several compartmental model systems.
3. Extend implications of the biological model to economic and sociological impacts. This action would present an integrated approach to a national coastal ocean management system.
4. Utilize techniques of dynamic programming to develop such a management system. This process will define those conditions which must be satisfied by an optimal time -- staged decision process. We will discover what conditions will result in a best strategy for monitoring ocean welfare.

The biological models are concerned with energy flow and yields. The ultimate operating model, however, will probably be the economic involving maximization of benefits. Research will be supported only from the political premise that assures certain things are being done to support a "status quo" of the environment.

In the absence of attitudinal studies toward the marine environment, we can infer public attitudes are derived from the common-property status of marine resources. Environmental requirements affecting water quality, resource abundance, palatability, and food chain continuity are paramount. Maintenance of the aesthetic impressions of a shoreline or fishing experience is also important. We must understand these as given rights and benefits to the community of citizens. Considerations, such as these will ultimately govern management actions.

Table 7. Summary of suitability of various statistical methods for each of the study disciplines. Coding as follows (1) definitely appropriate; and (2) possibly useful.

	Community.Diversity .Equitability	Seasonal Abundance	Succession	Anaerobes	Calorimetry	Physiological Activity	Parasites	Virology	Anomalies (gross-Histo)	Nutrient Bioassay	Uptake	Genetic	Petroleum Bioassay	Limiting Factors	Hydrocarbon	Benthic Respiration	Plankton Respiration	Primary Productivity	Nutrient Studies
Regression																			
Simple	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Multiple	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Correlation																			
Simple	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Partial	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Multiple	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Multivariate																			
Principal Component	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Discriminant Function	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Canonical Correlation	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Analysis of Variance	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Analysis of Covariance	2	2	2	2	1	1	2	2	1	2	2	1	2	1	2	1	1	1	1
Bioassay	2	2	2	1	2	1	2	1	2	1	1	1	1	1	1	2	2	2	1
Probit Analysis	2	2	2	1	2	1	2	1	2	1	1	1	1	1	1	2	2	2	1
Time Series	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Non-parametric	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Diversity	1	1	1	2	2	2	1	1	1	2	2	1	2	1	1	1	1	1	2
Similarity	1	1	1	2	2	2	1	1	1	2	2	1	2	1	1	1	1	1	2
Cluster	1	1	1	2	2	2	1	1	1	2	2	1	2	1	1	1	1	1	2

10. LITERATURE CITED

Afifi, A. A., and S. P. Azen. 1972. Statistical analysis: a computer oriented approach. Academic Press, New York.

Anderson, R. L., and T. A. Bancroft. 1952. Statistical theory in research. McGraw-Hill Book Co., Inc., New York.

Anderson, T. W. 1958. An introduction to multivariate statistical analysis. John Wiley and Sons, Inc., New York.

Bartlett, M. S. 1947. The use of transformations. Biometrics, 3:39-52.

Bliss, C. I. 1934. The method of probits. Science 79:38-39 and 409-410.

Bliss, C. I. 1935. The calculation of the dosage-mortality curve. Ann. Appl. Biol. 22:134-167.

Bliss, C. I. 1952. The statistics of bioassay. Academic Press, New York.

Bliss, C. I. and D. W. Calhoun. 1954. An outline of biometry. Yale Cooperative Corp. New Haven, Conn., 272 pp.

Boesch, D. F. 1977. Application of numerical classification in ecological investigations of water pollution. EPA-600/3-77-033.

Clifford, H. T. and W. Stephenson. 1975. An introduction to numerical classification. Academic Press, New York.

Cochran, W. G. 1942. The chi-square correction for continuity. Iowa State Coll. J. Sci. 16:421-436.

Cochran, W. G. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics, 3:22-38.

Cochran, W. G. 1952. The chi-square test of goodness of fit. Ann. Math. Stat., 23:315-345.

Cochran, W. G. 1954. Some methods for strengthening the common Chi-square tests. Biometrics. 10:417-451.

Cochran, W. G. 1977. Sampling techniques. John Wiley and Sons, Inc., Third Edition, New York.

Cochran, W. G. and G. M. Cox. 1957. Experimental design. Second Edition. John Wiley and Sons, Inc., New York.

Cochran, W. G., F. Mosteller, and J. W. Tukey. 1954. Principles of sampling. J. Am. Stat. Assoc., 49:13-35.

Conover, W. J. 1971. Practical nonparametric statistics. John Wiley and Sons, Inc., New York.

Cooley, W. W. and P. R. Lohnes. 1971. Multivariate data analysis. John Wiley and Sons, Inc., New York.

Cornfield, J. 1964. Comparative bioassays and the role of phamacology and experimental therapeutics. 144:143-149.

Cox, D. 1958. Planning of experiments. John Wiley and Sons, Inc., New York.

Croxton, F. E. and D. J. Cowden. 1939. Applied general statistics. Prentice-Hall, Inc., New York.

Dahlberg, M. L. 1969. A users guide for linear regression. M.S. Department of Forestry and Wildlife, Virginia Polytechnic Institute.

Davis, H. T. 1941. The analysis of economic time series. The Cowles Commission for Research in Economics. Monograph No. 6.

Dixon, W. J. 1974. BMD, Biomedical Computer Programs. University of California Press, 773 pp.

Dixon, W. J. 1977. BMD, Biomedical Computer Programs. University of California Press.

Draper and Smith. 1966. Applied regression analysis. John Wiley and Sons, Inc., New York.

Eisen, P. A., A. Sadler, Jr., and M. E. Scheffler. 1978. Data Administration for Marine Ecosystem Analysis. NOAA Tech. Mem. ERL MESA-36, 77 pp.

Eisenhart, C. 1947. The assumptions underlying the analysis of variance. Biometrics, 3:1-21.

Everitt, B. 1974. Cluster analysis. John Wiley and Sons, Inc., New York.

Federer, W. T. 1955. Experimental design. The MacMillan Co., New York.

Finney, D. J. 1964. Statistical method in biological assay. Second Edition. Charles Griffin and Co., Ltd., London.

Finney, D. J. 1965. The meaning of bioassay. Biometrics, 21(4):785-798.

Finney, D. J. 1971. Probit analysis. Third Edition. Cambridge University Press, Cambridge.

Fisher, R. A. 1947. The design of experiments. Fourth Edition. Oliver and Boyd, Ltd., Edinburgh.

Fisher, R. A. and F. Yates. 1964. Statistical tables for biological, agricultural and medical research. Sixth Edition. Oliver and Boyd, Ltd., Edinburgh.

Gonor, J. J. and P. F. Kemp. 1978. Procedures for quantitative ecological assessment in intertidal environments. EPA-600/3-78-087.

Grosslein, M. D. 1969. Groundfish survey program of BCF, Woods Hole. Comm. Fish. Rev. Vol. 31, No. 8-9, pp. 22-35.

Grosslein, M. D. 1971. Some observations on accuracy of abundance indices derived from research vessel surveys. Int. Comm. North Atl. Fish., Redbook, Part 3, pp. 249-266.

Gulland, J. 1966. Manual of sampling and statistical methods for fisheries biology, Part 1. Sampling methods FAO of the United Nations Manuals in Fisheries Science, No. 3.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow. 1953. Sample survey methods and theory. Vol. 1 and 2. John Wiley and Sons, Inc., New York.

Hennemuth, R. C. 1976. Variability of Albatross IV catch per tow. Int. Comm. North Atl. Fish. Res. Doc. 76/6/104, 18 pp.

Hollander, M., and D. A. Wolfe. 1973. Nonparametric statistical methods. John Wiley and Sons, Inc., New York.

Jacob, F. and G. C. Grant. 1978. Guidelines for zooplankton sampling in quantitative baseline and monitoring programs. EPA-600/3-78-026.

Kempthorne, O. 1952. The design and analysis of experiments. John Wiley and Sons, Inc., New York.

Kendall, M. G. and A. Stuart. 1961. The advanced theory of statistics. Vol. 2. Hafner Publishing Co., New York.

Kendall, M. G. and A. Stuart. 1966. The advanced theory of statistics. Vol. 3. Hafner Publishing Co., New York, 552 pp.

Li, C. C. 1964. Introduction to experimental statistics. McGraw-Hill Book Co., New York.

Miller, R. L. 1966. Simultaneous statistical inference. McGraw-Hill Book Co., Inc., New York.

Mearns, A. J. and M. J. Allen. 1978. Use of small otter trawls in coastal biological surveys. EPA-600/3-78-083.

Mendenhall, W. 1975. Introduction to probability and statistics. Duxbury Press, New York.

Natrella, M. G. 1963. Experimental statistics. National Bureau of Standards Handbook, No. 91.

Neyman, J. 1934. On the two different aspects of representative method: The method of stratified sampling and the method of purpositive selection. J. Roy. Stat. Soc. 109:558-606.

Pielon, E. C. 1974. Population and community ecology: principles and methods. Gorden and Breach Science Publishers, New York.

Pielon, E. C. 1977. Mathematical ecology. Second Edition. John Wiley and Sons, Inc., New York.

Ricker, W. E. 1973. Linear regressions in fishery research. J. Fish. Res. Bd. Canada 30:409-434.

Saila, S. B., J. B. Flowers, and R. Campbell. 1965. Applications of sequential sampling to marine resource surveys, ocean science and ocean engineering. Transactions of the Joint Conference. Marine Tech. Soc. and Am. Soc. Limnol. Oceanogr. 782-802.

Siegel, S. 1956. Nonparametric statistics for the behavioral sicences. McGraw-Hill Book Co., New York, 312 pp.

Steel, R. G. D. and J. H. Torrie. 1960. Principles and procedures of statistics. McGraw-Hill Book Co., Inc., New York, 481 pp.

Snedecor, G. W. and W. G. Cochran. 1967. Statistical methods. Sixth Edition. Iowa State University Press, 593 pp.

Stofan, P. E. and G. C. Grant. 1978. Phytoplankton sampling in quantitative baseline and monitoring programs. EPA-600/3-78-025.

Swartz, R. 1978. Techniques for sampling and analyzing the marine macrobenthos. EPA-600/3/78-030.

Walker, H. A., S. B. Saila, and E. L. Anderson. 1979. Exploring data structure of New York Bight benthic data using post-collection stratification of samples, and linear discriminant analysis for species composition. Estuarine and Coastal Marine Science (in press).

Wilson, E. B. 1952. An introduction to science research. McGraw-Hill Book Co., Inc., New York.

Yates, F. 1960. Sampling methods for censuses and surveys. Third Edition. Charles Griffin and Co., Ltd., London.

APPENDIX I.

Excerpts from a paper on the accuracy of abundance indices from research vessel surveys by Grosslein (1979) are relevant where analagous sampling conditions in the Ocean Pulse area prevail.

"In order to help evaluate the cost-benefit ratio of surveys it is necessary to have some idea of the magnitude of change in stock size that is considered significant, as well as the magnitude of change we are able to detect and with that probability." Clearly one of the most important questions is whether surveys can measure changes in abundance with sufficient accuracy to permit meaningful assessment of the short-term affects of fishing. However, it is important to remember that we are also concerned with long-term changes involving not just a few priority species but the entire groundfish community. In general, a lower level of accuracy probably would suffice for monitoring long-term changes than in the case of assessment on a year-to-year basis. My principal aim here is to provide some information on what accuracy is possible with catch-per-haul statistics from research vessel surveys.

"When considering accuracy of estimates, we must distinguish between statistical precision or sampling error (variance) and bias. That is, an estimate may be very precise in terms of a small variance but have a large bias, and therefore not be very accurate. In our problem we are concerned not only with precision but also with the possible biases in the survey abundance index (catchability coefficient) between the relative abundance index and the true absolute abundance of the stock. The next step is to estimate this coefficient so that we can estimate actual total numbers in the population."

Statistical characteristics of trawl catch data. "As is well known, trawl catches are highly variable even within relatively restricted areas because fish are not uniformly distributed; and random trawl hauls result in a frequency distribution of catches which is highly skewness is that the variance is generally much larger than the mean resulting in very imprecise (although unbiased in the statistical sense) estimates of the mean, and even less reliable estimates of the variance itself, except with very large sample sizes. That is, the standard error associated with the variance is particularly susceptible to departures from normality, and without a reliable estimate of the variance of course, it is not possible to calculate meaningful confidence limits about the mean...

"A standard approach to this general problem is to stratify the population to be sampled into high and low density units or strata, and then sample randomly within individual strata within each of which skewness is then reduced. Control of variability in this manner is one of the primary advantages to be gained from the technique of stratified random sampling. However, in the case of trawl catcher considerable skewness remains even after stratification...

"Another well known approach is to try to find a transformation which normalizes the frequency distribution of variables. We have found that on the average, stratum variances of trawl catches are approximately proportional to the square of the mean, i.e. the standard deviation is proportional to the mean...

"This relation indicates that a log transformation is appropriate and such a transformation tends to normalize the data and stabilize the variance (i.e. make means and variances independent). Also the log transformation converts multiplicative effects into linear additive effects. In terms of our problem

of estimating proportional changes in abundance, this means that linear changes on a log scale represent estimates of multiple or factor changes on the original scale. That is, the anti-log of the difference between two log means represents the proportionality constant relating means in the linear scale. The estimates unbiased in the statistical sense, but it should be noted that the re-transformed mean is a biased estimate of the true mean on the linear scale (an unbiased estimate is theoretically possible).

Calculation of stratified mean and variance

"The basic index of abundance dealt with here is the stratified mean catch per standard haul, calculated by weighting each stratum mean according to the proportional size (area) of the stratum relative to all strata in the set. The variance of a stratified mean is similarly derived by weighting each stratum variance in proportion to the stratum area and inversely according to the number of hauls in the stratum."

Examples of precision on log scale

On the log scale the variances are yearly stabilized and the CV's of stratified means are on the order of 10-15 percent for the same species and strata. However, note that now we are interested in the absolute rather than relative size of the standard deviation. For haddock ± 2 S.D.'s ($\pm$ .40) corresponds to ± 50 percent of the linear scale. Thus there is no great improvement in the size of difference (proportional change on linear scale) we are able to detect as compared with the non-transformed scale, but we have more efficient estimates of those differences over the range of abundance levels, and the estimated confidence limits more closely approximate true 95 percent confidence intervals.

The most significant feature of these data is that they indicate the present survey cannot detect with high probability, proportional changes in abundance which are less than a factor of about 2. That is, the log difference between the lower and upper limits of the 95 percent C.I. is about 0.7 corresponding to a factor difference of 2 on the linear scale; and to be very sure that two means are significantly different there must be no overlap in the 95 percent confidence intervals.

"The most serious biases in commercial abundance data arise from unknown changes in the effective unit of effort usually related to economic or technological factors. Even with standard gear however, bias can result simply from variations in availability of fish. With proper sample design the research changes in availability. For example the catchability coefficient for a given species and research trawl may change due to a change in vertical distribution of the species, in response to some environmental factor or even as a function type intuitively would seem to be much greater for a species for which the trawl has a very low efficiency."

Appendix II

		MDH x̄	MDH S	LDH	M/L	PK	PRO
78-4	11	1513	369.54	29.3	11.82	61.5	0.36 0.04
	N(5)	(2)77		(2)93	(6)51	(2)40	(4)94
Station	N(10)	5.19		15.73	16.38	3.10	1.23
		7.76	0.26	5.33	0.41	9.[illegible]	0.38	6.56	0.19	-1.03	0.11
		3.37		[illegible]		27.24		2.62		[illegible]	
	12	1715	412.38	935	204.4	70.51		5[illegible].3	100.79	0.37	0.04
	N(5)	(17)05		(8)11		612.05		(11)29		(7)67	
	N(10)	4.27		21.75		153.42		2.82		1.17	
	(13)	1843	2384.37	990	8.73	(6)3.61		[illegible]	144.65	0.32	0.06
		(17)39		(12)49		(11)2		(12)81		(14)06	
	N(10)	4.35		3.12		12.3		3.2		3.52	
	(15a)	1743	290.71	868	31.76	[illegible]	4.32	1003	[illegible]	0.38	0.05
	N(5)	(11)13		(5)85		(16)61		(22)51		(6)93	
	N(10)	2.95		12.9		4.15		5.63		1.73	

		MDH x̄	MDH S	LDH	M/L	PK	PRO
78-12	(3)	1435.2	657.55	734	39.22	(25.4)22.61		79.5	217.55	0.33	0.09
	N(5)	(9)15		(11)42		293.14		(30)2		(46)54	
	N(10)	22.87		28.55		73.35		7.53		16.74	
	(5)	645.4	111.71	(50.2)	18.01	(14.6)	6.5	587.4	155.46	0.32	0.06
	N(5)	(11)98		51.48		79.28		(25)02		(14)06	
	N(10)	3.60		12.89		19.82		7.04		3.52	
	(16)	853.5	205.25	(52.9)	17.65	(17.8)	6.35	863.6	190.2	0.37	0.04
	N(5)	(23)18		46.08		52.07		(19)4		(4)67	
	N(10)	5.79		11.52		15.01		4.85		1.17	
	(22a)	649.8	149.2	(54.3)	31.04	(14.2)	8.47	805.2	195.5	0.42	0.06
	N(5)	(21)03		248.16		142.31		(19)--		(8)16	
	N(10)	5.26		62.04		35.58		4.75		2.04	

APPENDIX III.

An interesting approach to analyses of benthic populations in relation to pollution paid in the New York Bight was presented by Walker, Saila and Anderson (1979). Their approach was to search for patterns among the physical variates and then search for related patterns among biological variates. Some of their rationale is relevant to Ocean Pulse research.

"The correspondence of geographical space and physical space assumed on the classical analysis tempts one to treat the station grid of the New York Bight exactly as if it were a cornfield. However, we are not sampling from a geographical space which is uniform except for externally imposed treatments, as in the cornfield example. Instead, we are searching for the effect of the input of various types of waste being dumped in the New York Bight, where biological variability seen as the result of the dumping is superimposed on substantial microenvironmental variability...

"We are in the position of the agricultural experimenter who is trying to determine the response of corn to multiple treatment inputs, the spatial extent of which is not known at the time of the sampling. It is as if corn were not planted on a field of uniform soil, but within the field there exists and unknown mixture of soil types. In addition, it is not possible to return to the same geographical position during each sampling interval and expect to find the same soil characteristics of fertilizer levels. Thus, all semblance of treatment plots have disappeared because there is no longer any correspondence between geographical position and treatment...

"Within the benthic sample data set, the way we have chosen to face these problems in analysis is to break up the continuous response variables of the sediments into discrete levels: Four for sediment mean grain size and two each for heavy metals and percent organic matter. In this way we can test

for the response of the organisms sampled to 16 combinations of the three variables (16 strata). We have defined a new set of 16 variates, and we examine the response of benthic macroinvertebrates to these variates...

"Because the physical variates are highly correlated, it is clear that we are not able to test for any main effects (i.e., the response of any organism to one of the variables independent of the other variables). Rather, we are limited to testing for the various combinations of the three variates which exist in the data set. Because of the high intercorrelation among the variables, many empty cells are to be expected within the cells of the strata thus defined. The geographical space of the cornfield has been replaced with a variable space. Given the assumption that the level of occurrence of specific benthic organisms is dependent upon these variates, we are in a position to test for the biological response of the system to physical and chemical surroundings...

"The discriminant functions are the best linear combinations for predicting strata from the biological information. For each species, the relative magnitude of the coefficients of the discriminant functions over strata indicates the relative importance of that species in predicting the various strata...

"There are several advantages inherent in this approach. (1) The problem of microhabitat variability is dealt with by stratifying on the basis of physical characteristics of each grab sample. Since the microhabitat variability presumably influences the variability in species abundances, estimates of species abundance which ignore microhabitat effects are apt to be much more variable than estimates which take microhabitat influences into account. As a direct consequence of judicious stratification the estimates of species abundance can be much more precise. However, the degree of information precision should be empirically tested. (2) It is posible to obtain estimates of known precision for strata of particular interest. Since the information in each

grab can be worked up in two steps, it is possible to allocate analytical effort much more efficiently. Species counts may be made for a few of the grabs in some strata, and many more grabs for strata of particular concern. For a particular value of variance for species abundance, increased sample size reduces the spread of confidence limits on a stable mean density estimate. (3) Even if it is realized that the monitoring program must fall short of the desired scope and precision, it can ofcus on a few key questions. Due to limited financial resources, it may not be practical to monitor the abundances of a large number of species. Rare of highly variable species may have to be ignored. Of the remaining list a few key species can be selected in order to monitor the influence of sludge dumping with sufficient precision to say something about shifts in species abundance over time.

In attempting to assess the stability of benthic faunal populations, several populations parameters are important. True insight will be possible when interpretation of density changes can be related to a detailed knowledge of life history, age or stage-specific fecundity and mortality, and survival strategies of species under consideration. For most benthic organisms this type of background information is sorely lacking, and as a result it is difficult to determine if density changes are due to natural variations in the population or the effects of a pollutant.

From data on abundance of a few common conservative species and their within strata variations over time, the analysis could move into a third step; that of size frequency (or age frequency) estimation. It is here that a real jump in information about the population stability of selected species might be expected."

This approach to the problem differs from tradition techniques which either search for patterns in biological variates and attempt to interpret them as responses to physical variables or search for patterns of relationship in two sets of variates simultaneously. This technique could be useful in present Ocean Pulse analyses.

Table 1. Stratum means (catch/haul, pounds) and viariaces for haddock in three sampling strata on Georges Bank. *Albatross IV* surveys.

	STRATUM 16				STRATUM 19				STRATUM 20			
CRUISE	No. hauls	Mean	Variance	Std. deviation	No. hauls	Mean	Variance	Std. deviation	No. hauls	Mean	Variance	Std. deviation
63-05	7	41	2,740	52	4	126	22,442	150	3	7	52	7
63-07	7	101	4,330	66	4	291	66,992	259	4	115	33,379	183
64-01	10	41	857	29	7	147	37,875	194	5	37	1,322	36
64-210	8	300	338,823	582	5	364	209,248	457	5	356	70,072	264
64-13	7	148	31,926	179	6	168	26,652	163	5	335	155,074	394
65-2	6	73	6,309	80	6	392	243,932	494	5	21	338	18
65-510	8	405	682,555	826	6	800	2,019,784	1421	5	618	188,942	435
65-14	7	78	3,266	57	5	171	14,377	120	5	332	160,830	401
66-601	7	73	17,357	132	6	49	6,058	78	5	43	1,243	35
66-614	7	62	1,423	38	6	54	15,495	124	5	126	11,584	108
67-721	8	14	564	24	9	52	4,096	64	6	37	4,140	65
68-803	9	49	5,533	74	8	42	1,189	34	6	13	351	19
68-817	8	19	2,850	53	9	0	-	-	6	25	3,574	60
69-902	14	71	26,570	163	8	45	1,831	43	6	3	41	6
69-908	10	7	185	14	9	6	124	11	6	23	2,610	51
69-911	12	4	117	11	9	7	413	20	6	16	1,137	34
70-703	10	130	120,926	348	8	11	409	20	5	5	76	9

Table 4. Stratified mean catch per haul (lb, $\log_e$ scale) and measures of precision for selected species. *Albatross IV* fall surveys, Strata 13-25.

Year	Mean	Variance	S.D.	S.D./ Mean	2 S.D.	Mean± 2 S.D.	Factor diff.
YELLOWTAIL							
1963	1.97	.026805	.1637	.08	.33	1.64-2.30	1.9
1964	1.41	.037142	.1927	.14	.38	1.03-1.79	2.1
1965	1.32	.029119	.1706	.13	.34	.98-1.66	2.0
1966	0.96	.025860	.1608	.17	.32	.64-1.28	1.9
1967	1.32	.027724	.1665	.13	.33	.99-1.65	1.9
1968	1.40	.038260	.1956	.14	.39	1.01-1.79	2.2
1969	1.35	.025200	.1587	.12	.32	1.03-1.67	1.9
1970	0.96	.0204	.1428	.15	.28	.68-1.24	1.8
HADDOCK							
1963	3.34	.052176	.2284	.07	.46	2.88-3.80	2.5
1964	3.86	.080315	.2834	.07	.57	3.29-4.43	3.1
1965	4.02	.042355	.2058	.05	.41	3.61-4.43	2.3
1966	2.43	.044512	.2110	.09	.42	2.01-2.85	2.3
1967	2.45	.052075	.2282	.09	.46	1.99-2.91	2.5
1968	1.15	.029587	.1720	.15	.34	0.81-1.49	2.0
1969	1.10	.021536	.1467	.13	.29	0.81-1.39	1.8
1970	1.35	.0345	.1857	.14	.37	0.98-1.72	2.1
COD							
1963	1.75	.084829	.2912	.17	.58	1.17-2.33	3.2
1964	1.29	.056270	.2372	.18	.47	0.82-1.76	2.6
1965	1.32	.041737	.2043	.15	.41	0.91-1.73	2.2
1966	1.20	.040673	.2017	.17	.40	0.80-1.60	2.2
1967	1.74	.047301	.2175	.12	.44	1.30-2.18	2.4
1968	1.04	.031888	.1786	.17	.36	0.68-1.40	2.1
1969	1.32	.025381	.1593	.12	.32	1.00-1.64	1.9
1970	1.35	.0332	.1822	.13	.36	0.99-1.71	2.1

APPLICATIONS

None of the quantites involved in Ocean Pulse research can be observed or measured throughout the whole population. Conclusions will be based on the attributes of samples considered representative. If the sampling is good the conclusions derived will differ little from reality. One of the tasks expected will be that of forecasting. In order to achieve this goal will require a thorough grasp of individual subjects and indices developed in allied fields. Some recognition of limitations is necessary for deriving projections of events. The correlation of time series will very likely be employed and the topic will be a part of the synthesis of research finding. At the present time the array of test species and disciplines is noted in Table I. Each of the 15 activities considered a promising arbiter of environmental quality. However, the efficiency, reproductibility and other attributes of the studies will remain to be evaluated in many cases. The selection of test species has been derived from the availability encountered in sampling gear during early cruises. The subjects range from phytoplankton, constituent chemicals and chlorophyll through particulate and filter feeding invertebrates to species used for harvest. The suitability of various statistical tests are arrayed below for each of the study disciplines. However, it must be remembered that after a basic series of results are abailable there will be material for determining time series variation and analysis of covariance between disciplines. The most powerful test of effects will prevail when reinforcements are found to occur between several studies.

SPECIES	Community Diversity Equitability	Seasonal Abundance	Succession	Anaerobes	Calorimetry	Physiological Activity	Parasites	Virology	Anomalies (gross-Histo)	Nutrient Bioassay	Uptake	Genetic	Petroleum Bioassay	Limiting Factors	Hydrocarbon Exposure
Water															
Temperature	X														
Constituents	X			X											
Chlorophyll		X													
Phytoplankton	X	X	X							X					
Sediments	X	X		X											
Benthos	X				X		X°		X°		X				
Nekton					X		X°		X°						
Mysids						X			X						
Isopods						X									
Euphaunds						X									
Crangon						X									0
Rock Crab															0
Lobster						X									0
Tellinoid Mussel						X	X		X						
Scallop						X			X						0
Spisula						X								X	0
Arctica						X									
Herring							X	X							
Flounders															
Winter												X**	**		
Yellowtail							X	X				X**	**		
Windowpane												X**	**		
Hake															
Red													**		
Silver							X	X							
Ammodytes									X						
Cod							X	X				X*	**		
Haddock							X	X				X*	**		

°- Crustacean
*- Fisfish eggs available
**- Fish larvae and gonad developement
O- Eggs and larval invertebrate

Table I

GLOSSARY OF SELECTED STATISTICAL TERMS*

Confidence Interval

If it is possible to define two statistics t_1 and t_2 (functions of sample values only) such that, Θ being a parameter under estimate,

$$P(t_1<\Theta_2) = \alpha$$

where α is some fixed probability, the interval between t_1 and t_2 is called a confidence interval. The assertion that Θ lies in this interval will be true, on the average, in a proportion α of the cases when the assertion is made.

Correlation

In its most general sense corelation denoted the interdependence between quantitative or qualitative data. The concept is quite general and may be extended to more than two variates.

The word is most frequently used in a somewhat narrower sense to denote the relationship between measurable variates or ranks.

Correlation, Coefficient of

A correlation coefficient is a measure of the interdependence between two variates. It is usually a pure number which varies between -1 and 1 with the intermediate value of zero indicating the absence of correlation, but not necessarily the independence of the variates. The limiting values indicate perfect negative or positive correlation.

If there are two sets of observations $x_1 \ldots x_n$ and $y_1 \ldots y_n$, and a score is alloted to each pair of individuals, say a_{jj} (for the x-group) and b_{jj} (for the y-group), a generalized coefficient of correlation may be defined as

$$r = \frac{\Sigma a_{ij}\, b_{ij}}{\sqrt{(\Sigma a_{ij}^2\ \Sigma b_{ij}^2)}}$$

where Σ is a summation over all values of i and j $(i \neq j)$ from 1 to n.

---

* Adopted from Kendall, M. B. and W. R. Buckland. A Dictionary of Statistical Terms. Hafner Publ. Co., 575 p. 2nd ed.

If positive values of one variate are associated with positive values of the other (measured from their means) the correlation is sometimes said to be direct or positive; as contrasted with the contrary case, when it is said to be inverse or negative.

There are numerous other correlation coefficients of a different character.

Degrees of Freedom

This term is used in statistics in slightly different senses. It was introduced by Fisher on the analogy of the idea of degrees of freedom of a dynamical system, that is to say the number of independent coordinate values which are necessary to determine it. In this sense the degrees of freedom of a set of observations is the number of values which could be assigned arbitrarily within the specification of the system; for example, in a sample of constant size n grouped into k intervals there are k-1 degrees of freedom because, if k-1 frequencies are specified, the other is determined by the total size n; and in a contingency table of p rows and q columns with fixed marginal totals there are $(p-1)$, $(q-1)$ degrees of freedom.

From a different viewpoint the expression "degrees of freedom" is also used to denote the number of independent comparisons which can be made between the members of a sample.

Eigenvalue

The characteristic root of a square matrix A is a value λ such that $[A-\lambda I] =)$, where I is the identity matrix. For a $p \times p$ matrix there are, ingeneral, p such roots. They are also known as Latent Roots and Characteristic Roots.

The corresponding row-vestors u or column-vectors v for which

$$uA = \lambda u \text{ or } Av = \lambda v$$

are called characteristic vectors.

Exponential Curve

A series of observations ordered in time which has a constant, or approximately constant, rate of increase can be represented over a long period by the curve:

$$y = ae^{bt}$$

where a and b are constants and t is time. This, or some simple transformation, is called the exponential curve. The fitting of an exponential trend of this form by the method of least squares is facilitated by transforming into the logarithmic form:

$$\log_e y = \log_e a + bt.$$

Goodness Fit

In general, the goodness of agreement between an observed set of values and a second set which are derived wholly or partly on a hypothetical basis, that is to say, derive from the "fitting" of a model to the data. The term is used especially in relation to the fitting of theoretical distributions to observation and the fitting of regression lines. The excellence of the fit is often measured by some criteria depending on the squares of differences between observed and theoretical value, and if the criterion has a minimum value the corresponding fit is said to be "best".

Graeco-Latin Square

An extension of the Latin-square. Formally, it is an arrangement in a square of two sets of letters (say A, B,... etc. and α, β,... etc.), one of each in each cell of the square, such that no Roman letter occurs more than once in the same row or column, no Greek letter occurs more than once in the same row or column, and no combination of the two occurs more than once anywhere. For example, a 4 x 4 square of this kind is

$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
$B\gamma$	$A\delta$	$D\alpha$	$C\beta$
$C\delta$	$D\gamma$	$A\beta$	$B\alpha$
$D\beta$	$C\alpha$	$B\delta$	$A\gamma$

The arrangement is used in experimental designs to allocate treatment of three factors so that all comparisons are orthogonal.

Latin Square

One of the basic statistical designs for experiments which aim at removing from the experimental error the variation from two sources, whicy may be identified with the rows and columns of the square. In such a design the allocation of k experimental treatments in the cells of a k by k (Latin) square is such that each treatment occurs exactly once in each row or column. A specimen design for a 5 x 5 square with five treatments, A, B, C, D, and E is as follows:

A	B	C	D	E
B	A	E	C	D
C	D	A	E	B
D	E	B	A	C
E	C	D	B	A

The earliest recorded discussion of the Latin square was given by Euler (1782) but it occurs in puzzles at a much earlier date. Its introduction into experimental design is due to R. A. Fisher.

Level of Significance

Many statistical tests of hypotheses depend on the use of the proability distributions of a statistic t chosen for the purpose of the particular test. When the hypothesis is true this distribution has aknown form (at least approximately) and the probability $P(t > t_1)$ or $P(t < t_0)$ can be determined for assigned t_0 to t_1. The acceptability of the hypothesis is usually discussed, in terms of the values of t observed; if they have a small probability, in the sense of falling outside the range t_0 to t_1 ($P(t \geq t_1)$ and $P(t \leq t_0)$ small) the hypothesis is rejected. The probabilities $P(t > t_1)$ and $P(t < t_0)$ are called levels of significance and are usually expressed as percentages, e.g. 5 per cent. The actual values are, of course, arbitrary, but popular values are 5, 1 and 0.1 per cent. Thus, for example, the expression "t falls above the 5 per cent level of significance" means that the observed value of t is greater than t_1 where the probability of all values greater than t_1 is 0.05; t_1 is called the upper 5 per cent significance point, and similarly for the lower significance point t_0.

Model

A model is a formalized expression of a theory or the causal situation which is regarded as having generated observed data. In statistical analysis the model is generally expressed in symbols, that is to say in a mathematical form, but diagrammatic models are also found. The word has recently become very popular and possibly somewhat overworked.

Nested Sampling

A term used in two somewhat different senses: (1) as equivalent to multi-stage sampling because the higher-stage units are "nested" in the lower-stage units; (2) where the sampling is such that certain units are imbedded in larger units which form part of the whole sample, e.g. the entry-plots of clusters are "nested" in this sense.

Precision

In exact usage precision is distinguished from accuracy. The latter refers to closeness of an observation to the quantity intended to be observed. Precision is a quality associated with a class of measurements and refers to the way in which repeated observations conform to themselves; and in a somewhat narrower sense refers to the dispersion of the observations, or some measure of it, whether or not the mean value around which the dispersion is measured approximates to the "true" value. In general the precision of an estimator varies with the square root of the number of observations upon which it is based.

Probit

The normal equivalent deviate increased by 5 in order to make negative values very rare. The word was suggested by Bliss (1934) as a contraction of "probability unit".

Random

This work may be taken as representing an undefined idea, or, if defined, must be expressed in terms of the concept of probability. A process of selection applied to a set of objects is said to be random if it gives to each one an equal chance of being chosen. Generally, the use of the work "random" implies that the process under consideration is in some sense probabilistic.

Regression

This term was originally used by Galton to indicate certain relationships in the theory of heredity but it has come to mean the statistical method developed to investigate those relationships.

If a variate y consists of two components, a variate and a systematic element $f(X)$ depending on a variable X,

$$y = f(X)+\varepsilon$$

then the regression of y on X is the equation

$$Y = f(X)$$

where it is supposed that ε has zero expectation. The definition remains valid, if X, instead of being a single variable, refers to a set of variables X_1, X_2, etc.

In particular, X itself may be given as the values of a variate, in which case the regression of y on x may be regarded as expressing the dependence of the mean of y (for given x) on the corresponding x:

$$E(y|x) = f(x).$$

The most frequently considered form of $f(x)$ is a polynomial, particularly a linear function, giving the regression of y on X

$$Y = \beta_0+\beta_1 X$$

or, for p variables

$$Y = \beta_0+\beta_1 X_1+\ldots+\beta_p X_p$$

Such regressions are called regression equations. The X's are called "independent", "predicated" variables, "predictors" or "regressions". y is called the "depenedent variate", "predictand" or "regressand".

Significance

An effect is said to be significant if the value of the statistic used to test it lies outside acceptable limits, that is to say, if the hypothesis that the effect is not present is rejected. A test of significance is one which, by use of a test-statistic, purports to provide a test of the hypothesis that the effect is absent. By extension the critical values of the statistics are themselves called significant.

Standard Deviation

The most widely used measure of dispersion of a frequency distribution. It is equal to the positive square root of the variance.

Variance

The variance is the second moment of a frequency distribution taken about the arithmetic mean as the origin namely

$$\int_{-\infty}^{\infty} (x-\mu_1')^2 dF$$

where μ_1' is the mean and F the distribution function. It is a quadratic mean in hte sense that it is the mean of the squares of variations from the arithmetic mean. It may also be regarded as one-half of the mean-square of differences of all possible pairs of variate-values.

Variance-Analysis

The total variation displayed by a set of observations, as measured by the sums of squares of devisions from the mean, may in certain circumstances be separated into components associated with defined sources of variation used as criteria of classification for the observations. Such an analysis is called an analysis of variance, although in the strict sense it is an analysis of sums of squares. Many standard situations can be reduced to the variance-analysis form.