United States
Department of
Agriculture

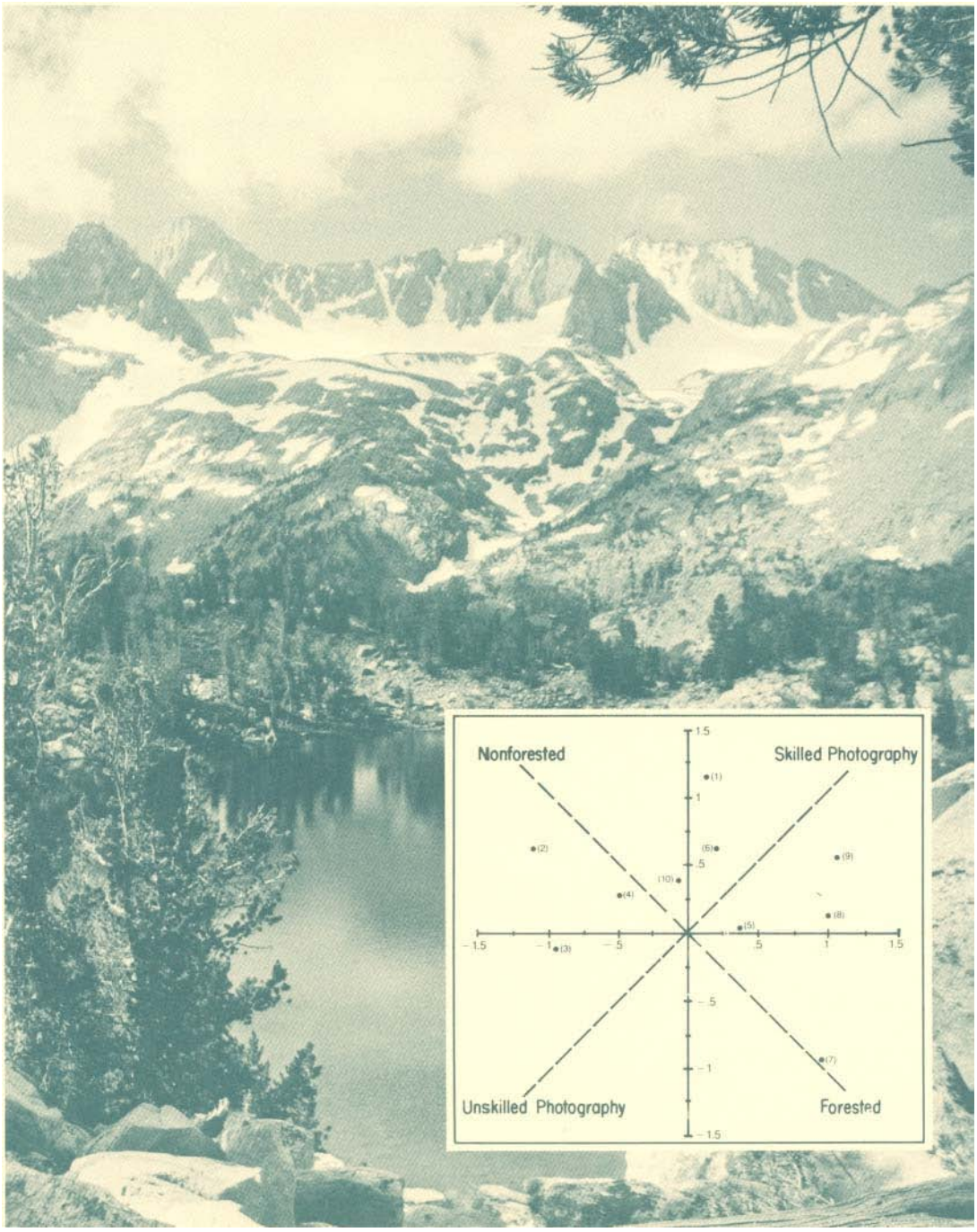
Forest Service

Pacific Southwest
Forest and Range
Experiment Station

General Technical
Report PSW-87

Evaluating Statistical Validity of Research Reports: a guide for managers, planners, and researchers

Amanda L. Golbeck



The Author:

AMANDA L. GOLBECK is Assistant Professor of Mathematics at San Diego State University. She earned degrees in anthropology: bachelor's (1974) at Grinnell College, and master's (1977) at the University of California, Berkeley. She also holds a master's degree (1979) in statistics and a doctorate (1983) in biostatistics from Berkeley. This report was prepared under a research contract while the author was a doctoral student.

Preface

Just as some managers are more professional, effective, and skilled than others, so are some researchers more competent than others in their use of statistics. And, just as managerial decisions have many dimensions, so do statistical decisions. However, decisions of a midlevel manager are subject to scrutiny of superiors only, but statistical decisions in a research report are subject to the scrutiny of all who read it. Consumers of research reports must not just accept conclusions, but must investigate the methods used to obtain them. I urge resource managers and planners to read research reports critically and to judge the choice of statistical method. In this report I explain the need to be critical and describe some ways in which conclusions can be evaluated. In addition, I recommend discussing the statistical validity of research reports with professional statisticians before applying the results to managerial or planning decisions.

Publisher:

Pacific Southwest Forest and Range Experiment Station
P.O. Box 245, Berkeley, California 94701

May 1986

Evaluating Statistical Validity of Research Reports: a guide for managers, planners, and researchers

Amanda L. Golbeck

CONTENTS

IN BRIEF ...

Golbeck, Amanda L. **Evaluating statistical validity of research reports: a guide for managers, planners, and researchers.** Gen. Tech. Rep. PS W-87. Berkeley, CA: Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture; 1986. 22 p.

Retrieval Terms: scaling of attitudes, statistical assumptions, ordinal data analysis, sampling biases

Inappropriate statistical methods, as well as appropriate methods inappropriately used, can lead to incorrect conclusions in any research report. Incorrect conclusions may also be due to the fact that the research problem is just hard to quantify in a satisfactory way. Publication of a research report does not guarantee that appropriate statistical methods have been used, or that appropriate methods have been used correctly. Publication also does not guarantee that actual measurements are reasonably close to the underlying concept.

You may not be able to tell if an appropriate statistical method was used correctly. You can, however, (with the help of a professional statistician) judge whether the choice of method was appropriate. You can also judge how close the actual measurements seem to the underlying concept that the researcher is studying.

Two steps are preliminary to judgments about whether the choice of method was appropriate. The first involves categorizing the study according to its primary statistical purpose in terms of examining variables. This purpose may be describing variables, testing hypotheses about variables, exploring relationships among variables, or building prediction models using variables. The statistical purpose of most studies of visual quality has been exploring relationships or building models.

The second preliminary step to evaluating the validity of a research report involves categorizing variables according to their level of measurement. This level of measurement may be (1) nominal, (2) ordinal, (3) interval, or (4) ratio. Most studies of visual quality involve several variables with different levels of measurement. One of these variables is usually an attitude or preference variable having an ordinal level of measurement.

Any given statistical technique presumes a specific level of measurement for the variable(s). Use of a statistical technique upon data that are at a lower level of measurement than what the technique presumes leads to results that are neither empirically nor semantically meaningful. Many visual quality researchers controversially have used statistical techniques that presume interval level measurements upon attitude or preference variables that have ordinal (lower) levels of measurement.

A variety of methods can be used to scale an attitude variable (such as Likert scaling and paired comparisons), all of which result in an ordinal level of measurement. Some researchers make psychological assumptions (e.g., invoke Thurstone's Law of Comparative Judgment) to claim that they have achieved an interval level of measurement for their attitude variables. If you are not willing to accept psychological theory as an ingredient to the determination of level of measurement, then you take the same position that professional statisticians do: no feasible method is available for deriving interval data from ordinal rating scale data.

When the statistical purpose of the study is that of exploring relationships among variables, several methods are available,-including simple correlation, factor analysis, and multidimensional scaling. Several measures of simple correlation are available, each presupposing a certain level of measurement for the two variables to be correlated. For hypothesis testing purposes, factor analysis presupposes an interval level of measurement and some types of multidimensional scaling presuppose an ordinal level of measurement.

When the statistical purpose of the study is to build prediction models using variables, the most commonly used methods are regression methods. These typically use a straight line to approximate the relationship between one dependent variable and one or more independent variables. For hypothesis testing, linear regression requires that variables meet five assumptions: fixed, independent, normal, equal variance, and linear. Standard linear regression methods presuppose an interval level of measurement. When the variables are nominally or ordinally scaled, log-linear models should be used in the place of standard linear regression methods.

Inappropriate sampling methods can also lead to incorrect conclusions in any research report. The planned introduction of chance or probability into a sampling method can minimize or eliminate the possibility of bias. Small convenience samples—consisting only of students, for example—cannot yield valid measures of general public attitudes and preferences in the area of visual quality or in any other research area.

INTRODUCTION

The dependency of conclusions upon the choice of statistical methods can be illustrated by an example. The Ecological Society of America surveyed members, applied certain statistical methods to the data, and concluded that "applied ecologists and other ecologists were in remarkable agreement" in most of their views on publishing (Ecological Society of America 1982, p. 27). But, by applying different statistical methods to the same survey data, another researcher arrived at the opposite conclusion: applied ecologists and other ecologists differed significantly in their views (Saunders 1982, p. 336).

The fact that publication of research results does not assure their correctness can be illustrated by another example. Schor and Karten (1966) studied the statistical methods used in a large series of medical studies reported in several journals, and found that only 28 percent of them were statistically acceptable. This finding led the American Statistical Association to raise the question of whether a code of principles can be maintained to assure basic levels of statistical competence, or whether formal certification is necessary to assure credibility of an author. At this writing, the question has not been resolved, and levels of statistical training and competence vary among the users of statistics.

In fields such as visual quality where the concepts in question are sometimes harder to quantify than those in medical research, and where less professional statistical input is employed, the problem is likely to be worse than that reported by Schor and Karten (1966). The fact is, despite its numerical base, statistics is art as well as science. Often the user of statistics must choose among methods, a somewhat subjective process, and may use the method in a subjective fashion.

This subjectiveness invalidates neither statistics as a science nor statistical methods. But, nonstatisticians should be aware of this "artistic side" of the discipline. Do not unthinkingly trust figures that are published, posted on a bulletin board, or used for political purposes, the way that—for example—statistics for cost-of-living and unemployment are sometimes used. Carefully examine how figures were derived before believing them.

Sections such as the abstract and management implications in reports are convenient. But beware! Don't accept the conclusions until you have investigated the appropriateness of the analytic methods. You can examine the methods section of a research report and judge the choice of statistical method; however, often it is difficult to tell if an appropriate method was used correctly. If you don't feel competent to judge the appropriateness of the analytic methods, the best thing to do is have some doubts, and ask a professional statistician for help.

This report shows the need to judge the statistical validity of research results, especially those involving many variables or theoretical concepts. It explains—at a level of complexity compatible with the statistics involved—how methods can be evaluated. This report is a statistical guide for resource managers and planners, as well as for physical and social scientists, to use while reading research reports. It should also prove useful to researchers in planning, conducting, and reporting their studies.

1. IDENTIFY VARIABLES

The fundamental element of statistical thinking is a *variable*. Defined in simplest terms, a variable is the object of interest that is measured or counted. It can be age of mother at birth of a child, decay time of an isotope, frequency of lung cancer, angular width of a panoramic scene, density of trees in foreground, or practically any numeric quantity.

Frequently the object of interest is a theoretical concept that is not directly measurable, scenic beauty, for example. In such a case, much of the basic research is devising methods of measurement. Breakthroughs in knowledge often occur when a measurement procedure is developed or discovered that allows previously immeasurable theoretical concepts to be quantified.

Two kinds of definitions are used in research: theoretical and operational. Researchers think with theoretical concepts. They conduct empirical research using operationalized concepts.

A *theoretical definition*, like most ordinary definitions, defines a concept in terms of other concepts which supposedly are already understood. In this type of deductive system, certain concepts are undefined or primitive, and all other concepts are defined in terms of these. An example may be

taken from Euclidean geometry, where the concepts of point and line are undefined. Other geometric concepts such as triangle and rectangle can be theoretically defined in terms of these fundamental concepts. An example relevant to visual quality is the concept of scenic beauty. A theoretical definition might be a natural scene that is pleasing.

Operational definitions of concepts include procedures for classifying and measuring them. The concept of scenic beauty has various operational definitions. A common one is "the quantification of public preferences for certain formal aspects of a landscape (e.g., color, line, form)."

This definition may be an imperfect indicator of the underlying concept. Operational definitions are, for this reason, often considered as indices. Three assumptions are implicit in the typical operational definition of scenic beauty: (1) the esthetic quality of a landscape is meaningfully correlated with certain preferences for that landscape, (2) those preferences are those of the general public, and (3) the esthetic quality of a landscape can be described in terms of formal aspects only (e.g., forms, lines, textures, colors). These assumptions clarify that the operationalized concept in this research is public preferences for certain formal aspects of a landscape (or, more directly, of a photograph), which may not be equivalent to the theoretical concept of scenic beauty (Carlson 1977).

Conclusions arising from quantitative research apply strictly to concepts as operationally defined. Propositions involving theoretically defined concepts cannot be empirically tested. The question arises whether, in practice, a particular operational definition is reasonable. Generally, is there any logical way to determine if an operational definition adequately measures the theoretically defined concept? Most researchers do not believe so. Instead, they rely on simple convention or general agreement that a given operational definition should be used as a measure of a certain concept. Such convention or agreement is based on the argument that the operations "seem reasonable" on the basis of the theoretical definition. That is, the operational definition seems reasonably close to the underlying concept.

The problem can and does arise of then having several different operational definitions or indices associated with each theoretical concept, each of which may produce significantly different results. For example, if there are two distinct operational definitions of scenic beauty or landscape preferences, two distinct hypotheses are being tested. Researchers may have to revise or clarify the theoretical definition when several scientifically acceptable operational procedures carried out under similar circumstances yield different results.

Recognizing the difference between a theoretical concept and its operational definition is largely a matter of common sense. When you read reports of studies on a theoretical concept like scenic beauty, ask these questions:

- What was studied?
- Can the concept(s) be measured directly? If not, what was actually measured?
- What assumptions had to be made to get back to the underlying concept?
- Are the assumptions acceptable?

2. DETERMINE STATISTICAL PURPOSE

Studies in any field have both theory and methods components. Both can be manipulated to support a desired conclusion. In terms of theory, a researcher may tend to be selective and may report only references supporting a favored position. Some researchers may also manipulate the conclusion by publishing only the supporting results. In terms of methods, the match between the research problem and the statistical method used may not be good, or appropriate statistical methods may have been used incorrectly.

Many statistical methods are available for describing and analyzing variables. Depending on the type and number of variables and the problem at hand, some statistical methods are much more satisfactory than others in producing a reliable conclusion. Before examining the conclusions of a study and the theory they support, try to categorize the statistical procedures, to get a sense of the plausibility of conclusions. The primary purpose of the statistical methods can be one of these:

- Describing variables
- Testing hypotheses about variables
- Exploring relationships among variables
- Building prediction models using variables.

The majority of studies of visual quality have done more than merely describe variables. In particular, a few studies have tested hypotheses about variables, and even more studies have either explored relationships among variables or built prediction models using variables.

2.1 Describing Variables

Descriptive statistics can be used to organize and summarize data. Such techniques help both researchers and readers of research reports to understand more readily the importance of the data.

The researcher begins with *raw data.* These are the values that are collected for each variable, unaltered by statistical or other manipulation. They are obtained by counting or measuring with a scale. For example, a sample of 100 people is taken to yield 100 values on the variable, scenic quality. These 100 observations or measurements are raw data.

The purpose of descriptive statistics is to examine the distribution of values for single variables in order to gain understanding of the research problem. The researcher attempts to condense the values (data) for a variable to a few representative numbers. For example, it is difficult for the researcher to comprehend the relation of 100 individual values for scenic beauty to the research problem. Therefore, some descriptive statistics can be computed to reduce the 100 values to one or two convenient summary measures.

Descriptive statistics can answer four crucial questions about a data set:

1. Where do the bulk of values fall? For example, how do most of a sample of observers rate the scenic beauty of a particular landscape?

2. What proportion of the values fall in a range of particular interest? For example, what proportion of the observers gave the landscape a positive scenic beauty evaluation?

3. What are the upper and lower extreme values? For example, what is the highest and lowest scenic beauty rating given to a particular landscape within a sample of observers?

4. What is the relationship of a particular value to the group of values? For example, do the sample of observers have widely differing evaluations of the beauty of a landscape?

One common type of representative number is a measure of location or central tendency. This number indexes the center of a distribution of a set of observed values for a variable. The most common measures of location are the *mean* (arithmetic average), the *median* (the value that divides the ordered observed values into two groups of equal size) and the *mode* (the value that occurs most often).

Measures of dispersion are also common. These measure the variability among values for a variable. Three common ways of measuring dispersion are the *range* (largest value in a set of observations minus the smallest value), the *sample variance* (sum of squared deviations of each observation from the mean, divided by the number of observations minus 1), and the *standard deviation* (square root of the sample variance). The latter two measures show how tightly packed the observations are about the mean.

Beware of the seductive ease of summary statistics. Many individuals, especially social scientists, are lured to "invent" new types of summary statistics for their research problems. Summary statistics can mask important differences within and between groups of subjects. For example, these two sets of measurements have the same mean (=20)

21, 22, 19, 18

1, 2, 3, 74.

These two sets of measurements are very different, but the reported summary statistic does not reflect this.

If a study was mainly descriptive, one or more of these measures will have been the statistical focus: mean, median, mode, range, sample variance, and standard deviation.

2.2 Testing Hypotheses

Whenever numerical results are subject to chance, the researcher can go beyond descriptive statistics in analyzing data. *Statistical inference* uses statistical methods designed specifically to assess the likelihood that research results are explainable by chance.

One type of statistical inference involves testing a *statistical hypothesis.* A statistical hypothesis is a statement concerning the distribution of probabilities for different values of a random variable. *A random variable* is a variable that has probabilities attached to specific numeric values. That is, there is a certain probability that a specific value will occur if only one observation is taken. For example, suppose the variables is preference for a particular type of landscape, and it can take on the value 1 (indicating low preference), 2, 3, 4, or 5 (indicating high preference). Preference would be a random variable because probabilities are associated with each possible value. Thus, if only one observation is obtained for the variable, the probability that the value equals 1(is low) exists.

The researcher may hypothesize that the preference variable has a uniform distribution. That is, the probability of taking on a specific value is the same for all five values of the variable, and the probability is equal to 0.2. The latter statement is a statistical hypothesis. The researcher then asks each person in a sample to rate their preference for the landscape. These (raw data) are then summarized into a relative frequency distribution, which indicates the proportion of people in the sample that gave a preference rating of 1, the proportion that gave a rating of 2. etc.

Testing compares the observed relative frequency distribution with the hypothesized probability distribution and answers the question: Do the relative frequencies differ significantly from 0.2? Two hypotheses are tested at a time: the *null hypothesis* and the *alternative hypothesis.* In the example, a null hypothesis would be that preferences are uniformly distributed, and the alternative hypothesis would be that preferences are *not* uniformly distributed. A *hypothesis test* is used either to reject or not reject the null hypothesis while knowing the probability that the decision is wrong (probability of error).

As can be seen from the example, the null hypothesis is an educated guess that the distribution of the observed values shows no basic difference from the assumed probability distribution. Not rejecting the null hypothesis then means that the data show no systematic difference from the assumed distribution, that is, no difference beyond that attributable to random variation. Rejecting the null hypothesis lends support to the alternative hypothesis. In such a case, indications are that a systematic difference exists between the observation and a theoretically derived standard. That is, a difference exists beyond that attributable to random variation.

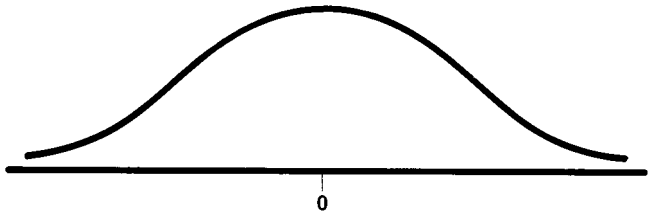
Statistical tests have four possible outcomes. Two are correct actions and two are possible errors. Consider an example concerned with landscape management, where the null hypothesis is that a road is visually subordinate to the characteristic landscape. The four outcomes are illustrated below:

Reality / Action	Null hypothesis is true (i.e., alternative hypothesis is false): The road is in fact visually subordinate to the characteristic landscape.	Null hypothesis is false (i.e., alternative hypothesis is true): The road does in fact visually dominate the characteristic landscape.
Reject the null hypothesis (i.e., accept the alternative hypothesis): the sample indicates that the road visually dominates the characteristic landscape.	Error	Correct action
Accept the null hypothesis (i.e., reject the alternative hypothesis): the sample indicates that the road is visually subordinate to the characteristic landscape.	Correct action	Error

Certain assumptions must be true for statistical tests to be appropriate. The nature of these assumptions depends on the particular test. If appropriate assumptions are not met, the results of statistical tests are invalid.

Statistical hypotheses can be tested in either a *univariate* or *multivariate* situation. Univariate situations involve only one variable; multivariate situations involve a number of variables operating simultaneously. Visual quality research usually involves multivariate situations. The assumptions necessary for multivariate tests are usually more difficult to satisfy than are those for univariate tests.

The F-test is an example of one of the statistical tests that has appeared in the literature on visual quality. To use the F-test, three assumptions must be met. When you see an F-test reported for research on visual quality, look for some evidence within the article that the assumptions described below are true of the data used (often you will be given no evidence—in that case, you are simply unable to judge whether the assumptions are met). (1) Each variable used in the test has its own underlying normal (i.e., bell-shaped) probability distribution.



The mathematical form for the *normal probability distribution* is given in the *glossary.* (2) The population variance of all variables used in the test is the same. (3) The values for a random variable are *statistically independent.* In a loose sense, the last assumption means that the value obtained for one observation does not affect the values that are likely for other observations. For example, if a random sample of people were asked to rate the scenic beauty of a landscape, the ratings between people are likely to be independent. But if a random sample of people were asked to give such a rating before and after a landscape intervention, then for each person the two ratings are not independent. Analyses using statistical inference are occasionally reported for visual quality research. The difficulties of conducting visual quality research, together with its multivariate character, seldom permit valid use of inferential statistics.

Knowing the names of three common tests—t-test, Z-test, F-test—will help you recognize when the statistical purpose of a study is testing statistical hypotheses about variables.

When reading reports of studies that statistically tested hypotheses, ask these key questions with the help of a professional statistician:

- What statistical assumptions are necessary for using that test statistic?
- Are these assumptions either true or closely approximated by the data?
- At worst, do the assumptions seem reasonable for the given set of data and for the study approach to this particular problem?

2.3 Exploring Relationships

Most data sets involve observations associated with more than one aspect of a particular background, environment, or experiment. Because of this, data are usually multivariate, as in visual quality research. The basic question in the multivariate situation is the following: If a large number of variables are characterized by complex relationships, what will make the problem easier to understand?

Several statistical methods are available to simplify a multivariate situation. One of the simplest methods begins with the concept of *association.* Two variables are highly associated if the value of one variable can be used to reliably predict the value of the other. For example, the researcher might find that distance to the back ridge was highly associated with scenic beauty rating. This could mean that either the closer an observer is to the back ridge, the higher the scenic beauty rating (positive association); or the closer an observer is to the back ridge, the lower the scenic beauty rating (negative association).

The different statistical measures of association—such as Pearson's product-moment correlation, Spearman's correlation coefficient, joint biserial correlation—are all summary measures and must be interpreted cautiously like any single number that summarizes an entire set of observed values. Also, certain assumptions must be met for a particular measure of association to be appropriate.

In the multivariate research problem, the researcher usually has a large number of variables. Describing the interrelations among them could go beyond the concept of simple association between two variables. In particular, the researcher could (1) study the associations of a large number of variables by clustering them into groups within which variables are highly associated; (2) interpret each group by studying the variables in it; and (3) summarize many variables by a few *post hoc* variables constructed to represent each group.

Several statistical techniques can be used to accomplish the above three goals. In general, these techniques reduce a complex data set into something that is easy to understand, visualize, and interpret, while retaining sufficient detail for adequate representation. They focus attentions on meaningful relationships between variables and uncover the hidden structure of a data base. Two techniques for exploring relationships among variables are factor analysis and multidimensional scaling. For more details, see *chapter 5.*

2.4 Building Prediction Models

No empirical problem is directly concerned with mathematical (in this case, formal probabilistic) concepts. The researcher needs to translate an empirical problem into formal probabilistic terms before using probability theory to analyze the problem. This translation amounts to building a *probability model* of the problem.

There are many different ways to build a probability model. Choosing the model that best fits the data is straightforward when the adequacy of a particular model can be tested empirically. But just as often it cannot. When a model cannot be tested empirically the researcher is forced to rely on intuitive judgment about the adequacy of how probability model components correspond to the phenomena being studied. Once again, the situation is one of "art plus science." An analysis of a problem based on a given model applies to the model, not necessarily to the phenomena. More precisely, the amount of correspondence between a mathematical description and the phenomena depends on the adequacy of the model. The model may be accurate but be impossible to use because of the difficulty of mathematical analysis. Alternatively, the model may work but be too simple to adequately represent the problem.

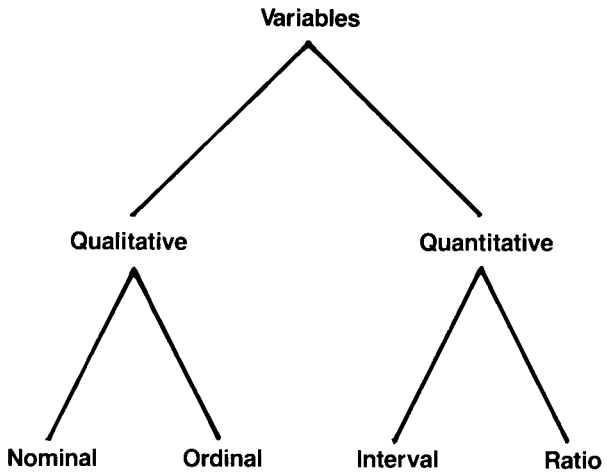
The researcher often can use regression techniques to build a linear mathematical model. Sometimes the model will be unrealistic, but acceptably so. In other words, it will give predictions that are not entirely accurate, but yet accurate enough to be useful. Such models aid in choosing the most salient group of variables and in understanding the interactive effects among them. Regression techniques for prediction of scenic beauty are discussed in detail in *chapter 6.*

3. COMPARE LEVELS OF MEASUREMENT AND ANALYSIS

To critically evaluate a research document in the field of visual quality, you need to appreciate the concept of level of measurement and the associated notion of types of variables. This is necessary because any given statistical technique presumes a specific level of measurement.

3.1 Precision of Measurements

In general, variables are either *qualitative* or *quantitative*. In turn, qualitative variables are either *nominal* or *ordinal*, and quantitative variables are either *interval* or *ratio*.



Each type of variable is a distinct level of measurement with statistical procedures that are also distinctly appropriate.

A quantitative variable measures things that are expressed as real numbers and have real-world (physical) counterparts. Examples would be temperature measured in degrees Celsius, the number of trees in a particular section of forest land, or material wealth in dollars. Qualitative variables are extensions of the concept of measurement that include certain categorization procedures ordinarily used in the social sciences. Examples would be species of trees, types of roads, or amount of vegetation cover measured in the three simple categories of low, medium, or high.

Nominal variables result from sorting things into homogeneous categories. An example would be 20 photographs of 20 individual trees sorted by species. The result might be one category of oaks and one category of pines, which could be labeled by an arbitrary number instead of by name (e.g., 1 = pines, 2= oaks). This is the simplest level of measurement. No assumptions are made about relationships between categories. As long as the categories are exhaustive (include all the photographs) and do not overlap (no photographs in more than one category), the minimal conditions are met for the application of certain statistical procedures.

Ordinal variables also result from sorting things into homogeneous categories, but the categories are also ordered or ranked with respect to the degree, intensity, or amount of something. An example is 20 photographs of 20 individual trees categorized by estimated age of trees. With five age categories, some of the photographs would fall in category 1 (youngest), some in 2, some in 3, some in 4, and some in 5 (oldest). One point needs to be understood about the ordinal level of measurement: it supplies no information about the magnitude of the differences between the categories. Ordinal measurements do not tell if the trees in category 5 were five times older than those in category 1, or two times older, or any other information about how many years of age were represented by each category of trees. The implication of this point will be discussed in the next section.

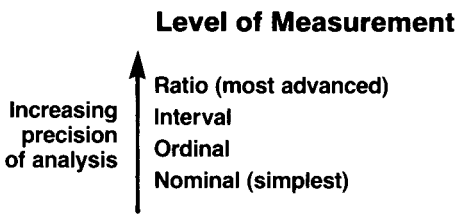
Interval and ratio scales differ from the other levels of measurement in that they both rank observations and indicate the exact distance between them. This is the true interval

level of measurement, which requires establishing some physical unit of measurement as a common standard. Examples are length measured in meters, time measured in seconds, and age of trees measured by cutting them down and counting the annual rings of each.

Ratio scales or levels of measurement differ from interval levels only in that ratio scales allow the location of an absolute or nonarbitrary zero point on the scale. Interval scales allow the arithmetic operations of addition and subtraction. Ratio scales go further to allow comparison of values by taking their ratios. The distinction between interval and ratio levels of measurement, however, is largely academic. Most real-world examples of interval scales are also ratio scales.

3.2 Sophistication of Analysis

Research methods and operational definitions determine a level of measurement. Then, statistical procedures are applied to what is measured. Any given statistical technique presumes a specific level of measurement. The more advanced the level of measurement, the more sophisticated the statistical techniques available.



It is always legitimate to use analysis techniques that presume levels of measurement that are one or more levels below the data. For example, interval level data may easily be collapsed into ordinal categories or ranks, and an ordinal level statistical procedure may be applied. The reverse, i.e., using an analysis technique higher on the scale of measurement than the data, is statistically invalid. For example, after collection, ordinal data can in no way be upgraded to interval measurements.

Consequently, the effect of applying an interval level statistical procedure to upgrade ordinal data is unknown. This unknown effect is, in fact, a major controversy concerning quantification in the social sciences. The following excerpt regarding a survey illustrates this controversy (Saunders 1982, p. 336):

> ... some of the statistical analyses are questionable and provide poor examples of data analysis .. .
>
> Data are usually placed in four types based on the criteria of identity, order, and additivity (Drew 1980). These four data types in order of increasing criterion properties are nominal, ordinal, interval, and ratio data. The Likert or 1 to 5 scale (l=strongly disagree, 3=neutral, 5=strongly agree, or 1=least desirable, 3=mixed or neutral, 5=most desirable) is an example of ordinal data. Inherent to this scale is the recognition that the differences between any two responses (e.g., 2 and 3, or 1 and 2) do not represent equal intervals. While ordinal data have the properties of identity and order, they lack the property of additivity, since the interval is unknown. Sokal and Rohlf (1969) refer to ordinal data as ranked variables.
>
> Because ordinal variables lack the property of additivity, the use of such central tendency measures as the mean and standard deviation is not possible. Some authors (Labovitz 1967, Nunnally 1967, and Borgatta 1968) argue that because emotions cannot be limited to five points on a scale, but rather are on a continuum, the Likert scale may be treated as an example of interval data. Interval data have the property of additivity. However, such logic does not answer the question of interval size and equality of intervals. Nunnally (1967) is a strong advocate of using the Likert scale as interval data, arguing that almost any parametric test can be applied to these data.
>
> Examination of the social and biological literature reveals acceptance of the Nunnally positions in certain journals, its partial acceptance in other journals, and its rejection in still other journals. Generally, the calculation of means for ordinal data are accepted, or at least done, because it may show data trends. However, since most response means are in the range of 2.75 to 3.75, a neutral rating says very little about the data. The same is true for the use of standard deviations, variances, and the t test. The use of frequency categories, an appropriate statistic, are more telling about the same data set. Chi-square analysis would be an appropriate inferential statistic for such ordinal data.

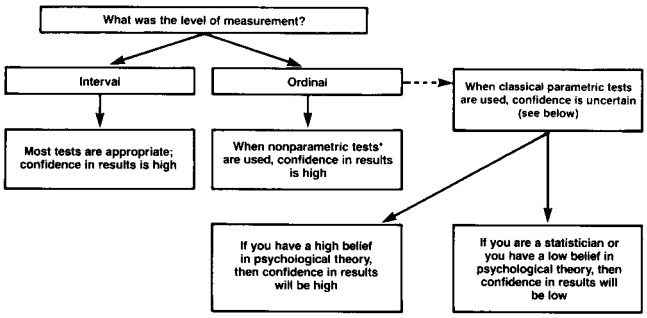
The statistical position is that using nonparametric rank-based statistical procedures (as opposed to classical *parametric statistical* procedures such as Pearson's correlation and factor analysis) for inference on ordinal variables is correct. Nonparametric procedures make no assumption of a probability model with finite numerical parameters. Classical parametric inferences should be restricted to interval level variables from the viewpoint of statistical theory, and without going into details, this position is hard to argue. Classical parametric procedures, however, are the common tools of psychological statistics. Like many areas of investigation, psychology has its own statistical peculiarities with nonstandard usage that is adapted to the prevailing practical situation. This situation in psychology is largely historical. Parametric tests were the first to be developed and still are the standard fare of introductory statistics courses. Several decades ago nonparametric tests (those appropriate for ordinal level variables) were relatively unknown to the average researcher.

Many psychologists and researchers with psychological training in environmental and other fields continue the tradition of using classical parametric statistical procedures with ordinal data. Statistical theory on the other hand dictates that using statistical procedures with an inappropriate level of measurement leads to conclusions that are neither empirically nor semantically meaningful. Unfortunately, the prevailing psychological orientation has been characteristic of the study of landscape quality and preferences.

The difference between parametric and nonparametric tests is in many cases only slight as to statistical power and statistical significance. In recent years the catalog of versatile and appropriate statistical procedures for ordinal level variables has been greatly extended to include the following (and others): median, Spearman's rank correlation, Kendall's tau, gamma correlation, Wilcoxon rank sum test, Kruskal-Wallis

test, multidimensional scaling, log-linear models, scoring based on preference pairs.

The following flow diagram may help you to judge the statistical support for or validity of a finding.



*Examples of nonparametric procedures include median, Spearman's rank correlation, Kendall's tau, gamma correlation, Wilcoxon rank sum test, Kruskal-Wallis test, multidimensional scaling, log-linear models, scoring based on preference pairs.

4. QUESTION ASSESSMENTS OF ATTITUDES AND PREFERENCES

The most common characteristic of techniques used to measure (landscape) preferences and perceptions is the lack of a clearly established metric base, i.e., lack of an unambiguous interval scale. There is no general agreement on an objective, physical instrument for measuring attitudes. Without such a measuring instrument, psychological assumptions must be made in order to presume an interval level of measurement. As was discussed in chapter 3 the validity of all such assumptions is questionable.

A variety of scaling techniques have been used for the measurement of preferences and perceptions in the scenic beauty evaluation literature. Several categories of techniques include ratings and transformed ratings, rank ordering (including paired comparison), Q-sort, and the semantic differential.

4.1 Likert Rating Scales

The most common type of rating scale in the literature on scenic beauty is the *Likert Scale*. The classic Likert Scale is formed as follows: subjects are presented with a list of statements (stimuli) on a single topic. Each item on this list is intended to measure the same attitude. Subjects are instructed to respond to each statement in terms of their degree of agreement, or disagreement, usually on a scale of 1 to 5. Responses for each subject are then summed over the questions to produce a single measure of attitudes on the corresponding topic.

An example is a Likert scale consisting of 10 statements on the subject of the respondent's self-esteem (Rosenberg 1965). Three of Rosenberg's statements illustrate the idea of Likert scaling:

1. I feel that I have a number of good qualities.
2. I wish I could have more respect for myself.
3. I feel I'm a person of worth, at least on an equal plane with others.

Each of the 10 statements used by Rosenberg had the following possible responses:

(5) Almost always true
(4) Often true
(3) Sometimes true
(2) Seldom true
(1) Never true

The sum of the response scores over all 10 questions produced a single measure called the Rosenberg Self-Esteem Scale.

The Likert method scales subjects, not stimuli. Consequently, all systematic variation in responses to stimuli is attributed to differences between the subjects. The major problem with Likert Scaling is its insensitivity to the location of individual items on an underlying attitude continuum. Therefore, an absolute interpretation of a person's score in terms of that continuum is not derivable.

In Likert Scaling, the recommended set of scale scores for each favorable statement is the set of successive positive integers, e.g., 0 = strongly disagree; 1 = disagree; 2 = undecided; 3 = agree; 4 = strongly agree. For unfavorable statements, this weighting scheme is reversed, e.g., 4 = strongly disagree. If each item is scored identically in this manner and responses for all items are summed, then the possibility that each item contributes equally to the total score is maximized. The Likert Scale was developed as an improvement over rank-ordered scales by introducing intensity-scaled responses for each item.

4.2 Rank-Ordered Scales

Rank-ordered scales operate on stimulus comparison data under the assumption that the subject can rank each item in a set. The method of *paired comparisons* is one type of rank-ordered scale. Stimuli are presented in combinations of pairs, and the subject is asked to judge each pair. For example, if the goal is to determine landscape preferences by viewing 10 photographs, subjects will be shown every possible pair of photographs and for each pair will be asked which of the photographs he or she prefers. Paired comparison is based on the law of comparative judgment (Thurstone 1927): for each stimulus there exists a most frequently occurring response. That is, a subject can discriminate the relative degree of an attribute, such as scenic beauty. Further, the degree to which any two stimuli can be discriminated is a direct function of their difference in regard to the attribute in question.

4.3 Q-Sort

Q-sort is. essentially a sophisticated method of rating and rank-ordering stimuli. Each subject is given a set of cards (stimuli) and asked to what extent each card characterizes the concept being evaluated. The subject is then instructed to sort the cards into a fixed number of piles in terms of the degree to which each stimulus represents the concept. These piles are taken to represent points along a continuum of representativeness of the stimuli to the topic. Determining the number of piles and the imposed frequency distribution of cards in piles is left to the investigator. The larger the number of piles the greater the potential for finer discrimination among items. An example of an imposed structure found by psychologists to be useful is the following:

Pile number:	1	2	3	4	5	6	7
Number of cards:	3	7	11	14	11	7	3

In this example the subject is instructed to place exactly 3 cards in pile 1, 7 cards in pile 2, 11 in pile 3, etc. Most statisticians recommend not forcing a distribution on the data, i.e., not predetermining how many cards should go into each pile.

4.4 Semantic Differential

The semantic differential uses sets of bipolar adjective pairs (e.g., warm and cold or hard and soft) to judge a concept. Subjects are asked to decide to what degree a concept is associated with each bipolar adjective pair. The scale for each bipolar adjective looks like a standard rating scale. Scale values are then factor-analyzed (see *chapter 5*) to answer questions regarding the number of dimensions underlying the concept. In comparison, the scaling models already discussed have all operated on the assumption of a unidimensional concept.

4.5 Two Examples of Visual Quality

A psychological orientation has been characteristic of the study of landscape quality and preferences. Two examples, one public evaluation approach and one professional evaluation approach, are given below.

4.5.1 Public Evaluation Approach

The public evaluation approach first purported to produce a separate quantitative measure of scenic beauty for a given landscape for each observer (Daniel and Boster 1976). It then produced a quantitative measure of scenic beauty for a given landscape across a sample of observers.

Data used to produce a quantitative measure for a given observer consisted of a set of preference judgments for that landscape. Specifically, the observer looked at a sample of slides taken at randomly selected locations and directions within the same area and rated each scene on a scale of, for example, 1to 10. Furthermore, the observer was told explicitly to use the full range of the scale, if possible. The *empirical distribution function* (cumulative relative frequency) of the observer's responses was computed at each category on the rating scale. These values were taken to be estimates of the probability distribution function, i.e., the cumulative probability that the landscape will be assigned a given rating by that observer. Standard scores associated with each cumulative probability were abstracted from a statistical table of normal theory *Z-scores*, and then an average Z-score was computed for this sample distribution.

The above procedure was repeated for the same observer for at least one other landscape. Let N denote the total number of landscapes; therefore, N was greater than or equal to two. Thus, a set of N average Z-scores was obtained for the observer, one for each landscape. One of these mean Z-scores was randomly selected to represent a referent landscape. Thus, if Zbar(1) denotes the mean Z-score computed from the referent landscape, then a scenic beauty estimate for the *i*th landscape was computed by

$$SBE(i) = [Zbar(i) - Zbar(1)] \times 100, \quad i = 2,3, ..., N.$$

Repeating the entire above procedure for a sample of observers allowed SBE's for a given landscape to be averaged across observers to obtain a mean SBE for that landscape. This method of estimating scenic beauty can be applied to one observer and at least two different landscapes, or to multiple observers and at least two different landscapes.

The basic scaling technique was a Likert rating scale modified for use with photographic stimuli from the same landscape (for discussion of Likert Scale, see *section 4.1*). The public evaluation approach described here attempted to improve on earlier approaches using Likert scaling by adjusting for each observer's idiosyncratic use of a rating scale, e.g., each observer's use of a different "underlying" scale.

From a statistical standpoint, this approach is nonstandard and knotty. For one thing, rating scale responses yield ordinal level measurements, which were treated as interval level measurements. The treatment is valid if each observer uses a scale of the same magnitude, that magnitude spans the entire range of response to the stimulus, and the categories are equally spaced. Because these specifications cannot be tested, they amount to psychological assumptions; the validity of results depends on the validity of these untestable assumptions.

Another problem with this public evaluation approach is that it did not follow statistical theory. By definition, the 50th *percentile* associated with a probability distribution is the specific value of a variable that corresponds to a cumulative probability of .5; the 35th percentile is the value that corresponds to.35, etc. Percentiles can be computed for any specified value of cumulative probability. The public evaluation

approach did the following for each landscape for each observer: it computed percentiles associated with a standard normal distribution for all values at which the empirical distribution function had been evaluated. But then it computed means and standard deviations of this set of standard normal percentiles. It is difficult to say exactly (or even approximately) what the resulting estimates of scenic beauty mean.

4.5.2 Professional Evaluation Approach

The professional evaluation approach used a descriptive inventory approach that relied on judgments of professionals rather than on public preferences (Sheppard and Newman 1979). The method purported to produce for a given landscape a quantitative estimate of the visual contrast created by a proposed intervention.

A detailed description of the landscape was prepared before the proposed modification. This was accomplished by breaking down the landscape into components (land and water, vegetation, and structure). Each component was described in terms of six visual elements: scale, color, line, form, texture, and space. A simulation of the landscape after the proposed intervention was then prepared, and as before, the simulation was described in terms of visual elements of landscape components. Contrast ratings were produced for landscape components to estimate the change in visual elements created by the proposed intervention. These were weighted to reflect their "relative importance," e.g., color has a weight of 3 and texture has a weight of 1. An intricate scheme was devised to produce a visual contrast rating for the landscape over all visual elements. Finally, the magnitude of the overall rating was used to determine whether the landscape intervention was approved.

Again, it is difficult to say exactly what the results mean. For one thing, I am not aware that color has been proven to be three times more important to perceptual discrimination than is texture. Where color is constant or color differences subtle, texture may prove extremely important. Also a line could be the most important visual element if it coincides with an edge (e.g., a skyline).

4.6 Magnitude Scales

4.6.1 Theory

One body of literature attempts to measure interval preferences for landscapes by using an approach called *magnitude scaling* (e.g., Daniel and Boster 1976, Buhyoff and Wellman 1980). In general, this type of approach was originally developed in signal detection theory for the ratio scaling of sensations that are physically measurable, such as heaviness of lifted weight, loudness of sound, brightness of light, and other perceptions of the five senses. Many social scientists claim that this psychophysical scaling technique solves the problem of ordinal-interval levels of measurement.

The paradigm for sensory psychophysical scaling follows. Subjects are presented with a series of sensory stimuli across a wide range (e.g., varying light intensities), one at a time in random order. Subjects are instructed to give numbers to the perceived brightness of each stimulus relative to the first light intensity, which is called the reference. So if a given light seemed 10 times brighter than the reference, the subject would give a number 10 times larger.

The results of hundreds of such numeric estimation experiments proved that humans are capable of using numbers to make proportional judgments of physical stimulation levels for virtually all of the five senses. These numeric estimates of the perceived strength of sensory stimuli were found to have a simple and regular mathematical relationship to the objectively measured stimulus values. The principle behind this mathematical relationship is that equal stimulus ratios produce equal subjective ratios. This principle is the essence of the psychophysical "power law" governing human impressions of most physical sensations and is probably the most strongly supported law of human judgments in psychology.

Early in the development of magnitude scaling, it was found that the empirically obtained mathematical function relating numeric estimation to each sensory modality varied reliably between sensations; specifically, different sensations were found to grow at different rates. These results, however, were challenged by critics who argued that the sole reliance on numeric estimation made verification of the power law impossible independent of numbers. Consequently, the following technique was used. Rather than match numbers to stimulus intensities, subjects would, for example, use force of hand grip to respond to the brightness of light. A basic conclusion was that the power law is not dependent on numeric estimation, it also occurs with other response modalities. This also cleared the way for cross-modality matching in which quantitative response modalities, each of which grows at a known characteristic rate, are matched to each stimulus. An example would be responding to the stimulus of light with both loudness of voice and force of hand grip. Within the cross-modality paradigm, the use of two responses allowed validation of the magnitude scale of impression.

Cross-modality matching allowed those who believed in the techniques of psychophysics to extend it to the magnitude scaling of social-psychological impressions by the simple substitution of social for physical stimuli. Usually, words or phrases denoting instances (items) on a social-psychological dimension take the place of the physical stimuli traditionally used in classic psychophysics. The reasoning behind this was straightforward; some researchers came to believe that estimates of the intensity of physical stimuli—impressions of the brightness of light, loudness of sound, heaviness of lifted weight—are indeed judgments, in part as a consequence of successful applications of the cross-modality matching paradigm to social stimuli. Several criterion tests (see Lodge 1981) were then developed to validate a magnitude scale of social judgments. If these tests are satisfied, then the derived scale is labeled a "psychophysically validated ratio scale."

This procedure was applied to social science data to surmount the difficulties involved with ordinal data. As such, it was a worthy effort. One criticism is an almost compulsive obsession with finding linear examples of regression. Linear relationships are not necessarily the ultimate in regression analysis (see *chapter 6*); for example, the data might be described better by a nonlinear function. But, the need was appreciated for "something" better than forcing ordinal data into interval statistical techniques for analysis of social science data.

4.6.2 Studies of Visual Quality

How are signal detection theory, psychophysics, and magnitude scaling used in the landscape literature? In terms of theory, they support the claim that a comprehensive stimulus-response function describing landscape preferences may exist. More specifically, the goal of this body of literature is to ". . . explore the possibility of the existence of a standard psychophysical or stimulus-response function that specifies *a priori* the shape and character of the relationship between preference and dimensions of the landscape" (Buhyoff and Wellman 1980, p. 259).

In terms of methods, however, this same body of literature should be examined critically. Three different landscape preference studies conducted over a 4-year period on a wide variety of subjects used paired comparisons of landscape slides (Buhyoff and others 1978, Buhyoff and Leuschner 1978, Buhyoff and Reiseman 1979). Paired comparisons produce ordinal measurements; however, in all three studies the results were assigned interval scaling scores by invoking Thurstone's law of comparative judgment. Complicated regression techniques were then used to search for a stimulus-response function to describe the data.

Invoking Thurstone's Law is equivalent to the need for belief in psychological theory to have confidence in results of classical parametric tests on ordinal data (see *section 3.2*). Those working in the psychophysical tradition of landscape preferences seem to have missed the point of signal detection theory, psychophysics, and magnitude scaling. The trend has been to use the psychophysical approach to justify the continued use of parametric procedures such as regression on ordinal variables. Why? Partly because of tradition, partly because of the perceived need to make longitudinal studies within some fields comparable, and partly because ordinal scaling is less expensive and less time consuming than magnitude-type scaling alternatives.

When you read studies that involve an attitude variable, ask these questions:

- What method (e.g., Likert scaling, paired comparisons, etc.) has been used in the study to scale the attitude variable?
- Does the author of the research report claim that this variable has an interval level of measurement?
- If so, what psychological assumptions does the author make to try to rationalize this claim? Does the author invoke Thurstone's law of comparative judgment, for example?

5. VERIFY ASSUMPTIONS OF CORRELATIONAL ANALYSES

When the methods component of a study includes statistical considerations, it will have a given purpose in terms of examining variables. Of the four purposes listed in *chapter 2,* one was that of exploring relationships among variables. Several common statistical methods are used for this purpose in evaluating landscape quality. The most important conceptually are *correlation, factor analysis,* and *multidimensional scaling.* These techniques seem complex, but they are useful tools that help to explain what is happening in a data set with a large number of variables.

5.1 Definitions

5.1.1 Simple Correlation

Correlation usually denotes the degree of strength of relationship between random variables taken two at a time. Even though a study involves a large number of variables, examining the correlation between each possible pair of variables is usually instructive. Correlation is particularly useful in exploratory studies: if the strength of the relationship is high (i.e., close to 1or close to -1), then for example we might be interested in trying to predict one variable from the other. Several measures of correlation for two variables are available. Examples are Pearson's product-moment correlation, Spearman's rank correlation, Kendall's tau, Phi-correlation, and intraclass correlation. The following chart summarizes the levels of measurement presupposed by these correlation measures:

Level of measurement of the first variable / Level of measurement of the second variable	Nominal	Ordinal	Interval
Nominal	Phi-correlation (two binary variables)		Intraclass correlation
Ordinal		Spearman's rank correlation; Kendall's tau	
Interval	Intraclass correlation		Pearson's product moment correlation

5.1.2 Factor Analysis

Factor analysis refers to a family of statistical techniques whose common objective is representing a set of variables in terms of a smaller set of hypothetical variables. It is based on the assumption that the smaller number of underlying factors are responsible for correlation among the variables. The simplest case is one in which one underlying common factor is responsible for the correlation between two observed variables.

For example, suppose 100 individuals are randomly selected from the population and their weight, height, blood pressure, etc., are measured. The measurements constitute observed variables and are interval-level measurements (see *section 3.1*). These basic data are then arranged systematically, in what is usually called a data matrix:

Entity	Variables		
	1	2	3. . .80
1	125	63	7
2	149	59	8
3	220	61	12
4	190	62	7
.			
.			
.			
100			

The data matrix has two dimensions. One is called the entity mode, which represents the cases—persons in this example—arranged as rows. The other dimension is the variable mode, which displays the observed measurements in columns. If 80 variables were measured for each of the 100 individuals and each measurement produced one value for a variable, then the data matrix would contain 100 people times 80 measurements or 8,000 numbers. The matrix would need to be simplified; it would contain too many numbers for easy comprehension.

The first step toward simplification is examining relationships among the variables. This can be done by forming a correlation matrix, in which the variables from the data matrix are both the rows and the columns. The values in the correlation matrix measure the association between each variable and each of the other variables (see *section 5.1.1*). A correlation matrix shows whether there are positive relationships among these variables (correlation values are greater than zero), negative relationships (correlation values are less than zero), and whether the relationships within some subsets of variables is stronger (correlation values are closer to 1 or -1) than that between the subsets.

Factor analysis is then used to address the question of whether these observed correlations can be explained by a small number of hypothetical variables, e.g., perhaps weight and height together tap a dimension having to do with body build. If the researcher has little idea as to how many underlying dimensions exist, factor analysis will uncover the minimum number of hypothetical factors that can account for the observed pattern of correlation, and allows an exploration of the data so that it can be reduced in size and analyzed economically. This is exploratory factor analysis. The majority of the applications in social science belong to this category. However, there is also confirmatory factor analysis. If the researcher, for example, believes at the start that different dimensions underlie the variables and that certain variables belong to one dimension or another, factor analysis confirms or tests these hypotheses. The division between the two uses is not always clear, and factor analysis has many methods and variants.

Factor analysis is virtually impossible to do without computer assistance. The usual approach is to input the correlation matrix into a factor analysis program and choose one of the many methods of obtaining the solution. (Several major alternatives are given in the literature, but the specifics can be safely ignored at this point.) The researcher specifies the number of common factors to be extracted or the criterion by which such a number can be determined.

Roughly, the program searches for a linear combination of variables (a factor) that accounts for more of the variation in the data as a whole than does any other linear combination of variables. The first factor is the single best summary of linear relationships exhibited in the data. The second factor is the second best linear combination of variables, given that it is not correlated with the first. That is, the second factor accounts for a proportion of the variance not accounted for by the first factor. Subsequent factors are defined similarly until all the variance in the data is exhausted.

The resulting set of factors is called the initial solution. A terminal solution—obtained by a complex procedure called rotation—may be a further simplification of the data, but it is beyond the scope of this report.

In general, in the fields of psychology and education, the main motivation behind use of factor analysis is finding the factor structure among a set of variables. This is not the motivation in other disciplines. Most other social science disciplines use factor analysis to simplify data by obtaining factor scales that can be used as variables in a different study. Factor scales are commonly analyzed along with other variables.

5.1.3 Multidimensional Scaling (MOS)

Multidimensional scaling is a set of mathematical techniques that enable discovery of the "hidden structure" of a data set. MDS operates on numbers that indicate how similar or different two objects are or are perceived to be. Such numbers are called *proximities.* Proximities can be ordinal-level measures. *A* correlation coefficient may be used as a measure of proximity. Proximities may be obtained during the data collection phases of a study by asking people to judge the similarity of a set of stimuli (e.g., photographs of landscapes). MDS results are displayed as a geometric pattern of points, like a map, with each point corresponding to one of the stimuli. This pattern is considered to be the hidden metric structure of the data. The greater the similarity between two objects, as measured by their proximity values, the closer they should be on the map. Generally, the most useful insights from MDS are gained by visually examining and interpreting the configuration. More complicated nonvisual techniques may also be used.

For example, suppose that each of 20 campers rated the degree of overall similarity between 10 landscape photographs on a scale ranging from 1 for "very different" to 9 for "very similar." No information would be given to the campers concerning the characteristics on which similarity was to be judged because the goal is to discover such information and not impose it. These data are input to an MDS computer

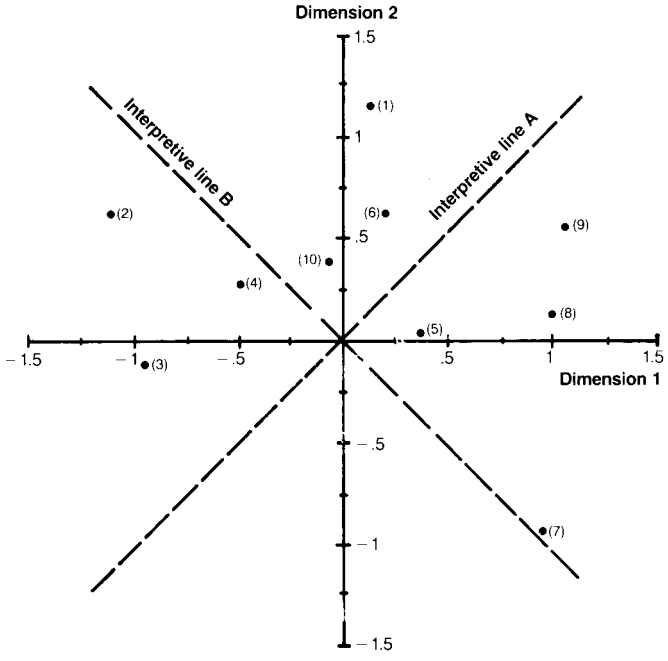
program. MDS calculations are complex, and even the simplest versions are always performed with the aid of a computer. Also, a large variety of different computational methods are used, some of which depend only on the rank order of the proximities.

The chief output of the MDS computer program is the map-like representation of the proximities data. The motivating concept is that the distance between the points should correspond to the proximities. In this sense, the output of MDS is not much different than a scatter diagram. MDS, however, can be mathematically complex. It is possible not only in two and three dimensions but also in four or more dimensions. This is impossible to visualize, but it can be dealt with mathematically.

Suppose, for the sake of simplicity, that the researcher starts with the results obtained from a two-dimensional analysis of these data. The computer output will include a list of coordinates for the landscapes and a plot of these values:

Stimuli (landscape)	Dimension 1	Dimension 2
1	0.15	1.12
2	-1.12	0.68
3	-0.90	-0.19
4	-0.50	0.29
5	0.36	0.02
6	0.19	0.64
7	0.96	-0.90
8	1.04	0.12
9	1.14	0.59
10	0.03	0.36

The most common way of interpreting such an MDS plot is looking for lines—possibly at right angles to each other—that would divide the data in some easily describable way. Interpretive divisions could be represented as dashed lines on the plot: Numbers in parentheses indicate which stimulus is plotted at that point.



Also suppose that the researcher knows something about these landscape photographs and might be able to interpret the results in a rough-and-ready fashion. For example, everything below line A is an example of a forested landscape. Everything above line A is an example of a nonforested landscape. Line B distinguishes both the forested and nonforested photographs in terms of how skilled the photographer was on the day(s) the photos were taken.

Researchers who use MDS virtually never stop at this simple level of analysis. Configurations are rotated (as in factor analysis), other statistical techniques may be applied for dimensional interpretation, etc. Analysis can get very complex indeed.

When reading a research report based on MDS, be aware that—because the techniques are relatively new and computer driven—users sometimes understand them less than they do other statistical procedures. Therefore, you should ask a statistician for help.

5.2 Assumptions and Pitfalls

The central aim of factor analysis and MDS is simplifying data without losing much information. Other factors aside, reliability of results can be judged by the degree to which the assumptions necessary for these techniques have been met.

Two basic assumptions underlie factor analysis. The first is that the observed variables are linear combinations of some underlying causal variables (factors). This assumption has to be substantiated by having some knowledge of the data and research problem to begin with. The second assumption involves the notion of parsimony: if both one-factor and two-factor models explain the data equally well, the convention is to accept the one-factor model. If either assumption is invalid, results can be fallacious or indeterminant. Factor analysis also requires interval level measurement of variables. When used with ordinal variables, several operations are not well defined in a statistical sense. The conservative approach dictates use of factor analysis on ordinal data for exploratory uses only and not for statistical inference.

The assumptions for MDS are less stringent. Interval level measurement is not required for some types of MDS. The only assumption required is that the subjects ranking the stimuli (photographs in the example) according to degree of similarity or difference know something about the items being ranked. This technique also does not require that the data have a multivariate normal distribution as does factor analysis. Therefore, even if it is little understood by most users, MDS is useful for describing the attitudes, opinions, or perceptions of the individuals doing the ranking. MDS can do what it is designed to do—including inferring a metric structure from nonmetric ordinal data. MDS, as developed by Roger Shepard (1962, 1963) at Bell Laboratories in the early 1960's, was partially a reaction against "Thurstonian" scaling procedures when used with psychological data.

The following tabulation summarizes the levels of measurement presupposed by factor analysis and multidimen-

sional scaling, according to whether these techniques are used for exploratory uses or for hypothesis testing.

Use	Level of measurement: Ordinal	Level of measurement: Interval
Exploring relationships	Factor analysis or MDS may be used	Factor analysis or MDS may be used
Testing hypotheses	Some types of MDS only may be used	Factor analysis or MDS may be used

• Is Spearman's rank correlation or Kendall's tau used to measure correlation between two ordinal-level variables? If so, the results are valid.

• Is Pearson's product-moment correlation used to measure correlation between two ordinal-level variables? If so, the results of the analysis are neither empirically nor semantically meaningful.

For the purpose of hypothesis testing (as opposed to exploratory use) ...

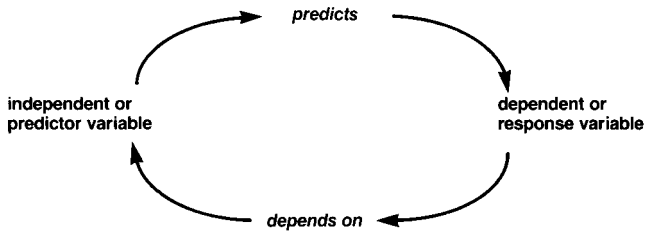
• Is an appropriate type of multidimensional scaling used on ordinal-level variables? If so, the results are valid.

• Is factor analysis used on ordinal-level variables for the purpose of hypothesis testing? If so, the results of the analysis are neither empirically nor semantically meaningful.

6. EVALUATE SUITABILITY OF PREDICTION MODELS

6.1 Simple Models

The simplest kind of prediction model involves two variables. One is called an *independent* or *predictor variable*, the other variable is called a *dependent* or *response variable*.

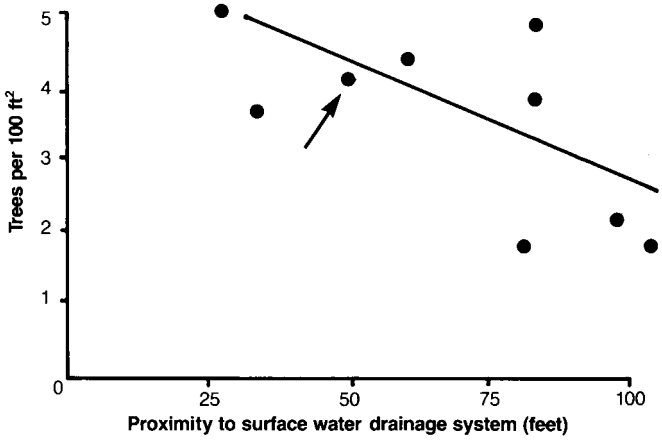


For example, a model might involve one variable measuring proximity to surface-water drainage system and a second variable measuring trees per 100 square feet. Proximity to surface-water drainage system is the independent variable, because the interest is in studying how and how well it may be used to "predict" the other variable. The dependent variable is therefore the number of trees per 100 square feet. Put another way, the object is to see how the number of trees per 100 square feet (the dependent variable) "depends on" proximity to surface-water drainage system (the independent variable).

6.1.1 Linear Models

Two variables can be related in literally hundreds of thousands of ways. One of the simplest is a *linear relationship*, meaning that it may be described by a straight line. For example, for every 25 yards closer to surface-water drainage system, the number of trees per 100 square feet will increase by one. Few real-world relationships are this simple. Sometimes, however, a straight line provides an approximation to a real-world relationship that is "good enough" to be useful.

The graph below is an example of data that are "approximately" linearly related:



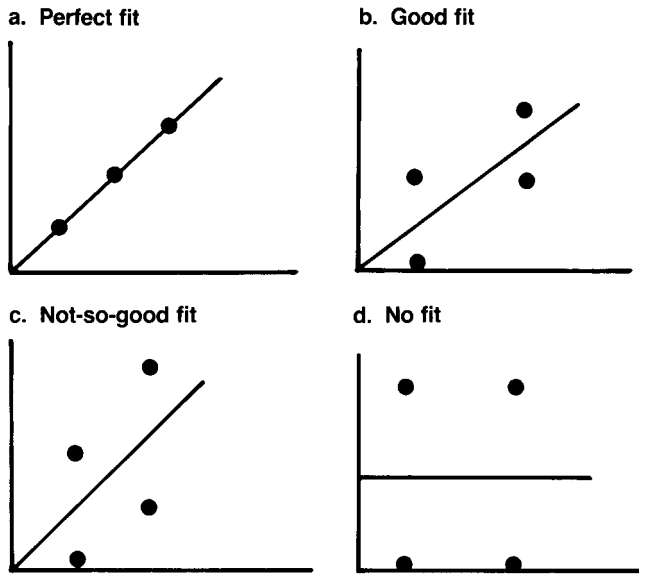
Proximity to surface water drainage system is plotted horizontally; number of trees per 100 square feet is plotted vertically. The sample consists of 9 pairs of observations. Each pair constitutes one sample point and consists of a measurement of each of the two variables. For example, 50 feet from surface water drainage system were four trees per 100 square feet (this point is marked with an arrow on the graph). The straight line in the plot is an "approximation" to the real data.

To describe the relationship between the two variables, the researcher no longer has to look at a list of data points. Such a list gets harder to "read" as the number of data points gets larger. Instead, a simple linear mathematical equation concisely summarizes the list of data points.

A central problem in using a straight line to approximate real data is finding the line that gives the best fit to the data. There are an infinite number of lines to choose from. The researcher will want to choose the straight line that is in some way "closest" to all the data points. Luckily, statisticians agree on the best method to use to choose the line: *least-squares estimation*. It is used universally in the computer programs

that researchers invoke to statistically fit straight lines to their data.

The procedure that produces the "best" straight line for a given set of data does not guarantee that a straight line will adequately describe the data. The graphs below illustrate lines that have been fit to data by least squares estimation.



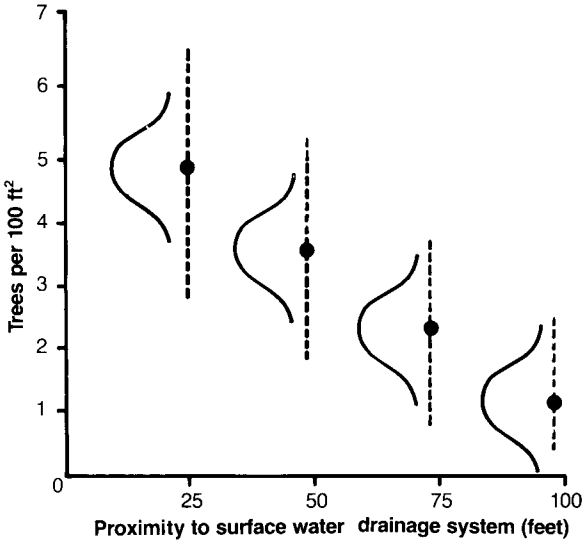
This "best" line is next to useless in summarizing the observed relationship between two variables in graphs c and d, because the data just are not linear!

I want to emphasize the difference between *model-building* and hypothesis testing (or inference). In model-building, the object is to describe a set of data points by a mathematical equation that provides a good approximate fit to these points. No assumptions are necessary because the researcher isn't saying anything about probabilities. In hypothesis testing, various assumptions are required because the researcher needs to know the probability of reaching a wrong conclusion. Five assumptions are necessary for inference with straight lines involving two variables. These assumptions involve how and where the data were collected.

Assumption 1. The values of the independent variable are interval level and are *fixed*, not random. That is, the researcher decides in advance of data collection upon the specific levels of the independent variable at which to take measurements on the dependent variable. This assumption would be satisfied for the example if—before collecting data—the researcher decided to measure number of trees per 100 square feet at each of the following distances from surface water drainage systems: 25, 50, 75, and 100 feet. The levels of the independent variable thus would be fixed at 25, 50, 75, and 100 feet. No data would be collected for other values of the independent variable.

Assumption 2. The values of the dependent variable for each level of the independent variable are interval level and are statistically independent random variables. If the researcher fixes levels of the independent variable, then the dependent variable will take on values at these levels with associated probabilities. That is, the values of the dependent variable for each level of the independent variable will be *random variables*. To test statistical hypotheses, the researcher must also assume that these values are statistically independent. Roughly speaking, statistical independence means that the occurrence of one observation in no way influences the occurrence of the others. Random sampling helps to assure the statistical independence of the measurements *(chapter 7)*.

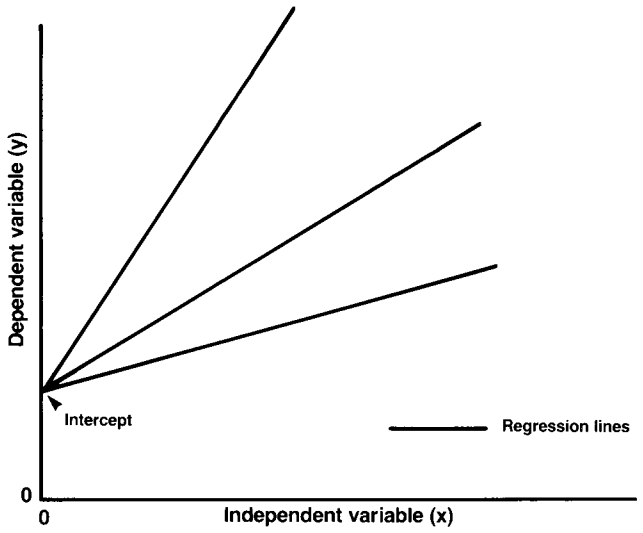
Assumption 3. For each value of the independent variable, the distribution of the dependent variable is normal. In the example, suppose the researcher works with four surface water drainage systems and therefore ends up with four values of the dependent variable at each level of the independent variable. Each set of values at each fixed level of the independent variable represent a set of observations sampled from a population consisting of many values. The assumption is that the population values have a normal distribution.



Assumption 4. The population variance of the dependent variables is the same for all values of the independent variable. Thus the normal distributions may all be situated in different places (have different means), but all have the same shape (have the same variance). This assumption and assumption 3 can be checked by past experience, or by present data if numbers of observations at each level of the independent variable are sufficient. Often, however, insufficient numbers of observations are collected, or else the researcher neglects to check.

Assumption 5. The regression is linear. That is, the means of the normal distributions described above lie over a straight line. These means are indicated by the points in the above graph.

The mathematical equation for a straight line involves two *parameters* (i.e., constants). These control where the straight line falls on the graph. One parameter is the intercept. It designates the point at which the line hits the vertical axis. The other parameter is the slope. It designates the change in the dependent variable that is associated with one unit of change in the independent variable:



The three regression lines in the above graph have the same intercept, but have different slopes (directions).

Several statistical hypotheses can be tested if the five assumptions above are met. These hypotheses include the following:

- The slope of the *regression line* is zero, indicating that the independent variable is not linearly related to the dependent variable.
- The slopes of two lines calculated from samples drawn from two independent populations are equal.
- The intercept of the regression line is equal to some specified value.

Often researchers will fit a regression line to data and proceed to test statistical hypotheses where one or more of the five assumptions are violated. What is the result?

The effect of departure from a specific assumption must usually be studied by analyzing specific cases. General conclusions of effects of departure from an assumption are therefore not rigorously derived mathematical results. Nevertheless, some general principles can be stated:

- The assumption of independence of the observations is vital to the accuracy of the analysis. Therefore, the researcher should strive to assure that the observations are independent.
- For modest sample sizes, moderate deviations from the assumption of normality have little effect on the accuracy of the analysis.
- Deviations from the assumption of equal variances have little effect on the accuracy of the analysis, if the numbers of observations occurring at each level of the independent variable are equal. Therefore, a researcher should strive for a research design with equal sample sizes for each value of the independent variable.

6.1.2 Log-Linear Models

Log-linear models are similar to linear models. A basic difference is that log-linear models are designed to operate on nominally scaled data (see *chapter 3*), whereas linear models are designed to operate on interval-level data. Data for a log-linear model are best viewed as arising from a table with several dimensions, one dimension corresponding to each variable. For example, a two-dimensional data table might look like this (one variable is sex, another is race):

	Race	
Sex	White	Non-White
Male	36	22
Female	54	23

Although log-linear models were originally formulated to operate on nominal data, recent focus is on developing models tailored specifically for use with *singly ordered tables*, i.e., one ordinal variable and *doubly ordered tables*, i.e., two ordinal variables. A two-dimensional singly ordered table might involve the two variables sex (nominal) and social class (ordinal). Categories of social class might be lower, middle, and upper. An example of a two-dimensional doubly ordered table might involve the two ordinal variables social class and preference. Categories of preference might be weak, medium, and strong.

The general log-linear approach involves an *a priori* assignment of scores to the categories of the ordinal variables. For the example of preference, scores might be 1—low preference, 2—medium preference, and 3—strong preference. The model could be specified with either equal or different (sometimes arbitrary) spacing between categories of the ordinal variable. Spacing parameters could also be estimated from the data by established statistical methods, which would be optimal from the standpoint of not requiring strong spacing assumptions.

The study of landscape quality and preferences often includes assessing whether intervention into a landscape (e.g., constructing a road) will affect public preferences for the landscape. Suppose the interest is not only in constructing a road, but also in choosing the type of road to construct: gravel, black-top, or cement. The intervention variable would have these three categories plus a fourth corresponding to no road. A log-linear model could be constructed for a singly-ordered table, in which one of the variables (type of road) is nominally scaled and the other (preference for the landscape) is ordinally scaled. Separate log-linear models could be applied to the same sample of individuals for different landscapes. Then, however, the results of the different models would have to be pieced together, which—as with general linear models—has not been systematized. That is, the manner in which results are pieced together remains arbitrary.

6.2 Complex Models

More complicated types of prediction models are possible by increasing the number of variables, or specifying a more complex relationship among the variables, or both. Suppose that instead of one independent variable the analysis includes two or more independent variables in conjunction with a

single dependent variable. Suppose the researcher wants to describe the relationship between the independent variables and the dependent variable as linear. Again, the method of least squares estimation can be used to find a straight line that best fits the data. Regression with two or more independent variables is called *multiple regression* (as opposed to simple regression). Multiple linear regression is basically similar to simple linear regression except for the complexity of the calculations. The same assumptions are necessary for inference.

For example, if data were collected on 20 independent variables, not all of them would be used in the final model, because the simplest model possible is desired to predict the dependent variable. Furthermore, a mathematical relationship other than a strictly linear one might be used. Whenever data on a set of independent variables is collected, two questions arise: (1) Which of the variables should be included in the regression model? (2) What mathematical relationship should be used to describe them?

Finding a subset of independent variables for the final model and a function between them that adequately predicts the dependent variable, is called model-building. The process destroys the inferential capabilities of the standard linear model. The reason for model building is not that a true model exists and just needs to be found. Instead, it is to predict approximately the dependent variable and to understand the phenomenon being studied. A simple model is needed and one that cannot be significantly improved.

First consider what mathematical relationship should be used to describe the relationship between the variables. Looking at the mathematical form of a simple line and two variables helps in exploring this question. Let Y denote the dependent variable, x denote the independent variable, a denote the intercept parameter, and b, denote the slope parameter. The linear equation is

$$Y = a + b_1x$$

The mathematical form of the relationship between the independent variable and the dependent variable can be changed by "transforming" terms in the equation. Some common transformations include taking the natural logarithm of one or more terms, taking the square of one or more terms, or raising one or more terms to some other power. One of the reasons for *transformations* is improving the fit of the model to the data.

Hull and Buhyoff (1983) fit six different mathematical functions to an independent variable measuring distance to a topographic feature and a dependent variable measuring scenic beauty. These included a variety of transformations:

$$\begin{aligned}
&\ln(Y) = \ln(a) + b_1\ln(x)\\
&Y = a + b_1(x^2)\\
&Y = a + b_1\ln(x)\\
&\ln(Y) = \ln(a) + b_1 x\\
&Y = a + b_1x + b_2x^2\\
&Y = a + b_1x
\end{aligned}$$

The dependent variable (scenic beauty) in this study was an ordinally scaled variable (see chapter 3) but it was treated as an intervally scaled variable and several statistical tests were performed for each of the six functions. While their model-building exercise was appropriate, the results of their hypothesis tests are suspect because ordinality has destroyed the inferential capabilities of the model.

Now consider which variables to include in the model. Virtually all regression analyses involving more than two variables are done by computer. A procedure that helps to select best candidates of variables and transformed variables is called *stepwise regression.* There is no unique best subset of variables. One reason to use transformations is to improve the fit of the model. Another reason is that transforming one variable will sometimes result in a simpler model—one with fewer terms.

Suppose a certain model (variables plus functional relationship between them) is appropriate to study certain phenomena. This appropriateness can be established by building the model with one sample of data and testing the fit of the model on a separate sample of data. Then the model can be used to predict the dependent variable from the independent variable. For example, suppose that in the study of the effect of proximity of surface-water drainage system (x) on number of trees per 100 square feet (Y), the researcher found the following functional relationship:

$$Y = 6 + (-.04)x$$

Then for x = 90, Y = 6 - .04(90) = 2.4. In other words, at 90 feet from the water drainage system, the researcher predicts 2.4 trees per 100 square feet.

Models with many independent variables give more information about potential causal factors for the phenomena being studied. While including variables in a regression equation does not imply causality, the inclusion of certain factors over others in the equation helps aid understanding of the mechanisms at work in the phenomena.

If the primary statistical purpose of the study is building prediction models using variables ask the following questions:

- Which variables are dependent (i.e., response) variables?
- Which variables are independent (i.e., predictor) variables?

If the researcher used simple linear regression (or multiple regression) and tested hypotheses, were the following assumptions met:

1. Values of the independent variables are interval level and fixed.

2. Values of the dependent variable for each level of the independent variable are interval level and are statistically independent random variables.

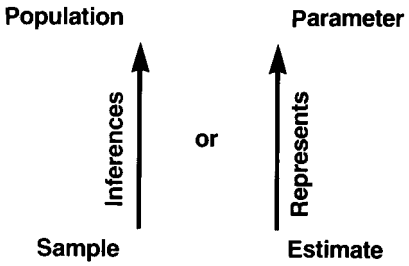
3. For each value of the independent variable, the distribution of values for the dependent variable is normal.

4. Population variance of values for the dependent vari-able is the same for all levels of the independent variable.

5. Regression is linear.

7. DETERMINE REPRESENTATIVENESS OF SAMPLE

Researchers usually want to generalize about a whole class or *population* of individuals or things; however, for reasons of cost, time, and practicality, this is not really possible. Only part of a population can be examined, and this part is called the *sample*. A researcher then makes generalizations from the part to the whole, or more technically a researcher makes inferences from the sample to the population:



What a researcher usually wants to estimate are certain numerical facts or parameters about the population of interest. An example is how many pines per acre grow in a National Forest. The forest is the population, and the number of pines per acre is the parameter. Because limited time and money prohibit counting all of the pines, the researcher samples a few acres of the forest, counts the pines on those acres, then estimates the number per acre in the population. Because parameters like this one cannot be determined with total precision, a major issue is accuracy—how close is the estimate from the sample to the actual parameter in the population?

Parameters are estimated by statistics—the numbers that can be computed from a sample. Statistics are known, parameters are unknown. Estimating the parameters of a population from a sample is justifiable if the sample represents the population. In general, (1) the method of choosing the sample determines whether the sample is representative; and (2) the best methods of choosing a sample involve the planned introduction of chance or probability. The main reason for probability samples is to avoid *bias,* which can be defined as systematic error. Many different types of bias exist. A few of these are discussed in *section 7.3*.

7.1 Nonrepresentative Samples

The estimate of a parameter will be fairly accurate if the sample is obtained by chance. If the sample is obtained by human judgment then the estimate will be biased, that is, it will systematically deviate from the true population parame ter. The reason is that human judgment is not impartial, but chance or probabilistic methods are impartial.

Convenience sampling and *quota sampling* are methods that involve human judgment. In convenience sampling, the units sampled are those readily available. An example would be choosing acres that are near roads in the National Forest, then counting the pines in those acres. Such a sample would not represent the population at large and would be biased, unless the researcher is lucky.

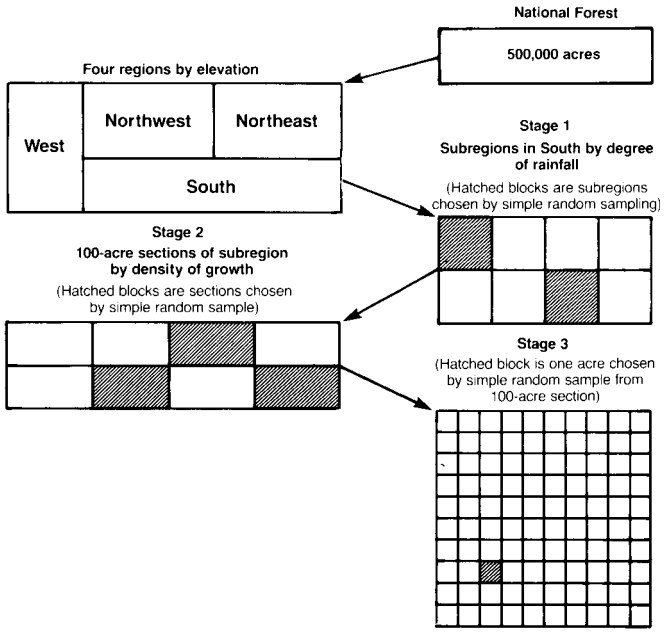
In quota sampling the National Forest would be canvassed. The goal is to find a perfect cross section of the forest on all the key variables that relate, say, to the existence of pine trees. Each person would be assigned a fixed number of acres to count pines, a fixed number of acres having pines of certain types, heights, etc. This type of sampling involves several difficulties. First, assigning quotas of type, height, or whatever about pines involves circular reasoning; the quotas are what you are trying to find out about the population (the parameter) and are not information that can be taken from other sources and used to construct a sampling procedure. Second, within the assigned quotas of acres, assistants are free to pick any acre they like. This leaves a lot of room for human choice, which is always subject to bias. In short, in quota sampling the sample is handpicked to resemble the population with respect to some key characteristics. This method seems sensible, but does not work well in practice. It was, for example, the sampling scheme used by Gallup, Roper, and others, that predicted Dewey would be elected president in 1948; Truman won the election. The interviewers used their own discretion to fill their quotas and chose too many Republicans, because they are marginally easier to interview in terms of having phones, living in nicer neighborhoods, or having permanent addresses.

7.2 Probability Samples

How is chance used to draw a sample? Assume that the National Forest in the example is well plotted in terms of acres and their boundaries, and that the forest is composed of 5,000 acres. Each acre is assigned a separate number, all 5,000 numbers are placed in a bin, and a statistician helps to decide that 200 of them should be drawn at random in order to have statistically significant results. Because there would be no point in counting the pines on the same acre more than once, the 200 draws are made without replacement. In other words, the bin is shaken well to mix up the plot numbers, and one is drawn out at random and set aside, leaving 4,999 in the bin. The bin is shaken again and the procedure repeated until 200 plot numbers have been chosen. These 200 plot numbers form the sample, called *a simple random sample.* The plot numbers simply have been drawn at random without replacement, and at each draw, every remaining plot number has an equal chance to be chosen. No human discretion is used, and the procedure is impartial—every acre has the same chance of getting into the sample.

For various reasons, simple random sampling often is impractical, and *multistage cluster sampling* is used instead. Suppose the National Forest is composed of 500,000 acres instead of 5,000 acres. Drawing plot numbers at random, in the statistical sense, becomes difficult. And because the forest is much larger, visiting the right acres becomes expensive.

In multistage cluster sampling, the forest would be separated into regions that are similar to each other. Suppose four regions are distinct in terms of elevation. Within each region, subregions could be grouped on the basis of another similarity, such as amount of rainfall. Then a simple random sample of these subregions would be taken. Only the selected subregions would be visited. This completes the first stage of sampling. Each subregion would then be divided into 100-acre sections on the basis of some other index of similarity—perhaps density of growth. At the second stage of sampling a simple random sample of these 100-acre sections would be drawn. At the third stage of sampling one acre would be drawn at random from the 100 acres in each selected section. The pines in this 1-acre plot would then be counted. This is a rather crude example of a somewhat complicated concept:



Multistage cluster sampling eliminates selection bias because it eliminates human choice. In summary, all probability methods for sampling have two critical properties in common with simple random sampling:

(1) A definite procedure exists for selecting the sample, and it involves the planned use of chance, and

(2) The procedure involves no human discretion as to who is interviewed or what part of the environment is sampled.

7.3 Possible Biases

In terms of sampling, selection bias is the most obvious type of bias. *Selection bias* is a systematic tendency on the part of the sampling procedure to exclude one kind of person or thing from the sample. For example, acres over a certain elevation might be excluded because of the difficulty of getting sampling personnel and equipment to the site. If a selection procedure is biased, taking a larger sample doesn't help, but just repeats the basic mistake on a larger scale.

A second common form of bias in terms of sampling is *nonresponse bias.* This occurs when a large number of those selected for a sample do not in fact respond to the questionnaire or interview. When counting pine trees this problem does not arise, but in other types of research it does. Nonresponders usually differ in terms of social class, education, values, etc., from responders. This means that the information obtained in the sample is distorted because it does not truly represent the whole population of interest. To compensate for nonresponse bias, a researcher can give more weight in the analysis to responses from people who were available but hard to get.

The basic formula presented so far concerning an attempt to choose a representative sample is

$$\text{estimate} = \text{parameter} + \text{bias}.$$

Any sample used to derive an estimate, however, is still only part of a population. Our estimate is likely to be a bit off in terms of accurately estimating the parameter. If a sample is part of a population that is chosen at random, then the amount by which an estimate misses a parameter is controlled by chance. A more accurate equation then is

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}.$$

Where the cost of a study fits into this equation is unclear. No straightforward relationship exists between cost and bias, for example. Depending on the type of study and the capability of the investigators involved, less expensive methods can be less biased than more expensive ones. Methods are often chosen by the researcher (not the statistician).

In general, examine any sampling procedure and try to decide if it is good or bad. Many, if not most, samples are unsatisfactory. If the sample is unsatisfactory, then the results of analysis must also be unsatisfactory. To decide if a sample is satisfactory, first examine how it was chosen. Were probabilistic methods used? If not, was there selection bias? Was there nonresponse bias? Are any other sources of bias obvious in the sampling procedure?

GLOSSARY

alternative hypothesis: the opposite of the null hypothesis; an educated guess that the distribution of the observed values does show a difference from the assumed probability distribution

association: two variables are highly associated if you can use the value of one variable to predict the value of the other with confidence

bias: any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth

convenience sample: the units sampled are those readily available (not a good sampling method)

correlation: a measure of association, usually between two random variables; correlation values close to1mean strong positive association; values close to-1 mean strong negative association; 0 means no association

dependent variable: a variable that "depends on" another variable, i.e., a variable that is predicted by another variable

descriptive statistics: techniques to organize and summarize data

doubly ordered table: a table with two ordinal variables

empirical distribution function (cumulative relative frequency): the proportion of responses up to and including a specified response category

factor analysis: a family of statistical techniques designed to represent a set of (interval-level) variables in terms of a smaller set of hypothetical variables

fixed values: values of the variable are fixed in advance of data collection

hypothesis test: a type of statistical inference that involves comparison of an observed value (or values) with a value (or values) derived from probability theory

independent variable: a variable used to predict another variable

interval and ratio scales: result from measuring things with a physical unit, such that the exact distance between things is established

least squares estimation: the method commonly used to choose the straight line that is closest to the data points

Likert Scale: uses multiple stimuli to produce a single attitude measure

linearity: (see linear relationship)

linear relationship: the relationship between a dependent variable and one or more independent variables may be described by a straight line

magnitude scaling: the attempt to measure interval-level preference by using psychophysical scaling methods

mean: arithmetic average

median: the value that divides the ordered observed values into two groups of equal size

mode: the value that occurs most often

model-building: finding a subset of variables and a function between them that adequately predicts a dependent variable(s)

multidimensional scaling: similar to factor analysis; "proximities" among objects are used as data

multiple regression: regression with two or more independent variables

multistage cluster sampling: a type of probability sample that is conceptually more complex than the simple random sample but on the other hand is more practical when the population is large

multivariate: involves two or more variables

nominal scale: results from sorting things into homogeneous categories

nonresponse bias: occurs when a large number of those selected for a sample do not respond to the questionnaire or interview

normal probability distribution: a symmetric distribution specified by two parameters and the following equation:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

The distribution is sometimes called an "error curve" or Gaussian distribution or "bell-shaped" curve. σ^2 is the population variance

null hypothesis: an educated guess that the distribution of the observed values basically does not differ from the assumed probability distribution

operational definition: definition of a concept that includes procedures for classifying and measuring the phenomenon

ordinal scale: results from sorting things into homogeneous categories that are ordered or ranked with respect to the degree, intensity, or amount of something they contain

paired comparisons: subjects are asked to judge pairs of stimuli

parameters: constants that determine the shape and location of a distribution; facts about the population of interest

parametric statistics: statistics that assume a parametric model, i.e., a model with finite numerical parameters or facts

percentile: the specific value of a random variable that corresponds to a given cumulative probability

population: a whole class of individuals or things

predictor variable: independent variable

proximity: a measure of how similar or how different two objects are

probability model: translation of an empirical problem into probabilistic terms

Q-sort: subjects are asked to sort stimuli into a fixed number of categories in terms of the degree to which each stimulus represents the concept

qualitative: extends the idea of physical measurement to include various categorization procedures

quantitative: involves physical measurement

quota sampling: assigning quotas to characteristics of the population and then sampling to achieve these quotas (not a good sampling method)

random variable: a variable that has probabilities associated with each possible numeric value

range: largest value in a set of observations minus the smallest value

rank-ordered scales: example is the method of paired comparisons (see above)

rating scales: example is the Likert Scale (see above)

ratio scales: (see interval and ratio scales)

raw data: the values collected for each variable, without any statistical or other manipulation having been done to alter or adjust them

regression line: a line fit to the data using least squares estimation and analyzed within the framework of statistical theory of regression

response variable: dependent variable

sample: part of the population

sample variance: the sum of squared deviations of each observation from the mean, divided by the number of observations minus 1

selection bias: a systematic tendency on the part of the sampling procedure to exclude one kind of person or thing from the sample

semantic differential: subjects are asked to decide to what degree a concept is associated with selected sets of bipolar adjective pairs

simple random sample: each sample unit has an equal chance of being chosen (best sampling method)

singly ordered table: a table with one ordinal variable

standard deviation: the square root of the sample variance

statistical hypothesis: a statement concerning the distribution of probabilities for different values of a random variable

statistical inference: statistical methods designed to assess the impact that chance variation has on research results

statistical independence: the value obtained for one observation does not affect the values we are likely to get for other observations

stepwise regression: a computerized procedure for selecting best candidates of predictor variables in a regression problem

theoretical definition: definition of a concept in terms of other concepts which supposedly are already understood

transformations: a function of a variable (any variable—dependent, independent, etc.—can be transformed)

univariate: involves one variable

variable: something that is measured or counted

Z-scores: normally distributed random variables that have been converted to units of standard deviations relative to the mean

REFERENCES

Buhyoff, Gregory J.; Lauschner, W. A. **Estimating psychological disutility from damaged forest stands.** Forest Sci. 24(3):424-432; 1978.

Buhyoff, Gregory J.; Reiseman, M. F. **Experimental manipulation of dimensionality in landscape preference judgments; a quantitative validation.** Leisure Sci. 2(3/4): 22a-238; 1979.

Buhyoff, Gregory J.; Wellman, J. Douglas. **The specification of** a **non-linear psychophysical function for visual landscape dimensions.** J. Leisure Res. 12(3): 257-272; 1980.

Buhyoff, Gregory J.; Wellman, J. D.; Harvey, H.; Frazer, R. A. **Landscape architect's interpretation of people's landscape preferences.** J. Environ. Manage. 6: 255-262; 1978.

Carlson, A. A. **On the possibility of quantifying scenic beauty.** Landscape Plann. 4: 131-172; 1977.

Daniel, Terry C.; Boster, Ron S. **Measuring landscape esthetics: the scenic beauty evaluation method.** Res. Paper RM-167. Fort Collins, CO: Rocky Mountain Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture; 1976. 66 p.

Ecological Society of America Committee, subcommittee on journal content: **final report.** Bull. Ecol. Soc. 63(1): 26-41; 1982.

Hull, R. Bruce, IV; Buhyoff, Gregory J. **Distance and scenic beauty:** a **nonmonotonic relationship.** Environ. Behav. 15(1): 77-91; 1983.

Lodge, Milton. **Magnitude scaling: quantitative measurement of opinions.** Beverly Hills, CA: Sage Publications; 1981. 87 p.

Rosenberg, M. **Society and the adolescent self-image.** Princeton, NJ: Princeton University Press; 1965. 326 p.

Saunders, Paul Richard. **Letters to the editor.** Bull. Ecol. Soc. Am. 53(4): 336-337; 1982.

Schor, S.; Karten, I. **Statistical evaluation of medical journal manuscripts.** J. Am. Med. Assoc. 195: 1123-1128; 1966.

Shepard, R. N. **The analysis of proximites: multidimensional scaling with an unknown distance function.** Psychometrika 27: 125-140, 219-246; 1962.

Shepard, R. N. **Analysis of proximities as a technique for the study of information processing in man.** Hum. Factors 5: 33-48; 1963.

Sheppard, Stephen R. J.; Newman, Sarah. **Prototype visual impact assessment manual.** Berkeley, CA: Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture; 1979; prepared as part of Cooperative Research Agreement PSW-62. 88 p.

Thurstone, L. L. **A law of comparative judgment.** Psych. Rev. 34: 273-286; 1927.

APPENDIX—ADDITIONAL READING

Chapter 1—Research Concepts

American Statistician. **Ethical guidelines for statistical practice: historical perspective, report of the ASA Ad Hoc Committee on Professional Ethics, and discussion.** Am. Stat. 37(l): 1-20; 1983.

Blalock, Hubert M. **Social statistics.** New York: McGraw-Hill; 1960. 583 p.

Huff, Darrell. **How to lie with statistics.** New York: W. W. Norton and Company; 1954. 142 p.

Tanur, Judith M., ed. **Statistics: a guide to the unknown.** (2d ed.) San Francisco: Holden-Day; 1978. 430 p.

Chapter 2—Statistical Purpose

Dixon, Wilfrid J.; Masey, Frank J., Jr. **Introduction to statistical analysis.** (4th ed.) New York: McGraw-Hill; 1983. 678 p.

Mendenhall, William. **Introduction to probability and statistics.** (6th ed.) Boston: Duxbury Press; 1983. 646 p.

Neyman, Jerzy. **First course in probability and statistics.** New York: Holt; 1950. 350 p.

Wonnacott, Thomas H.; Wonnacott, Ronald J. **Introductory statistics.** (2d ed.) New York: John Wiley & Sons; 1972. 510 p.

Chapter 3—Levels of Measurement

Blalock, Hubert M. **Social statistics.** New York: McGraw-Hill; 1960. 583 p.

Ellis, Brian. **Basic concepts of measurement.** London: Cambridge University Press; 1966. 219 p.

Ross, S. **Logical foundations of psychological measurement.** Copenhagen, Denmark: Munksgaard; 1964.

Stevens, S. S. **On the theory of scales of measurement.** Science 103: 677-680; 1946.

Stevens, S. S. **On the averaging of data.** Science 121: 113-116; 1955

Stevens, S. S. **Measurement, statistics, and the schemapiric view.** Science 161: 849-856; 1968.

Chapter 4—Assessing Attitudes and Preferences

Green, David M.; Swets, John A. **Signal detection theory and psychophysics.** New York: John Wiley & Sons; 1974. 497 p.

Likert, R. **A technique for the measurement of attitudes.** Archives of Psychology. New York: Columbia University Press; 1931. 55 p.

Marks, Lawrence E. **Sensory processes: the new psychophysics.** New York: Academic Press; 1974. 334 p.

Wilson, T. P. **Critique of ordinal variables.** In: Blalock, H. M., Jr., ed. Causal models in the social sciences. Chicago: Aldine-Atherton; 1971. 16 p.

Chapter 5—Multivariate Analysis

Kim, Jae-On; Mueller, Charles W. **Introduction to factor analysis: What is it and how to do it.** Beverly Hills: Sage Publications; 1978. 79 p.

Kruskal, Joseph B.; Wish, Myron. **Multidimensional scaling.** Beverly Hills: Sage Publications; 1978. 93 p.

Rummel, R. J. **Applied factor analysis.** Evanston: Northwestern University Press; 1970. 617 p.

Shepard, Roger N.; Romney, A. Kimball; Nerlove, Sara Beth, eds. **Multidimensional scaling: theory and applications in the behavioral sciences, volumes I and II.** New York: Academic Press; 1972. 584 p.

Chapter 6—Prediction Models

Draper, N. R.; Smith, H. **Applied regression analysis.** New York: John Wiley & Sons; 1966. 407 p.

Feinberg, Stephen E. **The analysis of cross-classified categorical data.** Cambridge: The MIT Press; 1977. 151 p.

Haberman, Shelby J. **Analysis of qualitative data.** Volume 1: Introductory Topics. New York: Academic Press; 1978. 368 p.

Neter, John; Wasserman, William. **Applied linear statistical models: regression, analysis of variance, and experimental designs.** Homewood, IL: Richard D. Irwin, Inc; 1974. 842 p.

Chapter 7—Sampling

Cochran, William G. **Sampling techniques.** New York: John Wiley & Sons; 1977. 428 p.

Freedman, David; Pisani, Robert; Purves, Roger. **Statistics.** Part VI. New York: W. W. Norton & Co.; 1978. 506 p.

Sackett, David L. **Bias in analytic research.** J. Chronic Dis. 32(1/2): 51-63; 1979.

Williams, Bill. **A sampler on sampling.** New York: John Wiley & Sons; 1978. 254 p.

The Forest Service, U.S. Department of Agriculture, is responsible for Federal leadership in forestry. It carries out this role through four main activities:

- Protection and management of resources on 191 million acres of National Forest System lands.
- Cooperation with State and local governments, forest industries, and private landowners to help protect and manage non-Federal forest and associated range and watershed lands.
- Participation with other agencies in human resource and community assistance programs to improve living conditions in rural areas.
- Research on all aspects of forestry, rangeland management, and forest resources utilization.

The Pacific Southwest Forest and Range Experiment Station

- Represents the research branch of the Forest Service in California, Hawaii, and the western Pacific.

Golbeck, Amanda L. **Evaluating statistical validity of research reports: a guide for managers, planners, and researchers.** Gen. Tech. Rep. PSW-87. Berkeley, CA: Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture; 1986. 22 p.

Publication of a research report does not guarantee that its results and conclusions are statistically valid. Each statistical method serves a particular purpose and requires certain types of data. By using this report as a guide, readers of research reports can better judge whether the statistical methods were appropriate, how closely measurements represent the concept being studied, and how much confidence to place in the conclusions. Descriptions of sampling methods and of possible biases show how results can be better evaluated with respect to sampling.

Retrieval Terms: scaling of attitudes, statistical assumptions, ordinal data analysis, sampling biases