CORNELL NUTRITIONAL SURVEILLANCE PROGRAM

Working Paper Series

Division of Nutritional Sciences
New York State Colleges of Human Ecology & Agriculture and Life Sciences
Statutory Colleges of the State University of New York
Cornell University
Ithaca, New York 14853

Cornell University is an equal opportunity, affirmative action employer and educator.

ANALYTICAL METHODS FOR NUTRITIONAL AND SOCIO-ECONOMIC DATA

by

Hamid Tabatabai

April 1983 No. 4

The CNSP is supported by Cooperative Agreement #AID DSAN CA-0240 between the Office of Nutrition, Bureau for Science & Technology, USAID, and the Division of Nutritional Sciences, New York State Colleges of Human Ecology & Agriculture and Life Sciences, Cornell University, Ithaca, New York 14853.

A report of research of the Cornell University Agricultural Experiment Station

ABSTRACT

This paper covers some major techniques of analysis of nutritional and socio-economic data. The choice of techniques is primarily based on the experience gained at the Cornell Nutritional Surveillance Program where a number of data sets arising from various nutritional and socio-economic surveys have been analyzed over the past year or two. Most of the illustrations come from this on-going research. The paper is directed in the main to practitioners involved with some aspect of nutritional surveillance although others may also find some of the material useful for other purposes. A rudimentary understanding of basic statistics and some familiarity with survey data is assumed. Relevant concepts and definitions are given first and elements of data processing and presentation are reviewed. The bulk of the paper discusses basic and advanced analytical techniques. Basic techniques include categorical data analysis, correlation analysis and analysis of variance. Advanced techniques consist of regression analysis, covariance analysis, models with binary dependent variable, and some less widely used methods of classification: discriminant analysis, principal components analysis and cluster analysis.

TABLE OF CONTENTS

		Page
	Abstract	
1.	Purpose	1
2.	Concepts and Definitions	3
	2.1 Types of Data and Variables	3
	2.2 Basic and Advanced Analyses	6
	2.3 Causality, Statistical Significance and "Importance"	7
3.	Processing of Data	12
	3.1 Hardware	12
	3.2 Processing	12
4.	Presentation of Data	19
5.	Basic Analytical Techniques	25
	5.1 Hypothesis Testing	26
	5.2 Categorical Data Analysis	28
	5.3 Correlation Analysis	39
	Correlation Coefficients	39
	Significance Tests for Correlation Coefficients	45
	5.4 Analysis of Variance	50
6.	Advanced Analytical Techniques	60
	6.1 Regression Analysis	62
	Basic Elements	62
	Estimation and Interpretation	65
	Some Illustrations	67
	Violations of Assumptions	72
	6.2 Covariance Analysis	75
	6.3 Models with Binary Dependent Variable	77
	6.4 Discriminant Analysis	81
	6.5 Factor Analysis and Principal Components Analysis	88
	6.6 Cluster Analysis	92
References		100

ACKNOWLEDGEMENT

This paper is a revised version of Chapter 9 of "Nutritional Surveillance: Review of Progress" by J. Mason, J-P. Habicht, H. Tabatabai and V. Valverde (1981). Special thanks go to Dr. John Mason for many helpful discussions and for contributing some material to the earlier sections of this working paper. He bears, of course, no responsibility for any errors of omission or commission that may remain.

1. PURPOSE

Satisfaction of basic needs requires, above all, the provision of adequate nutrition for all. To assist in meeting this challenge national and local nutritional surveillance systems are being set up in a growing number of developed and developing countries alike. The major function of these systems is to provide decision-makers with useful information that helps them to improve the selection, implementation, monitoring and evaluation of policies and programs that bear, directly or indirectly, on the nutritional conditions of a population. This function can be suitably discharged only when nutritional and socio-economic data are properly collected and analyzed.

Mason et al. (1982) have dealt extensively with the need for and the usual sources of statistical information in efforts directed at nutritional surveillance. They note that a great deal of routinely collected, relevant data often exist in many developing countries, not to mention developed countries, even if the institutional structures with direct responsibility for nutritional surveillance are only beginning to take shape. While it may on occasion be necessary to commit more resources to the collection of additional data, data accumulation has often received more attention than adequate analysis. Much of the available data have either not been analyzed and used at all or only partially analyzed and used in decision-making. Indeed it may well be that excessive accumulation of data has detracted attention from analysis not only by diverting the limited resources available, but also by overwhelming those that remain. Clearly there is little point in collecting more and better data unless they are intended to be analyzed and used to either describe the existing situation or make more informed decisions. This is the ultimate justification for any data collection.

The results produced from existing nutritional surveillance systems are largely confined to tabulations of single indicators by broad groupings of population, commonly by administrative area, and these are indeed the first results that are likely to be of interest. The use of multiple indicators tabulated by socio-economic grouping is also important for many purposes in the planning context (ibid, Table III.1) These purposes essentially involve: comparison of nutritional conditions (or other outcome indicators) of different population groupings at one point in time; detecting changes in these indicators over time; describing or accounting for both these types of differences in terms of the characteristics of the groups involved, and studying the possible causes of these

differences at one time or over time. The purposes of data analysis for adequacy evaluation (ibid, Chapter IV) are to follow nutritional conditions of program target groups and/or recipients over time and compare them with planned and baseline values. For impact assessment (ibid) it is necessary to investigate associations between indicators of program delivery and changes in outcome indicators, controlling statistically for other factors that may be associated with changes in outcome. Finally, in timely warning and intervention programs (ibid, Chapter V) the purpose of data analysis is to rapidly decide whether certain variables are showing significant changes, specifically whether they are crossing the threshold or trigger values that should lead to predetermined actions.

This paper seeks to provide some guidance on data analysis in these contexts. This is possible because, as the experience in data analysis at the CNSP indicates, there are in fact a limited number of widely applicable techniques that will cover most needs. Because of the important emphasis given to bringing out the meaning of data for decisions on policies and programs, the potential analytical techniques needed for nutritional surveillance require somewhat more systematic application than has been the case in analyses of nutritional and related data in the past. This is all the more feasible now with the wide availability of computer facilities and, more importantly, powerful programs that can routinely do analyses undreamt of (by the layman anyway) only a few years ago. While this paper covers a good many useful techniques no attempt is made to be comprehensive. Frequent references to standard texts are provided throughout.

This paper is directed to practitioners in the nutrition area who have had some prior training in statistical methods, albeit in the distant past, some experience of working with nutrition-related data and a desire to go beyond using only simple descriptive tools for presentation of data. The analytical techniques in this paper become increasingly more sophisticated but an attempt has been made to present their essential elements in simple terms and with frequent illustration. Simplification, however, may have at times been achieved at the expense of rigor. Hence, for confident application of these techniques it is advisable to consult the principal references given.

The rest of this paper consists of five sections. Section 2 presents concepts and definitions. Processing of data and their presentation will be briefly discussed in Sections 3 and 4. The last two sections constitute the core of the paper where basic and advanced analytical techniques are discussed.

2. CONCEPTS AND DEFINITIONS

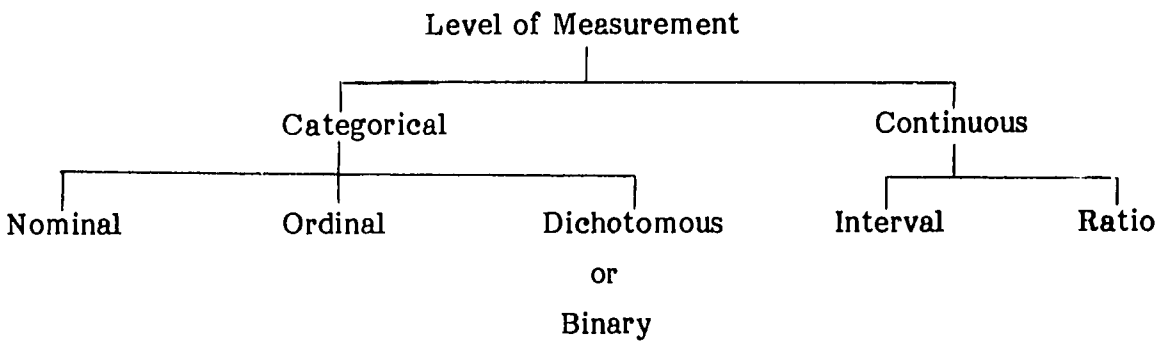
2.1 Types of Data and Variables

Types of data can be distinguished by their level of measurement, i.e. by the amount of information conveyed by their numerical values (Nie et al. 1975). The weakest or lowest level of measurement belongs to **nominal** data. The values of a nominal variable merely classify its different categories. The variable "sex", for example, takes on two values, say 0 and 1, depending on whether a subject is male or female. Other examples include the presence or absence of a certain disease or the type of water source where numerical values may refer to "well", "stream", "piped", etc. It is not possible to interpret these numerical values as reflecting the relative size or worth of the relevant categories.

Ordinal data provide more information by ranking or ordering the categories with respect to the degree to which the characteristic being measured by the variable is possessed. Grouping by attitude (strongly disagree, ..., strongly agree) and by income brackets are two examples of ordinal measurement. A third example is when it is possible to ascribe a ranking to otherwise nominal data — e.g. when housing can be ranked bad, medium or good; or when it is valid to postulate that a piped water source is better than a well, which in turn is better than a stream. This latter example converts nominal to ordinal data. **Categorical** level of measurement encompasses nominal data (unordered categories), ordinal data (ordered categories) and **dichotomous** or **binary** data (exactly two categories).

The third type of data is **interval.** Interval data additionally provide a measure of distance between categories. In contrast to the former two types of data which were discrete, interval data are continuous and are measured in terms of some well-defined and recognizable physical unit; height, weight, distance and income are a few of the more common examples of intervally - measured characteristics. Along with ratio data — which we do not need to discuss separately here - - interval data are also referred to as **continuous.**

A summary of classification of levels of measurement is as follows:



A higher level of measurement can always be reduced to a lower level at will. Continuous data, for example, can be made ordinal by the simple device of creating class intervals or brackets as is commonly done when households are grouped into various income brackets. By ignoring their ranks or order, ordinal data too can be reduced to become nominal. The reverse can happen as well. Nominal data can be made ordinal if the categories can be ordered on the basis of the degree to which a particular characteristic is present or absent. Ordinal data are also sometimes treated as, or converted to, continuous data in an ad hoc fashion. For instance, distance of villages to nearest road may have been given in four categories as:

Category	Walking distance to nearest road in minutes
1	0 - 50
2	50 - 100
3	100 - 200
4	> 200

One method of transforming this distance variable into a continuous variable is to assign the mid-point of each interval to all the villages in the category to which the interval pertains. All villages in the third category for example, may be assumed to be 150 minutes walking distance from the nearest road. Villages in the last category may be assigned a distance of, say, 240 minutes arbitrarily. A great deal of care must however be exercised when the level of measurement is artifically increased.

It may be useful to briefly define a few additional terms as well. A **variable** can assume any one of a number of possible values. If the values it assumes depend, at least in part, on those of one or more other variables, i.e if it

is "described" by other variables, it will be known as a **dependent** variable. If however it does not depend on other variables and is used instead to describe them, it will be known as an **independent** variable. Two or more variables that depend on one another simultaneously are referred to as **interdependent** variables. A **random variable**, also known as a **variate**, takes on different values with given probabilities that must sum to one. It may be **discrete** (when it assumes a finite or countable number of numerical values) or **continuous** (when it may assume any numerical value on a continuous scale). The **probability distribution** of a random variable provides a probability for each possible value (in the discrete case) and a probability that its value lies in any given range (in the continous case).

Since data from a particular sample survey will be extensively used in the examples and illustrations in this paper, it may be useful to provide a brief outline of its contents. The survey covered about 500 households but our interest focuses on 261 households with children 12-84 months of age, for whom data were complete. There were 428 such children belonging to 36 area clusters (villages) chosen in the first of a two-stage sampling process with probability proportional to the size of the cluster, i.e. to the number of households within each cluster. The entire population of the survey area consists of 20,016 households grouped into 694 area clusters. These clusters were ordered by altitude and the sample clusters were chosen by a systematic sampling design. The selection of sample households in the second stage was carried out with probabilities inversely proportional to those with which their corresponding clusters were chosen in the first stage. The upshot of this design is that every household in the population would have approximately the same probability of being included in the final sample. This is known as a **self-weighting design.** The number of children within the sampled clusters range from 2 to 25. For each child weight-for-age (WA), height-for-age (HA), and weight-for-height (WH) can be calculated using Harvard Standards (Hitchings/CBS 1979). Using these figures the mean values of these variables and their standard deviations can be found for each cluster and for the entire group of children in the sample. For the entire sample of 428 children these figures, in percent, are respectively 80.3, 90.3 and 94.0 for the means and 11.9, 6.6 and 8.5 for the standard deviations. These values were calculated <u>assuming</u> that the observations had been generated by simple random selection. Whether or not this assumption is justified will be considered later in Section 5.4. A partial list of the variables from this data set

is given in Section 3.2, Table 3.

2.2 Basic and Advanced Analyses

Data analysis is the process of applying statistical techniques to data in order to answer questions in which the analyst may be interested. Its purpose is to extract information which helps understanding, description, interpretation and decision making. The type of analyses or techniques to use depends on the questions asked, data available, resources at hand, time, assumptions made, level of confidence and precision required, etc. Our emphasis here is on analytical techniques that can profitably be applied to nutritional and socio-economic data. Analysis of such data is needed, for nutritional surveillance, to lead to decisions. In essence, this comes down to looking at two related aspects of data: differences, and correlations or associations. We are interested in differences among population groups and overtime; in characteristics of these groups and in factors associated with their differences. Conceptually differences and correlations are similar. A difference in nutritional status between groups defined by landholding size implies also a correlation or association between nutritional status and landholding. But a similar difference across areas, although still implying an association between nutritional status and area, is commonly studied as a difference rather than as a correlation since area is a nominal variable and correlations ordinarily require data measured at least ordinally. The basic analyses described here have the objective of studying differences and correlations whereas more advanced techniques deal with these at a more sophisticated level as well as with other questions such as modeling, data reduction, prediction, etc.

Basic analytical techniques are fairly simple and tend to be rather regularly employed. Advanced analytical techniques on the other hand are less well-known in nutrition-related literature and are thus less frequently used. By briefly discussing them in this paper and illustrating some of their applications it is hoped that they will be employed more regularly in the future. By encouraging the more widespread application of the advanced techniques we in no way wish to detract attention from the regular use of the basic techniques, particularly when time and expertise are in short supply. Indeed, it is worth emphasizing that prior use of the basic methods is often essential before proceeding to the more advanced analyses. A brief guide to the methods discussed in this paper is given in Table 1.

2.3 Causality, Statistical Significance and "Importance"

In making policy and evaluating programs it is essential to try to have as thorough an understanding as possible of the nature of interrelationships that link various factors or variables together. Data analysis serves this purpose by pursuing one or more of the following distinct, but related, objectives: it may try to establish the structure of causality within a given set of variables; it may seek to measure the strength of a relationship in alternative ways; and it may try to single out the more "important" variables for purposes of not only analysis but also of survey design. Conclusions on all these points can be derived from both basic and advanced techniques of analysis. Some elaboration is in order.

The statistical techniques of empirical research are inherently incapable of establishing a causal connection between two or more variables. They may suggest one, but the ultimate proof, if there exists one at all, could only come from some kind of theoretical construct. Short of that, experimental design techniques may provide the most convincing evidence of a possible structure of causality. Such causal structures are however often hypothesized and the object of empirical investigation then becomes to discover whether the available information corroborates the hypothesized causality or not. Indeed, in a strict sense, all that a statistical analysis demonstrates is whether the available evidence contradicts or rejects an a priori hypothesis with a given level of confidence. It is important to bear this fact in mind since all too often a strong correlation or the statistical significance of a relationship is presumed to offer evidence for the existence of a cause-and-effect relationship on the basis of which a particular program is advocated and justified. The reverse situation occurs just as often. Such inferences may be very misleading.

Assuming that the structure of causality among some variables has been satisfactorily established, or that the hypothesis of its existence has been statistically verified, it is useful to try to measure the relative strengths of such mutual interrelationships. Such measurements can be arrived at in a variety of ways. Correlation coefficients can be found and the variation in the "dependent" variable(s) "explained" or "accounted for" by the variation in the "independent" variable(s). These relationships, once tested and verified, can be employed for a variety of purposes including making predictions regarding changes in the dependent variable(s) assumed to result from changes in the independent variable(s).

TABLE 1

BRIEF GUIDE TO SELECTED ANALYTICAL METHODS

Name of Method	Dependent Variable Classification	Independent Variable Classification	General Purpose	Section
Basic				5
Categorical data analysis	Nominal	Mostly nominal, but sometimes ordinal	To describe the relationship between a nominal dependent variable and several nominal or ordinal independent variables, although applications to situations involving only dependent variables are possible	5.2
Correlation analysis	(The variables can be continuous or ordinal.)		To measure the degree of association between pains of continuous or ordinal variables	5.3
Analysis of variance	Continuous	All nominal	To describe the relationship between a continuous dependent variable and one or more nominal independent variables	5.4
Advanced				6
Multiple regression analysis	Continuous	Classically all continuous, but, practically, any type(s) can be used	To describe the extent, direction, and and strength of the relationship between several independent variables and a continuous dependent variable	6.1
Analysis of covariance	Continuous	Mixture of at least one continuous and at least one categorical variable	To describe the extent, direction and the relationship between a mixture of continuous and categorical independent variables and a continuous dependent variable	6.2
Logit analysis	Categorical (often dichotomous)	Any type(s), usually a mixture of continuous and categorical variables	To describe the extent, direction, and strength of the relationship between one or more independent variable(s) and a categorical dependent variable	6.3

TABLE 1 (Cont'd)

BRIEF GUIDE TO SELECTED ANALYTICAL METHODS

Name of Method	Dependent Variable Classification	Independent Variable Classification	General Purpose	Section
Discriminant Analysis	Nominal	Classically all continuous but practically can have a mixture of various types as long as some are continuous	To determine how one or more independent variables can be used to discriminate among different categories of a nominal dependent variable	6.4
Principal Component Analysis and Factor Analysis	(The variables used in these two techniques are classically continuous, but, practically, may be of any type. These variables are not clearly identifiable as being either dependent or independent, although the resulting factors may be used as dependent or independent variables in a later analysis.)		Variable construction or reduction using several variables to define one or more new composite variables called principal components or factors	6.5
Cluster Analysis	(The variables may be of any kind.)		To identify groups or "clusters" of entities having an interesting feature on the basis of observations on their characteristics	6.6

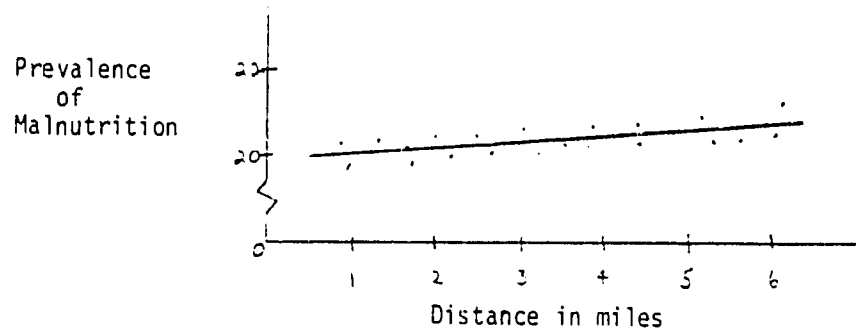
Source: Based in part on Kleinbaum and Kupper 1978, Table 2.1

Finally it is sometimes necessary to decide on the relative "importance" of variables. This is of extreme practical value. It is not uncommon for an investigator to be swamped with an overwhelming mass of statistical information. Indeed there are many examples, in the developing countries and elsewhere, of all kinds of expensively collected survey and non-survey data that are, because of their sheer volume, never analyzed. Such masses of data would first have to be reduced to manageable proportions in meaningful ways. The essential elements, or the important variables, should of course be retained. However these have to be identified first. Data reduction techniques can sometimes be helpful in deciding what is important and what is not. Such knowledge is invaluable in survey design if the available resources are to be employed optimally.

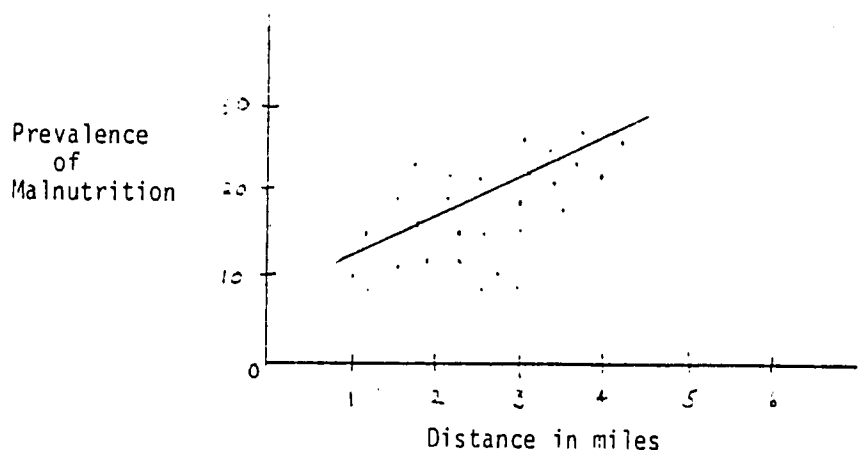
It should be noted that tests of "significance" in statistics, while often helpful, may not always be able to decide on the "importance" of a variable. In other words statistical significance should not be confused with the practical significance of a relationship. An example may be useful. It might be found that a highly significant correlation existed (in the sense that there was a high degree of confidence that the relationship was not due to chance) between, say, prevalence of malnutrition in different villages and the distance of a village to its water supply. This, however, would be of little practical importance if the prevalence rate changed little across villages. In Figure 1 note that the relationship between the two variables in the top panel is almost perfectly linear resulting in a high correlation coefficient. On the other hand the observations in the bottom panel are far more scattered and the relevant correlation coefficient is much smaller. Nevertheless a sizable upward trend is clearly evident. The rate of prevalence does appear to move up with increasing distance to water supply even though the correlation coefficient might be fairly small. Statistical significance is a strong function of, among other things, the number of observations (sample size) whereas the importance of a relationship is not; it has to do with the practical value of the information to increase understanding and improve decision making. An important relationship need not be statistically significant; neither is a statistically significant relationship necessarily important.

FIGURE 1

Significance versus Importance



Highly significant but not very important.



Less significant but more important.

3. PROCESSING OF DATA[1]

3.1 Hardware

Most processing of data, and hence application of analytical techniques, is presently done using computers. This may not always be the best solution, particularly (and paradoxically) if speed is of crucial importance. In this case simple hand-tallying and calculation may be more reliable and controllable. Moreover, too much reliance on computer processing can inhibit understanding of results. However, application of statistical techniques is almost always done with some electronic assistance - at a minimum from a programmable calculator. This section deals with some considerations when undertaking or planning analyses by computer.

The first consideration is, obviously, what computer set-up to use. For nutritional (or any social data of this type) the major constraint at present is suitable programs (computer equipment is usually known as **"hardware"**, programs as **"software"**). The availability of powerful package programs for use on large computers, with improved access to these installations through time-sharing, has made a lot of difference to the potential for data analysis. There are advantages, in terms of access, in considering smaller installations, but these lack suitable programs at present. It is urged that decisions on computer installations take program availability into account as a crucial determinant, as well as organizational requirements. In Table 2 we give a brief summary of the main choices available, with potential advantages and disadvantages of each.

3.2 Processing

The processing of data by means of a computer can be broken into four main steps: coding, data entry, editing and statistical analysis. While the use of statistical packages in the analysis may require special skills, it is generally the least time consuming of these main steps, often taking up less than a fifth of the total time. The total time will vary of course with the amount and complexity of the data and the analyses required.

Coding. The data are usually entered in the form of a matrix where the rows refer to the observations or cases, e.g. households, and the columns to the variables, e.g. nutritional indicators or economic status variables. It is easiest to begin by preparing a **coding manual** which lists the descriptions of variables,

---

[1] A portion of this section was first drafted by Cay Loria.

their names, the columns in which each variable can be found, format, the values which each variable can assume and the code for missing values. An example of portions of a simple coding manual is given in Table 3. If the design is relatively simple, much time can be saved by setting up the questionnaires so that the data can be entered into the computer directly from them.

It is also important to conceive the structure of the sort of data we are concerned with. This is easiest with respect to sample survey data, but actually applies to any type, even if collected from routine sources. Essentially, we have a number of variables collected per sample unit (or element) - household, child, etc. For each individual unit this can be conceived as a string of numbers, e.g.:

Identification			Measurements/Variables	
Region No. 6	Household No. 46	Child No. 2	Household size 6	Child's Weight 12.3kg

For many units, this would logically be arranged in the familiar form of a table or matrix:

Identification			Variables	
Region	Household	Child	Household size	Child's weight
6	46	2	6	12.3 ...
6	47	1	8	15.2 ...
6	47	2	8	9.3 ...

Note in passing that when this is computerized, the data are entered in one row, even if conceptually they remain as a series of rows. The section of the computer tape with these data, when printed out, might read:

/06046206123... /06047108152... /06047208093...

Data entry. The method of data entry depends on the available hardware. Keypunching cards has traditionally been the main technique; however, this mode is quickly becoming outdated. Typing the data directly into a computer file on a typewriter-like terminal has the advantage of allowing the input data to be seen on a screen or printer, thereby making it easier to catch any obvious errors. Such interactive data entry can be accomplished by means of a data entry program (or **editor**). Each operating system usually has an editor through which files can be created and changed. A program written specifically for entering

TABLE 2

Computers

System	Example	Characteristics	Advantages	Disadvantages
1. Mainframe Computer	IBM, etc. - main computer installations.	Large Capacity, speed. Expensive. Requires substantial organization. Time sharing usually available.	1. Ability to handle large amounts of data. 2. Can use powerful statistical programs, which are widely available. 3. Organization usually includes service for data entry from coded sheets. 4. Service back-up usually good.	1. User has little control over data handling. 2. Delays experienced in access to facility. 3. Skills needed to use programs; more so to program directly.
2. Minicomputers	WANG, HEWLETT-PACKARD, etc.	Cost range about $20,000 to $100,000. Capabilities approaching those of mainframe computers. Unavailability of software a chronic constraint.	1. Can be de-centralized, therefore easier access and control. 2. Capacity approaching that of mainframe computers, can handle large amounts of data. 3. Often compatible easily with mainframe computers. 4. Service back-up reasonable from major companies.	1. Virtually no suitable programs yet available for social data, therefore custom programming needed. 2. Still require organization, especially for data entry; established organization cannot always be used (except where compatibility with mainframe computer can be exploited). 3. Expense still substantial, so that a certain volume of use needed to justify. 4. Service back-up not always well-developed.

3.	Microcomputers	APPLE, PET, H-P 80, etc.	Cost range $5,000 to $20,000. Mass produced. Fairly portable. Virtually no software available. Capacity less than minicomputers, but increasing with present rapid technological development.	1. Cheap enough to be feasible for small unit to own 2. Control over installation and access. 3. Suitable for small sets of data with simple analyses. 4. Suitable for de-centralized use.	1. Virtually no suitable programs available. 2. Limited capacity so that analysis slow; this also may require more complex programming. 3. Not easily compatible with larger systems. 4. Since multiple terminals not used, data entry, which can only be direct, is slow and ties up machine for long periods. 5. Still experimental to some extent hence questionable reliability; also quickly obsolete. 6. Doubtful service back-up.
4.	Programmable Calculator	H-P, Texas Instruments, Casio, etc., etc.	Desk-top, portable, cheap. No capacity for data storage (or very limited). Many programs available.	Flexible, total control by user.	Unsuitable for handling any volume of data since data needs to be re-entered each time. Some recent models allow very limited storage.

TABLE 3

AN EXAMPLE OF PORTIONS OF A SIMPLE CODING MANUAL

Description of Variable	Variable Name	Columns	Format	Value Labels
CARD 1				
Identification				
FAO1	—	1-4	-	FAO1 (All codes)
Area number	AREA	5-6	2.0	1-19
Cluster number	CLUSTER	7-8	2.0	1-40
Household number	HOUSE	9-11	3.0	5-996
Area Characteristics				
Ecozone	ECOZONE	12	1.0	1: Plain; 2: Mountain
Distance to road category	ROADCAT	13	1.0	1: 0 - 50 mins. 2: 50 - 100 mins. 3: 100 - 200 mins. 4: > 200 mins.
Distance to road-absolute	ROADABS	14-16	3.0	10-240 mins.
	. . .			
Income characteristics				
Main occupation	MAINOCCI	27	1.0	1: Farmer 2: Other (does not exist) 3: Livestock (does not exist) 4: Cash employment 5: Self-employed 6: Other
	. . .			
Farmer data				
Landholding (area cultivated)-category	CULT	31	1.0	1: <.1 ha. 2: .1- .2 ha. 3: .2 - 1.0 ha. 4: 1.0 - 5.0 ha. 5: > 5.0 ha.
	. . .			

TABLE 3

AN EXAMPLE OF PORTIONS OF A SIMPLE CODING MANUAL

Description of Variable		Variable Name	Columns	Format	Value Labels
CARD 2					
Identification					
Same as Card 1, columns 1-11					
Family data					
Kid 2					
Birthdate	- day	DAY 2	12-13	2.0	Should be born
	- month	MONTH 2	14-15	2.0	before 03-03-80
	- year	YEAR 2	15-16	2.0	
Weight		WGT2	18-20	3.1	2.0 - 25.0 kg.
		. . .			
		. . .			
		. . .			
Edema		EDEMA 2	27	1.0	0: No; 1: Yes
		. . .			
		. . .			
		. . .			

data can prompt the user for each specific data item, ensuring it is entered in the correct format and position. These programs can also be made to check whether the value entered is within the correct range. Data entry programs are the most efficient especially when the data are complex; however, they are not widely available and must often be custom written. SPSS (Statistical Package for the Social Sciences) Conversational Statistical System (SCSS), the interactive version of SPSS, does have a prompting data entry system which can simultaneously edit the data.

Editing. Data editing or cleaning, is a combination of techniques which check for out-of-range values, inconsistencies, and structural errors (e.g. missing records). Inconsistencies are logical impossibilities for a specific variable which can be checked on the basis of another variable. For example, if the value for the variable sex is male, the number of miscarriages cannot be greater than 0. Once these kinds of errors are located, they can be corrected with an editor. The importance of cleaning is often overlooked; if insufficient time is devoted to cleaning, a disproportionate amount of time will be spent on determining the cause of the errors in the output that originate from errors in the data.

A few editing packages exist, such as UNEDIT, solely for purposes of cleaning data. UNEDIT is easy to use and will run on most small computers (Francis, 1979). Most general-purpose statistical packages can also be used to clean data. The ranges of variables can be checked by printing the frequencies and thus the values for each variable (e.g. PROC FREQ in SAS (Statistical Analysis System), general mode of the FREQUENCIES subprogram in SPSS and the HISTOGRAM command in MINITAB). With all of these packages, the values of the identification variable(s) for cases with errors can then be printed out. Inconsistencies can be found by using logical operators (=, $>$, $<$, $\neq$) and can be printed out in a number of ways, depending upon the package. For instance, in SAS one command to find the errors for the example used above would be: IF SEX = MALE AND MISCAR $>$ 0 THEN LIST; where MISCAR is the variable name for the number of miscarriages. In this case, the input line for any records containing these specific conditions would be printed. Some packages (e.g. PSTAT and SAS) can also check for missing records.

Statistical Packages. There are many statistical packages available to choose from (for a detailed review see Francis, 1979). These packages run mostly on mainframe computers. The three main kinds of packages are: general-

purpose ones; those with restricted applications (i.e. they perform only a few statistical techniques); and those which tabulate data and calculate a limited number of descriptive statistics. Choosing between them is largely dependent upon specific needs. The capabilities of each package, i.e. the level and quantity of statistical analysis and the size of data set which each is designed to handle, are of primary importance. Other major considerations are the ease of learning and using them, their reliability and ease of maintenance, and their portability, i.e. space and storage requirements (Francis, 1979).

Some examples of large-scale, general-purpose packages are PSTAT, SAS and SPSS. All of these perform a wide array of statistical techniques including multivariate and nonparametric techniques. These packages can easily manipulate files, especially those which are hierarchically structured (e.g. organized by households within villages within districts within states). MINITAB is an example of a smaller-scale version of general-purpose packages. It is designed for use with small to medium-sized data sets and although its limitations depend upon which version is being used, the maximum number of variables is usually one hundred. Also, with the exception of multiple regression and analysis of covariance, MINITAB will not perform multivariate techniques. MINITAB is easier to learn than some of the larger packages and can be installed on many small systems. It is presently suitable for minicomputers if these are programmable in FORTRAN or a FORTRAN complier is available. Tabulation packages, such as CENTS-AID and COCENTS, are generally limited to producing crosstabulation tables and simple descriptive statistics such as the mean or standard deviation. However, both packages have the advantages of being designed for use with census and household data, are very easy to use, and run on small computers.

4. PRESENTATION OF DATA

Just like beauty, the contents of a data set must be cleverly and clearly revealed to be appreciated. The revelation comes with the presentation of data, i.e. with how the informational content of data is conveyed clearly and effectively. Powerful presentation should also enable the reader to understand the data quickly and as easily as possible. This also implies that to the extent possible a particular presentation should be self-contained so that frequent references to the accompanying text would not be needed.

How the data are presented is not a particularly demanding task but a small dose of artistic talent can be quite valuable. Methods of presentation are varied and are largely governed by four characteristics of the data: the number of variables to be included together, their levels of measurement, the number of observations, and finally the aspect(s) of data that need to be highlighted.

A typical survey, for example, consists of a relatively large number of characteristics observed on a far larger number of elements, say households. Because of the number of variables involved and the variety of permissible household classifications it is possible to create an extremely large number of tables, correlation coefficients, etc. To actually do this is of course totally inappropriate even if possible. One can easily get lost in the resulting massive output and sensible interpretation would become all but impossible. It is necessary to theorize or to hypothesize on the basis of one's expectations, preferably even before the survey itself is conducted and data collected. While the prior knowledge of the relationships is of course very limited in most circumstances, it should always be possible to concentrate only on the more important relationships. Fishing expeditions are not recommended but selective choice of a relatively few relationships is both necessary and usually sufficient, particularly when they are being reported.

Table 4 shows three ways of reporting the number of households in four land categories and is based on our sample survey data. The first panel of the table records the frequency distribution in a one-way "table". Graphical demonstration in panels (B) and (C), in the forms of a **line diagram** and a **bar chart**, are designed to make visual impact and tend to stress the relative frequencies rather than their absolute values as in the first panel. These are some of the simplest kinds of data presentation that are quite common for single categorical or ordinal variables. Indeed, as in this example, the variable can be of interval type as long as class intervals are introduced to transform it into a categorical or ordinal variable. This appears to be the practical approach that is widely used in nutrition related literature as well as in many other fields. The number of classes may be as large as 20 or 30 or larger even though the presentation may increasingly become cluttered and awkward. The way out of this problem is to redefine the class intervals (or categories) to make them larger and fewer.

TABLE 4

ALTERNATIVE METHODS OF REPRESENTING THE DISTRIBUTION OF HOUSEHOLDS BY LANDHOLDING CATEGORY

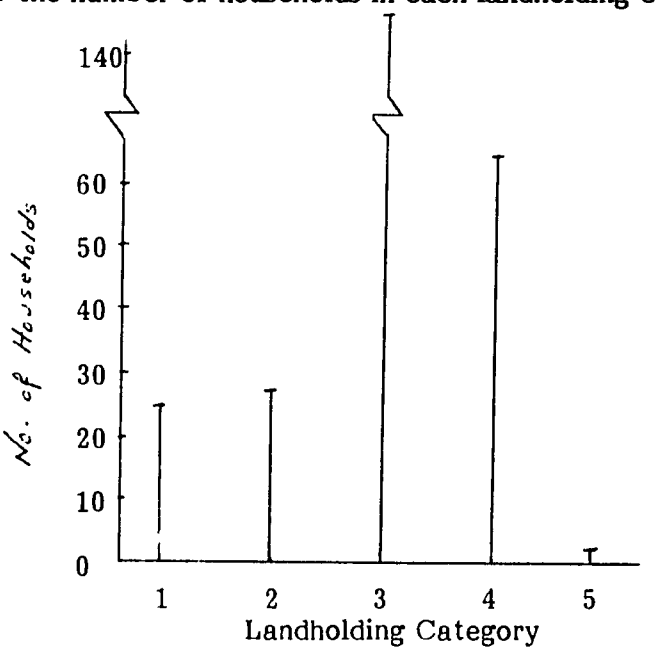
(A) Number and Proportion of Households by Landholding Category.

	Landholding Category					
	1	2	3	4	5	Total
No. of Households	25	27	145	62	2	261
Proportion of Households	.096	.103	.556	.237	.008	1.000

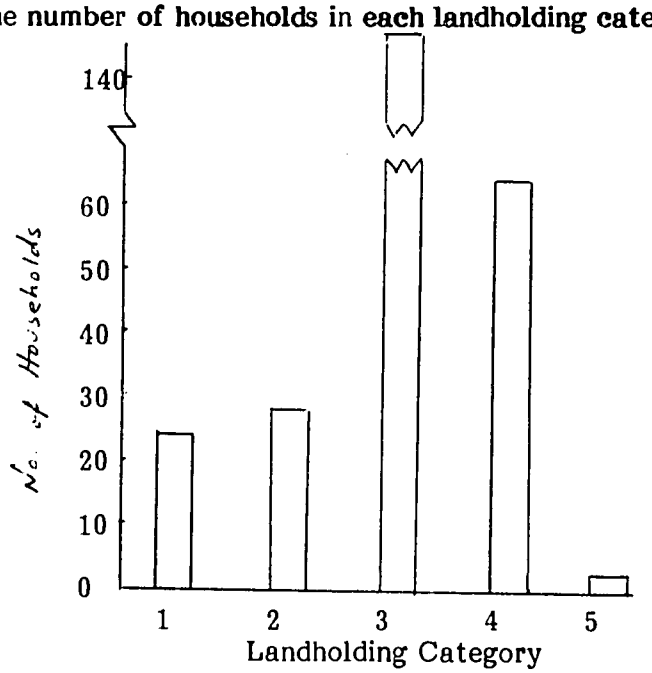
[a]1 Carreau = 1.29 hectares

Landholding Category	Size of Holding
1	<0.1 Carreau (Cx)[a]
2	.1 - .2 Cx
3	.2 - 1.0 Cx
4	1.0 - 5.0 Cx
5	≥5.0 Cx

(B) Line Diagram for the number of households in each landholding category.



(C) Bar Chart for the number of households in each landholding category.



A variation of the bar chart technique arises when a researcher is interested in the relative frequency of households in each class interval defined for an initially continuous variable. The bars can be placed adjacent to one another and the result is known as a **histogram.** The class intervals are represented by the widths of the individual bars and need not be equal. One advantage of presenting data in the form of a histogram is that it provides some understanding of the shape of the distribution of the variable thereby enabling the researcher to decide whether or not the distributional requirements of an analytical technique he may wish to employ are adequately satisfied. (For a formal way of testing this requirement see Section 5.2). There are other insights that may be gained in some situations. The histogram of age distribution in Table 5, for example, suggests that the "local peaks" at 6, 12, 18 and 24 months of age may be due to rounding of age by respondents. If the implications of this rounding are deemed to be serious it may be preferable to define class intervals of 2 or 3 months to remove the possible errors due to rounding. Another use of a histogram is in cleaning data (see Section 3.2) where peculiarities in the distribution of variables (such as the existence of **outliers** or observations which are very much out of the ordinary, genuinely or otherwise) may be traced to various errors in the earlier stages of data processing.

Data presentation of course need not be limited to a single variable at a time. Cross-tabulations, or contingency tables in the parlance of statisticians, are well-known and widely used. A two-way cross-tabulation, for instance, represents the joint frequency distribution of entities or cases classified according to two of their attributes. Each cell of such a table gives the frequency of observations having a particular combination of attributes. n-way cross-tabulations are possible even if they soon get out of hand with n above 3 or 4. Many computer packages are available (see Section 3.2) that provide such outputs cheaply, accurately, rapidly and neatly. An illustration of a 2-way table produced by SAS is given in Table 6 and is based on our sample survey. It gives the frequency of well-nourished (0) and malnourished (1) children in the sample by area along with three derived statistics. The four numbers in each cell are respectively: the number (frequency) of the children, its percentage in total number of children, percentage of total children in each row and percentage of total children in each column. The numbers in the margins give row and column totals as well as percentages relative to the total number of children.

TABLE 5

HISTOGRAM AGE DISTRIBUTION—CARE-PAKISTAN

MIDPOINT	HIST%	COUNT	FOR 3.AGE (EACH X= 1)
3.0000	.1	1	+X
9.0000	.2	2	+XX
10.000	1.0	13	+XXXXXXXXXXXXX
11.000	1.5	13	+XXXXXXXXXXXXXXXXX
12.000	3.9	48	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
13.000	2.0	25	+XXXXXXXXXXXXXXXXXXXXXXXXX
14.000	1.1	14	+XXXXXXXXXXXXXX
15.000	1.6	20	+XXXXXXXXXXXXXXXXXXXX
16.000	2.0	25	+XXXXXXXXXXXXXXXXXXXXXXXXX
17.000	1.2	15	+XXXXXXXXXXXXXXX
18.000	3.5	43	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
19.000	2.5	31	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
20.000	1.5	19	+XXXXXXXXXXXXXXXXXXX
21.000	2.3	28	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
22.000	3.3	41	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
23.000	2.9	36	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
24.000	3.7	46	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
25.000	1.4	17	+XXXXXXXXXXXXXXXXX
26.000	1.3	16	+XXXXXXXXXXXXXXXX
27.000	1.3	16	+XXXXXXXXXXXXXXXX
28.000	1.9	24	+XXXXXXXXXXXXXXXXXXXXXXXX
29.000	2.3	29	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
30.000	2.3	28	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
31.000	1.9	24	+XXXXXXXXXXXXXXXXXXXXXXXX
32.000	2.3	28	+XXXXXXXXXXXXXXXXXXXXXXXXXXXX
33.000	3.7	46	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
34.000	2.4	30	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
35.000	3.3	41	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
36.000	3.1	39	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
37.000	1.6	20	+XXXXXXXXXXXXXXXXXXXX
38.000	1.5	19	+XXXXXXXXXXXXXXXXXXX
39.000	1.5	19	+XXXXXXXXXXXXXXXXXXX
40.000	1.7	21	+XXXXXXXXXXXXXXXXXXXXX
41.000	2.3	28	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
42.000	1.7	21	+XXXXXXXXXXXXXXXXXXXXX
43.000	2.3	29	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
44.000	1.2	15	+XXXXXXXXXXXXXXX
45.000	2.5	31	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
46.000	2.6	32	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
47.000	3.0	37	+XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
48.000	1.9	23	+XXXXXXXXXXXXXXXXXXXXXXX
49.000	1.4	17	+XXXXXXXXXXXXXXXXX
50.000	1.7	21	+XXXXXXXXXXXXXXXXXXXXX
51.000	.8	10	+XXXXXXXXXX
52.000	.9	11	+XXXXXXXXXXX
53.000	1.4	17	+XXXXXXXXXXXXXXXXX
54.000	1.1	14	+XXXXXXXXXXXXXX
55.000	1.9	23	+XXXXXXXXXXXXXXXXXXXXXXX
56.000	1.7	21	+XXXXXXXXXXXXXXXXXXXXX
57.000	1.0	12	+XXXXXXXXXXXX
58.000	.7	9	+XXXXXXXXX
59.000	.2	2	+XX
60.000	.4	5	+XXXXX

(Drake et al. 1980, p. 101)

TABLE 6

DISTRIBUTION OF CHILDREN BY NUTRITIONAL STATUS AND AREA

	AREAS						
	1	2	3	4	5	6	TOTAL
Well nourished	36 8.41 17.22 45.57	34 7.94 16.27 42.50	34 7.94 16.27 72.34	51 11.92 24.40 48.11	24 5.61 11.48 44.44	30 7.01 14.35 48.39	209 48.83
Malnourished	43 10.05 19.63 54.43	46 10.75 21.00 57.50	13 3.04 5.94 27.66	55 12.85 25.11 51.89	30 7.01 13.70 55.56	32 7.48 14.61 51.61	219 51.17
TOTAL	79 18.46	80 18.69	47 10.98	106 24.77	54 12.62	62 14.49	428 100.00

Numbers in each cell, top to bottom, refer to:

number of children
percentage in total number of children
percentage in all well nourished children
percentage of all malnourished children

The clarity of presentation is of critical importance. A presentation that is difficult to comprehend creates alienation on the part of the reader. To minimize this possibility it is advisable that it be reasonably self-contained. To the extent possible a presentation should tell its own story without requiring frequent references to the accompanying text. It should also be precise enough to remove any possibility of misinterpretation.

Finally a word of caution is in order here: whereas our purpose in this section has been to stress the importance of the way in which data are presented, we should hasten to add that this aspect can easily be overemphasized. The truth of the matter is that how data are presented is never as important as the data being presented and their reliability. Elaborate presentation of unimportant and/or unreliable data may sometimes suggest the contrary. This possibility should always be guarded against.

5. BASIC ANALYTICAL TECHNIQUES

The earlier sections of this paper have dealt with some preliminary issues in data analysis, preparation of data and some aspects of their presentation. This section and the next take up data analysis proper. The term data analysis can refer to nothing more than mere inspection of raw data resulting in a few (often tentative) conclusions, on up to the application of the most sophisticated analytical techniques requiring great investments in time, equipment and expertise. The common thread that runs through the various techniques of this section and the next is an attempt to draw out the relationships that may exist between two or among a set of variables. Some of these techniques are relatively simple, widely known and quite powerful. Computing facilities for them are widely available and they can be profitably applied in varied circumstances. They are furthermore not very time consuming. We are referring to these techniques as basic to stress their importance and encourage their more widespread application. This section discusses basic analytical techniques. More advanced techniques will be taken up later. In Section 5.1 some elements of hypothesis testing will be briefly reviewed. Analysis of categorized data will then follow and the last three sections will deal with correlation analysis and analysis of variance. In most cases these analyses involve only two variables simultaneously.

5.1 Hypothesis Testing

The subject of **statistics,** according to Bhattacharyya and Johnson, "is comprised of the art and science of the collection, interpretation and analysis of data and the ability to draw logical generalities that relate to the phenomenon under investigation," (1977, p. 4). Statistical techniques can be of two kinds: **descriptive and inferential.** The primary purpose of descriptive statistics is to summarize the informational content of data in a way that attention can be focused on the essential features rather than on unimportant details. Summary measures of data such as the mean, median, mode, standard deviation, correlation coefficient, etc. are used in descriptive statistics to reduce unmanageable raw data to manageable dimensions. Descriptive techniques are not limited to such simple procedures and can include elaborate multivariate methods of data reduction and data classification. Some of these techniques will be discussed in the section on advanced analytical techniques.

The more important function of statistics however is to use methods of **statistical inference** to arrive at some generalized conclusions inferred about a population from the knowledge of data gathered by a sample. Inferences about one or more populations involve **estimation** and **hypothesis testing.** The estimation process is concerned with setting up the appropriate rules, or formulae, with the help of which estimates of population **parameters** are made in such a way that they would possess certain desirable properties. The term "parameter" refers to the actual value of a characteristic of the population which is to be estimated on the basis of sample data. The estimates can be either **point estimates** (i.e. a single value, e.g. a mean prevalence of 25%) or **interval estimates** (e.g. a prevalence of 20-30%) with an associated **level of confidence** (e.g. 95%).

Hypothesis testing is the more dynamic form of statistical inference and is concerned with procedures for deciding whether or not one can reject an assertion about a population on the basis of sample data. An assertion about a population is termed a **statistical hypothesis.** The assertion may be true or false and hence there are two complementary hypotheses:

H_0: the assertion is false;

H_1: the assertion is true.

The first hypothesis is usually called the **null hypothesis** and the second the **alternative hypothesis.** It is always assumed that the null hypothesis is true, i.e.

the assertion about the population is false, unless there is adequate evidence to the contrary. This is analogous to the null hypothesis that an accused person is "not guilty" (the assertion being that he is) until and unless substantial evidence to the contrary can be found (Bhattacharyya and Johnson 1977, p. 168). It is considered a more serious error to incorrectly reject the null hypothesis (e.g. to conclude that there is a difference in, say, prevalence rates when in reality there is none or that someone is guilty when in fact he is not) than to fail to reject it when it is in fact false (not to conclude that there is a difference when in fact there is or not to find someone guilty when in fact he is). The error made when a true null hypothesis is rejected is called a **Type I error. Type II error** refers to the failure to reject a false null hyothesis. In hypothesis testing the probability of committing a Type I error is termed the **level of significance** and is set by the researcher. Common levels are 5% and 1% even though it should be noted that there is nothing sacrosanct about these levels. Even much higher levels of significance (lower levels of confidence, where level of confidence = 1 - level of significance) may at times prove to be useful and adequate. The steps involved in hypothesis testing can be summarized as follows (Kleinbaum and Kupper 1978, p. 26):

1. Check the assumptions concerning the properties of the underlying variable(s) being measured, which are needed to justify the use of the testing procedure being considered.
2. State the null hypothesis H_0 and the alternative hypothesis H_1.
3. Specify the significance level, commonly denoted by α.
4. Specify the **test statistic** to be used and its distribution under H_0.
5. Form the **decision rule** for rejecting and not rejecting H_0 (i.e., specify the **rejection** and **acceptance regions** for the test).
6. Compute the value of the test statistic from the observed data.
7. Draw your conclusions concerning rejection or non-rejection of H_0.

Many examples of simple applications of hypothesis testing are available in all standard textbooks of statistics. These usually involve the comparison of means or proportions in two samples obtained in a variety of different ways (see, e.g. Lapin 1978). We will not dwell on such applications partly for reasons of space but also because they are generally well known. Instead we take up those

techniques that are likely to be of particular relevance to the type of data available from nutritional surveys even though these applications do not exhaust the areas of their usefulness.

5.2 Categorical Data Analysis

Nutritional surveillance frequently employs categorized data where the observations fall into a fairly small number of categories or groups. In the sample we use for our data base, for instance, most of the variables pertaining to the characteristics of the clusters (villages) and households have from two to five categories. Sometimes this is inevitable in that the only possible level of measurement for a variable may be nominal (e.g. area). At other times measurements could have been made as continuous variables, (e.g. distance to water supply), but are not for various reasons (such as ease of collection). It may also be the case that a variable is measured at more than one level. Distance to road in this sample is given both in four categories and by minutes. As noted earlier, a higher level of measurement can always be reduced to a lower level and since nominal measurement is the lowest possible, every variable can be so measured. This fact increases the potential usefulness of the types of analysis considered in this section. These types of analysis are not however the only ones the can be used with categorical data. References to some others will be given and indeed a few of the techniques to be discussed later in this paper allow for, sometimes in fact require, categorized data along with other data measured at higher levels. In this section however our interest lies in simple analytical techniques that are suitable for application to nominally measured variables or variables whose level of measurement is reduced to make them so. A more sophisticated method of dealing with the same type of data is to use log-linear models which will not be discussed here (see, e.g. Anderson et al. 1980 and Knoke and Burke 1980). We assume that our data consist of frequency counts for the categories defined by reference to one or more characteristics of entities under observation. The number of categories must be at least two.

The most widely used method of analysis of categorical data relies on **Pearson's chi-square statistic** that has the following general form:

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} \qquad (5.1)$$

where O_j is the actual count (frequency) of observations in group or cell j, E_j refers to its expected count and k is the total number of categories or cells. Three of the purposes for which this statistic can be used are the following:

(a) **Testing for "goodness-of-fit"** is the first area of the application of the **chi-square test.** Its purpose is to determine whether or not the frequency distribution of the sampled population follows a particular, hypothesized shape. Many statistical techniques for example require variables to have a **normal distribution** and the assumption of normality may need to be tested beforehand. This use of the test is however relatively rare.

(b) **Testing for the statistical independence** of two variables is the second reason for the use of this test and is far more common. The data pertaining to this application of the test are usually presented in the form of a cross-tabulation recording the actual frequencies in each cell. The null hypothesis of statistical independence allows the investigator to determine the expected frequency of each cell so that the value of the X^2 (chi-square) can be determined and the test performed.

(c) At times it may be desired to test whether the rates of prevalence of a condition, e.g. infant mortality, are the same in different groups of people. When only two groups are involved the elementary procedures of **testing for the equality of two proportions** can be applied (see, e.g. Lapin 1978, pp. 490-3). With more than two groups the appropriate test statistic is the Pearson's X^2. This use of the test is also very common.

In each of these cases the general procedure is to assume that the null hypothesis is true and to find the expected frequencies in all the cells. The null hypotheses in the three cases above are respectively: that a particular distribution, such as normal, uniform, etc., applies and the sample is drawn from a population distributed thus; that the two variables are statistically independent; and that the rates of prevalence, or proportions, are the same for all groups. The actual frequencies are of course given by the sample. With these pieces of information now given, a X^2 value can be found using the above formula. An associated number of **degrees of freedom** is also determined using

the appropriate rules (to be mentioned below). Given the number of degrees of freedom and the specified level of significance the critical value of X^2 can be found from the table of **X^2 distribution.** If the calculated X^2 exceeds this value the null hypothesis is rejected. The intuitive idea is that if actual frequencies deviate "too much" from the expected frequencies, i.e. if X^2 is "too large", the sample would be unlikely to come from a population for which the null hypothesis would be true. For an explanation of the rationale behind the test see any modern statistical text, e.g. Winkler and Hays (1971) or Bhattacharyya and Johnson (1977). We now proceed to three examples based on our data set to illustrate the three situations described above.

Example 1: Testing for Goodness-of-Fit

In hypothesis testing involving anthropometric indicators such as WA, HA and WH it is often assumed that the distribution of these indicators in a population is normal. This assumption is of course never satisfied precisely but slight deviations usually do not matter very much since many tests are quite **robust** and remain valid. Large deviations, of course, do sometimes create difficulties and to find out the validity of the assumption it would be necessary to test it once in a while.

The assumption that anthropometric indicators such as WA, HA and WH are distributed normally is a convenient one since it allows the use of many statistical hypothesis testing procedures for these indicators. But how good is this assumption? A visual inspection of the histograms of these indicators based on our data set appears to support it (Tables 7-9). These histograms are drawn with class intervals equal to one-fourth of the standard deviation of each indicator. This ratio is chosen somewhat arbitrarily but the fact that the class intervals are defined in terms of standard deviations is intended to facilitate subsequent testing of the hypothesis of normality.

The actual frequency, or the number of children, in each class interval is given in Tables 7-9. Under the null hypothesis of normality, the expected frequencies can also be easily calculated. This is done by first finding the probability, or the proportion, of observations that are expected to be in each class interval. These probabilities are calculated from a table of areas under the **standard normal curve.** Since distances from the mean in these tables are given in standard deviation units the definition of class intervals above makes it a simple matter to find the relevant probabilities. These probabilities can then be multiplied by 428, or the total number of children, to find the expected frequency for each class interval. If the null hypothesis of normality were correct it would not be very likely that actual frequencies deviate "substantially" from the expected frequencies. In terms of the X^2 test statistic a "low" value indicates that the null hypothesis (of normality) cannot be rejected and, on the contrary, a "high" vlaue is evidence that it should probably be rejected. The values of X^2 statistic for WA, HA and WH, as calculated from equation 5.2.1, are 20.0, 21.2, and 24.5 respectively. Note that the intervals at both ends of the distribution are combined to prevent any individual expected frequency from falling below 5 (see below).

Are these values of X^2 statistic "high" or "low"? The answer comes from a comparison of each with a **"critical" value** of the X^2 for a given level of significance and an appropriate number of degrees of freedom. Let us set the

TABLE 7

HISTOGRAM OF WEIGHT FOR AGE AND TESTING FOR THE NORMALITY OF THE DISTRIBUTION

S T A T I S T I C A L A N A L Y S I S S Y S T E M

FREQUENCY BAR CHART

MIDPOINT WA	FREQ	EXPECTED FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
45.94029	0		0	0.00	0.00
48.92220	1		1	0.23	0.23
51.90411	2	9.59	3	0.47	0.70
54.88602	4		7	0.93	1.64
57.86793	13	7.40	20	3.04	4.67
60.84984	9	11.43	29	2.10	6.78
63.83174	18	16.61	47	4.21	10.98
66.81365	20	22.73	67	4.67	15.65
69.79556	26	29.06	93	6.07	21.73
72.77747	43	35.22	136	10.05	31.78
75.75938	36	39.72	172	8.41	40.19
78.74129	50	42.24	222	11.68	51.87
81.72320	30	42.24	252	7.01	58.88
84.70511	46	39.72	298	10.75	69.63
87.68702	28	35.22	326	6.54	76.17
90.66892	34	29.06	360	7.94	84.11
93.65083	22	22.73	382	5.14	89.25
96.63274	15	16.61	397	3.50	92.76
99.61465	17	11.43	414	3.97	96.73
102.59656	6	7.40	420	1.40	98.13
105.57847	4		424	0.93	99.07
108.56038	2		426	0.47	99.53
111.54229	0		426	0.00	99.53
114.52419	1	9.59	427	0.23	99.77
117.50610	0		427	0.00	99.77
120.48801	0		427	0.00	99.77
123.46992	0		427	0.00	99.77
126.45183	1		428	0.23	100.00
129.43374	0		428	0.00	100.00

5 10 15 20 25 30 35 40 45 50

FREQUENCY

Results:

MEAN WA = 80.2

STD DEV = 11.9

χ^2 = 20.0

DEG of FREEDOM = 15

$\chi^2_{0.05}$ = 25.0

Do Not Reject Null Hypothesis of Normality.

TABLE 8

HISTOGRAM OF HEIGHT FOR AGE AND TESTING FOR THE NORMALITY OF THE DISTRIBUTION

STATISTICAL ANALYSIS SYSTEM

FREQUENCY BAR CHART

MIDPOINT HA	FREQ	EXPECTED FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
69.85854	0		0	0.00	0.00
71.49729	1		1	0.23	0.23
73.13605	0	9.59	1	0.00	0.23
74.77481	4		5	0.93	1.17
76.41356	7		12	1.64	2.80
78.05232	8	7.40	20	1.87	4.67
79.69108	12	11.43	32	2.80	7.48
81.32983	10	16.61	42	2.34	9.81
82.96859	19	22.73	61	4.44	14.25
84.60734	30	29.06	91	7.01	21.26
86.24610	30	35.22	121	7.01	28.27
87.88486	43	39.72	164	10.05	38.32
89.52361	50	42.24	214	11.68	50.00
91.16237	42	42.24	256	9.81	59.81
92.80113	40	39.72	296	9.35	69.16
94.43988	46	35.22	342	10.75	79.91
96.07864	31	29.06	373	7.24	87.15
97.71740	24	22.73	397	5.61	92.76
99.35615	11	16.61	408	2.57	95.33
100.99491	6	11.43	414	1.40	96.73
102.63366	1	7.40	415	0.23	96.96
104.27242	6		421	1.40	98.36
105.91118	2		423	0.47	98.83
107.54993	3		426	0.70	99.53
109.18869	0	9.59	426	0.00	99.53
110.82745	0		426	0.00	99.53
112.46620	0		426	0.00	99.53
114.10496	0		426	0.00	99.53
115.74372	0		426	0.00	99.53
117.38247	0		426	0.00	99.53
119.02123	0		426	0.00	99.53
120.65999	0		426	0.00	99.53
122.29874	2		428	0.47	100.00

5 10 15 20 25 30 35 40 45 50

FREQUENCY

Results:

MEAN HA = 90.3

STD DEV = 6.5

χ^2 = 21.2

DEG of FREEDOM = 15

$\chi^2_{0.05}$ = 25.0

Do Not Reject Null Hypothesis of Normality.

TABLE 9

HISTOGRAM OF WEIGHT FOR HEIGHT AND TESTING FOR THE NORMALITY OF THE DISTRIBUTION

S T A T I S T I C A L A N A L Y S I S S Y S T E M

FREQUENCY BAR CHART

MIDPOINT WH	FREQ	EXPECTED FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
58.99336	0		0	0.00	0.00
61.11044	1		1	0.23	0.23
63.22753	0		1	0.00	0.23
65.34422	0		1	0.00	0.23
67.46111	0		1	0.00	0.23
69.57799	0	9.59	1	0.00	0.23
71.69488	0		1	0.00	0.23
73.81177	1		2	0.23	0.47
75.92865	4		6	0.93	1.40
78.04554	7	7.4	13	1.64	3.04
80.16243	8	11.43	21	1.87	4.91
82.27932	10	16.61	31	2.34	7.24
84.39620	32	22.73	63	7.48	14.72
86.51309	34	29.06	97	7.94	22.66
88.62998	35	35.22	132	8.18	30.84
90.74687	51	39.72	183	11.92	42.76
92.86375	44	42.24	227	10.28	53.04
94.98064	40	42.24	267	9.35	62.38
97.09753	49	39.72	316	11.45	73.83
99.21441	26	35.22	342	6.07	79.91
101.33130	26	29.06	368	6.07	85.98
103.44819	16	22.73	384	3.74	89.72
105.56508	14	16.61	398	3.27	92.99
107.68196	8	11.43	406	1.87	94.86
109.79885	7	7.4	413	1.64	96.50
111.91574	6		419	1.40	97.90
114.03263	2		421	0.47	98.36
116.14951	2	9.59	423	0.47	98.83
118.26640	2		425	0.47	99.30
120.38329	0		425	0.00	99.30
122.50017	1		426	0.23	99.53
124.61706	1		427	0.23	99.77
126.73395	0		427	0.00	99.77
128.85084	1		428	0.23	100.00
130.96772	0		428	0.00	100.00

5 10 15 20 25 30 35 40 45 50

FREQUENCY

Results:

MEAN WH = 93.0

STD DEV = 8.5

X^2 = 24.5

DEG of FREEDOM = 15

$X^2_{0.05}$ = 25.0

Do Not Reject Null Hypothesis of Normality.

level of significance at 5 percent. Since the population values of the mean and standard deviation of these indicators have had to be calculated from sample data, the appropriate number of degrees of freedom is the number of class intervals for which expected frequencies are calculated (18 in this case) less 3 (for the underlying argument in this case, see Lapin 1978). The critical values of X^2 are read from the table of X^2 distribution which can be found at the end of almost every statistical textbook. With the chosen level of significance of 5 percent and with 15 degrees of freedom this critical value is 25.0. This value exceeds the calculated values of X^2 statistics for all three indicators. It follows then that the null hypothesis of normality of the distribution of these indicators cannot be rejected. This conclusion can be restated in this way: only in about five out of 100 cases where the population values of the indicator are in fact normally distributed are we going to have a X^2 value above 25.0; since this is a rare event we conclude that we can, with a high level of confidence, refuse to reject the assumption of normality. We say "refuse to reject" rather than "accept" to emphasize the purpose of a statistical test of significance which is to evaluate the falsity of a statement rather than to try to establish its truth.

The results above are reassuring since they imply that the common assumption of normality is a valid one. Some cautionary remarks are however in order. The calculated values of X^2 depend partly on the choice of class intervals. It is quite possible to obtain seemingly contradictory results by choosing different class intervals in a given situation. In these marginal cases it is probably safe to assume that the hypothesis of normality holds. Slight deviations are usually of little consequence since many statistical tests that are based on the assumption of normality are "robust" and still remain valid. Even if the choice of class interval happened to be irrelevant and the null hypothesis was still rejected, but only barely, it may not be totally inappropriate to still maintain the assumption. In the case of WH above the calculated value of X^2 was only just below the critical value of X^2 leading to the non-rejection of the null hypothesis. If it had been only just above, it would have led to its rejection. In practical terms however there is almost no difference between these two cases. The main point to remember is that statistical tests should not be applied and interpreted mechanically. Some exercise of judgement regarding their practical importance is to be recommended. Statistical tests of significance should be carried out and their results should be taken seriously, but not always too seriously.

Another point related to the definition of class intervals is that they need not be equal. Some statisticians, for instance, prefer to define the class intervals such that the expected frequency within each class would be equal. In this case it is often recommended that twenty unequal intervals be defined. The expected frequency for each interval in our case would then be 0.05 x 428 = 21.4. The actual frequencies for these unequal intervals can easily be determined by first identifying the limits of class intervals in terms of standard deviation units. For an assumed normal distribution, and with twenty equal areas (probabilities or expected frequencies) the interval limits are - 1.645, - 1.282, - 1.036, - 0.842, - 0.675, - 0.524, - 0.385, - 0.253, - 0.126, 0.000, + 0.126, + 0.253, + 0.385, + 0.524, + 0.675, + 0.842, + 1.036, + 1.282, + 1.645. The actual observations in each indicator are then to be standardized and the number of observations falling within each class interval counted. The usual procedure can then be followed to carry out the test.

Example 2: Testing for Independence

The previous example concerned a situation where the hypothesized distribution of a single variable was being statistically tested. Suppose now however that observations are grouped according to two variables. Such outputs are ordinarily presented in 2-way tables (see Section 4) and a primary question of interest with regard to such tables is whether the two variables are independent. The general approach is to set up a null hypothesis of independence between the two variables, use it to derive the expected frequencies of each cell and use the X^2 test of independence to determine whether the actual frequencies are sufficiently different from the expected frequencies to lead to the conclusion that, with a given level of confidence, the null hypothesis should be rejected.

To give an example of this application of the X^2 test, consider Table 10 which is based on our sample survey data. The question is whether sample data support the null hypothesis of no association between children being malnourished or not on the one hand and the area cultivated by their households on the other. If the null hypothesis is true, the expected frequency of each cell would be equal to the product of the **marginal totals** (row and column totals) pertaining to that cell divided by total frequency (see Lapin 1978). For the first cell, for instance, the expected frequency would be (208 x 38)/428 = 18.5. With the knowledge of expected and actual frequencies for all cells, the X^2 statistic can be computed from equation 5 .1 which in our example turns out to be 12.9 (this value can also be read from computer printout; the output of option ALL in SAS (1979) is reproduced at the bottom of Table 10 which reports the X^2 value as well as the values of other test statistics. Some of these are going to be discussed later in Section 5.3. The associated number of degrees of freedom in a two-way table is the product (r-1)(c-1) where r and c refer to the number of rows and columns of the table. Here we have (2-1)(4-1) = 3 degrees of freedom. With 95 percent level of confidence (level of significance of 5 percent) the critical value of X^2 is 7.8. Since the calculated value of 12.9 exceeds 7.8 it is possible to reject the null hypothesis of no association. It appears that there exists some association between children being malnourished and the size of holding of their families. The **p-value** for X^2 = 12.9 with 3 degrees of freedom is 0.005 implying that if a sample of 428 children was randomly drawn from a population for which the null hypothesis of no association held, there would be 0.5 percent chance that a sample would be chosen for which the X^2 value would be at least as high as 12.9. Since this is a very small probability, it is concluded that, with 95 percent level of confidence, the hypothesis of no association can be rejected. For 99 percent level of confidence the critical value of X^2 is 11.3 which again leads to the same conclusion, if only just so. This conclusion can also be immediately arrived at by noting that the p-value of 0.005 is less than the level of significance 0.01 corresponding to a level of confidence of 99 percent.

Example 3: Testing for Equality of Proportions

It is often desired to compare the rates of prevalence, or proportions, in two or more groups for equality. For instance one may be interested in the comparison of the proportions of children sick in regions 1 and 2. Suppose that out of a total of 428 children, 164 come from the first region and 264 from the second. There are 68 children in the first region and 165 in the second who were sick i.e. 41.5 and 62.5 percent respectively. Are these prevalence rates significantly different from one another?

One statistical approach to answer this question is to apply a X^2 test. This example applies the X^2 test to two regions but its extention to more than two

TABLE 9.10

RESULTS BY LAND HOLDING AREA

TABLE OF MALNOUR BY CULT

MALNOUR CULT AREA CULTIVATED CLASS

FREQUENCY / PERCENT / ROW PCT / COL PCT	0	1	2	3	TOTAL
0	17	20	102	69	208
	3.97	4.67	23.83	16.12	48.60
	8.17	9.62	49.04	33.17	
	44.74	51.28	42.32	62.73	
1	21	19	139	41	220
	4.91	4.44	32.48	9.58	51.40
	9.55	8.64	63.18	18.64	
	55.26	48.72	57.68	37.27	
TOTAL	38	39	241	110	428
	8.88	9.11	56.31	25.70	100.00

MALNOUR = 0, WA ≥ 80 CULT = 0 Area cultivated < 0.1 Cx
MALNOUR = 1, WA < 80 CULT = 1 0.1≤ " " < 0.2 Cx
CULT = 2 0.2≤ " " < 1.0 Cx
CULT = 3 1.0≤ " " Cx

RESULTS BY LAND HOLDING AREA

STATISTICS FOR 2-WAY TABLES

CHI-SQUARE	12.928	DF= 3	PROB=0.0048
PHI	0.174		
CONTINGENCY COEFFICIENT	0.171		
CRAMER'S V	0.174		
LIKELIHOOD RATIO CHISQUARE	13.020	DF= 3	PROB=0.0046
GAMMA	-0.203	ASE1= 0.081	
KENDALL'S TAU-B	-0.113		
STUART'S TAU-C	-0.124	ASE1= 0.050	
SOMER'S D C|R	-0.124	ASE1= 0.050	
SOMER'S D R|C	-0.103	ASE1= 0.042	
PRODUCT MOMENT CORRELATION	-0.097		
SPEARMAN CORRELATION	-0.120		
LAMBDA ASYMMETRIC C|R	0.000		
LAMBDA ASYMMETRIC R|C	0.139		
LAMBDA SYMMETRIC	0.073		
UNCERTAINTY COEFFICIENT C|R	0.014		
UNCERTAINTY COEFFICIENT R|C	0.022		
UNCERTAINTY COEFFICIENT SYM	0.017		

ASE1 IS THE ASYMPTOTIC STANDARD ERROR.
R|C MEANS ROW VAR DEPENDENT ON COLUMN VAR.

regions follows a similar procedure and is trivial. The basic idea is to recognize that this problem can be transformed such that the test of independence in Example 2 can be used.

With the information given above (the number of children and the rates of prevalence in each region) it is easy to set up a 2 x 2 (in general a 2 x n) table such as the following:

	SICK	NOT SICK	TOTAL
Region 1	68	96	164
Region 2	165	99	264
TOTAL	233	195	428

If the rates of prevalence in the population of children in both regions are equal, the variables region and health status must be independent of one another. On the other hand if prevalences are indeed different this implies an association between region and health status. The next step should now be clear: apply a X^2 test of independence to these data and interpret the results accordingly. The calculated X^2 value in this example turns out to be 18.0. The critical value of X^2 with a level of significance of 5 percent and (2-1)(2-1) = 1 degree of freedom is 3.8. Therefore the null hypothesis of independence must be rejected since the calculated value of the test statistic is by far higher than the critical value of the X^2. This is equivalent to the conclusion that the rates of prevalence in the two regions do indeed differ significantly. A level of significance of 1 percent would lead to the same conclusion since the corresponding critical value of 6.6 is still far lower than the calculated value of X^2.

When more than two regions are involved, and the null hypothesis of equal prevalences has been rejected, it would be interesting to attempt to identify the region or groups of regions that contributed to the finding of significant difference. This attempt usually involves the partitioning of all regions into two regional groups and comparing their prevalences in the usual way. The choice of the two regional groups can be either planned beforehand or suggested by the data, i.e. whether, say, the neighboring regions are grouped together, or the regions with "low" and "high" prevalences are grouped separately. There are slight differences in these two cases and the interested reader may consult Fleiss (1973), pp. 92 ff.

In using Pearson's X^2 test, be it for testing goodness-of-fit, statistical independence or sample homogeneity, a number of points need to be kept in mind about the assumptions on which it rests. The first is that this test is only an approximation. To have adequate validity the number of observations (total frequency) must be large enough so that the expected frequency in each category or cell would be no less than 5 where there is more than one degree of freedom (df > 1) and no less than 10 where df = 1. If these conditions are not met categories must be combined so that they would be. But this option may not always be open with small sample sizes or more categories. For the 2 x 2 case with a small sample size the **Fisher-Erwin Exact Test,** which uses the actual probabilities of getting the observed frequencies, should be used (see Bhattacharyya and Johnson 1977).

The second point to note is that the sample units must be independently and randomly drawn from the population. Thus there must be independence both within and between the groups or categories being compared. This assumption however is not often satisfied since a **simple random selection** (srs) is the exception rather than the rule. In the sample survey we use as our data base a two-stage selection procedure was involved and hence the application of the X^2 test as explained in the examples above is really inappropriate since the observations are not made entirely at random. Nevertheless most hypothesis tests are carried out as if the sample was drawn by simple random selection. In Section 5.4 the effects of clustering on sample statistics will be discussed and, based on that discussion, a suggestion or two will be made as to the modifications that need be made to the X^2 tests of this section.

5.3 Correlation Analysis

Correlation Coefficients

Correlation analysis is one of the most commonly employed analytical techniques. In a summary table of studies relating measures of nutritional status (Asrat 1980, Part II) correlation analysis is identified as the sole method of analysis in about half of them. In most of the rest, it was used along with at least one other analytical technique. Correlation analysis provides some measure of the degree to which two variables are associated. The concept is related to that of testing for independence of two variables as discussed in the previous section. Two causally related variables are expected, under most circumstances, to exhibit a relatively high degree of association but the reverse need not be the case, i.e. a high correlation is not necessarily an indication of a causal connection. Examples of "nonsense" or **spurious correlations** abound and any inferences regarding a possible cause-and-effect relationship must rely on more than merely a high degree of association.

The degree of association or correspondence between two variables is measured by a correlation coefficient lying in the range from -1 to +1. There are many different types of correlation coefficients. They are not always alternatives to one another and the choice of which one to use is largely determined by the level of measurement of the two variables. Our purpose here is to give brief descriptions of a few of them that are likely to be of some value in the study of nutritional status. Our illustrations will be based on the sample survey data we have used before.

The best known measure of correlation between two continuous variables is the **Pearson product moment correlation coefficient.** Suppose that paired observations are available on two random variables X and Y. The **z-score** of each of these variables is a linear transformation such that the mean and the standard deviation of the transformed variables are 0 and 1 respectively. The Pearson correlation coefficient, r, between X and Y is the mean of the z-score products of the two variables. Hence,

$$r = \frac{1}{n} \sum^{n} z_x z_y = \frac{1}{n} \sum^{n} \left(\frac{X-\bar{X}}{s_x} \right) \left(\frac{Y-\bar{Y}}{s_y} \right)$$

where z_x and z_y are the z-scores, $\bar{X}$ and $\bar{Y}$ are the means of n sample observations and s_x and s_y are their standard deviations. A value of zero for the coefficient indicates a lack of correlation between the two variables. The values +1 and -1 signify perfect positive or negative correlation, respectively. Positive

correlation indicates that the two variables tend to move together either up or down and negative correlation implies a tendency to move in opposite directions.

The correlation coefficient measures the degree of linear association between two variables. If the two variables move in perfect correspondence but in a non-linear fashion, e.g. on a parabola, the coefficient would be different from unity. Furthermore, it is a pure (dimensionless) number and its statistical significance depends partly on the sample size, i.e. the number of observations. With a small sample size even a large coefficient may be entirely lacking in statistical significance. The square of the coefficient, r^2, is often referred to as the **coefficient of determination** (on which more later) and measures the proportion of **variation** of one of the two variables accounted for by the other (this interpretation is standard in regression analysis (see Section 6.1).

A straightforward extension of the concept of correlation coefficient is the **correlation matrix** which records all correlation coefficients between all pairs of a given set of variables. The matrix is symmetrical and has unit elements in its principal diagonal. The correlation matrix in Table 11 is an example based on our data set. As expected the correlation coefficient between WA and HA is quite high (0.80) and that between WA and WH (0.58) demonstrates their relatively strong association. The correlation coefficient between HA and WH is however practically zero and for this reason the negative sign is of no importance, or interest. (For a discussion of the correlations between these three pairs of indicators, see Hitchings/CBS 1979.) Also note that the above statements about correlation coefficients are not made based merely on the size of their estimates but also on their statistical significance. The statistical significance of a correlation coefficient is discussed later.

As noted earlier, the Pearson correlation coefficient requires intervally measured (continuous) variables. Many variables are however not measured on an interval scale. When at least one of the two variables is ordinal, the **Spearman rank correlation coefficient** could be used instead. It is denoted by r_s and is merely the Pearson correlation coefficient of the ranks of the variables rather than their values. For example if the children in the sample survey were ranked from 1 to 428 according to their WA, HA, and WH, the matrix of Spearman rank correlation coefficients would be as given in Table 12. This is nothing more than the matrix of Pearson correlation coefficients calculated from the ranks of children rather than from their actual measures of WA, HA

TABLE 9.11

S T A T I S T I C A L A N A L Y S I S S Y S T E M

VARIABLE	N	MEAN	STD DEV	SUM	MINIMUM	MAXIMUM
WA	428	80.23224299	11.92763456	34339.40000000	47.70000000	127.00000000
HA	428	90.34299065	6.55502542	38666.80000000	70.80000000	121.80000000
WH	428	93.92219626	8.46754918	40196.70000000	60.70000000	129.20000000

CORRELATION COEFFICIENTS / PROB > |R| UNDER HO:RHO=0 / N = 428

	WA	HA	WH
WA	1.00000 0.0000	0.79901 0.0001	0.57636 0.0001
HA	0.79901 0.0001	1.00000 0.0000	-0.00976 0.8401
WH	0.57636 0.0001	-0.00978 0.8401	1.00000 0.0000

TABLE 9.12

S T A T I S T I C A L A N A L Y S I S S Y S T E M

VARIABLE	N	MEAN	STD DEV	MEDIAN	MINIMUM	MAXIMUM
WA	428	80.23224299	11.92763456	79.44999695	47.70000000	127.00000000
HA	428	90.34299065	6.55502542	90.34999084	70.80000000	121.80000000
WH	428	93.92219626	8.46754918	93.39999390	60.70000000	129.20000000

SPEARMAN CORRELATION COEFFICIENTS / PROB > |R| UNDER HO:RHO=0 / N = 428

	WA	HA	WH
WA	1.00000 0.0000	0.81606 0.0001	0.62809 0.0001
HA	0.81606 0.0001	1.00000 0.0000	0.12770 0.0082
WH	0.62809 0.0001	0.12770 0.0082	1.00000 0.0000

and WH. It is seen that the rank correlation coefficients exhibit the same pattern as do Pearson correlation coefficients but are slightly larger.

Kendall tau rank correlation coefficient is the third in our list of correlation coefficients. It is an alternative to Spearman rank correlation coefficient even though its rationale is completely different. Its main advantage over its alternative is that it can be generalized to a **partial correlation coefficient** (the correlation coefficient between two variables when the effect of one or more other variables is controlled) whereas Spearman's cannot. The mechanics of its computation are somewhat more involved and with more than about 200 observations it becomes fairly expensive to compute (see SAS 1979, p. 175). For an explanation of the computational procedure, see Nie et al. (1975) or Roscoe (1975). An example of its application to a study investigating the causal factors of nutritional status may be found in Sanjur and Romero (1972).

The correlation coefficients discussed so far can be used in situations where paired observations are available on two variables. When observations are arranged in a bivariate frequency table, or a cross-tabulation, other correlation coefficients may be appropriate. As a possible example we consider the **Goodman-Kruskal gamma coefficient** which provides a measure of the association between two discrete, ordinally measured variables in a cross-tabulation. The simplest case is that of a 2 x 2 table in which H and L refer to high or low values that two variables can take.

In the following table

		First Variable	
		H	L
Second	H	a	b
Variable	L	c	d

a, b, c, and d are cell frequencies. There are "a" entities which show high scores on both variables, "b" entities that have a low score on the first variable but a high one on the second, etc. The Goodman-Kruskal gamma measures the extent to which high values or low values tend to happen together across the sample. In this case it is defined as:

$$\gamma = \frac{ad - bc}{ad + bc}$$

When high values of one variable pair only with high values of the other and its low values only with the low values of the second, b and c would both be zero and $\gamma = +1$. On the other hand when high values of one variable pair only with the low values of the other $\gamma = -1$. Generalization to more than two categories for each variable is straightforward but rather messy. An example from the sample data we use relates to association between four distance-to-town categories and five land-holding categories as measured by gamma (Table 13). The value of 0.12 indicates that among farmer households there is some association between proximity to town and wealth as measured by the amount of land a farmer household cultivates.

A correlation coefficient appropriate for a situation where one of the variables is continuous such as HA, and the other is dichotomous, e.g. whether a child was sick the week before or not, is the **point biserial correlation coefficient.** It is given by

$$r_{pbi} = \frac{M_1 - M_0}{\sigma_x} \sqrt{pq}$$

where x = the continuous variable, a child's HA

M_1 = the mean of the continuous variable for the group scoring one on the dichotomy, the child being sick the week before

M_0 = The mean of the continuous variable for the group scoring zero on the dichotomy, the child not being sick the week before

σ_x = the standard deviation of the continuous variable

p = the proportion of the total group scoring one on the dichotomy, the proportion of children sick last week

q = the proportion of the total group scoring zero on the dichotomy, the proportion of children not sick last week.

On the basis of our data $M_1 = 90.09$, $M_0 = 90.64$, $\sigma_x = 6.55$, $p = 0.54$, and $q = 0.46$. Hence the point biserial coefficient in this case is 0.04, indicating almost no correlation at all.

A simple extension of the above is to a situation where both variables are dichotomized, i.e. when a 2 x 2 table is involved. The appropriate correlation coefficient in this case is the **phi coefficient.** Suppose that we wish to find the degree of association between the WA of a child being below or above 80 percent

TABLE 13

CROSS TABULATION OF CHILDREN BY DISTANCE TO TOWN CATEGORY AND LANDHOLDING

DISTANCE TO TOWN (in mins.)	AREA CULTIVATED (in Cx = 1.29 hectares)					
	0.1	0.1-0.2	0.2-1.0	1.0-5.0	> 5.0	Total
0-50	2	12	36	16	1	67
50-100	21	9	56	30	2	118
100-200	12	13	106	43	0	174
> 200	3	5	43	18	0	69
Total	38	39	241	107	3	428

of standard and a morbidity measure such as whether or not he was sick last week. The cross-tabulation is the following:

Distribution of Children by Malnourishment and Morbidity

	WA > 80%	WA < 80%	TOTAL
Sick last week	a = 97	b = 99	196
Not sick last week	c = 111	d = 121	232
TOTAL	208	220	428

The formula for the calculation of the phi coefficient is:

$$phi = \frac{bc - ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

which gives a value of 0.016 indicating practically no correlation between malnourishment and morbidity. No significant phi correlation could be found between wasting and morbidity either. As a rule of thumb a phi coefficient of less than .30 or .35 may be taken to indicate no more than trivial association (Fleiss 1973, p. 60). It should also be pointed out that an exact relationship exists between phi and the X^2 for the same table which is of the following form

$$phi = \sqrt{\frac{X^2}{n}}$$

where n is total frequency. The relative advantages of X^2 and phi are discussed in several sources including Fleiss (1973). Briefly we point out that X^2 is very sensitive to the total frequency whereas phi is not.

There are a host of other measures of correlation in addition to those mentioned above. For a more thorough discussion of those above and some others the reader may consult Hays (1973) and Roscoe (1975). A summary of the correlation coefficients introduced here, with their uses, is given in Table 14.

Significance Tests for Correlation Coefficients

Correlation coefficients typically lie within a range from - 1 to + 1. Suppose that the data from a particular sample result in an estimate of 0.7 for the Pearson correlation coefficient in the population. Does this estimate necessarily imply that the population value is significantly above zero? The

Table 14

Summary of Uses of Different Correlation Coefficients

CORRELATION COEFFICIENT	TYPES OF VARIABLES	EXAMPLE: Correlation Between	COMMENTS
1. Pearson's product moment	Both continuous	HA and WA	The most common type of correlation coefficient; measures linear association.
2. Spearman's rank	Both ordinal	rankings of HA and WA	Finds Pearson correlation coefficient of rankings of two variables.
3. Kendall's tau rank	Both ordinal	rankings of HA and WA	Alternative to Spearman's rank correlation coefficient; can be generalized to a partial correlation coefficient.
4. Goodman-Kruskal	Ordinal data in bivariate frequency tables	location and farm size	Popular with sociologists; the only correlation coefficient here designed for use with discrete ordinal data.
5. Point biserial	One continuous, other dichotomous	WA and sick/well	Invariant with respect to actual values of dichotomous variable.
6. Phi	Both dichotomous	Wasted/not wasted and sick/well	Extension of point biserial coefficient; not sensitive to total frequency.

answer is of course that it does not. Correlation coefficient is a sample statistic with a sampling distribution which depends not only on the magnitude and sign of the correlation in the population, but also on the sample size. Just as for any other statistic, the statistical significance of a sample coefficient of correlation can and, sometimes, should be tested.

Consider the Pearson correlation coefficient first. If its value in the population is zero and the joint distribution of the two variables is bivariate normal in form, the sampling distribution of its sample estimate would be symmetrical and approximately normal. With increasing sample size the approximation improves, albeit rather slowly. When population value of the correlation coefficient is different from zero, however, the sampling distribution of the sample estimate would be skewed to the right for negative correlation and to the left for positive correlation. As the population value increases, the degree of skewness increases as well. The difference in the sampling distribution between the two cases when the population value is zero and when it is not suggests that it is simpler to carry out significance tests for the correlation coefficient in the first case than it is in the second. This is indeed the case.

The null hypothesis of interest for the moment is $\rho = 0$, i.e. population value of the correlation coefficient is zero. For relatively small sample sizes the relevant test statistic is the following:

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

with N - 2 degrees of freedom. r is the sample estimate of the correlation coefficient. The comparison of the calculated value of t from the above equation with the critical value for a given level of significance and N - 2 degrees of freedom (read from the table of t-distribution) results in the rejection or non-rejection of the null hypothesis. When N is large the normal distribution can be used instead of the t-distribution above. For instance the correlation between WA and HA was calculated for a sample size of 428 to be 0.8. The value of the above statistic is

$$0.8\sqrt{\frac{428-2}{(1-0.8)^2}} = 27.5$$

which is considerably higher than the critical value read from the normal distribution table for a level of significance of, say, 0.01. Therefore the

population value of the correlation coefficient between these two indicators is not zero.

This conclusion is hardly striking, or interesting. Indeed in many tests of significance the null hypothesis that the population value is zero should be of little interest. Its popularity stems from the simplicity of the testing procedure even though it often adds little to the knowledge already available. When this is the case a null hypothesis that the population value is some number different from zero is of considerably more interest. For instance one may wish to test whether the population value in this case is 0.7. Alternatively one may wish to construct a confidence interval for the population value. This is likely to be a more useful result than that the population value is significantly different from zero.

Let us find the confidence interval for the correlation coefficient between WA and HA. Since the population value is significantly above zero, the sampling distribution of the sample estimate of the correlation coefficient, r, is highly skewed and despite the large size of the sample, the usual procedure of establishing confidence intervals using normal distribution is not applicable. Fisher has however shown that, for samples of at least moderate size, the usual procedure can be applied to a particular function of r given by

$$Z = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$$

where $\log_e$ indicates the natural logarithm of the expression in parantheses. Once the confidence interval for Z is found it can be converted to that for r with a reverse transformation using the above relationship (see Winkler and Hays 1971, pp. 609 ff.)

Since the sample estimate of correlation coefficient r is 0.8 the corresponding value of Z can be calculated to be 1.099. The approximate sampling variance of Z is given by

$$var(Z) = \frac{1}{N-3}.$$

The confidence interval for Z can be written as:

$$Z - z_{(\alpha/2)} \sqrt{\frac{1}{N-3}} \leq \zeta \leq Z + z_{(\alpha/2)} \sqrt{\frac{1}{N-3}}$$

where $Z_{(\alpha/2)}$ is 1.96 for 95 percent level of confidence and N = 428. is the Greek letter zeta and refers to the population value of Z. Therefore the 95 percent confidence interval for ζ is

$$1.099 - 1.96\sqrt{\frac{1}{428-3}} \leq \zeta \leq 1.099 + 1.96\sqrt{\frac{1}{428-3}}$$

or

$$1.004 \leq \zeta \leq 1.194$$

Using these confidence limits for and the relationship between Z and r, the 95 percent confidence interval for the population correlation coefficient of WA and HA, ρ, is given by:

$$0.763 \leq \rho \leq 0.832.$$

This result is of considerable interest. It clearly points out that the sample estimate of 0.80 is very accurate. This is a consequence of the large sample size used in this case. There is a problem however in that the sample units were not drawn randomly as the assumptions of this test would require. We show in the next section that the violation of this assumption can be partly compensated for by using an effective sample size instead of the actual sample size. The effective sample size is the size of an equivalent simple random sample that yields the same degree of precision as that offered by the actual clustered sample. It will be shown there that the effective sample size in the case of our data set is roughly 60 percent of the actual sample size of 428, or about 260. Using this value for N in the above procedure results in a 95 confidence interval that is slightly larger than the one derived above. It turns out that the new confidence interval is:

$$0.751 \leq \rho \leq 0.840.$$

Having a clustered sample does not appear to involve much "cost," in terms of precision, in this particular case.

This discussion is intended to emphasize two points. First, the calculation of a confidence interval for a population correlation coefficient is far more useful than that of merely a sample estimate since its precision is also taken into consideration. Second, and more importantly, significance tests involving the null hypothesis of zero population value are often useless since, in many

instances, they add nothing to the current knowledge. Because of the difficulty of computing confidence intervals they need not of course be calculated for every sample estimate even if it were possible. But for a few important estimates the extra effort is likely to be well worth it.

The arguments above were made with reference to Pearson correlation coefficient, but they are equally valid for the point biserial correlation coefficient as well. For tests of significance of the other correlation coefficients, which are incidentally simpler, see Roscoe(1975), Chapter 31.

5.4 Analysis of Variance

Consider a situation where at least two groups of individuals, households, entities, etc. are identified by the different values, or combination of values, of one or more categorical variables. It is sometimes desired to carry out statistical tests of hypotheses involving the means of a continuous variable across these groups. A familiar example is where interest lies in comparing the means of two groups for equality. The usual procedure in this simple case is to apply the **t-test.** Where more than two groups are involved, and where there may be more than one categorical variable defining the groups, the appropriate procedure is to apply the techniques of analysis of variance.

The term analysis of variance (ANOVA or AOV) is something of a misnomer since these techniques are in fact concerned with the simultaneous comparison of means of one variable across several groups. Such comparison however requires the examination of ratios of sample variances and this fact explains the origin of the term. The categorical variable(s) that distinguish the groups from one another are sometimes referred to as the independent variable(s) whereas the continuous variable is commonly known as the dependent variable. Examination of the effects of a single categorical variable on the dependent variable is referred to as a **one-way** ANOVA. When simultaneous effects of two or more categorical variables are studied, the ANOVA is called **two-way** or **n-way** respectively. A **fixed-effects** or **linear hypothesis model** is used in ANOVA where classes or groups of entities are fixed (i.e. remain the same) over repeated experiments. In contrast, a **random-effects** or **variance component model** is used where classes or groups included are a random sample of a larger number of possible groups. The key difference between these two models is whether inferences are being made about only groups presently included in the study, or a set of all possible groups of which only a sample is being studied. A combination

of fixed- and random-effects models is called a **mixed-effects model.** Most statistical packages are limited to performing the ANOVA using a fixed-effects model even though, since computations are in most cases parallel, a random-effects interpretation is usually possible, if appropriate. The techniques of ANOVA are discussed in all elementary statistical texts. Interesting and fairly comprehensive treatments are given in Blalock (1972), Winkler and Hays (1971), Hays (1973), Nie et al. (1975), and Snedecor and Cochran (1980).

The one-way ANOVA test is based on estimates of variances. The total variation over all groups under study can be partitioned into two parts: one is specifically due to the variation between the group means under comparison (commonly referred to as the **between sum of squares** or SSB) whereas the second part is the variation within groups not accounted for by the variable defining the groups (**within sum of squares** or the **mean square error,** MSE). If there is little or no difference between the group means, i.e. if the null hypothesis is true, the ratio of these two estimates will be approximately equal to one. On the other hand if the groups do indeed differ with respect to the dependent variable, i.e. the null hypothesis is false, the ratio will be larger than one. The F-distribution can then be used to determine how large a difference is tolerable, i.e. how large the difference can be and still have the groups come from the same population.

While quite powerful, ANOVA models are based on somewhat limiting assumptions: sample observations are to be statistically independent and randomly drawn; populations must be normally distributed and their variances must be equal. The process of randomization, of course, ensures that the first assumption is satisfied. The normality assumption may not always be justified and partly because of this, nonparametric alternatives of ANOVA are sometimes preferred (e.g. using Kruskal-Wallis test in place of one-way ANOVA). The homogeneity assumption, or the assumption of equality of variances, appears to be not too restrictive since it has been established that choosing samples of equal size goes a long way towards preserving the robustness of the ANOVA models. If it is desired to test this assumption, a simple method, i.e. t-test, can be used in the two-sample (group) case. The ratio of the sample variances, S_1^2/S_2^2, approximates an F-distribution with n_1-1 and n_2-1 degrees of freedom. Thus this ratio can be compared with the appropriate critical value read from an F table in order to decide whether the hypothesis of equal variance should be rejected. Several methods are suggested in textbooks on general statistics for testing the equality of variances for more than two groups. Barlett's Test for Homogeneity

of Variance and the maximum variance/minimum variance ratio are often recommended although their limitations should be borne in mind when interpreting the results (Snedecor and Cochran 1980). Blalock (1972) offers three strategies for dealing with unequal variances. If the departures from homogeneity are moderate, transformations such as using the logarithm of each value may help resolve the problem. If only one group is widely different from the rest, it can be omitted from the ANOVA test. Alternatively the nonparametric analog, the Kruskal-Wallis test, can be used. For a discussion of the assumptions underlying the ANOVA models and methods of resolving the difficulties that arise when they are violated see Hays (1973).

Let us turn now to the illustration of the ANOVA technique. The first illustration involves a typical application: we wish to test the null hypothesis that the mean WA of children is the same in all four distance to town categories.

Distance to town in mins.	No. of children	Mean WA
0-50	67	85.8
50-100	118	80.8
100-200	174	78.7
> 200	69	78.2

Note that there are four distance to town categories and the number of children in each is fairly large. Note also that there appears to be a negative relationship between distance to town and mean WA. With ANOVA however we are not examining the nature of this relationship but only whether any relationship exists between WA and distance to town. The null hypothesis to be tested therefore is whether the mean WA is the same across distance categories. Some differences may be expected among pairs of category means due to sampling variability. The question of interest is whether at least one of these differences is large enough to be statistically significant at a given level. The results of ANOVA are given in Table 15. The F value of 6.82 corresponds to a p-value of 0.0002 indicating that category means are very significantly different. The existance of such a difference is indicative of a relationship of some sort between the dependent variable WA and the independent variable distance to town. The nature of this relationship can be examined with other techniques such as regression analysis.

TABLE 15

ANALYSIS OF VARIANCE OF WEIGHT FOR AGE

ACROSS DISTANCE TO TOWN CATEGORIES

Source of Variation	df	Sum of Squares	Mean Squares	F	p
Between distance categories	3	2798.68	932.89	6.82	0.0002
Within distance categories	424	57999.37	136.79		
Total	427	60798.05			

The second illustration is provided in the context of a problem already mentioned in an earlier section, namely, the effect of clustering in sample design on the standard errors of estimates of means and rates of prevalence. In our data set there are 36 clusters that are selected in the first stage of a two-stage sampling design. Clustered designs have a particularly attractive feature: by selecting clusters of sample elements that are usually close geographically, they substantially reduce the cost of field work and alleviate problems associated with inadequate or unavailable sampling frames. The advantage however comes at a price: the standard errors of the estimates of population totals, means, prevalence rates, etc., increase compared with those of a simple random sample of equal size. While in the analyses of survey data this effect is often ignored and standard errors are calculated as if the sample was drawn at random, this practice can sometimes be inappropriate since it may lead to substantial errors regarding the reliability or precision of estimates. The error will be insignificant if the elements of each cluster are as variable as those in the population at large. Ordinarily however, because of the proximity of the cluster elements, there exists some degree of homogeneity within each cluster. The more homogeneous the elements of a cluster are, the less the additional information that is obtained by including an additional element in the sample. It follows that a clustered sample should usually be larger than a simple random sample from the same population if the same level of precision in the estimates is desired.

The degree of homogeneity within a cluster is measured by **intraclass correlation coefficient** "Rho."In terms of the components of variance in a clustered sample it can be defined as:

$$Rho = \frac{\sigma_a^2 - \sigma_b^2/(B-1)}{\sigma^2}$$

where σ_a^2, σ_b^2 and $\sigma^2 = \sigma_a^2 + \sigma_b^2$ are respectively variance between clusters, variance within clusters and total variance. B is the number of elements within a cluster (see Kish 1965, p. 170). The estimate of Rho is denoted by rho and can be obtained as

$$rho = \frac{s_a^2 - s_b^2/b}{\hat{s}^2}$$

where $\hat{S}^2 = S_a^2 + \frac{b-1}{b} S_b^2$ and S_a^2 and S_b^2 can be taken from analysis of variance Table 16.

Now consider the following question: should one worry about the clustering effect? It depends. There are numerous sampling designs used in practice and the exact calculation of standard errors can be all but impossible. Usually approximate formulas are used to find them. One such approximation that is often employed assumes simple random selection with the hope that the calculated standard errors would not be too different from their correct values. If the intraclass correlation coefficient is not significantly different from zero this assumption would be as adequately valid as any. To test it, an ANOVA table such as that in Table 17 is produced. If the value of F calculated from such a table is not significantly above 1, one could conclude that the intraclass correlation coefficient is not significantly different from zero and the assumption of simple random sampling would be valid. If however the F value is above 1 and statistically significant it can be concluded that the intraclass correlation coefficient is positive and significant. F values below 1 are possible but rare. They imply that intraclass correlation is negative or that the elements of a cluster are more variable than one would expect to find by random selection. This case is of no practical interest.

To illustrate these ideas let us consider our sample data and ask the question whether the intraclass correlation coefficient in this case is sufficiently close to zero for the assumption of simple random sampling to be valid. The table of analysis of variance corresponding to HA is given in Table 17. The sums of squares are divided by their respective degrees of freedom to arrive at mean squares. The ratio of the mean squares yields an F value of 1.57. Considering the probability of the occurence of this value under the null hypothesis of no difference in the cluster means (or zero intraclass correlation coefficient) which is 0.0235 we conclude that at five percent level of significance the means of HA across clusters are significantly different and that the intraclass correlation coefficient significantly exceeds zero. The assumption of simple random sampling would be inappropriate in this instance. It turns out that the same conclusion is reached in the cases of WA and WH. It is important to remember that since we are interested in all three of these variables the same test has to be carried out for each individually. It may

TABLE 16

ANALYSIS OF VARIANCE FOR TWO STAGES OF RANDOM SELECTIONS

Source of Variation	Degrees of Freedom	Sums of Squares (SQ)	Mean Squares	Expected Values of Mean Squares
Between clusters	$a - 1$	$\frac{1}{b}\sum_{\alpha} y_{\alpha}^2 - \frac{y^2}{ab}$	bs_a^2	$\left(\frac{B-b}{B}\right)S_b^2 + bS_a^2$
Within clusters	$a(b-1)$	$\sum_{\alpha}\sum_{\beta} y_{\alpha\beta}^2 - \frac{1}{b}\sum_{\alpha} y_{\alpha}^2$	s_b^2	S_b^2
Total	$ab - 1$	$\sum_{\alpha}\sum_{\beta} y_{\alpha\beta}^2 - \frac{y^2}{ab}$	s^2	

In the language of the analysis of variance the above would often be called a *finite* model of a *nested* or *hierarchal* situation. Note that when the *between* and *within mean squares* are equal, then $s_a^2 - s_b^2/b = 0$. In expected values we have $S_a^2 - S_b^2/B = 0$, whether we have complete clusters ($b = B$) or subsampling ($b < B$). These are the conditions for Roh = 0 (5.6.16). Hence the usual test to decide whether the ratio $F = bs_a^2/s_b^2 = 1$ or >1, also serves as a test to decide whether Roh = 0 or >0. Less frequently one may find genuine cases of $F < 1$, hence, negative Roh.

(Kish 1965, p. 173)

TABLE 17

ANALYSIS OF VARIANCE FOR HEIGHT FOR AGE BY CLUSTER

Source of Variation	df	Sum of Squares	Mean Squares	F	p
Between Clusters	35	2253	64.38	1.57	0.0235
Within Clusters	392	16094	41.06		
Total	427	18347			

well turn out that the hypothesis of no clustering effect is valid for some but not others of the variables. Another point to note is that at one percent level of significance the null hypothesis of equal cluster means (or zero intraclass correlation) for HA cannot be rejected. This fact implies that we are dealing with a borderline case and the conclusion depends on the arbitrary significance level chosen. However we maintain the significance level of 5 percent to demonstrate the procedure to be followed when intraclass correlation coefficient is found to be significantly above zero.

What is the value of rho in this case? For HA we have from Table 16 and the formulas above

$$b s_a^2 = 64.38$$
$$s_b^2 = 41.06$$
$$b = 428/36 = 11.89$$
$$s_a^2 = 5.41$$
$$\hat{s}^2 = 43.02$$

and finally rho = 0.05 or about 5 percent. Similar calculations are done for WA and WH and the corresponding rho's are both 0.06. These estimates of intraclass correlation are relatively close and fairly small even though they are all significantly above zero. Some idea of their importance can be gained by noting that the ratio of the variance of a cluster sample to that of a simple random sample of the same size, sometimes referred to as the **design effect,** is given by [1 + rho (b-1)] where b is the average number of sample elements per cluster. For HA, WA and WH the values of design effect are 1.50, 1.62 and 1.69 respectively. These results imply that the clustering effect of the sample design of our data set increases the variance, as compared to that of a simple random sample, by 50, 62 and 69 percent respectively. Since the variances for a simple random sample can be obtained very easily, it is now possible to estimate the variance of the mean HA, WA and WH for our cluster sample using these estimates of design effect. Disregarding the design effect leads to the underestimation of the standard errors by 18, 22 and 23 percent. These in turn imply that the confidence intervals based on standard errors calculated with the assumption of simple random selection would be too narrow for a given level of confidence, or that the actual level of confidence is smaller than intended.

A further point needs to be made. The unit of analysis in the above example was a child. At the household level where HA, WA and WH refer to the means of these indicators based on child level values within each household the clustering effect may be different than at the child level. Table 18 sets out the values of F, p, intraclass correlation coefficient and the design effect for the three indicators at both child and household levels. It is seen that at the household level the design effect for HA and WA is not significant; hence the assumption of simple random sampling for these indicators can be justified at household level.

The concept of design effect lends itself to an alternative interpretation. The effect of clustering is to increase the variance of an estimate relative to a situation where a simple random sample of the same size is employed. It follows then that for a given level of precision, the size of a clustered sample would have to be larger than that of a simple random sample by a multiplicative factor given by the design effect. (We ignore a rare exception when the design effect is less than one requiring a smaller clustered sample, see Hansen et. al. 1953.) When a clustered sample is being designed this factor is of course unknown. An informed guess based on past experience needs to be made. For instance, the Center for Disease Control uses a design factor of two for this purpose (see, e.g. Liberia, 1979). The reverse situation is when one knows the size of the clustered sample and wishes to find that of a simple random sample giving the same estimates of standard errors. Here the clustered sample size should be divided by the design factor. The result is sometimes referred to as the <u>effective</u> sample size, or **effective n** for short (Kish 1965, p. 162). The usefulness of the concept of effective n derives from the possibility of improvement that can be gained by substituting it for the actual n in some tests of significance. For instance in using the t-test for the equality of two means, effective n can be substituted for the actual n in the calculation of the standard error of the difference. In establishing a confidence interval around a mean one could multiply the critical value of t for a given level of significance by the square of the design effect to enlarge the interval (Kish 1965, Chapter 14). When X^2 test is being used too, the use of effective n reduces its calculated value, thereby decreasing the probability of the rejection of the null hypothesis. (Note that both the numerator and denominator of the X^2 statistic are reduced. But since the numerator is a squared term and the denominator is not, the resulting value for X^2 is reduced by the same factor as the design effect, or by the same factor

TABLE 18

ANOVA TESTS FOR PRESENCE OF DESIGN EFFECT

Dependent Variable	Level	No. of Cases	No. of Clusters	F	p	Interclass Corr. Coeff.	Design* Effect
HA	Child	428	36	1.57	0.024	0.046	1.50
WA	Child	428	36	1.73	0.01	0.058	1.63
WH	Child	428	36	1.80	0.004	0.063	1.69
MEANHA	Household	261	36	1.14	0.282	----	----
MEANWA	Household	261	36	1.27	0.15	----	----
MEANWH	Household	261	36	1.76	0.008	0.095	1.60

* Factory by which the sample size should be divided to give the "effective" sample size.

that the effective n falls short of the actual n.) These are ad hoc measures that partially take into account the clustering effect but they are adequate in most cases particularly when it is not very large.

The analysis of sample surveys ordinarily involves the estimation of many statistics. Computing standard errors for each can be a demanding task despite the availability of modern computers. It complicates presentation of the results of the survey and can overwhelm the human resources available particularly if considerations such as clustering effect are to be properly taken into account. For these reasons standard errors are in most cases not computed at all or computed only in the simplest cases and based on the simplest assumptions. The fact that the theoretical issues have not always been satisfactorily resolved cannot but add to the difficulty of an attempt at comprehensive analysis. Given this state of affairs, rules of thumb abound. Since this is not intended to be a sampling manual we are not going to dwell too much on these issues (see Kish 1965, Chapter 14). A few additional guidelines relevant to situations that arise often in the analysis of sample surveys may however prove to be useful.

Every effort should be made to provide correct standard errors for at least a few of the most important estimates. Correct standard errors need not be, and usually are not, exact. They are only the best approximations to the true values. The validity of the assumption of simple random sampling should always be examined if at all possible. A test such as that used above can be employed. If clustering effect cannot be assumed away it should be calculated and used to provide better standard errors than those obtained with the assumption of simple random sampling. As a rule of thumb, when a relationship is highly significant, adjusting for the design effect may not appreciably reduce its significance. It only needs to be taken into account when the significance is marginal. Calculation of design effect for each variable may not be needed if there is any reason to believe that they are more or less similar. Even if several are calculated it may be possible to find some average of them that will be used in further analysis. In the case of our data set it is probably fair to assume that the effect of clustering is to increase variances of simple random selection by an average of about 60 percent.

6. ADVANCED ANALYTICAL TECHNIQUES

The basic techniques of the previous section are used in a wide range of circumstances and can answer many important questions of interest to an

investigator. Furthermore, data analysis often involves some experimentation at first to provide the investigator with some feel for his data. At this preparatory stage he develops an impression as to the descriptive characteristics of his variables, acquires an understanding of the relationships that may exist among them and formulates simple hypotheses which he could then test. All these are crucial elements of data analysis and basic techniques are particularly suited for these purposes. There are however other considerations.

The basic techniques are typically limited to the consideration of two variables at a time. The simultaneous impact of several variables on one or more than one other variable is however also of great importance. In addition, the basic techniques are of little value insofar as prediction and forecasting are concerned. Even though partial correlation coefficients and 2- or n-way analyses of variance can statistically control the variation due to one or more variables, they are not prediction tools. Similarly these and other basic techniques are able to go some way towards identifying appropriate classification criteria but more formal techniques designed specifically for this purpose are needed as well.

This section considers a number of multivariate techniques. The choice is based primarily on their potential usefulness in analysis of nutrition-related data. Another consideration is that there must exist at least one example of their application in the relevant literature. Regression analysis is the subject of Section 6.1. This well-known multivariate method is highly versatile and is likely to be used more regularly in future nutritional studies. While the discussion of regression analysis here is at too basic a level to be of immediate practical value on its own, it does bring out certain important concepts of the subject. Analysis of covariance models and models with binary dependent variable are extensions of regression techniques that allow for categorical independent and dependent variables respectively. These extensions are discussed in Sections 6.2 and 6.3. The fundamental ideas of principal components analysis, discriminant analysis and cluster analysis are intuitively discussed in Sections 6.4 - 6.6 with some illustrations. These are essentially classification techniques that can also be useful for purposes of analysis. They have so far been rarely used in nutrition-related literature and their discussion here is intended to highlight their useful features and promote their more widespread application when appropriate. Note in passing that all the advanced techniques discussed here involve only one independent variable at a time.

6.1 Regression Analysis

Statistical techniques of regression analysis are well developed, powerful and flexible. With computing facilities now widely available, regression analysis has become the most popular method of multivariate analysis. Other multivariate techniques that are becoming increasingly popular are often extentions of regression analysis developed for specific objectives and types of data available in particular circumstances. This section reviews some elements of the basic theory of regression and provides illustrations of its use in the analysis of data pertaining to nutrition and related factors. For a gentle introduction to the topic of regression analysis the reader may consult Lewis-Beck (1980). More comprehensive treatments are available in many textbooks of which one of the best is Gujarati (1978).

Basic Elements

Broadly speaking, the purpose of regression analysis is to explain or account for the variation in a dependent variable by the variation in one or more independent variables presumed to be responsible for it.

Consider the following mathematical relationship between one possible **outcome indicator** of nutritional status such as a child's height-for-age Y (the dependent variable) and a number of **explanatory variables** $X_1, X_2, \ldots, X_k$ (the independent variables) presumed to account for its changes:

$$Y_i = a + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} \qquad (6.1)$$

where i refers to a particular unit of observation, in this case a child. There are a number of points to be made with regard to this relationship:

(a) The structure of causality is assumed a priori: it runs from the right-hand side (RHS) to the left-hand side (LHS). RHS variables change independently of the LHS variable thereby making it a dependent variable. This knowledge usually comes from theory, experience, previous analysis, or in the worst case, it may be no more than a pure guess. The important point to remember is that there is nothing in statistical techniques of regression analysis, or in any other statistical technique for that matter, that can "prove" or establish the causality. At the most, they may provide evidence of strong association. In epidemiology certain rules are applied to support the hypothesis of causality (see, e.g.,

Mausner and Bahn 1974, Ch. 5) but these do not have an equivalent for regression techniques.

(b) The number of factors that can be responsible for at least some change in a dependent variable is ordinarily large. For a number of reasons, not the least of which is simplicity, it is necessary to limit the independent variables to those that are judged to be of more importance. Statistical analysis can be of help in making such judgements.

(c) The assumption of linearity of the relationship is just that, an assumption. Many functional relationships are nonlinear and whether or not a nonlinear specification is called for, and if so, of what form, is itself to be studied and tested. Much of regression analysis is based on the assumption of linearity. Fortunately this is not as restrictive an assumption as it may appear at first since it refers to the linearity in parameters of the population (the b's), not in the variables obtained from a sample (the X's). Many types of nonlinearity in variables can be accomodated through some suitable transformation.

(d) As it stands equation (6.1) is an exact, or **deterministic,** or algebraic relationship. The actual data almost never fit this relationship exactly. There are many reasons why this must be so. Some factors may have been excluded because they may not be known to influence the dependent variable, because they may not have been measured or because they may not be measurable. The assumption of linearity may not be valid. Measurements are never exact, and then there is the inherent randomness in such relationships. All of these reasons provide a rationale for the introduction of a random **disturbance term** u_i to the RHS of the relationship. The result is then:

$$Y_i = a + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + u_i \qquad (6.2)$$

This is now a **statistical relationship** and the purpose of regression analysis is to use the available data on X's and Y to "estimate" parameters a, b_1, b_2,..., b_k as well as to "test" for their "significance". The results are then interpreted. Note that the function of the random disturbance term u_i is to take up the slack left after the influences of the X's on Y are accounted for. For all the reasons stated above the true values of this disturbance term for each and every observation cannot be known. In practice however they are usually assumed to follow a normal distribution with a zero mean.

Why would one want to estimate such a statistical relationship? There are two primary reasons: to explain an outcome and/or to predict it. To explain an outcome it is necessary to identify the nature of the relationship that relates it to its determinants. In practice, the feasible objective often turns ut to be no more than to identify only some of the major determinants of an outcome, or rather factors associated with it, in the hope that by manipulating them through intervention policies and programs the outcome can be influenced in the desired fashion. This objective often entails the separation of the individual effects of major outcome determinants. Since this cannot be done experimentally by holding all outcome determinants fixed except for the one whose influence is being estimated, regression techniques are often employed as a second-best alternative to control the influence of **confounding factors.** Confounding factors are those that influence the outcome but are also distributed differently (as compared with the distribution of the outcome variable) across the population of interest. Consider the following two examples. In a rural development project it is important to know whether coffee-producing farmers have better nutritional status than other farmers because of what they produce and the relatively high income coffee production generates or because they happen to live near roads and have better access to markets and services. If after statistically controlling for the possible confounding influence of the distance to road factor (which is assumed to vary across households) coffee producers still appear to have better nutritional status the plausibility of the hypothesis or theory that coffee production helps nutritional status is somewhat increased. There are of course ordinarily many potential confounding factors of which no more than a few may be statistically controllable in any given situation. Similarly in evaluating the nutritional effects of a food stamp program it is important to distinguish between a situation where the nutritional status of recipients is worse than that of non-recipients because the program has been ineffective and another where other factors account for the observed difference even though the program may have been highly effective.

An investigator's major interest however may not be in explaining an outcome but rather in predicting it with the use of available data on some of its potential determining factors. For prediction purposes a statistical relationship needs to be developed between the outcome and its potential determinants such that a major portion of the variation of the outcome variable is accounted for by the relationship. This objective is usually difficult to attain and prediction of, say, nutritional status on the basis of data commonly available at the individual,

household and community levels is often too imprecise to be of much value. Regression techniques can be used for both the explanation of an outcome and its prediction but the illustrations provided here have primarily the first objective.

Estimation and Interpretation

The estimation of the parameters (a, b_1, etc.) of equation (6.2) and the tests of significance conducted on the estimates involve a number of assumptions. These assumptions ensure that the **estimators** (formulas or rules employed to arrive at the estimates) have certain desirable properties (see, e.g. Gujarati 1978). The disturbance term u_i is assumed to be a random variable with zero mean and constant variance. It should also be uncorrelated with other disturbance terms and with the independent variables. The X's are assumed to be nonrandom, or at least uncorrelated with one another in a linear fashion. The elements of the sample from which the data are generated are also assumed to be randomly and independently selected to allow inferences to be made about the population.

Given these assumptions it can be shown that the dependent variable Y and parameters a, b_1, b_2,...,b_k are all random variables with associated probability distributions. These distributions would all be normal if a further assumption of the normality of the distribution of the disturbance term is added. The mechanics of how these coefficients are estimated need not concern us here. Computer packages automatically calculate and print them. Briefly, it may be stated that the usual approach to the estimation of a regression equation is to use the method of **ordinary least squares** (OLS). This method is based on the principle of minimizing the **sum of squared residuals,** $\sum^{n} e_i^2$, where $e_i = Y_i - \hat{Y}_i$ or the difference between the actual observation on the dependent variable Y_i and its estimate $\hat{Y}_i$ based on the corresponding value(s) of the independent variable(s). Those interested in pursuing this further may consult any introductory text on econometrics, in particular Gujarati (1978), and Wonnacott and Wonnacott (1979). We concentrate here on the interpretation of regression results.

If all the assumptions stated above hold, the interpretation presents little difficulty. The estimate of, say, b_2 is commonly denoted by b_2 and represents the expected amount of change in the dependent variable resulting from a unit increase in the value of the corresponding independent variable X_2 <u>when the values of other X's are held constant or fixed.</u> In controlled experiments the

fixity of these other X's can be assured by the investigator as he changes X_2 by a unit. The type of data that a social scientist investigates however is generally non-experimental, i.e. he is usually not in a position to fix all but one of the independent variables. To insure that the interpretation of regression coefficients would still be valid the assumption of independence, or at least zero correlation, among pairs of independent variables would have to be made at the outset. When this assumption is not valid the investigator faces the problem of **multicolinearity** and has to be very careful in interpreting the significance of individual regression coefficients. Multicolinearity will be briefly discussed later.

It was noted earlier that given the underlying assumptions of the regression model, each regression coefficient is a random variable and follows a normal distribution. The standard deviation of the sampling distribution of a coefficient is commonly referred to as **standard error** and is of paramount importance. It provides an indication of the degree of precision of the estimate of the coefficient. A "large" standard error may suggest that the estimate is quite unreliable and vice versa. The size of the standard error is gauged relative to the size of the corresponding estimated coefficient. The ratio of an estimated coefficient to its standard error is termed **t-ratio** which is normally provided as part of the output of a regression program. The t-ratio is used to test whether a parameter is significantly different from zero. If it turns out that it is not (when t-ratio is too low), one may be justified in concluding that the relevant variable has little or no effect on the dependent variable. On the other hand if the t-ratio is "high enough", the estimate of the coefficient would be significantly different from zero, i.e. the independent variable is likely to be a determinant of the outcome variable under consideration.

The RHS variables in an estimated regression collectively explain or account for a portion of the variation in the dependent variable. The ratio of the explained portion of the variation to total variation is termed **coefficient of determination** and denoted by R^2. This is a measure of the goodness of fit of the estimated regression model and may have any value in the range from 0 to 1. Ordinarily one would like to be able to explain a high proportion of the variation in the dependent variable. This can be done by choosing the "right" independent variables as well as by increasing their number. Addition of an independent variable never decreases R^2. This practice is not always to be recommended. One should never lose sight of the fact that regression analysis is a statistical

method of finding evidence for or against a presumed causal relationship. Hence the inclusion of variables on the RHS must be guided primarily by the knowledge of or hypotheses concerning this presumed relationship.

Regression analyses of sample survey data which seek to explain variation in nutritional status variables by the differences in socio-economic variables typically result in R^2 values that are fairly low, often as low as 0.05 or 0.1. Regression models involving economic variables, by contrast, frequently show R^2 values as high as 0.8 or 0.9 or higher. This characteristic difference between nutritional surveillance data and economic data explains why regression analysis as a predictive tool is much more suitably applied in the context of economic data than in that of nutritional surveillance data. Recall that the unit of analysis in nutritional studies is typically an individual or a household. At the present stage of knowledge and data availability, the prediction of nutritional status of such small units using regression techniques is likely to be highly unreliable and is best avoided. As an explanatory device however, regression analysis is finding increasing acceptance in nutritional studies particularly as relevant data on determinants of nutritional status become more widely available.

In the rest of this section and in the following two sections some additional features of regression will be discussed in the context of a few illustrations, violation of the underlying assumptions and the use of categorical variables as both independent and dependent variables.

Some Illustrations

With this much theoretical background behind us, we can now examine a number of estimated regression equations using our sample data. This sample, as already discussed, is selected in a two-stage process that, while guaranteeing equal probability of selection for each household, nevertheless violates the assumption of independent selection. Obviously this is in violation of one of the assumptions underlying regression theory. Hence even though the estimates of parameters remain relatively unbiased, the usual tests of significance may lose their validity. While this may sometimes present a serious problem, in our case the problem is not very serious since the clustering effect, while present, is not substantial (see Section 5.4). Where this is not the case, or when the significance is marginal, the interpretation of the results becomes more difficult and uncertain; some guidance on coping with this problem is given in Kish (1965), Chapter 14.

Our interest in this example lies in trying to relate one anthropometric indicator, HA, to household- and village-level characteristics. Our first decision has to do with the choice of the unit of analysis (child, household, or village) for the dependent variable. The difficulty with child-level analysis is that selection probabilities are not the same for children. Furthermore the set of independent variables would be identical for all the children in the same household thereby weighting households in proportion to their number of children. This is inconsistent with the fact that household selection was made with equal probability. Taking a household as the unit of analysis presents another problem however: what is the height-for-age associated with a household? There are several alternative answers: HA for a household could be taken as the mean of this measure for all the sample children in the household or the value of this measure for the worst-off child, for the oldest, or for the youngest. We choose to work with the mean HA for a household partly because the same approach could be adopted at the cluster level as well. The choice of other approaches may be equally valid, and it may be useful to undertake a comparative study of these alternatives in future research.

It is always useful to begin with hypotheses or likely guesses, about at least the direction of the influence of individual independent variables on the dependent variable. For instance, we might expect that the prevalence of malnutrition in more isolated areas would be higher than in less isolated areas The degree of isolation may be measured with distance to town and a simple regression of HA on the absolute measure of distance to town appears to confirm this hypothesis. The two **simple regressions** (only one independent variable) at the cluster level and at the household level are:

Cluster level	HA =	92.642	-	0.018 TOWNABS	$R^2 = 0.22$
					$F = 9.70$ ($p < 0.01$)
t-ratio		(126.29)		(-3.11)	$n = 36$
Household level	HA =	92.164	-	0.011 TOWNABS	$R^2 = 0.02$
					$F = 4.86$ ($0.01 < p < 0.05$)
t-ratio		(124.14)		(-2.20)	$n = 261$

We analyze these two regression equations somewhat more fully than those that follow in order to highlight the common approach to the analysis of regression results and some recurring features of our equations. While both the dependent and independent variables are the same in both equations, the units of

observation are not. A cluster consists of anywhere from two to twenty-five children and the HA variable at the cluster level is found as the mean of the values pertaining to individual children in each cluster. TOWNABS is a continuous variable defined for a cluster. Its values for households within a given cluster are the same. The sign of the estimated coefficient of this variable is negative in both equations as expected. This fact however would be of little value if the estimated coefficients of this variable were not statistically significant. As it happens, they both are at five percent level of significance since their corresponding t-ratios in absolute value (3.11 and 2.20) are "high enough". The second however is not high enough to be significant at one percent level of significance even though the first still is. As a rule of thumb, a t-ratio of about 2 or more indicates statistical significance at the level of, at least, five percent.

Another way of arriving at the same conclusion is to inspect the value of F-statistic reported to the right side of both regressions. Note that, apart from rounding errors, the F values are equal to the square of the t-ratio of the coefficient of the independent variable in each equation. This is a mathematical property in simple (one independent variable) regressions but has no counterpart in multiple regressions (more than one independent variable). The p-value would be exactly the same for simple regression equations regardless of whether the t-ratio of the independent variable or the F value of the equation is used. Computer outputs determine the appropriate number of degrees of freedom for both and print the p-value. The t-ratio of the intercept terms in both equations are extremely large and as a result the estimates of these intercept values are highly significant.

The value of $R^2 = 0.22$ in the first equation indicates that some 22 percent of the variation of HA at the cluster level has been explained by the variable on the RHS, distance to town. The R^2 of the second equation suggests that only about 2 percent of the variation of HA at the household level can be accounted for by distance to town. A striking fact about the values of this statistic in the two equations is that they are very different. In fact however they should not be compared at all since the units of analysis are different. The fact that the R^2 of the first equation is more than ten times that of the second does not necessarily imply that the regression at the cluster level is in any way

superior to that at the household level. The variation of HA at the cluster level is by far smaller than that at the household level since intra-cluster differences of HA are removed by aggregation. The two R^2 values are not comparable.

In simple regression the R^2 is the square of the simple correlation coefficient between the dependent and independent variables. The sign of the correlation coefficient is also identical to the sign of the coefficient of the independent variable. It will be recalled (Section 5.3) that the usual test of significance of correlation coefficient seeks to establish whether the coefficient is significantly different from zero. The exact equivalent of this test in simple regression is to establish whether R^2 is significantly above zero (R^2 is always positive). The relevant test statistic is the F value which is always provided as part of the regression output. A significant F value indicates that the independent variable (or the group of independent variables collectively in multiple regression, see below) explains a significant portion of the variation in the dependent variable, which is another way of saying that the two variables are correlated. It can be seen that all the information provided by simple correlation analysis is also provided by simple regression analysis. Regression analysis however provides some additional information as well. For example the coefficient of the independent variable relates the amount of change in the dependent variable to that in the independent variable, a piece of information not provided by correlation analysis. It is in this sense that regression analysis can be a prediction tool whereas correlation analysis cannot. This advantage is particularly useful when there are several independent variables as in multiple regression.

The two regressions discussed above are examples of <u>simple</u> regression where there is only one independent variable. <u>Multiple</u> regresion involves more than one independent variable. To illustrate it consider a regression similar to the cluster level regression before when a new independent variable, altitude, is added which may explain a little more of the variation in HA. The estimated regression is:

Cluster level	HA =	92.751	- 0.016 TOWNABS	-0.002 ALTI	$R^2 = 0.23$
					F = 4.9
t-ratio		(120.84)	(-2.40)	(-0.55)	$(0.01 < p < 0.05)$
					n = 36

The addition of a second independent variable increases the explanatory power only marginally, 23 percent against 22 percent. This variable however is not significant as its t-ratio is small, but distance to town retains its significance as does the regression itself (F value is still relatively high). Little seems to be gained by including altitude as an explanatory variable and it may be better not to do so for the sake of simplicity. Other variables could also be added and they may turn out to be significant but it should be emphasized again that the choice of independent variables must as far as possible be guided by theoretical expectations and availability of data.

Note that neither of the above conclusions (concerning the significance of the two coefficients) is ever certain. The idea is to assess the plausibility of the existence of a relationship. Such an assessment presumes that the assumptions made about the regression model are at least approximately valid. It also rests on a prior specification of what constitutes statistical significance. The level of significance is commonly set at one or five percent. In many circumstances however a higher level may still prove to be quite useful. There are no hard and fast rules for setting this level but it helps to understand why a coefficient may not be significant since it would then be possible to better appreciate the tentative nature of statistical significance.

Clearly, a higher level of significance (e.g. 5% against 1%) increases the chances of finding statistically significant coefficients but it also increases the likelihood of committing Type II error, i.e. concluding that a relationship exists when in fact it does not. This level should be set before the regression results are available. If a coefficient is statistically not significant the most likely conclusion is that the corresponding independent variable probably has no bearing on the outcome variable and does not cause it to change. However, it may be due to certain statistical artifacts as well. The standard error of a coefficient is calculated from the following formula:

$$S_b = \sqrt{\frac{\sum e_i^2}{n-k} \frac{1}{\sum (x-\bar{x})^2}}$$

where $\sum e^2$ is the sum of squared residuals, n is the sample size, k is the number of RHS variables and X refers to the particular RHS variable with mean $\bar{X}$. Lack of statistical significance arises from a high standard error relative to the estimate of the coefficient b. S_b may be relatively high because the sample size n may be low, or because the dispersions or variances of X-values may be

limited. If sample size and/or the dispersion of the values of the independent variables are high enough it is always possible to find statistical significance at any chosen level. This fact has to be borne in mind but it is also important to point out that the degree of control that can be exercized over these factors is usually fairly limited. For example sample size is often determined largely by the amount of resources that may be available for data collection. When sample size has to be small because of limited resources, it may be advisable, ceteris paribus, to allow the level of significance to be set at a higher level and vice versa.

From a theoretical standpoint the number of RHS variables in multiple regression could be as high as the number of observations minus one (in which case the fitted regression would be perfect and $R^2 = 1$). In practice however a regression with more than 5-10 independent variables is likely to give rise to problems of interpretation and violations of underlying assumptions. The inclusion of too many independent variables often signifies a lack of adequate understanding or clear hypotheses regarding the relationship. Addition of RHS variables has an apparent comforting effect in that it raises (or at least does not diminish) the coefficient of determination R^2 or the portion of explained variation in total variation of the dependent variable. This however need not always be a desirable property in that the incremental increase in the R^2 due to a particular added variable may be insignificant or, even if significant, may not be very important from a practical point of view. Aside from the notion that the inclusion of independent variables should be guided as far as possible by theoretical considerations and experience regarding the objectives of the analysis, statistical procedures are available to help in choosing what to include. Some of these will be briefly mentioned in the following pages.

Violations of Assumptions

Some of the difficulties encountered in proper use of regression techniques stem from the fact that on many occasions one or more of the underlying assumptions may be violated. If these violations are serious they may result in wrong signs, imprecise estimates or estimates that lack desirable properties. Much of the art of applying regression techniques consists of an ability to detect these violations, to understand their implications and to find ways of overcoming them.

When one or more assumptions are violated the usual approach is to revise the data or the model or, sometimes, the estimation procedure in order to reduce the extent of violation or to minimize its impact on the desirable properties of regression estimates. In this section we discuss some of the more elementary approaches to coping with these problems. For a comprehensive treatment see, e.g. Gujarati (1978).

There is precious little that can be done when the sample is non-random. The best advice in this case is probably not to use regression tools at all. When they are employed anyway, as is sometimes the case, statistical tests of significance should either not be performed or if performed, the results must be interpreted very carefully.

There are cases where the sample is indeed random but the sample selection procedure is not the simple random selection that ensures both equal probability of selection of the observations and their independence. Clustering methods of sample selection result in lack of independence although they may be self-weighting in that the probability of selection of each element is the same. This is how our sample data were collected. In this situation it may be possible to apply some simple methods of statistical correction (see e.g. Section 5.4 and Kish 1965, Chapter 14) to reduce the impact of lack of independence on the accuracy of the tests of statistical significance.

Sample data may also have been obtained through a selection procedure that does not ensure equal probability of selection. It may also be the case that all observations may not have been made with equal degree of precision or that they may not be equally important. For example, in the cluster level regression discussed above each cluster constitutes a unit of analysis, but the precision of the cluster level dependent variable HA is not the same. Some sample clusters had more sample households than others. Since the values of cluster level variable HA were derived on the basis of different numbers of households in clusters it may be contended that these values are not equally precise, i.e. the HA mean for a cluster with 20 sample households is likely to be more precise than a HA mean for a cluster with only 2 sample households. In these and similar cases it may be argued that a more precise observation, a more important one or one with a smaller probability of selection should receive relatively more weight than others with smaller precision, less importance or higher selection probability. If a proper weighting scheme can be found it may be useful to employ **weighted regression.** In weighted regression the (weighted) sum of

squared residuals, consists of a weighted addition of individual squared residuals, not of a simple addition of individual squared residuals of each observtion as is the case in ordinary least squares method. Computer packages normally provide for this option.

One feature of the regression analysis under discussion here is its linearity, i.e. the dependent variable must be more or less linearly related to each and every one of the independent variables. When this assumption is violated, the regression equation has a poor fit and the R^2 will have a lower value than it would if the nature of the relationship could be specified more accurately. It is advisable to plot the dependent variable against each independent variable to examine this assumption before estimating the regression equation. The nature of the relationship between the dependent variable and a particular independent variable often suggests the best way of incorporating the independent variable. A transformation may be required. For example if the dependent variable is number of calories per day consumed by an adult equivalent and the independent variable is household income it is likely that they are not linearly related. Calorie intake however may well be more or less linearly related with logarithm of income variable or some other transformation of it. The independent variable should in such cases be transformed to arrive at a more reliable specification of the regression model.

The RHS variables are, ideally, uncorrelated with one another. Some small amount of correlation is almost always present and is "acceptable" since its effects on the properties of the estimation of coefficients are usually negligible. High correlation in at least one pair of RHS variables gives rise to the problem of multicolinearity and can become troublesome. Multicolinearity leads to an inability to differentiate the individual influences of the correlated variables. It manifests itself in different ways but the most common sign of its presence is a situation where the entire regression is significant (as evidenced by a significant F-value) but individual coefficients are not. There are several methods of detecting and correcting for multicolinearity but none is free of criticism (see any textbook on econometrics or applied regression analysis, e.g. Pindyck and Rubinfeld 1976, Gujarati 1978, Draper and Smith 1980). Dropping one of the pair of correlated variables is one of the most commonly used devices to remove or reduce multicolinearity. This method may or may not be appropriate in a given situation but it does point to a larger problem: How does one decide if a variable "belongs" to, or should be included, in a particular regression?

The answer to this question can be sometimes very difficult. There are no hard and fast rules. The "kitchen sink" approach throws in just about any variables one could lay his hands on in an attempt to maximize R^2. This is almost always a bad approach. Seeking a theoretical basis for the inclusion of a variable is usually preferable except that in many cases it may not be strong enough. A formal method of approaching this problem is to use stepwise regression. In this method one proceeds by either introducing the RHS variables one by one (stepwise forward regression) or by including all the independent variables at the beginning and rejecting them one at a time (stepwise backward regression). The decision to include or drop a variable is usually made on the basis of its contribution to the explained sum of squares as judged by the F-test. This test is a formal measure of whether a variable is "worth" being included or not. A variation of the stepwise forward procedure allows for the possibility that a variable included in earlier steps may need to be rejected if those entering later make it superfluous (see Nie et. al. 1975 and SAS 1979). These methods are often employed at the exploratory stages of data analysis to gain an understanding of the data and possible relationships. They should not however be applied indiscriminately. Unfortunately the ease with which these methods can be applied (given the options available in widely used package programs) carries with it the danger that they may be overused, misused or abused. They should complement, not replace, theoretical formulation of relevant hypotheses and model specifications.

6.2 Covariance Analysis

The examples of simple and multiple regression discussed above involved only continuous independent variables. It will be recalled that analysis of variance, by contrast, required only categorical independent variables. Covariance analysis bridges the gap between these two techniques by allowing a mixture of continuous and categorical variables to be used on the RHS. The dependent variable is still required to be continuous. The following illustration involves two continuous independent variables (ROADABS: distance to road and DEMOG: the proportion of children younger than 6 to the size of the household) and two categorical variables (WEALTH which reflects housing condition and SANDUM which refers to distance to water supply reflecting sanitary conditions). The categorical independent variables are both dichotomous and are often referred to as **dummy variables.** Dummy variables usually take values 0 or 1 depending on the absence or presence of a certain characteristic. In our

example WEALTH = 0 refers to poor housing and WEALTH = 1 to improved housing conditions. In turn SANDUM = 0 refers to a household far from a source of water whereas SANDUM = 1 indicates a household with water supply nearby. These values are chosed such that if, as expected, improved housing condition and nearness to water supply are associated with better nutrition, the coefficients of these dummy variables would turn out to be positive. The standard of comparison is a household with poor housing and distant water supply. The estimated regression is as follows:

Household level	HA =	94.32	- 0.017 ROADABS	- 0.207 DEMOG	+ 1.608 WEALTH
t-ratio		(75.31)	(-3.10)	(-1.43)	(2.05)
		+0.111 SANDUM		$R^2 = 0.06$	
				$\bar{R}^2 = 0.04$	
t-ratio		(0.13)		$F = 3.86$ ($p < 0.01$)	
				$n = 261$	

This equation is highly significant as the p-value corresponding to F = 3.86 is lower than 0.01. The portion of total variation of the HA that is explained however is only about 6 percent. All the coefficient signs are in the expected direction although two of the independent variables (DEMOG and SANDUM) are not significant at 5 percent. The inclusion of DEMOG variable was intended to examine whether nutritional status would be poorer in households with relatively larger number of dependent younger children. This does not appear to be the case once the more important factors of accessibility and weath are accounted for. Similarly, distance to water supply does not seem to have any bearing on nutritional status in the presence of other included variables. The other dummy variables (WEALTH) is fairly significant. Its coeffcient of 1.6 indicates that the average HA in households with improved housing exceeds that in households with poor housing by about 1.6, all other things remaining equal.

The value of the R^2 indicates that only about 6 percent of the variation of the HA is being explained and **adjusted coefficient of determination,** $\bar{R}^2$, is somewhat less. The $\bar{R}^2$ is often more meaningful in multiple regression than R^2 as it "corrects" or "adjusts" for the inflation of the R^2 when more and more independent variables are added. Such additions never decrease R^2 but they may well reduce $\bar{R}^2$ if the "cost" of the addition of an unimportant variable is more than its "benefits". The $\bar{R}^2$ may be negative and is related to R^2 by the

following relationship:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

where k is the number of parameters in the equation including the intercept term (e.g. in simple regression k=2). Choosing the "best" model by maximizing $\bar{R}^2$ is only slightly less objectionable than doing so by maximizing R^2. Let us emphasize one point again: The purpose in regression analysis is to develop a model that is logical, theoretically sound and useful. The main concern is the relevance of the explanatory variables to the explanation of the dependent variable. The statistical significance of coefficients is of greater importance than the values of R^2 or $\bar{R}^2$. High values for these statistics are desirable but not desirable enought to be had at the expense of theoretical expectations and findings of previous empirical research.

6.3 Models with Binary Dependent Variable

In classical regression analysis a continuous dependent variable is regressed on one or more independent variables which are also continuous. As we saw in Section 6.2 the use of dummy (0-1) variables allows categorical independent variables to be used as well with the usual OLS method of estimation. The extention of the model to situations where the dependent variable is categorical is the subject of this section.

Consider the case of a random sample of children of whom some are stunted (with height-for-age below 90%) and others are not. The dependent variable here, stunting, is dichotomous (0: not stunted, 1: stunted) and can be related to a set of independent variables such as socio-economic characteristics of a child's household or his community, his health condition, his die., etc. If the relationship between the set of characteristics and stunting is sufficiently strong, it may be possible to use information on such characteristics to predict with some accuracy whether or not a child is stunted.

As another example (from Edmonston et al. 1981) consider a sample of 0-10 year old Bangladeshi children of whom a small number have died. The purpose is to examine whether child mortality depends on any of a number of maternal and child specific characteristics. Here, as in the previous example, the dependent variable is dichotomous (0: alive and 1: dead) and the explanatory variables consist of child's sex, child's parity (a function of household size), mother's and

child's ages, interaction between sex and age and prior birth interval. (The dependent variable may have more than two categories; see, e.g. Pindyck and Rubinfeld 1976.)

The usual OLS estimation procedure can be used to derive the estimated regression of the dichotomous dependent variable (stunting in the first example and mortality in the second) on a given set of independent variables. The result is a **linear probability model.** There are however three reasons which make it unsatisfactory:

(a) The error term is itself dichotomous and hence not normally distributed as required for testing hypotheses;

(b) The variance of the error term is not constant as assumed in classical regression, i.e. the the error term is **heteroskedastic;**

(c) The predicted values of the dependent variable can in principle have any value, not just 0 and 1 as is the case with actual observations on the dependent variable. What does a predicted value of 2 mean? Of 0.47? Of -1.5?

If the predicted value lies in the range 0-1 the interpretation is relatively simple: such values can be taken to represent the probability of the presence of the characteristic, for example, stunting or mortality. This is a natural interpretation since our information is almost never sufficient to make a perfectly certain prediction of whether or not a child is stunted. The problems however still remain of some predicted values lying outside the 0-1 range and of finding an estimation procedure that can cope with (a) and (b) above. One way to cope with these problems is to employ **probit analysis** which will not be discussed here (see, e.g. Pindyck and Rubinfeld 1976). Another, more appealing way to get around these problems is to use **logit analysis.** When P_i is the probability that a child is stunted, the **logit of stunting** can be defined as

$$\log\left(\frac{P_i}{1-P_i}\right)$$

Note that the expression $P_i/(1-P_i)$ is the ratio of the probability of stunting to that of not stunting and is known as the **odds** that a child is stunted. The logit model regresses the logarithm of the odds in favor of a characteristic on the independent variables.

The usefulness of this transformation of a simple probability P_i to the logit of stunting is that when P_i varies from 0 to 1 the value of the logit of stunting

takes values from $-\infty$ to $+\infty$, i.e. the entire range of real numbers. The logit model is then of the following form:

$$\log\left(\frac{P_i}{1-P_i}\right) = a + b_1 \log x_1 + b_2 \log x_2 + \cdots$$

The problem of predicted values outside the 0-1 range does not happen anymore. The logit transformation is only one of many possible ways of dealing with this problem although it is one of the most convenient, not least from a computational standpoint.

The parameters of logit model cannot be estimated by the usual OLS method since when P_i is either 0 or 1 (which is always the case in actual observations) the expression $\log[P_i/(1-P_i)]$ is undefined. There are several alternatives (see, e.g. Pindyck and Rubinfeld 1976, Andersen et al. 1980 and Morris and Rolph 1981) but the maximum likelihood procedure is the most common and in large samples its estimates will be approximately unbiased. An alternative procedure is available if subjects can be grouped by the values, or the combination of values, of one or more independent variables. This usually requires large samples and categorical independent variables. A third method of approaching the same estimation procedure relies on linear discriminant analysis which looks at the problem somewhat differently and which will be discussed in some detail in the next section.

As an illustration of the application of the three models above the results of the estimation of a linear probability model, a probit model and a logit model are shown in Table 19. The data came from a child-year file pertaining to Bangladeshi children and the estimated models along with two others, are reported in Edmonston et al. (1981). The dependent variable in the OLS model is the probability of death of a child, P_i, whereas in the probit model it is Z_i where $Z_i = F^{-1}(P_i)$ is the inverse of the cumulative normal probability function and in the logit model it is the logit of mortality or $\log\left(\frac{P_i}{1-P_i}\right)$ from which

$$P_i = \frac{1}{1 + e^{-(a + b_1 \log x_1 + b_2 \log x_2 + \cdots)}}$$

TABLE 19

PROBABILITY OF INFANT OR CHILD MORTALITY

Bangladesh 1966-1976

	OLS Model		Probit Model		Logit Model	
Variable	Coefficient	t-ratio	Coefficient	t-ratio	Coefficient	t-ratio
Sex (1=male)	.02106	2.74	.17584	1.70	.05203	2.79
Parity	.00125	.75	.01340	.46	-.01337	.25
Mother's Age in years	-.000056	1.13	-.00801	.91	-.00257	1.63
Child's Age in years	-.00842	7.68	-.21783	7.26	-.22924	5.66
Interaction of Sex and Child's Age	-.00343	2.21	-.04167	.98	-.12434	2.14
Prior Birth Interval in Months:						
<12	.02797	1.98	.46518	2.24	.36444	.90
12-23	.01918	3.24	.37324	3.61	.56908	3.20
24-35	.01062	1.86	.26929	2.55	.51196	2.90
36≤	----	----	----	----	----	----
Constant	.06124	5.26	-1.4395	7.28	-1.98067	5.66
	R^2 = .03567		Log likelihood= 587.6639 (pseudo R^2=.0414)		Log likelihood=-630.3.93	
	Mean of dependent variable = .0286		Chi-squared = 228.5481, df = 8		(-2) (log likelihood ratio) = 143.24, df = 9	
	Standard error of regression = .1639		Sample size = 5404 (155 deaths)		Sample size = 5404 (155 deaths)	
	Sample size = 5404					

Source: Edmonston et al. 1981, Tables 1 - 3.

The results of the models are broadly similar. The signs of the coefficients are in most cases in the expected direction. Except for parity, which is never significant, coefficients have the same signs across models and their levels of significance are about the same. Male children initially have a higher probability of death but the negative interaction between sex and age indicates that their rate of mortality declines faster with age. Both mother's age and child's age are negatively associated with mortality. Prior birth interval as it increases, is expected to reduce the probability of death and the first two models agree with this expectation although the logit model does not.

The agreement among the three model is reassuring although it is not unexpected. The OLS model suffers from statistical limitations — some of which were noted briefly above — which makes it theoretically somewhat unappealing, particularly since the estimated coefficients are very sensitive to the fact that there is a preponderance of observations in which a child lives (i.e. the dichotomous dependent variable is usually 0 and only in very few cases assumes a value of 1). However it is much cheaper to estimate than the alternative models and is "useful for an initial scanning of the relative importance of major variables," (Edmonston et al. 1981, p.6). The probit model is usually less costly than the logit model although the latter is often preferred because of its theoretical appeal (see Pindyck and Rubinfeld 1976).

6.4 Discriminant Analysis

Assume that a particular effort at collecting nutritional data in a certain country has had as its objective the collection of weight-for-age data on all preschoolers. These data could allow indicators of nutritional status to be derived for all the districts (about 1500) in the country. This information, apart from having a descriptive value, is hoped to be used by the government for purposes of targetting and possibly of evaluation. The coverage of the weighing program is uneven. For over half the districts no weight-for-age data are reported. For some of the remaining districts it is believed that the reported figures may not be very reliable if their coverage, i.e. the proportion of children weighed, is either too low or too high. When the coverage is too low, say below 20 or 30 percent, the sample may be "unrepresentative" of the population of children in a district. On the other hand a coverage of more than about 90 percent of the children may suggest that not all the children for whom weight-for-age data were reported were actually weighed. By specifying a range of

acceptable levels of coverage it is possible to identify several hundred districts whose weight-for-age data can be assumed to be adequately reliable. Along with this information, there exist some descriptive data on various characteristics of all districts in the country, much of it derived from a previous census. These include total population, population density, number of villages, climatic conditions, information on major crops, total number of physicians, etc. A question of major interest can now be posed in the following general terms: given the information on nutritional status and descriptive data for several hundred districts is it possible to determine the likely nutritional status in other districts with the knowledge of their descriptive characteristics? The answer is indeed yes and a formal technique that helps in this regard is discriminant analysis.

In the situation above, and in many real-life situations that are similar, it is desired to find a set of characteristics that enable a researcher to distinguish, or discriminate, between two or more groups. These groups are defined a priori, i.e. before the techniques of discriminant analysis are used. For example, in the situation above we may distinguish two groups of districts: those with more than, say, 20 percent malnourished children and those with less than 20 percent. We may of course define more than two groups: districts with less than 10 percent malnourished children, those with between 10 and 20 percent, between 20 and 30 percent, etc. These groups of districts might exhibit different characteristics that can be used to tell them apart. Except for the criterion used in defining them, percent of malnourished children, it is unlikely that any single characteristic, say their size or income per capita, could completely distinguish them from one another. A collection of characteristics usually works better; even then it would be a rare instance if all districts could be correctly identified as belonging to a particular group. The purpose of discriminant analysis is to find a set of characteristics, or variables, that when used in combination, can best distinguish one group from another. To explain the essentials of the discriminant analysis it would be useful to confine ourselves to a situation with three groups: for example, districts with more than 45 percent malnourished children (worst affected), those with between 20 and 45 percent (moderately affected) and those with less than 20 percent (least affected). (For an introductory treatment of the two-group discriminant analysis, Nie et al. (1975) or Kleinbaum and Kupper (1978) may be consulted. More advanced exposition of the general technique is available in most books on multivariate analysis including Kendall 1975 and Green 1978.)

Suppose data are available on several **"discriminating variables"**, x_i, for each member of the three groups, such as on some district characteristics. Two (in this case) **"discriminant functions"**, L_1 and L_2, can be formed by linearly combining the standardized values of these discriminating variables:

$$L_i = b_{i1} x_1 + b_{i2} x_2 + \cdots + b_{ip} x_p \qquad i = 1, 2$$

where p is the number of variables. If the values of b's were known, the above discriminant functions could be evaluated for each member of each group. There are as many values of L_1 and L_2 in each group as there are members and these values will in general be different from one another. However if the functions are to be able to best discriminate among the three groups on the basis of values of L_1 and L_2, the b's should be chosen such that:

(a) The means of L_1 values in the three groups are as different as possible;

(b) L_1 and L_2 are uncorrelated;

(c) The three group means of L_2 values are as different as possible.

The discriminant functions above include all the variables on which data are available. This need not always be the case. It may well happen that only a few variables would be necessary, and additional variables would not add significantly to the discriminating power of the functions. Computer programs such as SPSS provide options for the concurrent inclusion of all the variables or for several stepwise procedures that in general bring the variables into the functions one by one in the order of their contribution, measured according to one of several possible criteria. (For a discussion of alternative stepwise procedures, see Nie et al. 1975.)

If the underlying assumptions of the method of discriminant analysis are met, statistical testing of the derived coefficients is possible. A discussion of these issues is not attempted here. However, there is a simple method of partially evaluating the results of a discriminant analysis. Once the discriminant model is specified, it can be used to classify those members already known to belong to each of the three groups. A measure of the probability of assigning a member to the wrong group is given by the error rate or the mis-classification rate. There are two such error rates, one for each group. These error rates usually overestimate the discriminating power of the analysis since they are

after all based, through the discriminant functions, on the knowledge of the actual classification. When the function is used to determine the groups to which unclassified members belong it is likely that the error rates would in fact be higher.

One of the results of the discriminant analysis is a classification table which provides a measure of how well the model predicts. It can be seen in Table 20 that only 35 percent of the 815 districts whose nutritional group is known are correctly classified. This is not a very high percentage. The model has a tendency to allocate districts to extreme nutritional groups (worst and least affected). However, the rate of correct classification of worst effected districts is 76 percent which is quite high. This model therefore is much more useful if it is used for the identification of the worst affected districts rather than of all types of districts. The model is relatively poor primarily because the correlation between the discriminating variables and prevalence of malnutrition in a district are not very high. There are 683 cases whose actual prevalence grouping is not known but they are predicted to belong to one of the three groups. The breakdown is given in the last row of the table.

Another example comes from a rare application of this technique in the literature concerned with the nutritional status of children. In a study based on data from the nutrition module of the Kenyan Integrated Rural Survey (Hitchings/CBS 1979), Jon Hitchings attempts to identify the food group that best predicts which children had low, medium and high values of HA or of WH within each agroecological zone. Note that this application of the discriminant analysis has a different purpose than that described earlier. The purpose is not to classify districts in nutritional groups, but to find the most important food group that can discriminate among groups of children already known to belong to one or the other. The purpose is analysis, not classification.

There are 1372 children in this study with their HA and WH calculated previously. These two indicators are used in turn to define three groups of children with low, medium and high nutritional status in the following way:

Definition of Groups of Children by Nutritional Status

		Low	Medium	High
First case	HA	<90%	90-95%	>95%
Second case	WH	<90%	90-100%	>100%

TABLE 20

CLASSIFICATION RESULTS OF DISCRIMINANT ANALYSIS

Actual Nutritional Group	No. of Cases	Predicted Nutritional Group		
		Worst Affected	Moderately Affected	Least Affected
Worst affected	66	50 (76%)	8 (12%)	8 (12%)
Moderately affected	650	236 (36%)	187 (29%)	227 (35%)
Least affected	99	22 (22%)	31 (31%)	46 (47%)
Unknown	683	246 (36%)	188 (28%)	249 (36%)

Proportion of known cases correctly classified $= \frac{50 + 187 + 46}{66 + 650 + 99} = 35\%$

The agroecological zones, the number of children in each zone, food groups and the average monthly frequency of their consumption by the children in each zone are given in Table 9.21.

To gain an understanding of the problem more readily let us concentrate on the first zone (Tea West of Rift Valley) with 174 children. These children on the average consume cereals more frequently (48 times per month) than other food groups. Vegetables and milk are consumed about as often as each other but somewhat less frequently than the cereals. "Potatoes/Cassava", "Bananas", "Beans", and "Meat/Fish/Eggs" food groups are however rarely used, on average only about one to two times a week. It is of course clear that, as averages, these frequencies mask differences that may exist among the children in this zone. Using HA as an indicator, these children can be divided up into three groups with low, medium and high values of this ratio as defined above. It may be postulated that children in each group may have on average a specific consumption frequency pattern that is significantly different from the consumption patterns of the other two groups. For example those with low HA may tend to consume meat, fish and eggs less frequently than the children in the medium and high HA groups but rely more frequently on cereals for subsistence. The question of interest is to find one food group whose frequency of consumption best predicts the nutritional group to which a given child is likely to belong.

In principle this question can always be answered for each zone and for each indicator, HA and WH. One may wish however to set a priori guidelines that result in statistically significant results and also results that can be easily interpreted. Hitchings chose to consider only the first food group with the highest discriminatory power in each zone which satisfies the following three criteria:

1. The average frequency of consumption in the low, medium and high groups follows a consistent pattern, i.e. the trend is either decreasing or increasing. Patterns like 5, 20 and 10 are not allowed;
2. The average difference in the consumption rates between the high and low groups is more than 10 times per month, i.e. a pattern such as 5, 7 and 9 is ignored;
3. The overall F-statistic for the food group exceeds 2.5.

TABLE 21

AVERAGE MONTHLY FREQUENCY OF FOOD GROUP CONSUMPTION

BY ECOLOGICAL ZONE (PROBABLY UNDERESTIMATED)

Ecological Zone	Zone Code	n	Cereals	Potatoes Cassava	Bananas	Beans	Vege-tables	Meat Fish Eggs	Milk (Cow & Goat)
Tea West of Rift Valley	1	174	48	7	5	7	37	5	36
Coffee West of Rift Valley	2	201	41	9	5	4	29	7	30
Upper Cotton West of Rift Valley	3	246	52	13	6	3	24	8	23
Tea East of Rift Valley	4	93	33	22	15	30	12	5	32
Coffee East of Rift Valley	5	202	37	27	14	30	21	8	36
Upper Cotton East of Rift Valley	6	60	39	17	15	23	13	8	19
Lower Cotton East of Rift Valley	7	152	53	7	4	32	7	8	36
High Altitude Grass-lands	8	114	41	30	3	6	30	5	35
Coast	10/11	81	43	4	4	8	11	8	27
COMBINED	ALL	1372*	45	11	7	17	20	6	35

*Note: Includes some zones not shown.
(Hitchings/CBS 1979, p. I-5-10)

The final results of the discriminant analysis are given in Table 9.22. Note that not all agroecological zones appear in the table. Furthermore, although both indicators HA and WH were considered in each zone, in no zone could food groups be found that satisfied the above criteria in the case of both indicators. It is also interesting to note that the data discriminated more often groupings defined according to WH than HA. Hitchings interprets this result as a reflection of the fact that consumption data and WH both pertain to recent events whereas HA reflects past nutritional status. Zones 10 and 11 were excluded since it was felt that the samples of children in these zones were too small.

6.5 Factor Analysis and Principal Components Analysis

In multivariate analysis a typical data set consists of n observations on p variables arranged in the format of an n x p matrix. An observation pertains to an entity such as a village, a household, an individual, etc. The variables refer to the measures of different characteristics of these entities. For example a figure of 32 in row 18 and column 4 of a data matrix might refer to the age of a 32-month-old child who is 18th in the listing of children. Age is the fourth variable recorded. These variables can be viewed as the elements of a p-element **vector variable.** The number of parameters associated with this vector variable is $2p + p(p-1)/2$ of which p parameters refer to the means of the variables, another p parameters to their variances and the rest, $p(p-1)/2$, to the covariances, or more familiarly, correlations among all possible pairs of them. These correlations are ordinarily not zero since the variables are usually interdependent. It can easily be seen that the number of parameters to be estimated and interpreted rises extremely rapidly as the number of variables, p, increases. It is also clear that the primary reason for this phenomenon lies in the existence of correlations among the intercorrelated variables. Given that such intercorrelations are common, how can one transform the data so as to construct an uncorrelated vector variable having 2p (or even fewer) parameters, p means and p variances?

To give some intuitive meaning to this question consider a case where data on some 15 variables are available on many individuals. The choice of these 15 variables is governed by a variety of considerations such as ease of measurement, usefulness for targetting purposes, etc. It often happens that several of these variables may actually measure more or less the same thing, say wealth of an individual. Type of housing, possessions in the household, education

TABLE 22

SUMMARY OF DISCRIMINANT ANALYSIS
OF FOOD GROUPS BY ECOLOGICAL ZONE

Ecological Zone	Nutrition Variable	Food Group	Mean Monthly Consumption Frequency of Subgroups			F Statistic	Food Group Rank
			Low	Medium	High		
1	HA	Vegetables	27	35	43	5.2	1
2	WH	Cereals	33	41	48	8.6	1
3	WH	Vegetables	34	25	19	6.8	1
4-7	None	-	-	-	-	-	-
8	WH	English Potatoes	19	33	33	3.2	2
10/11	Omitted	-	-	-	-	-	-

NOTES:

1 The 'low', 'medium' and 'high' subgroups for HA were less than 90 per cent, 90-95 per cent and more than 95 per cent. The 'low' group was the same for WH. The other groups were 90-100 per cent, more than 100 per cent. Children in zone 1 having an HA below 90 per cent took vegetables an average of 27 times a month.

2 A food group rank of 1 means that no other food group in the zone had a higher overall F-statistic with respect to that nutrition variable.

3 The tubers in zone 8 are known to be English potatoes, not cassava or sweet potatoes.

(Hitchings/CBS 1979, p. I.5-10)

of the household head, etc. are normally expected to be highly correlated not only among themselves but also individually with wealth. In one way or another they are all measures of wealth. Several others of the 15 variables may in turn be various measures of another **"factor"** such as household composition which may be deemed to be essentially independent of the other "factor", wealth. It would be helpful if these factors could be identified. The advantages of transforming the original variables in this manner are that, first, the number of variables could be reduced substantially to a smaller number of important "factors" and, second, the intercorrelations among pairs of original variables can be removed by finding relatively independent factors.

Factor analysis (FA) consists of a set of techniques that can sometimes provide useful transformation of the original data resulting in new insights into their structure. It is based on the premise that a few **underlying factors** account for the covariances (or correlations) among the observed variables which are comparatively more numerous. Its purpose is to identify these underlying factors from the knowledge of the correlations among the observed variables. An attempt is then made to understand and interpret the nature of these factors as well as to name them. This is usually the end of factor analysis. The results however can be used in further analytical stages employing other techniques.

Identifying and interpreting underlying factors is easier said than done and, besides, the theoretical underpinnings and usefulness of factor analysis have been questioned (see, e.g., Williams 1971 and the references given therein). However the idea of trying to reduce the **dimensionality** of data by combining variables to construct fewer, relatively uncorrelated, composite variables can be of great practical value. This brings us to another multivariate technique, the principal components analysis (PCA), which has a lot in common with factor analysis. In fact the PCA is one means of finding the factors in factor analysis. The purpose of PCA is to find p uncorrelated linear functions of the original variables called **"principal components"**. There are as many principal components as there are original variables and their advantage over the original variables stems from the fact that they are constructed in such a way as to be independent of one another. Furthermore, the "first" principal component is chosen so as to maximize the amount of variation in the original data that it can account for. The "second" principal component is sought such that, subject to being uncorrelated with the first component, explains as much of the remaining variation in the original data as possible. The "third" principal component is uncorrelated with the first two

and maximizes the amount of variation explained of the remainder, and so on. The pth principal component accounts for the rest of the variation and is uncorrelated with all the other components.

It may be useful to point out some areas of application of principal components analysis. It should be noted that the number of uncorrelated principal components is always equal to that of the original observed variables. Since the correlations in the observed variables are removed by the construction of the principal components the number of parameters to be found is considerably reduced. But this is almost never the end of the story. The main purpose of the PCA is to reduce the dimensionality of the original data. Suppose that the first few principal components account for much of the variation in the original data. For instance, consider a situation where the first two or three principal components explain about as much of the variation in HA as do the 10 or 15 observed variables from which the principal components are derived. It would then be possible to discard, at the expense of a little loss in accuracy, all but the first few principal components which are then taken to represent a larger number of correlated observed variables. This procedure is, for example, sometimes recommended to deal with the problem of multicolinearity in regression analysis. This use of the technique may, however, give rise to some difficulties in interpretation. A principal component is a composite variable that lumps together variables with possibly different units. Altering the units by which the observed variables are measured can furthermore result in different sets of principal components which is an obviously serious drawback. One can usually get around these problems by standardizing all the original variables before any principal components are extracted. Standardized units are, however, sometimes awkward to work with.

The choice of the first few principal components can have more useful applications. We cite two examples from the literature relating to nutritional status and basic needs. The first pertains to a suggestion by E.P. Lozy (1972a) to the effect that the first two principal components derived from anthropometric and/or biochemical tests could be interpreted as "size" and "shape" factors, respectively. The size factor conveys an overall picture of the nutritional status, whereas the shape factor might provide an indication of whether a malnourished child is heading along the marasmus or kwashiorkor "line of development". In another short paper (1972b) he examines the first three principal components extracted from a correlation matrix derived from

measurements of eight anthropometric variables by Robinow and Jelliffe (1967). He finds that the first component is the sum of eight observed variables with almost identical weights. This component is the "size" factor above and explains some 60 percent of the variation in the measured variables. The second and third components account for 13 and 9 percent of the variation, respectively, and are taken together to represent the "shape" factor.

The second example relates to a study of basic needs performance in 110 less developed countries by Sheehan and Hopkins (1979). Data available to them pertain to eight basic needs output indicators: calorie consumption, protein consumption, population per doctor, population per nurse, life expectancy, infant mortality, literacy and primary school enrollment rate. They used the principal components technique to seek out the "best" indicators of basic needs performance from among the above eight indicators. Their results are reproduced in Table 23. The first two components account for slightly more than three-fourths of the total variation in basic needs output indicators. In the first the three indicators life expectancy, infant mortality and literacy rate figure prominently. The second component is heavily influenced by calorie and protein consumptions. These two being highly correlated, they conclude that the four indicators of life expectancy, infant mortality, literacy rate and calorie consumption are sufficient to represent the totality of output indicators, albeit not perfectly. This reduction in the number of indicators is a simplifying step to reduce data to manageable proportions. These four indicators are used in their work as input into a further stage utilizing discriminant analysis.

6.6 Cluster Analysis

Consider a multivariate n x p matrix wherein observations on p variables referring to a collection of n individuals, households, objects, or simply entities, are recorded. The number of entities n, is usually far larger than the number of variables, p. A typical sample survey, for example, might provide data on several characteristics of a few hundred localities or a few thousand households. In reporting a summary of such data, or for purposes of analysis, observed entities are ordinarily classified or grouped according to one or several given criteria. The choice of such criteria is naturally guided by the type of information it is desired to convey. In this sense the choice is fairly arbitrary. One may be interested, for instance, in grouping the prevalence of second and third degree malnutrition by location (municipality, province, etc.), by income

TABLE 23

Results of principal component analysis of basic needs output indicators

Variable (all were 1970)	Factors 1	2	3	4	5	6	7	8
Calorie consumption	.24	.86	-.17	-.13	.22	.01	.08	.31
Protein consumption	.17	.96	-.01	-.10	.04	.03	-.01	-.17
Population per doctor	-.35	-.09	.90	.15	-.18	-.03	-.03	-.02
Population per nurse	-.17	-.14	.13	.96	-.10	-.01	-.01	-.01
Life expectancy	.77	.33	-.29	-.11	.25	.36	.07	.00
Infant mortality	-.87	-.17	.27	.23	-.20	.15	.14	-.01
Literacy	.77	.25	-.26	-.12	.34	.06	.38	.05
Primary enrolment rate	.51	.19	-.26	-.17	.78	.03	.04	.02
Eigen value	4.87	1.22	.79	.45	.30	.17	.13	.07
% of variation explained	60.9	15.3	9.8	5.6	3.7	2.2	1.6	0.9

(Sheehan and Hopkins 1979, p.55)

bracket, by household size, etc. or by a combination of such criteria. Sometimes, however, it may be interesting to ask whether, given the data on the measured variables, there exist identifiable or "natural" groupings or **"clusters"** among the observed entities. In other words, we may wish to find out if the collective information on the observed entities can lead to the emergence of a (new) classification or pattern of clusters having an interesting feature. Techniques of cluster analysis are designed to respond to this question.

There is a multitude of clustering methods proposed in the literature (see Everitt 1974 and 1977). This is in large part due to the fact that no consensus exists as to what constitutes a cluster. Roughly speaking, a cluster is a collection of more or less **homogeneous** entities. Most methods of clustering are able to form clusters of similar entities, even from very large data sets, but there is no guarantee that they all yield same result. This can be something of a nuisance since the choice of an approriate technique is by no means straightforward. We shall, however, ignore such problems here since our main concern is only to introduce the subject of cluster analysis.

The available methods of cluster analysis are, in most cases, nonstatistical in the sense that no distributions or inferences are involved and no hypotheses are tested. Instead, the application of some technique or other to the data, often treated as the population rather than a sample, is expected to generate one or more hypotheses which will subsequently be tested. As such, cluster analysis can serve as a preliminary step to organize, summarize and/or describe the data so that statistical techniques of analysis can then be used.

The initial stage in most clustering methods is to seek a measure of **distance,** or alternatively of **similarity,** between pairs of observed entities on the basis of the values of their characteristics. A common example of the former is **Euclidean distance,** d_{ij}, defined as:

$$d_{ij} = d_{ji} = \left\{ \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right\}^{1/2}$$

where x_{ik} is the value of the kth characteristic for the ith entity. This means, in its simplest form, the difference between values in two cells of a matrix. The larger the "distance" between two entities, the greater is the **dissimilarity** between them, and the smaller the "distance", the more alike they are assumed to be. Alternatively one could measure the same thing through some kind of

similarity coefficient, ordinarily assumed to take a value in the range 0 to 1. Pearson's product moment correlation coefficient is, for example, a well-known coefficient of similarity. From one such measure one can form an n x n matrix of pairwise distance or similarity coefficients (a correlation matrix). If the number of entities, n, is not large, clusters may be formed merely by inspection of the distance or similarity measure (for an example see Kendall 1975, pp, 34-5). Usually, however, the task is performed by a computer. But it has to be told first how to do it.

To appreciate better the concept of cluster analysis let us take a simple illustration of a particular clustering technique known as **group average clustering.** Consider the following 6 x 3 data matrix which refers to six entities or individuals on whom the values of three characteristics are recorded (e.g. size of household, number of cows, monthly income or some such variable):

$$x = \begin{bmatrix} 4 & 9 & 0 \\ 1 & 2 & 5 \\ 9 & 7 & 8 \\ 3 & 0 & 0 \\ 6 & 6 & 2 \\ 9 & 8 & 1 \end{bmatrix}$$

Our first job is to compute a matrix of Euclidean distances between all possible pairs of these six entities. As an example, the distance between the second and last entities is calculated as follows:

$$d_{26} = \sqrt{(1-9)^2 + (2-8)^2 + (5-1)^2} = 10.77$$

An observant reader may object that we are here mixing apples and oranges. To deflect this criticism each variable can be standardized first to remove the effect of the measurement units (the usual approach) or a weighting scheme introduced or both. To get the main point across and to keep the numbers simple these corrections are ignored for the moment. The resulting distance matrix, D_1, is then:

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 0 & 9.11 & 9.64 & 9.06 & 4.12 & 5.20 \\ & 0 & 9.90 & 5.74 & 7.07 & 10.77 \\ & & 0 & 12.21 & 6.78 & 7.07 \\ & & & 0 & 7.00 & 10.05 \\ & & & & 0 & 3.74 \\ & & & & & 0 \end{bmatrix} \end{array}$$

To begin with, assume that each entity constitutes a cluster. There are six of them now. Our immediate purpose is to join the two clusters that are closest to each other, i.e. the two clusters that are more similar than any other two clusters. Excluding diagonal zeros which represent the distance of a cluster to itself, the two closest clusters are 5 and 6 since their distance of 3.74 is the smallest number in D_1. Joining these two reduces the number of clusters by one from six to five.

Next it is necessary to compute a new distance matrix D_2 pertaining to the five clusters we now have. The distances between any pair of the first four clusters are still the same since they are single-membered. They can be lifted directly from D_1. We have to compute, however, the distances between each of these four and the last cluster containing two members. In group average clustering the rule is to take the arithmetic average of distances from a single-membered cluster to each of the members of the multi-membered cluster to represent the distance between the two cluster. For instance, the distance between the second cluster and the last cluster containing entities 5 and 6, is given by:

$$d_{2,(56)} = 1/2\,(d_{25} + d_{26}) = 1/2\,(7.07 + 10.77) = 8.92$$

Similar calculations give the distances between each of clusters 1, 3 and 4, and the last two-membered cluster. The new 5 x 5 distance matrix D_2 is given by:

$$D_2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ (56) \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & (56) \end{array} \\ \begin{bmatrix} 0 & 9.11 & 9.64 & 9.06 & 4.66 \\ & 0 & 9.90 & 5.74 & 8.92 \\ & & 0 & 12.21 & 6.92 \\ & & & 0 & 8.52 \\ & & & & 0 \end{bmatrix} \end{array}$$

Since the smallest number in D_2 is 4.66 a new three-membered cluster is formed by joining the first and the last clusters together. A new 4 x 4 distance matrix D_3, is formed along the same lines as set out above and the result is:

$$D_3 = \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ (156) \end{array} \begin{array}{c} \begin{array}{cccc} 2 & 3 & 4 & (156) \end{array} \\ \begin{bmatrix} 0 & 9.90 & 5.74 & 8.98 \\ & 0 & 12.21 & 7.83 \\ & & 0 & 8.70 \\ & & & 0 \end{bmatrix} \end{array}$$

The next new cluster is formed by joining single-membered clusters 2 and 4 since their distance, 5.74, is the smallest number in D_3. We now have three clusters, namely:

	Cluster 1	Cluster 2	Cluster 3
Members	2,4	3	1,5,6

The corresponding distance matrix, D_4, is computed to be:

$$D_4 = \begin{array}{c} \\ (24) \\ 3 \\ (156) \end{array} \begin{array}{c} \begin{array}{ccc} (24) & 3 & (156) \end{array} \\ \begin{bmatrix} 0 & 11.06 & 8.84 \\ & 0 & 7.83 \\ & & 0 \end{bmatrix} \end{array}$$

where, for example, $d_{(24),\,(156)}$ is calculated in the following way:

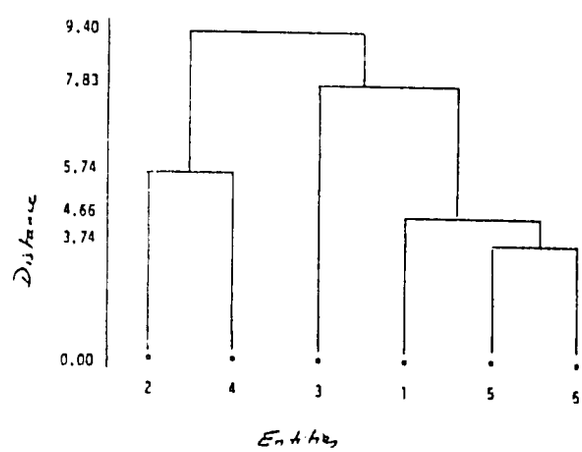
$$d_{(24),\,(156)} = 1/6\,(d_{21} + d_{25} + d_{26} + d_{41} + d_{45} + d_{46}) =$$

$$(1/6)(9.11 + 7.07 + 10.77 + 9.06 + 7.00 + 10.05) = 8.84$$

In the next step, the cluster containing entity 3 joins the last cluster and the final step fuses the two remaining clusters into one containing all six entities.

The series of steps above can be visualized with the aid of a **dendrogram** which is a diagram showing the fusion of clusters at each step of the procedure. For our example of group average clustering the dendrogram is shown below:

Group Average Clustering Dendrogram



The group average clustering is one example of a large number of **hierarchical methods** which begin by considering each entity as a single-member cluster. At each stage of the procedure the two closest clusters are joined together until a single cluster encompasses all entities. There are n - 1 such stages. Hierarchical techniques can also be worked backwards. Initially, a single cluster contains all entities. Successive stages then break it down into smaller and smaller clusters. The former are referred to as **agglomerative methods** and the latter are known as **divisive methods.**

In contrast to the hierarchical methods, **optimization methods** of clustering attempt to partition the observed entities in such a way as to optimize a given clustering criterion. Many such criteria have been proposed and one common example is the minimization of the total of sums of squares within all clusters.

Clearly the variety of possible criteria that one can conceive, the number of clusters to be had, the rules of aggregation of clusters and a host of other related factors give rise to a large number of techniques of cluster analysis. No comprehensive review of the comparative advantages and disadvantages of these techniques seems to be available yet.

The use of cluster analysis as a tool for aggregating observations in nutrition related studies has so far been infrequent. One such application is that by Crawford and Thorbecke (1978). They make an attempt to identify regional poverty groups in Kenya using the data for a sample of 139 sublocations, each with 12 households, from the Integrated Rural Survey carried out in 1974/75 and published in 1977. The variables of interest to them were only two: total household consumption and geographical location. However, other variables such as indicators of household amenities and access to government services could have been included as well. Their purpose is to find the grouping scheme that minimized the within-cluster variance by comparison to the variance between clusters. A fixed number of clusters ranging from five to twelve were set in each run. The geographical location is identified by latitude and longitude and a scaling factor of 1:50:50 was applied to the consumption, latitude and longitude respectively. In their words, "in spite of the heavy weight given to the geographical variables, the program generated a set of clusters that were relatively homogeneous in terms of consumption levels, but not in terms of geographical location" (ibid, p. 15). To them this result points to the diversity of rural Kenya, with rich and poor found in the same region. The regional clusters, while reflecting rather homogeneous consumption levels are, they point

out, "inconveniently shaped from the standpoint of targetting programs to meet the needs of their inhabitants". The pattern of regional distribution of welfare levels as a result of the application of cluster analysis does not appear to suggest any underlying causal mechanism but they regard the method of cluster analysis "promising in general".

It may be helpful here to contrast the methods of factor analysis and cluster analysis. Consider a multivariate data matrix with the sample or population entities appearing in rows and the observations on variables relating to them in columns. From the discussion of these two techniques it is seen that factor analysis attempts to find a useful grouping of the variables (columns), whereas cluster analysis seeks to find useful groupings of the entities on which observations are made. While the rationale for these two techniques are entirely different it should be clear that there are some similarities between them. In essence they are both methods of classification, one of variables and the other of observed entities. The parallel between them may suggest that it should be possible to use factor analysis to group observations and cluster analysis to find groupings of variables. This is indeed the case although the interpretation of the first is not clear. An example of the application of cluster analysis to group variables is given by Kendall (1975). The problem is usually much simpler since the number of variables is often fairly small and the mere inspection of the correlation matrix is sufficient to give the required groupings.

REFERENCES

Anderson, S., Auquier, A., Hauck, W.W., Oakes, D., Vandaele, W. and Weisberg, H.I. 1980. Statistical Methods for Comparative Studies: Techniques for Bias Reduction. New York: John Wiley & Sons.

Asrat, D. 1980. Some notes on the problems associated with the use of various indicators of nutritional status. Paper II. Mimeo.

Bhattacharyya, G.K. and Johnson, R.A. 1977. Statistical Concepts and Methods New York: John Wiley & Sons.

Blalock, H.M. 1972. Social Statistics, 2nd ed. New York: McGraw-Hill.

Crawford, E. and Thorbecke, E. 1978. The analysis of food poverty: an illustration from Kenya. Mimeo.

Drake, W.D.; Miller, R.I.; and Humphrey, M. 1980. Final Report: Analysis of Community-Level Nutrition Programs. Project on Analysis of Community Level Nutrition Programs. Vol. I. Washington, D.C.: USAID - Office of Nutrition.

Draper, N.R. and Smith, H. 1981. Applied Regression Analysis, 2nd ed. New York: John Wiley & Sons.

Edmonston, B., Green, W. and Smith, K. 1981. Multivariate Analysis of Survival Data: An Appraisal Using Bangladeshi Mortality Data. Working Paper No. 19. Infant and Child Mortality Project, International Population Program, Cornell University.

El-Lozy, M. 1972a. The assessment of nutritional status by composite measurements. The Journal of Pediatrics and Environmental Child Health, 18.1:3-4.

El-Lozy, M. 1972b. Multivariate analysis in nutritional anthropometry (An analysis of the Robinow-Jelliffe Correlation Matrix). The Journal of Pediatrics and Environmental Child Health, 18:132-3.

Everitt, B.S. 1974. Cluster Analysis. London: Heinemann Press.

Everitt, B.S. 1977. Cluster analysis. In Exploring Data Structures, C.A. O'Muircheartaigh and C. Payne, eds. Vol. I, chap. 3. New York: John Wiley & Sons.

Fleiss, J.L. 1973. Statistical Methods for Rates and Proportions. New York: John Wiley & Sons.

Francis, I., ed. 1979. A Comparative Review of Statistical Software. Voorburg, Netherlands: International Association for Statistical Computing.

Greene, P.E. 1978. Analyzing Multivariate Data. Hinsdale, Illinois: Dryden Press.

Gujarati, D. 1978. Basic Econometrics. New York: McGraw-Hill.

Hansen, M.H.; Hurwitz, W.N.; and Madow, W.G. 1953. Sample Survey Methods and Theory, Vol. 1 and 2. New York: John Wiley & Sons.

Hays, W.L. 1973. Statistics for the Social Sciences. 2nd ed. New York: Holt, Rinehart and Winston, Inc.

Hitchings/Central Bureau of Statistics, Kenya. 1979. Child Nutrition in Rural Kenya. Nairobi: UNICEF.

Kendall, M.G. 1975. Multivariate Analysis. New York: Hafner Press.

Kish, L. 1965. Survey Sampling. New York: John Wiley & Sons.

Kleinbaum, D.G. and Kupper, L.L. 1978. Applied Regression Analysis and Other Multi-variable Methods. North Scituate, Mass.: Duxbury Press.

Knoke, D. and Burke, P.J. 1980. Log-Linear Models. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-020. Beverly Hills and London: Sage Publications.

Lapin, L. 1978. Statistics for Modern Business Decisions. New York: Harcourt Brace Jovanovich, Inc.

Lewis-Beck, M.S. 1980. Applied Regression: An Introduction. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-022. Beverly Hills and London: Sage Publications.

Liberia. 1976. Liberia National Nutrition Survey. Washington, D.C.: USAID - Office of Nutrition.

Mason, J.B., Habicht, J-P., Tabatabai, H. and Valverde, V. 1981. Nutritional Surveillance: Review of Progress. Cornell Nutritional Surveillance Program.

Mason, J.B., Habicht, J-P., Tabatabai, H. and Valverde, V. 1982. Nutritional Surveillance. Cornell Nutritional Surveillance Program.

Mausner, J.S. and Bahn, A.K. 1974. Epidemiology: An Introductory Text. Philadelphia: W.B. Saunders Company.

Morris, C.N. and Rolph, J.E. 1981. Introduction to Data Analysis and Statistical Inference. Englewood Cliffs, N.J.: Prentice-Hall.

Nie, N.H.; Hull, C.H.; Jenkins, J.G.; Steinbrenner, K.; Bent, D.H. 1975. Statistical Package for the Social Sciences. New York: McGraw-Hill.

O'Muircheartaigh, C.A. and Payne, C. eds. 1977. Exploring Data Structures, Vol. I. New York: John Wiley & Sons.

Pindyck, R.S. and Rubinfeld, D.L. 1976. Econometric Models and Economic Forecasts. New York: McGraw-Hill.

Robinow, M. and Jelliffe, D.B. 1967. In Proceedings of the Seventh International Congress of Nutrition, Hamburgh 1966, Vol. 4, 144.

Roscoe, J.T. 1975. Fundamental Research Statistics for the Behavioral Sciences 2nd ed. New York: Holt, Rinehart and Winston, Inc.

Sanjur, D. and Romero, E. 1972. Conceptual levels of dietary indicators as predictors of nutritional status. In Proceedings of the 9th International Congress on Nutrition, Mexico, Vol. 2, pp. 214-22.

SAS Institute. 1979. SAS User's Guide. Cary, North Carolina: SAS Institute Inc.

Sheehan, G. and Hopkins, M. 1979. Basic Needs Performance. Geneva: ILO.

Snedecor, G.W. and Cochran, W.G. 1980. Statistical Methods. 7th ed. Ames, Iowa: Iowa State University Press.

Williams, K. 1971. Do you Sincerely want to be a Factor Analyst? AREA, Institute of British Geographers, Vol. 3, No. 4.

Winkler, R.L. and Hays, W. 1971. Statistics: Probability, Inference, and Decision. New York: Holt, Rinehart and Winston, Inc.

Wonnacott, R.J. and Wonnacott, T.H. 1979. Econometrics. New York: John Wiley & Sons.

107

CNSP WORKING PAPER SERIES

Currently available:

1. Nutritional Surveillance: A synopsis. June 1981.
2. Nutritional Surveillance. August 1981.
3. Surveillance Summaries. January 1982.
4. Analytical Methods for Nutritional and Socio-economic Data. April 1983.
5. Principles for Evaluation of On-going Programs. May 1982.
6. An Outline Plan for Developing a Nutritional Surveillance System in New York State. May 1982.
7. Initial Assessment for Nutritional Surveillance in the Philippines. Forthcoming.
8. Preliminary Analysis of Nutritional and Agricultural Data, Eastern Province, Kenya. August 1982.
9. Static and Dynamic Poverty Indicators: An Indonesian Example. May 1983.
10. Reducing Malnutrition Through Rural Development: A test of Planning Methods in Haiti. June 1983.
12. Malnutrition by Socio-Economic Classification in Palawan, the Philippines. June 1983.
13. Changes in Nutritional Status in the Philippines from 1979 to 1980. Forthcoming.
14. National Health Information Systems and their use in Nutritional Surveillance. January 1983.
15. On Sampling Variance of the Sensitivity of an Indicator. January 1983.
16. Note on a Framework for Monitoring and Evaluation of UNICEF/WHO Nutrition Improvement Programmes. March 1983.
17. On Error in the Estimate of Malnutrition Due to Bias and Random Error in Anthropometry and Age. June 1983.
18. Age Misstatement for Children: A Problem for Interpreting Anthropometric Measures in Bangladesh. June 1983

Additional Documents Available from CNSP:

Report of Workshop on Social and Nutritional Surveillance in Eastern & Southern Africa. May 17-19, 1982.

Background Papers for Workshop on Social and Nutritional Surveillance in Eastern & Southern Africa. May 17-19, 1982.

8/83