| BIBLIOGRAPHIC DATA SHEET | 1. CONTROL NUMBER PN-AAK-483 | 2. SUBJECT CLASSIFICATION (695) PC00-0000-0000 |
|---|---|---|

3. TITLE AND SUBTITLE (240)

applied survey methods for development projects

4. PERSONAL AUTHORS (100)

5. CORPORATE AUTHORS (101)

AID/PM/TD/MD

| 6. DOCUMENT DATE (110) 1981 | 7. NUMBER OF PAGES (120) 98p. | 8. ARC NUMBER (170) |
|---|---|---|

9. REFERENCE ORGANIZATION (130) PM/TD

10. SUPPLEMENTARY NOTES (500)

11. ABSTRACT (950)

12. DESCRIPTORS (920)

statistical analysis
statistical sample
project management
project analysis
AID
surveys
methodology

13. PROJECT NUMBER (150)

---

14. CONTRACT NO.(140) PM/TD

15. CONTRACT TYPE (140)

16. TYPE OF DOCUMENT (160)

AID 590-7 (10-79)

c.1

PN-AAK-483

# APPLIED SURVEY METHODS FOR DEVELOPMENT PROJECTS

Training and Development Division
Office of Personnel Management

**The U.S. Agency for International Development**
**Washington, D.C. 20523**

**September 1981**

# FOREWORD

AID Project Officers, and other decision makers use data as a basis for establishing targets, monitoring progress, taking corrective action, and evaluating overall project performance. Thus, analysis of data is an important aspect of any project officer's job, whether in the Design, Implementation or Evaluation stage. Most of the data they use -- whether qualitative or quantitative -- is (or can be reduced to) numerical or categorical entities for statistical manipulation.

Simple arithmetic tallying or percentage computations of available data will usually not suffice, however. There are inherent dangers associated with "number crunching" which can trap, or trip, the unwary. Knowing what is available is only the tip of the iceberg. Knowing what is needed, where and how it was obtained, and how to interpret it, are all closely interlinked functions; even when not performed by the same individual.

Generally, organizations spend a large portion of their effort gathering data, but very little in analyzing it once it has been obtained, other than the tallying and percentage computations. The principal reasons for this seem to be

1) lack of awareness of the hidden value in numbers

2) lack of skill in performing basic statistical analysis.

To improve our effectiveness in Development Project Administration, it is desirable that all those responsible for project formulation, monitoring and evaluation be familiar with the underlying principles and process of data gathering and analysis. Such an awareness will make them particularly cognizant of the fact that scientific sampling methods can obviate the need for much of the detailed 100% record-keeping, reporting and accounting that is currently done, while also being more useful.

All too often, essential aspects are presented in formidable statistical texts, which although technically complete, fail to communicate with the non-mathematically inclined. Hence, in too many instances, scientific statistical analysis has not been utilized, nor the benefits obtained by management. This booklet has therefore been prepared as a step-by-step illustrative guide, with the emphasis upon transmitting knowledge and creating understanding for subsequent application to the types of problems which are encountered in some aspects of AID agricultural development projects. However, the principles apply equally well to other fields and disciplines. No attempt has been made to provide an exhaustive treatment, nor to cover all the fine points of those topics selected. However, I have endeavored to familiarize the reader/user with the subject, and provide him/her with a working knowledge of the various concepts, together with "cook-book" examples to follow.

The basis for the text was initially developed during a tour in the Philippines, several years ago, and was used extensively in many different situations, both to train counterparts and their staffs in Agricultural, and Rural Development Programs, as well as for general, formal academic instruction. More recently, it was updated for the Office of Rural Development, USAID/Indonesia, to meet some of their program needs, and to incorporate more detail on "cluster sampling" as well as a few other minor modifications. Thus it has had extensive "field testing".

While the emphasis in these early developments has been to meet the needs of foreign national participants in AID-assisted projects, it has also become apparent that there is another large audience to be served within our own AID staff. It is an old, but nevertheless appropriate adage that "the cobbler's children are often without shoes"! In my current assignment with AID's Management Training Office in Washington, monitoring the Project Design and Evaluation, and Project Implementation Courses, I have become aware of the large number of project officers, and others, who could benefit by a more detailed (or updated) exposure to statistical analysis. This booklet is a step towards filling that need. First by familiarizing them with some of the tools; and second, by providing encouragement that mastery is not beyond their grasp.

I hope this will result in the application of better quantitative analysis in future Project Design, Implementation and Evaluation.

Kenneth F. Smith
Project Manager, PM/TD/MD
AID Washington Training Center
Washington, D.C. 20523

September 1981

When you can measure what you are speaking about, and express it in numbers, you know something about it.

When you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind. It may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the stage of science.

Lord Kelvin

CONTENTS

1 Definitions
2 Data Collection Methods
3 Comparison of Census & Sampling Data
4 Judgement Sampling
5 Scientific Sampling
6 Basic Statistical Concepts
7 Mean
8 Median and Mode
9 Range
10 Standard Deviation (Ungrouped Data)
11 Example of Calculating Standard Deviation
12 Frequency Distributions
16 Histograms
18 Calculating Standard Deviation (Grouped Data)
20 Normal Distribution Curve
22 Cumulative Probability
23 Percentages and Rank Ordering
24 Percentage Frequency Distributions
25 Determining Sample Size
26 Optimum Sample Size for Mean (Random Sampling)
27 Scientific Sampling Methods - Simple Random Sample
28 Standard Error of Mean
29 Confidence Interval and Standard Error of Mean
30 Systematic Random Sampling
31 Stratified Random Sampling
32 Weighting
33 Standard Error of Mean for Stratified Random Sample
34 Cluster Sampling
35 Selecting Samples with a Deck of Cards
36 Standard Error of Mean for Cluster Samples
41 Comparison of Cluster Sample with Random Sample of Same Size
42 Optimum Sample Size for Estimating a Percentage
43 Standard Error of a Percentage (Random Sample)
44 Confidence Interval and Standard Error of Percentage
45 The "T" Test for Estimating Confidence Intervals from Small Samples
46 Correlation
47 Linear Rank Order Correlation of Two Variables
48 Non-normal Distribution
49 Significance
50 The "Z" Test - Significance Testing for a Mean
52 TYPE I and TYPE II Errors
54 Significance Testing for a Percentage
56 The "Chi Square" Test for entire sample frequency distribution
62 Chi Square in Non-Normal Distributions
67 Regression Analysis
70 Designing the Survey
74 General Principles of Questionnaire Design
76 Preliminaries -- Before conducting the survey
80 Conducting the Survey
82 Analyzing the Data
85 Presentation of Results
86 Major Points in Writing Survey Reports
87 Briefings
88-93 APPENDICES -- Tables

# SOME DEFINITIONS OF TERMS

CENSUS A complete count or measurement of every unit in the population under study.

POPULATION The total collection of units of interest in a particular area under study. (Not necessarily "people".) For example: -

The number of households in a particular village in January 1979.

The number of cows in a particular Kabupaten in November 1980.

The number of Cangkul in Desa Salulo in December 1980.

SAMPLE A partial count or measurement of selected units (variables) in a population under study.

VALIDITY refers to the degree with which a measure or indicator does what it purports to do. Several of its characteristics are: -

Relevance A significant, substantial and demonstrable similarity with the variable being measured.

Bias A predisposition towards certain aspects which impairs the impartiality of the sample unit as representative of the population.

RELIABILITY is the extent to which a measurement produces stable, consistent results when repeated by different observers or under different conditions. Factors contributing to unreliability include: -

the Nature of the Variable being measured. For instance, measuring the length of a road should be relatively reliable, whereas attempting to measure attitudes of a target population will be less reliable.

Measuring instruments which are inappropriate for the task, or have not been adequately refined/calibrated/adapted.

Inadequate Sample Selection such as too small a sample, or a sampling of unrepresentative elements from the total population, could distort the conclusions drawn from the data.

MEASUREMENT The act or process of ascertaining the dimensions, capacity, amount or direction of some variable on a quantitative or qualitative scale.

## DATA COLLECTION METHODS

Even where statistical services are not very well developed, there are likely to be substantial sources and amounts of data. Such data must be used cautiously however. The major difficulty in using such data (particularly self-reporting of progress) is that it may be invalid or unreliable because of the built-in biases of the collectors and/or the manner in which it was collected or recorded. Furthermore the scope of the data may differ from the requirements for a particular project.

There are three basic methods for obtaining data about a project:

- -- Direct Measurement
- -- Observation, and
- -- Interrogation.

Direct Measurement is the most accurate, when it can be done. Devices, such as scales, yardsticks, surveying equipment, etc., are usually used and measurements recorded in standardized units. Thus, relatively little judgement is involved. Accuracy depends upon the recorder's familiarity with the measuring device, its appropriateness for the particular situation, and the care with which the results are recorded. A major limitation is that people under study are usually very conscious that measurements are being taken about them, their property or their productivity, and often react in some manner which tends to bias the results, positively or negatively.

Observation relies much more heavily upon the ability of the observer to perceive facts, patterns and relationships. Observation can be done either in a participatory or non-participatory manner, and is a reasonably unobtrusive method for collecting data. However, it tends to be more subjective than direct measurement.

Interrogation can be done in many different ways to gather a wide variety of information. Interrogation relies heavily upon the willingness of the respondent to cooperate, as well as the skill of the questioner in asking and interpreting the responses.

Other problems exist which hamper the collection of data. For instance, the invasion of family privacy may be resented, and different languages or dialects may compound mistrust and miscommunication during interviews. It is often difficult to find trained interviewers; travel may be extremely time consuming and difficult because of terrain, poor roads, lack of vehicles, and lack of lodging facilities for field workers. Often there are travel restrictions in certain areas, or bureaucratic protocol clearance procedures which slow down interviewers in attempting to get data. Each of these aspects must be considered in planning to obtain information.

## COMPARISON OF CENSUS & SAMPLING DATA

Error in measurement is unavoidable, however careful measurement can reduce the degree of error to acceptable levels.

There is an intuitive feeling on the part of many people that census data (i.e. 100% counting or measurement) is bound to be more accurate than sample data (i.e. partial counting of selected variables). However, this is not usually the case.

Because census taking requires more gathering and processing of data than sampling, it also provides more opportunities for errors to occur. This is especially true when the population is large, the type of information sought is somewhat subjective, different people are involved in various aspects of collecting and tabulating the data for analysis, and the information can reflect on the performance of the individual.

Although sampling is, by definition, incomplete; the accuracy (and probable error) of properly drawn samples can be estimated. For most analyses, this is preferable to census data where error is also likely, but where the extent of accuracy is unmeasurable.

Sometimes the sheer size of the population makes it prohibitively costly and/or time consuming to conduct a census, or analyze the results. Furthermore, often because of the nature of the information required (such as in destructive testing) 100% census testing is impractical because the product is consumed during the inquiry. For example, the only way to tell whether a match is a good one or not is to attempt to light it. If it lights, it is "good"; if it fails to light, it is "bad". However, once the match has been lit, it is no longer of any use as a match! Thus, selective sampling and testing is the preferable substitute for determining the probabilities of a given batch being "good" or "bad". Similarly, sampling is often the most practical method for gathering information under many different circumstances, as indicated below: -

| CENSUS (100% Count/Measurement) | SAMPLING (partial count/measurement) |
|---|---|
| -- Costly to obtain data | -- Relatively Inexpensive to obtain data |
| -- Time Consuming to obtain data | -- Can be done relatively quickly |
| -- Difficult to obtain data | -- Relatively easy to do |
| -- Often Impossible to obtain data | -- Usually possible |
| -- Usually takes a long time to produce results | -- Results can be obtained relatively quickly |
| -- Degree of Accuracy/Error unmeasurable | -- Probable error can be estimated in scientific samples |

## JUDGEMENT SAMPLING

Although accepting the notion that sampling is preferable to 100% census taking, many people nevertheless have reservations about the necessity for the procedures associated with scientific -- "random" -- sampling.

They question

> "Why should I bother to go through statistical mumbo-jumbo procedures in order to gather and analyze data? I know my technical field, I have a "feel" for the situation in my area, and I know where to go to ask questions that will supplement my own personal knowledge. How can outsiders who select names from a book of numbers, or a deck of cards, instead of going to the places and people I recommend, possibly come up with findings better than mine?"

Although this may not be said aloud, be sure the thought is there!

It is essential, therefore, that such doubts be overcome before plunging into extensive selection and gathering of data from respondents.

To the busy administrator, an occasional field trip to see conditions for himself (or herself) is an invaluable way of keeping informed about a project's progress. There is no substitute for actually seeing what is happening to gain a mental picture which reinforces written reports. However, there is a danger in drawing the wrong conclusions from such "spot checks" because what is seen, although accurate in itself, may not be truly representative of the larger project picture. All too often, where to go, and what to see is limited and established on the basis of convenience, willingness of recipients of the project to participate, time and accessibility, or the desire to depict extreme situations (good or bad) on the part of local administrators. Furthermore, conditions in a large project may change at differential rates, and where once"X" and "Y" may have been similar, after a time they may no longer be so. Over reliance on the same sources for information may blind you to other factors which warrant your attention.

Such purposive, or judgement, sampling can be useful in keeping the field managers alert to the possibility of inspection by higher headquarters, and for showing VIP's around, but it is not the most effective means for obtaining management data for making programmatic decisions.

For such purposes, Scientific "Random" Sampling is far superior.

## SCIENTIFIC SAMPLING

Scientific Sampling is the use of efficient and effective systematic methods for collecting, interpreting and presenting data in a quantitative manner to facilitate understanding. Scientific sampling is not infallible, but bias can be eliminated to a great extent, and the probability of being correct ascertained.

The prime purpose of scientific sample surveying is to assist program management and policy decision making. If sufficient secondary data[1] relevant to the problem is already available, it may be used as the basis for decision making. If secondary data is unavailable, or insufficient for the purpose, primary data[2] should be collected. Thus the need for a survey is created.

### COMPARISON OF SCIENTIFIC WITH NON-SCIENTIFIC (OR JUDGEMENT) SAMPLING

| Principal reasons for Scientific Sampling | Disadvantages of "judgement" Sampling |
|---|---|
| 1. Bias and subjectivity in selecting sample units can be minimized. | 1. Although seemingly logical, personal biases can severely limit data collected; the findings may be invalid; and subsequent utilization can lead to gross errors in policy and program management. |
| 2. Precise quantitative statements can be made regarding how representtative the sample is of the population from which it is drawn. | 2. The validity of "judgement" data cannot be estimated. |
| 3. The probability of being correct (or incorrect) can be estimated. | 3. The degree of accuracy of "judgement" data cannot be quantified. |
| 4. Scientific sampling is efficient, effective and economical, since the smallest sample size necessary to meet management's specifications can be calculated. | 4. The sample drawn by a "judgement" may be much larger than necessary to do the job (and consequently wasteful of resources); or too small to reflect the situation accurately, which in addition to wasting resources will also fail to provide management with an adequate assessment. |

In short, the validity of a "judgement" sample is generally limited to the sample itself, and cannot be applied to a larger population with any degree of confidence.

Furthermore, because there are many different sources of errors in mass data, sampling is generally more accurate than 100% enumeration and much more practical. For example, varying interpretations by many people of a common guideline, incomplete responses, errors in processing the data, delays in processing because of the volume. Such errors are not easily controlled; hence the smaller the sample, the less opportunity for mistakes to enter. Thus, a carefully controlled sample, even though small, is an invaluable aid in project management, and policy making.

---

1. Data originally gathered by someone else for another purpose
2. New and original data

## SOME BASIC STATISTICAL CONCEPTS

### AVERAGES

The most frequently used statistical measure for describing masses of data is the average, because it reduces many measurements to a single figure, and makes it possible to generalize about the situation. An average is a single value derived from a group of values, which is used to typify the entire group. However, since it is a single value, it should be remembered that it does not accurately reflect the standing of every item in the group. It merely provides a means to generalize about the group as a whole. This is often misunderstood, because the variation around the average is ignored. For example, if we were to state that the average padi production in rainfed areas of West Java is 22 quintales per hectare and further assume that 22 quintales/hectare enables a farmer to meet expenses and make a reasonable income, it does not follow that all farmers in the rainfed areas of West Java make a reasonable income; only that the average, or typical farmer did. Sole use of the average tends to disguise the fact that many farmers did not attain this level. Furthermore many farmers may have smaller plots of land, and while attaining a good rate of production, may not produce sufficient net quantity.

A further problem that is often encountered is that one statistical average may be used to represent groups of situations which are in fact dissimilar. Although the resulting mathematical computations may be correct, the average derived may not present an accurate or useful picture of the group. For example, given that Java is experiencing heavy rainfall and flooding while N.T.T. are having a drought, it would not be accurate to state that the overall rainfall level for those regions that time was "satisfactory" or "normal"! In calculating an average, a preliminary step is to separate the various groups into similar clusters or divisions, where known, and calculate separate averages for each of them.

There are several different types of "average" in common use -- the "Mean", "Median" and "Mode"; each of which has a special purpose.

## MEAN

The "Arithmetic Mean" , usually called simply a "Mean" is probably the most useful and commonly used average. It represents the summation of the values of a group, divided by the number of items in the group. Thus, it is a mathematical "balance point".

Thus it can be calculated from the formula: -

$$\overline{X} = \frac{\Sigma x_i}{n}$$

where

$\overline{X}$ = Mean

$\Sigma$ represents "the sum of"

$x_i$ = the value of item "i" in the group

n = the number of items in the group

A mean can be readily obtained from a series of data as follows:

| Data Item | Data Value $x_i$ |
|---|---|
| 1 | 20 |
| 2 | 32 |
| 3 | 44 |
| 4 | 68 |
| 5 | 68 |
| 6 | 80 |
| 7 | 92 |
| 8 | 104 |
| 9 | 116 |
| n = 9 | $\Sigma x_i = 624$ |

$$\overline{X} = \frac{624}{9} = 69.33$$

## MEDIAN

The median is the "mid-point" in the range of values in a data series. In the preceding example, the 5th item, where x = 68 would be the median value.

Since this data series has an odd number of items (9) the median value is the actual value of the middle item in the series after it has been ranked from low to high, or high to low.

Where the data series has an even number of items, the median value is calculated by taking the mean of the two values in the middle of the series. Thus if the data series was

2,4,6,8,10, and 12

the two middle values "6" and "8" would be used to calculate the median as follows: -

$$\frac{6 + 8}{2} = \frac{14}{2} = 7$$

The median is a useful average to employ when dealing with frequency distributions where the first and/or last grouping is open-ended (or the mid-points of these groups cannot be reasonably estimated) since the actual values of the end groups is not required. Furthermore, where there are extremely high or low values in a data series clustered around the extremes, use of the median will tend to overcome this distortion since only the value of the mid-point, or two mid-points, is significant in determining the median.

## MODE

The mode is a "concentration point" -- the most frequently occuring value in the data series. In the data series of nine items illustrating the mean, the mode is also "68" because "68" occurs twice whereas all other values occur only once.

The mode is often used when dealing with non-continuous variables since the average that results is a value that actually exists rather than a physically impossible "calculated" value such as 5.3 children per family, or 1.2 kerbau per farm.

The mode is also often used when the data is ungrouped, to avoid the burden of grouping and calculating the various data values in order to derive a "typical" data value. The various data values are merely scanned in order to detect the most frequently occuring value.

It should be remembered that none of the above averages is "more accurate" than the others. Each is a measure of "central tendency" that can be used under certain circumstances to assist in generalizing about a group of data; and in any given situation, the most appropriate one for the situation should be used.

## RANGE

The range is the spread between the maximum and minimum values in a series of data.

For example, given the data series

25, 50, 60, 80, and 90

the range would be 90 - 25 = 65

The range is often useful to know in planning to gather data, or in implementing a program and estimating its possible effects.

## THE STANDARD DEVIATION (FOR UNGROUPED DATA)

Averages such as the Mean, Median and Mode are measures of "central tendency" where a single value is used to typify the group of data in the series. We recognized that the variation around that average was frequently ignored, which could lead to distorted impressions of the true situation.

The range is one measure of "dispersion" which is helpful in analyzing data; however it is limited in that it does not take into account the intermediate values.

For program analysis and management purposes, the extent of the differences may be extremely significant. Therefore in addition to the average and the range, another unit of measurement is necessary which provides a quantitative measure of dispersion. This is the "Standard Deviation".

The usefulness of the Standard Deviation is that if we know how many standard deviations a data value is from the mean, we can estimate how likely it is to occur in the population. Thus the standard deviation is a very powerful concept in data analysis, for project management.

The formula for calculating the standard deviation from simple random samples for ungrouped data is as follows: -

$$S = \sqrt{\frac{\Sigma d^2}{n}}$$

where

S = Standard Deviation

$\Sigma d_2$ = Sum of the squared differences from the mean

n = number of data items in the group

## EXAMPLE OF CALCULATING THE STANDARD DEVIATION

The use of this formula can be illustrated with an example, as follows:

Find the Standard Deviation of the following group of five numbers: -

10, 20, 25, 40, 80

By addition, the sum of the data series is 175 and

the mean is $\frac{175}{5} = 35$

To calculate the sum of the squared differences from the mean, the data is arrayed in a table as outlined below.

| A | B | C | D | E |
|---|---|---|---|---|
| Item # | Item Value $x_i$ | Mean | Difference from Mean(d) | Difference Squared ($d^2$) |
| 1 | 10 | 35 | - 25 | 625 |
| 2 | 20 | 35 | - 15 | 225 |
| 3 | 25 | 35 | - 10 | 100 |
| 4 | 40 | 35 | + 5 | 25 |
| 5 | 80 | 35 | + 45 | 2025 |
| n = 5 | $\Sigma x_i = 175$ | $\frac{175}{5} = 35$ | | $\Sigma d^2 = 3000$ |

The difference of each value from the mean is shown in column D; to eliminate the influence of the $\pm$ signs, the difference is squared, and then later the square root is taken. Thus, by substituting in the formula, the standard deviation is calculated.[1]

$$S = \sqrt{\frac{3000}{5}} = \sqrt{600} = 24.5$$

Since the mean of the distribution was 35, this new measure tells us that one standard deviation less than the mean is 35 - 24.5, or 10.5; while one standard deviation greater than the mean is 35 + 24.5, or 59.5. We will use such measurements later in analyzing frequency distributions of data series.

---

1 Note: this is for illustrative purposes only. Actually "n-1" is used instead of "n" for data series of less than 30 items.

## FREQUENCY DISTRIBUTIONS

When data has been obtained from a survey, it is "ungrouped" in a disorganized state. The principal task before analysis then is to organize it into meaningful groupings.

Each question to be anlayzed must be extracted from the individual survey form and tabulated with all the other responses to that particular question.

For example, if we obtained the following data series from fifty responses and wanted to determine the Mean, Range and Standard Deviation we should proceed as follows:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data Series: | 68, | 97, | 15, | 45, | 66, | 81, | 99, | 105, | 26, | 60, | 78, |
| 47, | 55, | 72, | 78, | 130, | 85, | 74, | 57, | 86, | 77, | 102, | 47, | 52, |
| 73, | 69, | 57, | 88, | 73, | 69, | 45, | 101, | 93, | 54, | 65, | 92, | 77, |
| 85, | 60, | 65, | 58, | 72, | 64, | 73, | 79, | 36, | 83, | 96, | 96, | 67. |

We could calculate the mean directly, and by searching we should be able to identify the range. Without resorting the data, however, about all we could tell from this mass of data is that the responses vary. By resorting from high to low, as follows,

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 130 | 97 | 88 | 81 | 77 | 72 | 67 | 60 | 55 | 45 |
| 105 | 96 | 86 | 79 | 74 | 72 | 66 | 60 | 54 | 45 |
| 102 | 96 | 85 | 78 | 73 | 69 | 65 | 58 | 52 | 36 |
| 101 | 93 | 85 | 78 | 73 | 69 | 65 | 57 | 47 | 26 |
| 99 | 92 | 83 | 77 | 73 | 68 | 64 | 57 | 47 | 15 |

a pattern begins to emerge. The range is readily identifiable (a span of 115, from 15 to 130), and even without calculating it, it appears that the mean will be in the low 70's.

We could proceed to calculate the standard deviation at this stage, however the number of items to put in the table is large and the number of calculations to make would be very large and laborious, also. It is preferable for large series of data to reduce the number of data items in the table to about 10 to 15. (Since the objective is to reduce the amount of arithmetical calculation and reveal any meaningful pattern in the data, convenience rather than mathematical precision is the primary consideration.)

For this example, I establish ten (10) separate groupings.

Class Interval

The data in our example covers a span of 115, ranging from 15 to 130. Therefore our groupings should also cover that range. Thus the span of each individual group should be at least

$$\frac{130 - 15}{10} = \frac{115}{10} = 11.5$$

Generalizing, you determine the minimum span of each grouping as follows: -

$$\text{Minimum Group Span} = \frac{\text{Range of data series}}{\text{Number of groupings}}$$

This span is called the "class interval".

Again, class intervals do not have to be precisely calculated. They can be established in convenient numbers such as multiples of "5". or "even" numbers.

Thus, in the example above, I will establish a class interval of "12".

To summarize the example then, we have ten groupings with a class interval of twelve (for a total span of 120) which is enough to handle the data range.

Mid-point, Limits & Range of Class Interval

Because we are clustering our data into smaller groups (in our example, from 50 to 10 groups) for further calculations we will use the mid-point of each class interval to represent that group. To avoid cumbersome arithmetic, we should try to establish an easy number to work with -- preferably multiples of "5" (if the class interval is set that way) or "even" numbers.

At the same time, we must set the limits of the class interval. Starting with the lower end of the range, the likely candidates for the lower limit for the first class interval are determined as follows:

1) Calculate the "span of all class intervals" and subtract the "span of the data series"

   Thus, from our example, 120 - 115 = 5 (A)

2) Calculate one-half the class interval, which again from our example is

   ½ x 12 = 6 (B)

3) Use the smaller of (A) or (B) above in the following formula:

   Lower Limit of 1st class interval = Lowest number in data series − A (or B)

Thus, from our example the lower limit of the 1st class interval is determined as

Lower Limit of 1st class interval = 15 - 5

= 10

We can therefore establish the lower limit of the 1st class interval as any number between 10 and 15, bearing in mind that we want the mid-point of the class interval to be an easy one to manipulate. Because our class interval is 12, we cannot use multiples of 5 as mid-points. Therefore we will look for an even number.

Since 1/2 the class interval is 6; the lower limit of the class interval is between 10 and 15; and we want an even number, the following mid-points are "available" to select from.

6 + 10 = 16 6 + 12 = 18 6 + 14 = 20

I will select 20 as the mid-point of the initial class interval, with the lower limit to be 20 - 6 = 14, and the range 14 to 26.

---

1 Occasionally this is not possible because some items may approach infinity. In such instances, the first and/or last groups may be left "open-ended" i.e. "below 10" or above "150".

A fine, but significant point should be noted here. Data can be either "Continuous", or "non-continuous". It is continuous if within the range, any value is possible, if a more refined or sophisticated measuring device were used. It is non-continuous if the items only come in discrete intervals. For convenience in everyday life, we usually treat data as non-continuous, rounding off and using integers for our unit measures. However, in calculating statistical frequency distributions and class intervals, we should really consider the range throughout the whole grouping as continuous. Thus, with the lower limit at 14, and a class interval of 12, the range in the initial class interval is 14 through 26. The second class interval will be 26 through 38, the third 38 through 50, the fourth 50 through 62, etc. until we reach the final class interval of 122 through 134.

In making discrete groupings out of a continuous distribution however, confusion will arise as to which class interval data at the edges of the class interval should properly belong. For instance, the question would immediately arise whether 26 would be assigned to the first or second class interval, or both. Actually there is no overlap. In a continuous distribution, each integer includes all the values up to the next integer. Thus 14 includes 14.1, 14.2, 14.3, etc., etc., up to 14.9, 14.99 or however precisely you wish to refine and measure the process. In the above example, for instance, since our data is in integers in the initial class interval the lower limit would be set at 14, with the upper limit at 25.9 (rather than 26). We would however retain the mid-point at 20 for computational purposes.

We can now prepare a frequency distribution table with the class intervals, mid-points and frequency for our example as follows: -

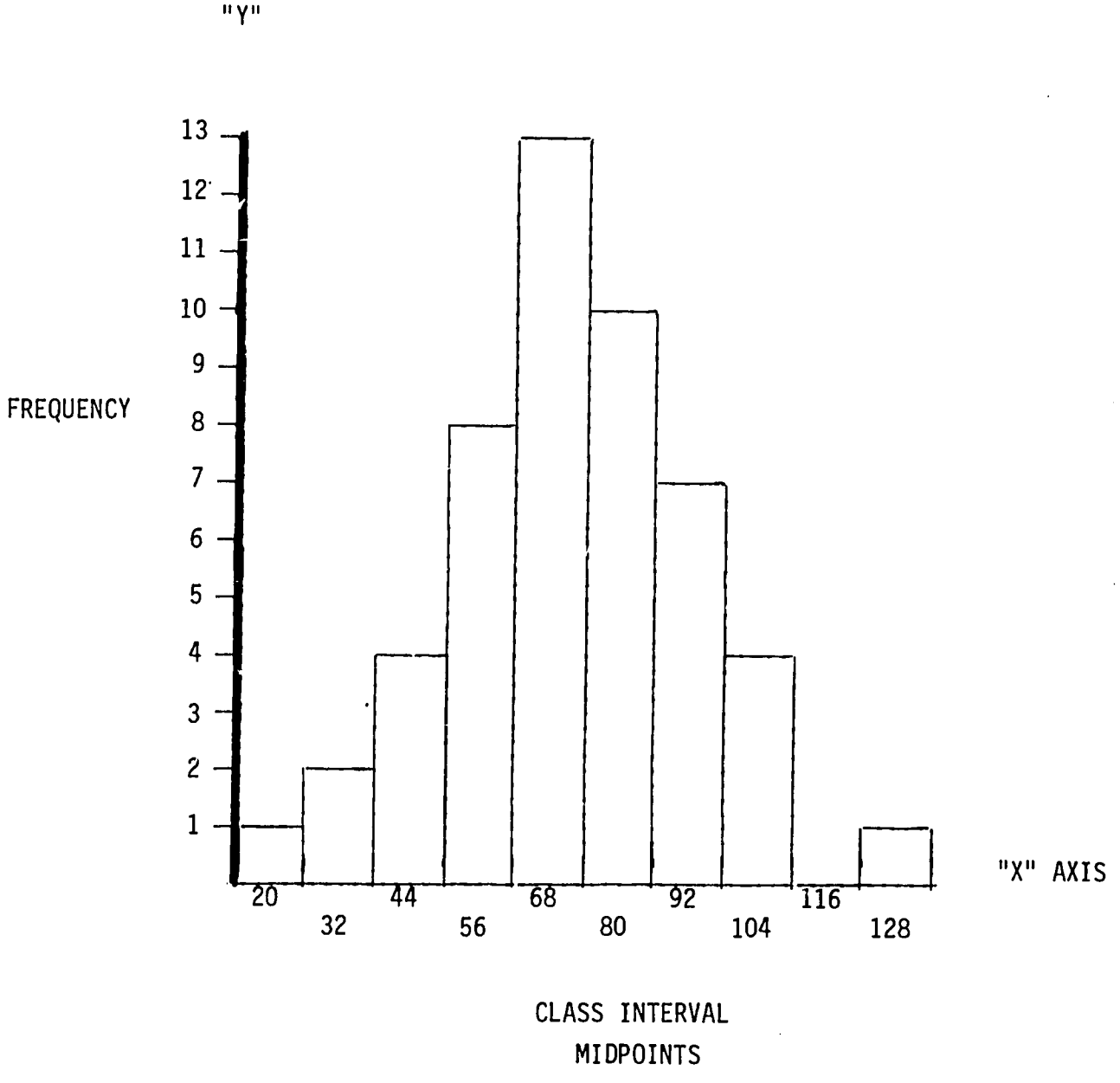
| Lower and Upper Limit | Mid-point | Frequency |
|---|---|---|
| 14 -- 25.9 | 20 | 1 |
| 26 -- 37.9 | 32 | 2 |
| 38 -- 49.9 | 44 | 4 |
| 50 -- 61.9 | 56 | 8 |
| 62 -- 73.9 | 68 | 13 |
| 74 -- 85.9 | 80 | 10 |
| 86 -- 97.9 | 92 | 7 |
| 98 -- 109.9 | 104 | 4 |
| 110 -- 121.9 | 116 | 0 |
| 122 -- 133.9 | 128 | 1 |

With a continuous distribution from 14 to 133.9, subdivided into 10 groups, (Class intervals) with even numbers for mid-points, and assurance that none of our data will overlap the limits of the class intervals, we are now ready for data analysis.

## HISTOGRAMS

Often, it is useful to display the data developed in a frequency distribution in chart form.

By using the class intervals as the points along the "X" axis, and the frequencies along the "Y" axis, a meaningful picture of the data series becomes apparent: -



Mear

A mean can be readily obtained from the data in a frequency distribution table as follows:

| A<br>Mid-point | B<br>Frequency | C = A x B<br>Values |
|---|---|---|
| 20 | 1 | 20 |
| 32 | 2 | 64 |
| 44 | 4 | 176 |
| 56 | 8 | 448 |
| 68 | 13 | 884 |
| 80 | 10 | 800 |
| 92 | 7 | 644 |
| 104 | 4 | 416 |
| 116 | 0 | 0 |
| 128 | 1 | 128 |
| | n = 50 | $\Sigma$ = 3580 |

$$\text{Mean} = \frac{3580}{50} = 71.6$$

It should be remembered however that although 71.6 is a precise looking number, it is the average of the group of 50 items using the mid-points of the class interval; not the average of the actual 50 items. By reducing our data to a frequency distribution to make analysis easier, we have lost the detail and the precision of the raw data. In this particular instance, it is not too difficult to calculate the mean of the entire series. (71.84) but it is not a practice that should be adopted. All analytical techniques follow this trend of reducing data to make analysis easier but losing a little in the process. It is something that management must learn to live with.

## CALCULATING THE STANDARD DEVIATION FROM GROUPED DATA

When the data has already been grouped by uniform class intervals, an adjustment must be made to the standard deviation formula to allow for the "compaction" of varying data into clusters.

$$S = i\sqrt{\frac{\Sigma f(d)^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2}$$

Where

S = Standard Deviation
i = Size of the class interval
f = Frequency of occurrence of data in the class interval
d = Difference of the class interval from the "origin" - an arbitrary selected class interval.
n = Number of items in the distribution

Any of the class intervals can be selected as the "origin" [1] and the difference from this point can be measured in class intervals. Then columns E,F, and G are calculated.

| A CLASS INTERVAL Lower Limit ---- Upper Limit | B MIDPOINT | C FREQUENCY (f) | D DIFFERENCE FROM "ORIGIN" (d) | E (= CxD) FREQUENCY x DIFFERENCE (fd) | F DIFFERENCE SQUARED ($d^2$) | G (= CxF) FREQUENCY x DIFFERENCE SQUARED ($f(d)^2$) |
|---|---|---|---|---|---|---|
| 14 ---- 25.9 | 20 | 1 | - 4 | - 4 | 16 | 16 |
| 26 ---- 37.9 | 32 | 2 | - 3 | - 6 | 9 | 18 |
| 38 ---- 49.9 | 44 | 4 | - 2 | - 8 | 4 | 16 |
| 50 ---- 61.9 | 56 | 8 | - 1 | - 8 | 1 | 8 |
| 62 ---- 73.9 | 68 | 13 | 0 | 0 | 0 | 0 |
| 74 ---- 85.9 | 80 | 10 | + 1 | +10 | 1 | 10 |
| 86 ---- 97.9 | 92 | 7 | + 2 | +14 | 4 | 28 |
| 98 ---- 109.9 | 104 | 4 | + 3 | +12 | 9 | 36 |
| 110 ---- 121.9 | 116 | 0 | + 4 | 0 | 16 | 0 |
| 122 ---- 133.9 | 128 | 1 | + 5 | + 5 | 25 | 25 |
| | | N = 50 | | $\Sigma fd = +15$ | | $\Sigma f(d)^2 = 157$ |

Note from the above table that $\Sigma f(d)^2$ and $(\Sigma fd)^2$ are not the same!

$\Sigma f(d)^2 = 157$ whereas $(\Sigma fd)^2$ is $15^2 = 225$

Thus:

$$S = 12 \times \sqrt{\frac{157}{50} - \left(\frac{15}{50}\right)^2}$$

$$= 12 \times \sqrt{\frac{157}{50} - \frac{225}{2500}}$$

$$= 12 \times \sqrt{3.14 - 0.09}$$

$$= 12 \times \sqrt{3.05}$$

$$= 12 \times 1.7464$$

$$= 20.957 \text{ or } 21 \text{ rounded off}$$

---

1 This is much simpler than using the mean, as in the ungrouped formula.

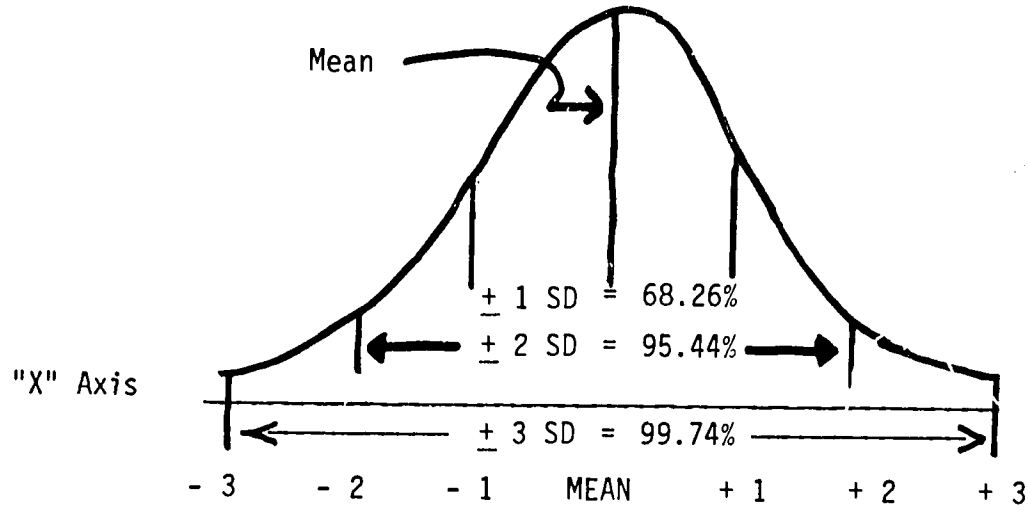
## THE "NORMAL DISTRIBUTION CURVE"

No two situations are ever exactly alike. Nevertheless, statisticians have discovered that the frequency distributions of processes that are repeated many times under similar conditions tend to form a general symmetrical "bell-shaped" distribution. This pattern is known as the "Normal Distribution Curve". Each occurence can be affected in minor ways by natural common factors and/or change. It is inappropriate to attempt to explain the statistical basis for the normal distribution in this booklet. Suffice it to state that many frequency distributions developed in analyzing project situations approximate the normal curve. It is thus a useful statistical concept whose properties we can employ.

### Probability of Deviation from the Mean

A major feature of the normal curve is in determining the extent to which any data value differs from the mean. This is done by measuring the area under the curve, from the mean to the value of the data item in question.

The distance from the mean to any point can be measured in terms of the Standard Deviation. Because of its shape, the proportions under the curve in terms of standard deviations are constant, regardless of the actual data values. For example 1 SD ± mean covers an area of 68.26% of the total area under the curve. Similarly, the areas under the curve at ± 2 and 3 standard deviations are standardized percentages as indicated below. A more complete range of values is indicated in Table 2.




Note that the shape of the normal curve is such that it approaches, but never touches the "x" axis; but for practical purposes, it is not necessary to go beyond 3 standard deviations in either direction.

Applying the normal curve to a rice growing situation where the mean of the distribution is 21.6, and given that one standard deviation is 2.3, it can be estimated that for

| | | |
|---|---|---|
| 1 Standard Deviation | 68.26% | of farmers should obtain a harvest between 21.6 ± 2.3 q/ha; or 19.3 to 23.9 q/ha. |
| 2 Standard Deviations | 95.44% | of farmers should obtain a harvest between 21.6 ± 4.6 q/ha; or 17 to 26/2 q/ha. |
| 3 Standard Deviations | 99.74% | of farmers should obtain a harvest between 21.6 ± 6.9 q/ha; or 14.7 to 28.5 q/ha. |

Although the probabilities have only been shown for ± 1, 2 & 3 standard deviations, by use of TABLE 2 (Appendix), the range for any desired probability can be determined; or by using TABLE 3 (Appendix), the probability for any range.

Example 1. Using Table 2, the range for 39% probability is ± .51 SD. Given the mean of 21.6 and a standard deviation of 2.3,

since 2.3 q/ha = 1 SD  2.3 x .51 = 1.17 q/ha for .51 SD.

Therefore the appropriate range for 39% of farmers around the mean is

21.6 ± 1.17 = 20.43 to 22.77 q/ha.

This is an extremely useful feature in analyzing sample data.

Example 2. Given the above mean of 21.6, and standard deviation of 2.3, what is the probability that farmers will get (or what percentage of farmers are likely to get) between 20 and 25 q/ha?

20 is 1.6 below the mean, or $\frac{-1.6}{2.3}$ = - .7 SD units

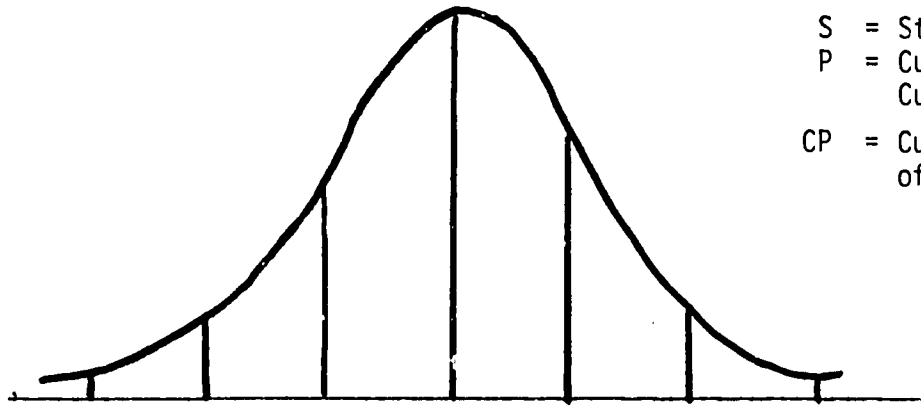
Similarly, 25 is 3.4 above the mean, or $\frac{3.4}{2.3}$ = 1.5 SD units

From TABLE 3, .7 SD Units = 25.8 %
(and) 1.5 SD Units = 43.32%

Thus the specified range encompasses 69.12% probability.

## CUMULATIVE PROBABILITY

Another utility of the normal distribution is that the probability of occurrence of any item in a distribution can be determined, given the distribution's mean and standard deviation.



S = Standard Deviation
P = Cumulative area under Curve
CP = Cumulative Probability of occurrence

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S: | -3 | -2 | - 1 | Mean | +1 | +2 | +3 |
| (+)P: | .13% | 2.28% | 15.87% | 50% | 84.13% | 97.72% | 99.87% |
| (-)CP: | 99.87 | 97.72 | 84.13 | 50 | 15.87 | 2.28 | .13 % |

This is done by expressing the value of the variable in terms of its standard deviation, then measuring the percentage of the area under the curve from either the extreme left (or right) of the curve (depending upon the situation) to the value of the variable in question.

The probabilities are shown above for several selected standard deviations. However, these values can be calculated for any value between -3 and +3 standard deviations, from TABLE 3, with the following modification:

When using the "P" Scale, if the sign is + add 50.
if the sign is - subtract from 50.

When using the "CP" Scale, change the sign from + to -, or - to +, then proceed as above.

For example, given a mean of 21.6 q/ha for farmers producing rice, and a standard deviation of 2.3, the probability of a farmer in this group obtaining not more than 16 q/ha could be determined from the "P" Scale, as follows:

$$16 \text{ q/ha in S.D. Units} = \frac{16 - 21.6}{2.3} = \frac{-5.6}{2.3} = -2.43$$

From Table 3, 2.43 SD units = 49.25%; however since the sign is negative, subtract from 50, or

50 - 49.25 = .75 % probability, (Less than 1%.)

The probability of farmers in the group getting at least 16 q/ha is thus determined using the "CP" Scale, changing - 2.43 to +2.43 and adding: -

50 + 49.25 = 99.25% probability.

## PERCENTAGES AND RANK ORDERING

Many management problems can be analyzed merely by the use of percentages. A percentage reduces figures to a standardized scale of 100, thereby making comparisons easy, particularly between two or more series of raw data drawn from different bases. The formula is: -

$$\% = \frac{DV}{\Sigma DV} \times 100$$

where

% = percentage
DV = data value
100 = constant (100)
$\Sigma$ DV = sum of data values

Thus, if we were to review the data indicated below from six equal areas, of the number of farmers using tractors, the use of the percentage would be more meaningful than the raw data, highlighting the differences and simplifying comparisons and rank ordering.

| AREA | No. Farmers Interviewed | No. Farmers Using Tractors | % Using Tractors | Rank Order |
|---|---|---|---|---|
| A | 86 | 8 | 9.3 | 5 |
| B | 80 | 7 | 8.8 | 6 |
| C | 60 | 7 | 11.7 | 3 |
| D | 40 | 5 | 12.5 | 2 |
| E | 20 | 3 | 15.0 | 1 |
| F | 9 | 1 | 11.1 | 4 |

Rank Ordering is the final step to provide the information "who is first" and "who is last". In comparing many series of data, often the rank ordering is of more importance to management than the actual technical data itself.

Note, however, that rank ordering merely indicates the sequence; it does not indicate the magnitude, or spread differences between each rank.

A fine point in rank ordering -- when there are "ties" for any position, the rank order should be arithmetically averaged (Mean) rather than assigning the most favorable appearing rank. Thus subsequent ranks are unaffected. See the table below for further clarification of this point.

| Percentage Score | Correct Rank Order | Examples of Incorrect Rank Ordering | |
|---|---|---|---|
| 80 | 1 | 1 | 1 |
| 65 | 2.5 | 2 | 2 |
| 65 | 2.5 | 2 | 2 |
| 60 | 4 | 4 | 3 |
| 40 | 5 | 5 | 4 |

## PERCENTAGE FREQUENCY DISTRIBUTIONS

Frequency distributions converted to percentages, are extremely useful for comparing two or more sets of data.

For example, in examining the production of rice farmers under an agricultural credit project, comparing the yield of a sampling of farmers who received credit with those who did not; the raw data is not directly comparable until it is converted to a percentage frequency distribution. To do this, the total number of farmers in each category (181 for borrowers, 129 for non-borrowers) is used as the base. The raw data, and percentage frequency distribution derived from it are shown below: -

| YIELD q/ha | NUMBERS OF | | PERCENTAGE OF | |
|---|---|---|---|---|
| | Borrowers | Non-Borrowers | Borrowers | Non-Borrowers |
| 0 - 3 | 13 | 8 | 7 | 6 |
| 3 - 6 | 7 | 7 | 4 | 5 |
| 6 - 9 | 9 | 12 | 5 | 9 |
| 9 - 12 | 16 | 11 | 9 | 9 |
| 12 - 15 | 16 | 4 | 9 | 3 |
| 15 - 18 | 20 | 13 | 11 | 10 |
| 18 - 21 | 26 | 18 | 14 | 14 |
| 21 - 24 | 13 | 19 | 7 | 15 |
| 24 - 27 | 18 | 13 | 10 | 10 |
| 27 - 30 | 18 | 6 | 10 | 5 |
| 30 - 33 | 11 | 11 | 6 | 9 |
| 33 - 36 | 13 | 4 | 7 | 3 |
| 36 - 39 | 1 | 3 | 1 | 2 |
| TOTAL | 181 | 129 | 100 % | 100 % |

When converting raw data to percentages, as above, some loss of precision will occur if the values are "rounded off". For instance, in the first category where yields are 0 - 3 q/hectare,

$$\frac{13}{181} \times 100 = 7.1823204\ \%$$

Whereas

$$\frac{8}{129} \times 100 = 6.2015503\ \%$$

This generally should not be cause for concern. Of course in some situations, fine measurements are essential, and slight variations in data values can be very significant. Often however the purpose of data reduction is to facilitate analysis and highlight gross differences. In such circumstances, no useful purpose is served by greater precision, and, in fact, visibility is often hindered by the additional "data clutter", and much extra preparation time is also required.

## DETERMINING SAMPLE SIZE

As a general rule, statistical techniques can usually be effectively applied when at least 30 measurements are obtained at random. This is usually insufficient however to present findings with any quantifiable degree of confidence.

Time, money and effort can be wasted if the sample is either larger or smaller than required to meet the specified needs of management. More samples than required waste resources, while fewer samples than necessary give results with less than the required reliability.

Two popular, but erroneous misconceptions should be reviewed; that:

(1) a sample should be some percentage (say 5% or 10%) of a population under study.

(2) a large sample should be taken from a large population, and a small sample from a small population.

Neither of these is correct!

In determining the size of a sample the actual numerical size is usually far more important in determining the reliability of the results than the percentage size.

Furthermore, the size of the population is a minor factor in determining the size of the sample.

The results of a survey are applicable to the total population from which the sample was drawn. Therefore it is economical to sample from as large a population as possible, given the limitations of homogeneity.

The most important criteria for determining the size of a sample are:

1. VARIABILITY in the population under study
2. ERROR that will be tolerated in the findings.
3. CONFIDENCE desired when presenting the findings, that the data is accurate.
4. RESOURCES available to obtain the data, conduct the survey and process the findings.

The first three of these criteria can be used directly in a formula to determine sample size. The fourth is a factor at management's discretion to modify its specifications of "2" and "3".

In order to determine the appropriate size of a sample, you must first establish the type of situation to be studied. One of two formulas can be used, depending upon whether you are seeking your answer in terms of an average or a percentage.

## CRITERIA FOR DETERMINING SAMPLE SIZE

VARIABILITY. This is the "Standard Deviation" of the population under study calculated from the formula below. For practical purposes it is based on the first thirty samples drawn. As previously discussed, the formula is

$$S = \sqrt{\frac{\Sigma d^2}{n}}$$

Where

S = Standard Deviation

$\Sigma d^2$ = Sum of squared differences from the mean

n = Number of items in the group

A more expedient (but less accurate) method for assessing Variability is to use one-sixth of the estimated range, based on historical data, experience in similar situations, or local "expert" opinion. In general, the greater the variability in the population, the larger the sample must be.

TOLERABLE ERROR. Any findings developed from a sample survey will only be approximations, no matter how scientifically they were obtained. Management must specify how precise it wants the answer to be -- within 1, 5, 10 or more units (or percentage points). In general, the greater the desire for accuracy, the larger the sample must be.

CONFIDENCE Desired. When presenting the findings, how sure do you want to be that the answer is within a particular range? It is never possible to be 100% sure, when dealing with samples. Generally, to increase the Confidence in an estimate, a larger sample must be taken.

## OPTIMUM SAMPLE SIZE FOR ESTIMATING A MEAN (RANDOM SAMPLE)

To determine the appropriate sample size, use the following formula:

$$S = \frac{D^2}{\left(\frac{E}{K}\right)^2}$$

Where

S = Optimum Sample Size

D = Standard Deviation of data in the population

E = Size of the mean error that management will tolerate

K = Confidence with which you wish to present the findings

| Selected Values of K (Standard Error)* | Confidence as a Percentage (%) | Numerical "odds" |
|---|---|---|
| 1 | 68.26 | 2:1 |
| 2 | 95.44 | 20:1 |
| 3 | 99.74 | 369:1 |

* Other values of "K" and their related confidence percentages can be determined from Table 2. Generally, however, it is not usually necessary to have a higher value for "K" than "2":- 95.44% confidence.

## SCIENTIFIC SAMPLING METHODS

Once you have established "How Many" samples to draw from a population; the next important issue is "which ones?" Random sampling is selection of items from a given population so that each item has an equal chance of being selected.

There are several methods for drawing samples, each of which has certain advantages, depending upon the circumstances. If each item in the population is considered to have equal importance, you can take either a "SIMPLE" or a "SYSTEMATIC" RANDOM SAMPLE. If on the other hand you know that the characteristics of the items in the population differ markedly and it is possible to classify them, you might want to select samples from each of these groupings in order to improve the validity of the survey. This more sophisticated approach is known as "STRATIFIED RANDOM SAMPLING." Finally, because of the difficulties in field travel in some situations, and/or in order to reduce travel time and costs, "CLUSTER" SAMPLING may be the only practical means for conducting the survey.

## SIMPLE RANDOM SAMPLING

A good "scientific" method in simple random sampling, to assure an equal chance for every item in the population to be selected, is to use a table of random digits, such as TABLE 1 (Appendix). These tables are carefully constructed to use the digits 0 - 9 in an unsystematic "random" manner, with each digit occurring with about the same frequency. The process is as follows: -

First — Obtain a count of the total population under study.

Second — Use the total size of the population to determine the grouping of digits in the table. (For example, if the population is between 10 and 99, use groupings of two digits; between 100 and 999, use groupings of three digits, and so forth).

Third — Assign sequence numbers to the population under study. (If you can't do this, you'll have to use another method.)

Finally — Select any point in the table to start, grouping as explained above, and proceed in any systematic manner (i.e. down, across, up, etc.) selecting and recording those numbers within the population range (disregard numbers outside the range), until the total designated sample size has been selected. Note: Do not skip around on the table. You must proceed in a "serpentine" manner i.e., Down-Up-Down, or Left-Right-Left for example.

What if you draw the same number twice? Treat that individual as two separate samples! (If the sample size is less than 5% of the population, you can disregard the second number and select another.)

An important aspect of using a random digit table is that by recording your working method, and including the particular table used, with the survey results, any charge of bias can be disporved. Hence the objectivity, relative validity and reliability of the survey is assured. This is especially important in highly controversial or crucial policy situations.

---

1 The term Population is used in statistics to signify the total number of items from which you are going to draw samples.

## STANDARD ERROR OF THE MEAN
## (SIMPLE RANDOM SAMPLES)

Because we are working with sample data, rather than the actual total population, the mean that we have derived is only a mean of the sample, rather than the true mean. It is important that this difference be taken into consideration. Otherwise our findings will be limited to only the sample population itself and we will have derived no benefit from sampling. The normal distribution can be used to estimate the likelihood that the true mean lies within a given range of the sample mean. By use of the following formula, we calculate the Standard Error of the Mean: -

$$SEM = \sqrt{\frac{S^2}{n}}$$

Where

SEM = Standard Error of the Mean
S = Standard Deviation of the Sample
n = Size of the Sample

In essence, the standard error is a standard deviation which measures the extent to which values estimated from samples differ from the true population value.

Thus given a sample mean of 71.6; sample size of 50; and sample standard deviation of 21, the Standard Error of the Mean is: -

$$SEM = \sqrt{\frac{21^2}{50}} = \sqrt{\frac{441}{50}}$$

$$= \sqrt{8.82} = \underline{\underline{2.97}}$$

The Magnitude of the Maximum Possible Error can be expressed by dividing the "Standard Error of the Mean" by the Mean itself, and describing it as a percentage, thus: -

$$\text{Magnitude} = \frac{SEM}{M} \times 100$$

Where
M = mean

which in this case is $\frac{2.97}{71.6} \times 100 = 4.15$ or about 4 percent

Note: Although we can compute the error due to sampling, we cannot calculate the error introduced through bias. Therefore careful attention must be given to sample randomization, rather than substituting "judgement" samples in the field to "make things easier" (i.e. walking that extra kilometer to interview a randomly selected farmer, rather than substituting and interviewing another farmer who happens to be nearby.)

## CONFIDENCE INTERVAL AND STANDARD ERROR OF THE MEAN

The significance of calculating the Standard Deviation and the Standard Error is we can now apply the findings from the sample survey data to the total population and be confident (within specified limitations) that it is an accurate representation of the true situation.

Since the Standard Error is a special case standard deviation, its probabilities are determined from the normal curve in the same manner as the standard deviation previously described. Thus $\pm 1$ standard error represents a probability (or confidence) of 68.26% that the true mean lies within this range of the sample mean. In our example where the sample mean is 71.6 and the standard error of the mean 2.97, therefore we can state with a confidence of 68.26% that the true mean of the population lies between

$$71.6 \pm 2.97, \text{ or between } \underline{\underline{68.63 \text{ and } 74.57}}$$

### To Obtain the Range

Depending upon the confidence with which we wish to express our findings, the number of standard errors of the mean to utilize can also be determined from the "Normal Curve and Related Probability Table" Table 2.

For example, if we wish to have a confidence of 99.5%, from table 2, a range of 2.81 standard errors of the mean would be necessary.

In the example, since 1 standard error of the mean = 2.97
$\therefore$ 2.81 standard errors of the mean would be 2.97 x 2.81 = $\pm$ 8.35
from the sample mean of 71.5, or between $\underline{\underline{63.25 \text{ and } 79.95}}$

### To Obtain the Confidence Level

Alternately, if management specifies the range within which it wishes the data presented, we can indicate the confidence that we have in that range by calculating as follows:

$$\frac{\text{Management tolerated error}}{\text{1 standard error}} = \text{number of standard errors of the mean utilized}$$

For example, in the above situation, if management wanted the answer within 1, our confidence would be calculated as follows:

$$\frac{1}{2.97} = .337 \text{ or rounded off } .34 \text{ standard errors of the mean which}$$

from the table gives us a probability of $\underline{\underline{26.62\%}}$.

## SYSTEMATIC RANDOM SAMPLING

This method purposely selects items from all parts of the population in a systematic manner, without bias, rather than attempting to pick items at random.

To use this method: -

1. Assign one sequence number to each item in the population.

2. Determine the "skip interval". Divide the number of units in the population by the sample size.

$$i = \frac{P}{S}$$

Where

i = skip interval
P = Population Size
S = Sample Size

3. Select a random starting point from the population (Use a random digit table)

4. Include that item in the sample, and every "i"th item thereafter, until the total sample has been selected.

Caution: Sometimes, items in a population are arranged in a particular order or pattern which may be repetitive or cyclical. If this is so, and the skip interval is on the same cycle, your sample items may not be representative of the total population but instead may all have the same charateristic.

For instance, you might decide to survey work activity in field offices using particular times of the day for sample observations. If you should happen to select a 2 hour skip interval, and start at 8 a.m. -- with a sampling of activity at 8 a.m. , 10 a.m., 12 noon, 2 p.m., 4 p.m., and 6 p.m., you might draw the conclusion that there is very little work going on since at most times people were arriving, on break, or leaving the office to go home!! This is an obvious case of using the skip interval inappropriately, but many other situations may be less obvious.

## STRATIFIED RANDOM SAMPLING

If it is known ahead of time that the characterisitics of some items in the population differ markedly; that these differences are significant to the problem being surveyed; and it is possible to classify these items on the basis of their characteristics; you can get a more accurate picture of the total population by selecting a random sample from each group.

For example, if we were studying the yields of rice farms in a province, it might be useful to stratify the farms by "irrigated", "rainfed" and "upland" since these characteristics are already known, can be classified, and are significant factors in determining yields. The result would be much more meaningful than merely selecting farms at random without regard to stratification.

Whenever possible, the sample size drawn from these stratifications should be proportionate to the size of the group, as this reduces the analytical problems in evaluating the results. For instance, if we wanted to take a sample of 200 hectares from a province stratified as indicated below, the sample size for each category should also be based on the same percentage. Thus: -

| Stratification | Hectares | Percentage | Sample Size |
|---|---|---|---|
| Irrigated | 35,000 | 46.5% | 93 |
| Rainfed | 31,228 | 42.2% | 84.4 |
| Upland | 8,500 | 11.3% | 22.6 |
| TOTAL | 75,228 | 100.0% | 200.0 |

Sampling within each stratum can then be done by any of the other methods discussed.

## WEIGHTING

Whenever a survey is conducted on a Stratified sample basis, the raw data responses must be "weighted" after the data has been collected. This is done to avoid distortion during evaluation, because the number of responses gathered in each stratum usually differs from the original sampling scheme.

For example, we might have planned a survey of rehabilitation efforts in a province, stratified according to the reported flood damage; with a sample size of 3600. Because of time and distance limitations, it may not have been possible to contact as many of the farmers (and hectares) as originally intended in some areas, while in other areas over-sampling of hectares might have occurred.

To "normalize" the data, a weighting factor is developed by dividing the original area designated to be surveyed in each instance, by that actually surveyed. Thus:

$$\text{Weight} = \frac{\text{Original stratification size}}{\text{Actual survey sample size}}$$

An Example:

| A<br>Province<br>Area | B<br>Hectares<br>Damaged | C<br>% | D<br>Stratification<br>(Ha to be surveyed) | E<br>Ha Actually<br>Surveyed | F<br>Weight<br>(D/E) |
|---|---|---|---|---|---|
| A | 2,500 | 5.438 | 196 | 350 | .56 |
| B | 9,000 | 19.565 | 704 | 400 | 1.76 |
| C | 9,000 | 19.565 | 704 | 1,060 | .66 |
| D | 15,000 | 32.609 | 1,174 | 980 | 1.20 |
| E | 3,500 | 7.609 | 274 | 270 | 1.01 |
| F | 7,000 | 15.217 | 548 | 690 | .79 |
| TOTALS | 46,000 | 100% | 3,600 | 3,750 | |

Thus, an adjustment must be made to the raw numbers in each tabulation to reflect the normalizing effect appropriate for that stratum. If this were not done, some areas would be overrepresented (such as "A") and others underrepresented (such as "B") in the final result.

## STANDARD ERROR OF MEAN FOR STRATIFIED RANDOM SAMPLE

The formula for calculating the standard error of a mean obtained through a stratified random sample is a little more cumbersome. It is a weighted standard error, since we must take into account the fact that each stratum (stratified grouping) has its own standard error.

First, the mean and standard error of each stratum is calculated as for a simple random sample; then the overall standard error is calculated from the following formula: -

$$= \sqrt{\frac{\sum \left[ SEM_i^2 \ (P_i^2) \right]}{10{,}000}}$$

Where

$SEM_i$ = Standard error of mean of stratum "i"

$P_i$ = Weighted percentage of stratum "i"

10,000 = constant

For example, in the following situation: -

| A<br>Province Area | B<br>Hectares Damaged | C<br>Percentage | D<br>Standard Error |
|---|---|---|---|
| A | 2,500 | 5.4 | 3.1 |
| B | 9,000 | 19.6 | 2.3 |
| C | 9,000 | 19.6 | 1.4 |
| D | 15,000 | 32.6 | 0.8 |
| E | 3,500 | 7.6 | 2.7 |
| F | 7,000 | 15.2 | 2.1 |

(Note: "D" would be obtained from survey data and calculated separately for each strata - calculations not shown here)

The standard Error of the Stratified Mean is calculated as follows: -

$$\sqrt{\frac{(3.1^2 \times 5.4^2) + (2.3^2 \times 19.6^2) + (1.4^2 \times 19.6^2) + (.8^2 \times 32.6^2) + (2.7^2 \times 7.6^2) + (2.1^2 \times 15.2^2)}{10{,}000}}$$

$$= \sqrt{\frac{(9.61 \times 2.916) + (5.29 \times 384.16) + (1.96 \times 384.16) + (.64 \times 1062.76) + (7.29 \times 57.76) + (4.41 \times 231.04)}{10{,}000}}$$

$$= \sqrt{\frac{28.02 + 2032.2 + 753.0 + 680.17 + 421.07 + 1018.89}{10{,}000}}$$

$$= \sqrt{\frac{4933.35}{10{,}000}} = \sqrt{.4933} = \underline{\underline{.702}}$$

## CLUSTER SAMPLING

Where time or budgetary limitations, shortage of personnel, and/or difficult field travel conditions make it impossible to obtain data any other way, cluster sampling is often resorted to as the only practical means to gather data. For example, it may take two or more days for an interviewer to obtain responses from ten farmers through simple random sampling methods if the farmers are scattered all over the province, because this may mean extensive travel from one remote village to another for each interview.

With cluster sampling, instead of selecting data from many different geographical locations, more respondents are queried at fewer locations, or "clusters".

Because the samples will be drawn from a more limited cross section of the total population, a greater possibility of bias and sampling error exists. It is therefore desirable to go beyond the minimum sample size specified for a simple random sample.

The first priority is to select as many clusters as can be accomodated by the time/budget/personnel limitations. (The clusters should still be selected on a scientific rather than a judgement basis to reduce bias.) Clusters should be approximately the same size.

Secondly, to reduce error, interview as many of the target population within the cluster as possible -- for instance rice farmers in a selected village. If sampling is done rather than the entire population within the cluster, the respondents should be selected by random methods.

Thus, by randomly selecting two villages, and interviewing as many farmers as possible within those villages, many more farmers may be contacted in a much shorter time period (and for much less cost and travel effort) than obtaining 10 random samples of farmers throughout the province.

Because the data is obtained from only two clusters rather than 10 random samples, however, it will not be as statistically useful, and will have a greater sampling error, even though many more farmers may have been interviewed in each cluster than the 10 specified by the random sample method.

This difference is taken into account in the data analysis, which is also more complex than analyzing a random sample.

If there is any administrative flexibility, it is preferable to obtain data from more clusters, rather than devoting the time and effort to obtaining larger samples within the cluster.

## SELECTING SAMPLES WITH A DECK OF CARDS

A practical field method for drawing samples on a random basis from a population is to use an ordinary deck of playing cards. Here you have a systematic 2, 4, 13 or 52-base selection pool, using the whole deck[1], or any intermediate size population, by eliminating some cards or disregarding and reselecting, if drawn. The deck of numbers is easily "randomized" by shuffling, cutting and drawing. As in using random digit tables, sequence numbers must be assigned to the population. This method is particularly useful for selecting clusters.

For populations larger than 52, you must employ a "multi-stage" method. To do this, initially sub-divide the group, and make a few preliminary eliminations before sequence numbering and selecting actual samples from each final group and/or sub-group.

This procedure introduces some problems. Unless you are careful it may not be as objective as a random digit table.[2] Nevertheless, it has certain practical advantages. It is a readily available and employable method under most field conditions, particularly where random digit tables are difficult to apply or cannot be employed because of the laborious (and often impossible) task of sequence numbering every item in a vaguely defined population. With cards you can work quite flexibly and rapidly to select clusters for further investigation where the total population is not masterlisted or well defined.

Psychologically, the attempt to eliminate subjectivity and the concept of chance can also be more appreciated by the people you are surveying. After you have chosen their area to be surveyed by a previous sub-grouping, it is a useful "ice-breaker" to have the field management staff "participate" in the selection of farmers to be interviewed by cutting and selecting cards for you.

For example, although you may know in gross numbers how many farmers are enrolled in a national program by province, you will not usually know their names.[3] Thus it would not be possible to select which farmers to visit. However, by a preliminary drawing you may select several provinces to survey. Upon arrival at each province, you may further select several municipalities to visit, and upon contact with the municipal management team, several villages. Ultimately, several farmers can be selected from the farm management technician's master-list.

---

1 2 - Red/Black; 4 - Heart, Club, Diamond, Spade; 13 - Ace through King regardless of color or suite; 52 - Hearts 1 - 13, Clubs 14-26, Diamonds 27-39, and Spades 40-52.

2 If the groupings, and divisions into sub-groupings are not equal and symetrical, the individual items in the population will not have an equal chance of selection

3 Nor should you. Generally it is not necessary nor desirable to accumulate masses of detailed data at higher management levels.

## STANDARD ERROR OF THE MEAN FOR CLUSTER SAMPLES

The possibility of sampling error is much greater in cluster sampling than with simple random sampling. However, this may be the only practical means of sampling in many situations due to time/budget/personnel limitations. You should therefore be prepared to analyze data gathered from cluster sampling.

Since the process is more complicated than simple random sampling, the formula for analyzing the data is also more complicated. However, if absorbed step by step, it is not as formidable as it first appears!

The total Formula is as follows:

Standard Error of Mean for Cluster Samples

$$= \sqrt{\left(\frac{N}{M}\right)^2 \sum \frac{M_i^2 (\bar{Y}_i - \bar{Y}_c)^2}{n(n-1)} \left(1 - \frac{n}{N}\right) + \frac{\left(\frac{N}{n}\right) \sum M_i^2 \left(S_{\bar{Y}_i}^2\right)}{M^2}}$$

However, in many instances, -- where the sampled clusters represent less than five percent of the total number of clusters, or if the samples are done "with replacement" (i.e. each item has the possibility of being selected more than once) -- the formula can be simplified to: -

$$= \sqrt{\left(\frac{N}{M}\right)^2 \sum \frac{M_i^2 (\bar{Y}_i - \bar{Y}_c)^2}{n(n-1)}}$$

The explanation of the symbols, and the process for computing the Standard Error of the Mean is outlined on the following pages.

| Symbol & Meaning | Example |
|---|---|
| $N$ = Number of potential clusters | Number of Villages in a Kabupaten |
| $n$ = Number of clusters sampled | Number of villages selected and sampled |
| $M_i$ = Total "population" in cluster "i" | Total number of farmers in village "i" |
| $M$ = Total "population" of all clusters. If notknown, is estimated as $\frac{N}{n} \sum M_i$ | Total number of farmers in all villages in Kabupaten |
| $m_i$ = Total "population" sampled (i.e. sample size) in cluster "i" | Number of farmers in Village "i" who were interviewed |
| $\overline{Y}_i$ = The mean of the variable sampled in cluster "i" | Mean number of quintales harvested by farmers interviewed in village "i" |
| $T_i = M_i\overline{Y}_i$ = The estimate of the total of the variable sampled in cluster "i" | Estimated Total quintales harvested by all farmers in village "i" |
| $\overline{Y}_c = \frac{\sum T_i}{\sum M_i}$ = The estimated mean of the variable for the population for all clusters (sampled and unsampled) | Estimated mean number of quintales harvested per farmer in the kabupaten |
| $S_{\overline{Y}_i}$ = Standard error of $\overline{Y}_i$ in the "i"th cluster, calculated from $\sqrt{\frac{S_i^2}{m_i}} \times \sqrt{1 - \frac{m_i}{M_i}}$ | Standard error of the mean number of quintales harvested by farmers interviewed in village "i" |
| $S_i$ = Standard deviation of $\overline{Y}_i$ in the "i"th cluster calculated from $\sqrt{\frac{\sum d^2}{m_i}}$ | Standard Deviation calculated for the cluster as a simple random sample. |

## Procedure for Computing Standard Error of Mean for Cluster Samples

**Example:** Given:
- project area encompassing 50 villages (N)
- 4 villages selected as clusters to sample (n)
- 8 farmers are interviewed per village ($m_i$)
- yields of individual farmers as shown in table below:
- Total number of farmers in each village ($M_i$) below

| Village | Farmer (& Yield in quintales) 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total Farmers ($M_i$) | Average Yield ($\bar{Y}_i$) |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 22 | 21 | 23 | 20 | 24 | 22 | 19 | 25 | 100 | 22 |
| B | 23 | 21 | 23 | 23 | 23 | 25 | 24 | 22 | 120 | 23 |
| C | 20 | 18 | 19 | 19 | 19 | 20 | 18 | 19 | 96 | 19 |
| D | 19 | 20 | 20 | 21 | 21 | 19 | 20 | 20 | 103 | 20 |

**Required to Find:** Average yield of farmers in the project area, in quintales; by determining average yield of farmers sampled and estimating the standard error of the mean.

*1. Calculate the data required for the simplified formula.

| Village | $M_i$ | $\bar{Y}_i$ | $T_i$ | $\bar{Y}_c$ | $\bar{Y}_i - \bar{Y}_c$ | $(\bar{Y}_i - \bar{Y}_c)^2$ | $M_i^2$ | $M_i^2(\bar{Y}_i - \bar{Y}_c)^2$ |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 22 | 2200 | 21.1 | .9 | .81 | 10,000 | 8100 |
| B | 120 | 23 | 2760 | 21.1 | 1.9 | 3.61 | 14,400 | 51984 |
| C | 96 | 19 | 1824 | 21.1 | 2.1 | 4.41 | 9,216 | 40643 |
| D | 103 | 20 | 2060 | 21.1 | 1.1 | 1.21 | 10,609 | 12837 |
| | $\sum M_i = 419$ | | $\sum T_i = 8844$ | | | | | $\sum M_i^2(\bar{Y}_i - \bar{Y}_c)^2 = 113564$ |

**Average Yield** $= \bar{Y}_c = \frac{8844}{419} = \underline{\underline{21.1}}$ $\qquad M = \frac{50}{4} \times 419 = 5237.5$

*2. Substitute in simplified formula to find Standard Error of Mean

$$\sqrt{\left(\frac{50}{5237.5}\right)^2 \frac{113564}{4(4-1)}} = \sqrt{.0096^2 \times 9463.7}$$

$$= \sqrt{.00009 \times 9463.7} = \sqrt{.872} = \underline{\underline{.934}}$$

Alternately, to calculate the other data required for the complete formula:

(1) Calculate Standard Deviation of $\overline{Y}_i$ for each cluster.

Village A

| A<br>Item #<br>(Farmer) | B<br>Item Value<br>(Yield q/ha) | C<br>Mean<br>(q/ha) | D<br>Difference<br>from Mean (d) | E<br>Difference<br>Squared ($d^2$) |
|---|---|---|---|---|
| 1 | 22 | 22 | 0 | 0 |
| 2 | 21 | 22 | 1 | 1 |
| 3 | 23 | 22 | 1 | 1 |
| 4 | 20 | 22 | 2 | 4 |
| 5 | 24 | 22 | 2 | 4 |
| 6 | 22 | 22 | 0 | 0 |
| 7 | 19 | 22 | 3 | 9 |
| 8 | 25 | 22 | 3 | 9 |
| | | | | $\sum d^2 = 28$ |

$$\text{Standard Deviation} = \sqrt{\frac{28}{8}} = \sqrt{3.5} = 1.87$$

The standard deviations for the other villages are calculated in a similar manner, with results as follows:

| Village | Standard Deviation ($S_i$) |
|---|---|
| B | 1.118 |
| C | 1.118 |
| D | .707 |

(2) Calculate Standard Error of $\overline{Y}_i$ for each cluster.

Village A Substitute available data in the formula

$$S_{\overline{Y}_i} = \sqrt{\frac{1.87}{8}} \times \sqrt{1 - \frac{8}{100}}$$

$$= \sqrt{.234} \quad \sqrt{.92}$$

$$= .484 \times .959$$

$$= .464$$

The standard errors for the other villages are calculated in a similar manner, with results as follows:

| Village | Standard Error |
|---|---|
| A | .359 |
| B | .359 |
| C | .285 |

(3) Calculate the last expression in the formula $M_i^2(S\overline{Y}2_i)$

| Village | $M_i^2$ | $S_{\overline{Y}_i}$ | $S_{\overline{Y}_i}^2$ | $M_i^2\ (S_{\overline{Y}_i}^2)$ |
|---|---|---|---|---|
| A | 10,000 | .464 | .215 | 2150 |
| B | 14,400 | .359 | .129 | 1857.6 |
| C | 9,216 | .359 | .129 | 1188.9 |
| D | 10,609 | .285 | .081 | 859.3 |

$$\sum M_i^2(S_{\overline{Y}_i}^2) = 6055.8$$

(4) Substitute all the values in the formula to find the Standard Error of the Mean.

(Note: The first part of the formula was previously calculated for the simplified version, and determined as $\sqrt{.872}$

$$\sqrt{.872\left(1-\frac{4}{50}\right) + \frac{\frac{50}{4}(6055.8)}{5237.5^2}}$$

$$= \sqrt{.872(.92) + \frac{12.5\ (6055.8)}{27431406}}$$

$$= \sqrt{.802 + \frac{75697.5}{27431406}}$$

$$= \sqrt{.802 + .0028} = \sqrt{.805} = \underline{\underline{.897}}$$

Comparison with Random Sample of Same Size

A random sample of 32 farmers obtaining the same average yields would have produced the following Standard Error:

| (f) # Farmers | ($X_i$) Yield | (FX) | Mean | (d) Difference | $Diff^2$ | f x d | $fd^2$ |
|---|---|---|---|---|---|---|---|
| 2 | 25 | 50 | 21 | 4 | 16 | 8 | 32 |
| 2 | 24 | 48 | 21 | 3 | 9 | 6 | 18 |
| 5 | 23 | 115 | 21 | 2 | 4 | 10 | 20 |
| 3 | 22 | 66 | 21 | 1 | 1 | 3 | 3 |
| 4 | 21 | 84 | 21 | 0 | 0 | 0 | 0 |
| 7 | 20 | 140 | 21 | -1 | 1 | - 7 | 7 |
| 7 | 19 | 133 | 21 | -2 | 4 | -14 | 28 |
| 2 | 18 | 36 | 21 | -3 | 9 | - 6 | 18 |
| n = 32 | | $\sum X_i = 672$ | | | | $\sum fd=0$ | $\sum fd^2=126$ |

$$\text{Mean} = \frac{672}{32} = 21$$

$$\text{Standard Deviation} = \sqrt{\frac{126}{32}} = \sqrt{3.9375} = 1.98$$

and the Standard Error of the Mean =

$$\sqrt{\frac{1.98^2}{32}} = \sqrt{\frac{3.9375}{32}} = \sqrt{.1230} = \underline{\underline{.35}}$$

Thus, the standard error for a random sample is much less (i.e., a random sample is more precise) than a cluster sample.

However, under most circumstances, it would be considerably easier, faster and less expensive to take cluster samples than random ones and to take <u>more</u> samples within the clusters.

## OPTIMUM SAMPLE SIZE FOR ESTIMATING A PERCENTAGE

If an answer sought is to be a percentage rather than a mean, use the following formula to determine the appropriate sample size for a random sample; where

$$S = \frac{(100 - P) \times P}{\left(\frac{}{k}\right)^2}$$

S = Optimum Sample Size

100 = Constant (100) in all equations

P = Preliminary Estimated Percentage

E = Size of the percentage error that management will tolerate

K = Confidence with which you wish to present the findings.

Preliminary Estimated Percentage. Similar to the need to determine the variability of the population ("D") in the 'optimum sample size for a mean' formula, in this situation you must select a percentage between 1 and 99. (0 and 100 do not compute!) you should be aware of the following general trends.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Where "P" = | 0 | 1 | 10 | 20 | 30 | 40 | 50 |
| or | 100 | 99 | 90 | 80 | 70 | 60 | |
| Then "(100 -P) x P" = | 0 | 99 | 900 | 1600 | 2100 | 2400 | 2500 |

Thus, if you have no"feel" for the situation, and can get no expert opinion, you can play safe by using 50%, as this gives the largest possible result.

Practically, you should increase the actual sample size over the optimum size to protect against possible error in estimating the standard deviation, to allow for some non-response during data gathering, errors in compiling data, and other loss because of inaccessibility, etc. Additional samples will increase the reliability of the estimate, while fewer samples than specified will lessen its reliability and perhaps fail to meet management's requirements.

## STANDARD ERROR OF A PERCENTAGE
(Random Sample)

The concepts of probability are equally applicable to the percentage. For instance, management might wish to know the extent to which low productivity was a problem in rainfed padi areas in percentage terms.

If we make an assumption that 25 q/ha is the satisfactory cut-off point, and we observe that 13 from sample of 50 (13/50, or 26 percent) fall in the problem area; from this sample information, what inference can then be drawn about the population?

First, we must determine the probable sampling error in the estimated percentage. The formula for this is as follows: -

$$\text{Standard Error of a percentage} = \sqrt{\frac{(100-P) \times P}{N}}$$

Where

SEP = Standard Error of a Percentage
100 = Constant (100)
P = Sample Percentage
N = Sample Size

Thus, substituting our data in the above

$$= \sqrt{\frac{(100 - 26) \times 26}{50}}$$

$$= \sqrt{\frac{74 \times 26}{50}}$$

$$= \sqrt{\frac{1924}{50}}$$

$$= \sqrt{38.48}$$

$$= 6.2$$

To get the Magnitude of the Possible Error, divide the Standard Error of the Percentage by the Sample Percentage; and express it as a percentage as follows:

$$\text{Magnitude} = \frac{SEP}{P} \times 100$$

Thus the error in this case could be as much as $\frac{6.2}{26} \times 100 = 23.85$, or almost 24%

## CONFIDENT INTERVAL AND STANDARD ERROR OF A PERCENTAGE
(Random Sample)

The confidence associated with the sample percentage can be calculated and applied to the true percentage.

For example, where the sample percentage is 26% and the standard error of the percentage 6.2%, we can state with a confidence of 68.26% (1 standard deviation) that the true percentage lies between

$$26 \pm 6.2, \text{ or between } 19.8 \text{ and } 32.2 \text{ percent}$$

By reference to table 2 the number of standard errors of the percentage to utilize can be determined for any desired confidence.

For example, to determine the minimum percentage with a confidence of 99.5% from table 2, 2.81 standard errors of the percentage would have to be subtracted from the sample percentage.

Since 1 standard error of a percentage = 6.2 percent
2.81 SEP = 6.2 x 2.81 = 17.42 or a
minimum of 26 - 17.42 = 8.58 percent.

By the same token, it could b as much as 26 + 17.42 = 43.42 percent.

Alternately, if management wanted the answer with a range of only 5 percent we could provide that answer, with the reservation that our confidence was not very high. Thus,

$$\frac{\text{Management tolerated error}}{\text{1 Standard error of percentage}} = \text{number of standard errors of the percentage utilized.}$$

For example a range of 5 percent represents 2.5 percent on each side of the sample percentage; thus

$$\frac{2.5}{6.2} = 0.4 \text{ standard errors of the percentage}$$

From table 2, this converts directly to a confidence level of 31.08%.

These concepts are first used to determine the appropriate size sample to be taken, using best guesses for the mean and the standard deviation with specified tolerances. Once the sample has been taken, we reverse the process, using the actual sample to determine that which we had previously only guessed.

## THE "T" TEST FOR ESTIMATING CONFIDENCE INTERVALS FROM SMALL SAMPLES

In the discussion of sample size, I indicated earlier that in general, at least 30 measurements should be drawn from a population to make a useful quantitative analysis. In some situations however, it may be impractical to draw this many samples, but nevertheless an analysis is still called for. What can one do?

One correcting feature to offset the small sample size is to use "N-1" rather than "N" in the various equations. A problem remains in calculating confidence estimates however. Generally a small frequency distribution tends to be much more widely dispersed than the normal distribution of the population from which it is drawn. As the samples become smaller, the difference between them and the true population tend to become greater.

Fortunately, for our purposes, a distribution has been calculated, -- known as the "Student's T", -- which we can utilize to arrive at a statement of confidence. The procedure is somewhat different from the foregoing however.

1. We calculate the Standard Error as before.
2. Then the "T" Table (Table 5) is used to obtain the value for "T" for different sample sizes, for any specified level of confidence.
   (Note: Instead of Sample Size (N), the column is headed "Degrees of Freedom". For our purposes here this is "N-1".)

   For example, if we only had a sample size of 15 and desired to present our findings with a confidence of 95%, the "T" value would be 2.145, corresponding to 14 degrees of freedom and 95% probability from the table.

3. To obtain the Range within which the true mean lies, associated with any given confidence level and sample size.

   Multiply the Standard Error by T

   Thus, given a standard error of 2.97 and a sample mean of 71.6, with 15 samples, the range would be 71.6 $\pm$ 6.37

   (2.97 x 2.145 = 6.37) or

   65.23 through 77.97 for 95% confidence level

4. To obtain the Confidence Level, associated with any range, the procedure is reversed, thus

$$T = \frac{\text{Range}}{\text{Standard Error}}$$

   which must then be looked up in the table for the appropriate sample size.

   Thus given a sample size of 11, a standard error of 2.244 and management's desire for an answer within $\pm$ 5, the value of T is

$$\frac{5}{2.244} = 2.228$$

   which corresponds to a probability of 95%.

   If this all sounds complicated, the way to avoid it is to take larger samples!!

## CORRELATION

In program management, recommendations are often made to adopt certain practices in order to improve results. For example, credit is often seen as a major factor which could increase farmers yields.

Whenever possible, recommendations are made on the basis of carefully evaluated experiments, particularly technical recommendations such as appropriate amounts of fertilizer per hectare. Sometimes, however when we want to change policies, we often have nothing better to go on than intuition and common sense. At other times, the need to do something is so great that there is no chance for pre-testing.

In these circumstances, it is appropriate that the impact of the recommended changes be evaluated as soon as practicable to determine whether the change was in fact beneficial, and thus should be continued, or whether it was insignificant, or even detrimental; in which case management would want to rescind it.

For example, Pairs of data might be obtained for (a) amount of credit and (b) yield from a sample number of farmers.

1. From these paired sets of data values, a Coefficient of Correlation "r" is calculated. This is then compared against a scale ranging from - 1.0 to + 1.0, interpreted as follows:

| COEFFICIENT OF CORRELATION | INTERPRETATION |
|---|---|
| - 1.0 | Perfect "Negative Correlation" (i.e. As "X" increases, "Y" decreases proportionately) |
| 0 | No correlation discernable. |
| + 1.0 | Perfect "Positive Correlation" (i.e. As "X" increases, "Y" also increases proportionately) |

2. By squaring the coefficient of correlation, the amount of variation attributable to the "independent" variable can be calculated. Thus

$$\text{Percentage of Variation of Y Attributable to X} = 100\ r^2$$

3. Alternately, the percentage of unexplainable variation can also be identified

$$\text{Percentage of Variation of Y Which is not attributable to X} = 100\ (1 - r^2)$$

The magnitude of these measurements provide management an indication whether further investigation is called for.

This is quite a complex area for analysis, and generally beyond the scope of this limited text. However, just to whet the appetite, an example is provided of the simplest of these correlation analysis techniques - linear relationship between two variables.

## LINEAR RANK ORDER CORRELATION OF TWO VARIABLES

A simplified approach is to rank order each data pair and then compare the rank orders using the following formula (known as the Spearman Rank Order Correlation).

$$r = 1 - \left( \frac{6 \sum d^2}{n(n^2-1)} \right)$$

Where
- $r$ = coefficient of correlation
- $1$ = constant 1
- $6$ = constant 6
- $\sum d^2$ = Sum of the squared differences between X and Y
- $n$ = number of pairs

For example

Management wanted to know whether the availability of credit has any impact upon yields. Sample data revealed the following:

| Independent Variable X Loans (000's of Rupiahs) | Rank Order X | Dependent Variable Y Yields (q/ha) | Rank Order Y | Difference Between Rank Orders X & Y | Difference Squared |
|---|---|---|---|---|---|
| 110 | 9 | 18 | 8 | 1 | 1 |
| 210 | 8 | 17 | 9 | 1 | 1 |
| 370 | 7 | 20 | 7 | 0 | 0 |
| 420 | 6 | 23 | 5 | 1 | 1 |
| 560 | 5 | 26 | 4 | 1 | 1 |
| 640 | 4 | 22 | 6 | 2 | 4 |
| 770 | 3 | 33 | 2 | 1 | 1 |
| 850 | 2 | 30 | 3 | 1 | 1 |
| 900 | 1 | 35 | 1 | 0 | 0 |
| | | | | | $\sum d^2 = 10$ |

Substituting

$$r = 1 - \left( \frac{6 \times 10}{9(9^2 - 1)} \right) = 1 - \left( \frac{60}{9(81-1)} \right)$$

$$= 1 - \left( \frac{60}{729-9} \right) = 1 - \frac{60}{720} = 1 - .083 = \underline{\underline{.917}}$$

Rank ordering considerably simplifies computation but it also is less accurate than using the actual data. It is a useful technique therefore when "probing" to determine whether a correlation might exist.

A caution when doing correlation analysis -- very often a high correlation may exist between two variables, but this does not necessarily mean that there is a "cause - effect" relationship between them. The correlation may be coincidental, or "spurious". High correlation does tend to reinforce intuition, and common sense; but a healthy measure of skepticism must be used also. Consider whether there are any other plausible factors which might have produced the result.

## NON-NORMAL DISTRIBUTION

Even if a series of data is not distributed in a normal fashion, calculation of of the standard deviation can still prove useful for management analysis. Regardless of how a series is distributed, the following formula can be used to determine the minimum percentage of probability of items that will be included in a given range.

$$MP = \left(1 - \frac{1}{NS^2}\right) \times 100$$

Where

NS = number of standard deviations from the mean

MP = Minimum percentage of items, or probability that items will be included within the range

Alternately, the number of standard deviations can be determined, given the percentage or probability desired, from the following formula

$$NS = \sqrt{\frac{1}{1 - \frac{MP}{100}}}$$

Some useful reference points derived from the above formulae are tabulated below:

| Number of Standard Deviations from the Mean (NS) | Minimum Probability (%) that items will be included in the range (MP) |
|---|---|
| .99 | 2.03 |
| 1.1 | 17.36 |
| 1.22 | 32.81 |
| 1.41 | 50 |
| 1.5 | 55.56 |
| 1.73 | 66.6 |
| 2 | 75 |
| 2.5 | 84 |
| 3 | 88.89 |
| 3.16 | 90 |
| 4 | 93.75 |
| 4.47 | 95 |
| 5 | 96 |

Thus, in any distribution, data beyond 5 standard deviations is highly unlikely and would be considered "statistically significant".

## SIGNIFICANCE

Sample surveys are often requested by management because they want information about an area of interest on which, for one reason or another, little or no data exists. For example, to assess the impact of a typhoon on rice plantings and/or harvestings which are underway. Other times new data may be required for an important program or policy decision -- such as whether to change the rate of fertilization for a particular seed variety during the dry season. Sometimes sampling is the most efficient method of gathering regular series of data - such as a Bureau of Agricultural Economics Quarterly Survey on Rice Production.

Often however, sample surveys are conducted to assist the program manager in identifying strong and weak areas, and to monitor the degree to which the program is living up to expectations. When regular program reports are received on key indicators from "interested" practitioners, periodic sampling of data in the field by "objective" evaluators can give indications as to the quality of those reports. For instance, does the sample survey indicate the same level of production as is being reported, or does it differ? If there is a difference, is it worth worrying about? i.e., is it "within the ballpark"? We can improve upon this subjective question by asking "is the variation statistically significant?"

The size of the Standard Deviation is a useful indicator of the quality of program implementation. Since the sample data should have been gathered in a random fashion from a relatively homogeneous population, the actual spread of the data should not vary much in absolute amount if all aspects of the process are well managed. A small standard deviation represents a narrow range and a relatively tightly managed program. A large standard deviation represents a wide data range and consequently much wider tolerances, pointing the need for follow-up and improvement. Of course, "Small" and "Large" are relative terms depending upon the subject under study. In agriculture, carefully controlled experimental plots may produce consistently good yields; but many individuals with different mental attitudes, farming under varying physical conditions can produce widely varying results. Nevertheless, the distribution should follow a normal pattern under most circumstances.

When results occur which are unlikely to have happened by chance, they are labelled "Statistically Significant". Statistical significance is based upon probability. When statistically significant data are identified in program analysis, this is an indication to management that something unusual happen, the significant difference may be good news. If we are not, it indicates that something is wrong; Either there is an anomaly in program implementation which requires remedial action, or the data reported is in error. In any event, we should make management aware that something unusual is happening.

Before raising alarms however, the initial assumption of a homogeneous population grouping (and thus the expectation of a normal distribution pattern) should be verified. For added confidence in searching for false/erroneous data reports, the data should be checked as to whether it is below the minimum expectations for a "non-normal distribution".[1]

There are several tests which can be applied to data to determine their significance, depending upon the situation. Some of them will be discussed on the following pages.

---

1 See next page

# THE "Z" TEST

## SIGNIFICANCE TESTING FOR A MEAN

Periodically, management should evaluate the quality of its regular progress reports from the field, particularly where the field workers report on their own performance. Even when no vested interests are involved, census-type reporting is rarely 100% accurate, and sampling cannot give absolute certainty either. However, sampling results can be expressed in terms of probabilities, and by using the "Z" test, the accuracy of the reported data can be judged.

The procedure for "Z" Significance testing is as follows: -

1. Assume that there is no "statistically significant difference" between the sample mean and the reported mean. This is known as the "Null" hypothesis. In other words, even though the "reported" and "sample" means are not exactly the same, management is willing to accept them as "close enough" if both the "reported" and "sample" data could have been drawn from the same population at least 5 times out of 100, for instance.

2. Determine management's minimum criteria for significance. Generally, the null hypothesis is rejected if the probability (due to sampling variability) of a result occurring is five times or less out of a hundred. The 5% (5/100) can also be written as 0.05 and is usually described as the .05 Level of Significance. Higher or lower levels of significance can be established by management for particular situations.

A Higher level of significance means that if the result occurs, it is more significant. A result is more significant if it has a lower probability of occurring. Thus a .03 level of significance (i.e. probability of occurring only 3 times out of 100) is of a higher level of significance than a .05 level.

3. Test the hypothesis.

   a. Calculate "Z" from the formula:

$$Z = \frac{\text{Sample Mean - Reported Mean}}{\text{Standard Error of the Sample Mean}}$$

   b. Look up the value for "Z" in Table 4

The "Z" value indicates the probability (percentage of occurrence, i.e. 40%; or 40 chances out of 100) that the sample mean and the reported mean could have come from the same population.

NOTE: If the reported mean could reasonably have been expected to be either higher or lower than the sample mean, MULTIPLY THE "Z" VALUE BY 2. If the reported mean could reasonably have been expected to be only higher (or lower) than the sample mean, USE THE "Z" VALUE DIRECTLY.

   c. IF THE "Z" VALUE IS EQUAL TO OR GREATER than management's minimum criteria, THE HYPOTHESIS IS ACCEPTED, and we conclude that there is NO SIGNIFICANT DIFFERENCE.

   IF THE "Z" VALUE IS LOWER than management's minimum criteria, THE HYPOTHESIS IS REJECTED, and we conclude that there IS A SIGNIFICANT DIFFERENCE.

NOTE: We cannot absolutely prove, or disprove a hypothesis, statistically. We can only indicate the probability of it being as stated; the higher the probability, the more likely the hypothesis is correct.

An example should clarify this. A province reports that the average padi yield is 28 quintales/hectare, but it is suspected that this report is somewhat inflated. A sample survey is conducted in that province, which indicates that the average yield is 25 q/ha, with a standard error of the mean 1.63.

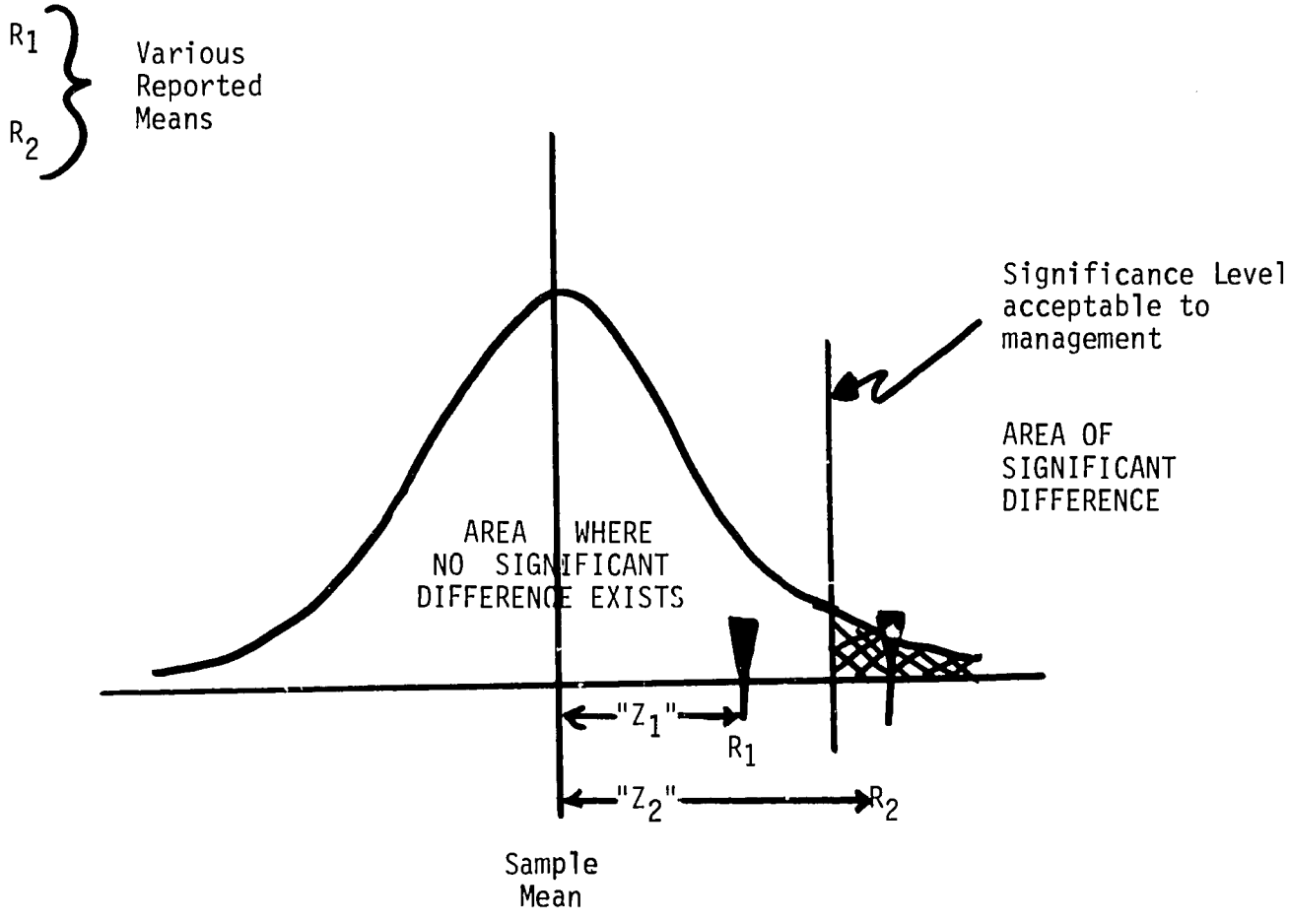
Management establishes the null hypothesis that there is no significant difference between the reported mean of 28 and the sample mean of 25, and is willing to accept a significance level of 5%. Then

$$Z = \frac{25 - 28}{1.63} = \frac{-3}{1.63} = -1.84$$

From Table 4, a "Z" Value of 1.84 corresponds to a probability of only 3.29%.

In other words, only in less than 4 cases out of 100 could the reported and sample means be from the same population. Since 3.29% is lower than the 5% management was willing to accept, the hypothesis is rejected, and we conclude that there IS a statistically significant difference between the two. Another team is therefore sent to the province to discuss reporting procedures and the need for improvement!

Study the sketch below to make sure you understand this concept.



## TYPE I AND TYPE II ERRORS

Because we don't have perfect information, by relying upon significance tests management runs the risk of making errors in judgment. These can be of two kinds and are known as Type I and Type II errors. For instance: Given a Management Minimum of 5% and "Z" value of 4%:-

| TEST INFERENCE AND ACTION | ACTUAL SITUATION | NET EFFECT |
|---|---|---|
| There is a significant difference. The Hypothesis is rejected. | 1. There IS a significant difference. | Correct Inference |
| | 2. There really is NO significant difference | Incorrect Inference TYPE I ERROR MADE |

Management is too "uptight".

The risk management takes by running a "tight ship" is to criticize the reporters unjustly, and/or look for problems where none exist. The chances of making such an error can be reduced by raising the level of significance (i.e. lowering the minimum acceptable probability). For instance, in the above example there is no significant difference at the 3% level.

If no significant difference is indicated, and the hypothesis is acceptable management faces another risk, known as a TYPE II error. For instance: Given management minimum of 5% and a "Z" value of 6%

| TEST INFERENCE AND ACTION | ACTUAL SITUATION | NET EFFECT |
|---|---|---|
| There is NO significant difference. The Hypothesis is accepted. | 1. There is NO significant difference. | Correct Inference |
| | 2. There IS a significant difference. | Incorrect Inference TYPE II ERROR MADE |

Management is "too lax".

The risk management takes by being lenient is to overlook poor reporting, and fail to take corrective action where it is needed. The chances of making such an error can be reduced by lowering the level of significance (i.e. raising the minimum probability acceptable). For instance if management's minimum acceptable probability had been 8% in the above example, a significant difference would have been observed Thus management should consider whether it is more important to avoid Type I errors, or Type II errors, or whether both are equally as critical.

It is important to recognize that it is not possible conclusively to prove a hypothesis on the basis of logic. It is possible only to increase the degree of confidence in a hypothesis.

However, it is possible to disprove a hypothesis on the basis of logic by creating a null hypothesis as the negative complement of the original hypothesis and then trying to disprove or discredit that; thus increasing confidence in the original hypothesis. For example if our hypothesis is that production of high protein wheat by small farmers will result in improvements in their children's health:- Our null hypothesis would be that there is no causal relationship between increased production and increased health. If it can be disproven or discredited, the original hypothesis gains in credibility.

Discrediting the null hypothesis is only one means for increasing confidence in the original hypothesis. It is necessary also to eliminate or reduce the credibility of other possible explanations. For instance preventive health services dispensed by mobile health clinics might have caused children's health to improve. Or the improvement might have been due to hygiene and sanitation instruction in school.

Several aspects of the hypothesis testing process which require special note are:

- -- The analytical study design must be carefully disciplined and systematic, with appropriate provisions for cross checking and verification.
- -- The hypothesis must be narrowly and specifically drawn to disengage or eliminate a variety of related social, economic, cultural and other factors.
- -- The testing must recognize the possibility of unanticipated causes and effects.
- -- The testing must be concerned not only with the independent (causal) variables and dependent (effects) variables, but also with the nature of the treatment, its characteristics and components.

Because of errors in measurement, and because the variable factors themselves may not be stable, the evaluator needs assurance that apparent effects (outcomes) are real, Statistical significance assumes great importance. Statistical significance is a measure which compares the observed magnitude of an effect to the amount of random variability/error inherent in the data. Thus statistical significance estimates the likelihood that an observed effect is not due to chance.

## SIGNIFICANCE TESTING FOR A PERCENTAGE

Significane testing for a percentage employs the Z-test in much the same way as for a mean. There are two principal differences however.

1. The Z-test only gives accurate results when the percentage and/or the number of samples is relatively large. The rule of thumb is to utilize the Z test when a combination of

   number of samples x reported percentage* = 500, or more

   For example 30 samples x 20 percent = 600

Otherwise the distortions are too great and a more exact method must be used.

2. In calculating the Standard Error of Percentage the "reported percentage" is used instead of the "sample percentage".

The formula is:

$$Z = \frac{\text{Sample Percentage - Reported Percentage}}{\text{Standard Error of Percentage}}$$

For Example, a province reports that 85% of its supervised farmers are being visited by the extension technician during the month. However, a sample survey of 25 farmers indicates that only 60% were visited.

STEPS:

1. Test whether the Z test is appropriate. Either (25 x 85) or 25 x (100 - 85) should equal at least 500. 25 x 85 = 2125, 25 x (100 - 85) = 375. Therefore the Z test is appropriate.

2. Establish the null hypothesis that there is no statistically significant difference between the sample percentage and the reported percentage.

3. Management establishes the minimum acceptable significance level at 5%.

4. Calculate Standard Error of Percentage using "reported percentage".

$$SEP = \sqrt{\frac{(100 - P) \times P}{n}}$$

Where

P = Reported Percent = 85
N = Sample Size = 25

$$= \sqrt{\frac{(100 - 85) \times 85}{25}} = \sqrt{\frac{15 \times 85}{25}}$$

$$= \sqrt{\frac{1275}{25}} = \sqrt{51} = \underline{\underline{7.14}}$$

---

* or (100 - report percentage)

5. Calculate Z

a. $$Z = \frac{60 - 85}{7.14} = \frac{-25}{7.14}$$

$$= \underline{\underline{-3.5}}$$

b. From Table 4, a Z of - 3.5 is less than .14 percent.

How much less, we cannot determine, since it is off the Table.

Even allowing for the possibility that the reported percentage could have been higher <u>or</u> lower than the sample, the percentage of occurence (i.e. the probability) would not be more than .28%.

Since this probability is lower than management's minimum acceptable level of 5.0% the likelihood of 60% and 85% being in the same general "ballpark" is very remote, and the hypothesis is rejected.

We conclude that there IS a significant difference.

## THE "CHI SQUARE" TEST

Where the observed data is classified into two or more categories, a "Chi Square" (pronounced "Kye Square") test can be used to determine whether the entire sample's frequency distribution is significantly different from another frequency distribution -- a total population, a reported distribution, or even an assumed distribution.

The Chi Square test is particularly useful because it is equally applicable to both normal and non-normal distributions, whereas the "Z" test is based on an assumption of normal distribution. The use of the Chi Square test is as follows: -

NORMAL DISTRIBUTION Assuming the population is normally distributed, the null hypothesis would be that there is no significant difference between the observed sample and the total population from which it was drawn. Thus the expected frequencies should not differ from which it was drawn except for the variability due to sampl ng error, and/or the limited size of the sample.

For example; to determine whether the yields of 210 sample farmers observed in a province were representative of the population (i.e. normally distributed) or had been specially "preselected" to impress the project manager during a VIP inspection visit,

1. First, set up the frequency distribution by class intervals, as follows:
   Table 1 Observed Frequency Counts of Yields (q/ha) of 210 sample farmers in Province A.

| CI of Yields: | Below 17 | 17 - 19 | 19 - 21 | 21 - 23 | 23 - 25 | 25 - 27 | Above 27 | TOTAL |
|---|---|---|---|---|---|---|---|---|
| # of Farmers: | 13 | 32 | 46 | 60 | 41 | 17 | 3 | 210 |

   Note: Any column with a frequency count of less than "5" should be eliminated. This can be done either by changing the size of the class interval, or by combining columns, as in this case. Thus: -

   Table 2

| CI of Yields: | Below 17 | 17 - 19 | 19 - 21 | 21 - 23 | 23 - 25 | Above 25 | TOTAL |
|---|---|---|---|---|---|---|---|
| # of farmers: | 12 | 32 | 46 | 60 | 41 | 20 | 210 |

2. Using the data above and the methodology indicated earlier in the text, the mean, standard deviation and probabilities for the mid-point of the above class intervals should be calculated. The results are as follows: -

TABLE 3

| Midpoint | Frequency | M x Frequency | Diff | F x D | $D^2$ | F x $D^2$ |
|---|---|---|---|---|---|---|
| 16 | 13 | 208 | - 2 | - 26 | 4 | 52 |
| 18 | 32 | 576 | - 1 | - 32 | 1 | 32 |
| 20 | 46 | 920 | 0 | 0 | 0 | 0 |
| 22 | 60 | 1320 | + 1 | 60 | 1 | 60 |
| 24 | 41 | 984 | + 2 | 82 | 4 | 164 |
| 26 | 20 | 520 | + 3 | 60 | 9 | 180 |
| | N = 210 | $\Sigma$MF = 4528 | | $\Sigma$fd = 144 | | $\Sigma fd^2$ = 488 |

$$\text{Mean} = \frac{4528}{210} = 21.56$$

$$\text{SDEV} = 2 \times \sqrt{\frac{488}{210} - \left(\frac{144}{210}\right)^2} = 2 \times \sqrt{2.32 - .6857^2}$$

$$= 2 \times \sqrt{2.32 - .47} = 2 \times \sqrt{1.85}$$

$$= 2 \times 1.36 = \underline{\underline{2.72}}$$

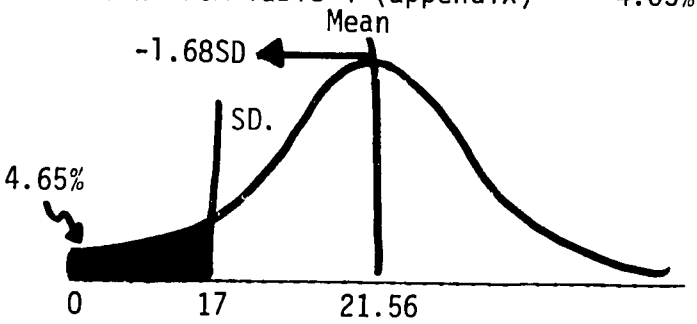
3. Calculate the probability of a "normal" yield distribution in each of the class intervals, given a mean = 21.56 and a standard deviation = 2.72 from the formula

$$\text{Standard Deviation Units} = \frac{\text{Data Value} - \text{Mean Value}}{\text{Standard Deviation Value}}$$

(a.) Below 17 q/ha

$$\frac{17 - 21.56}{2.72} = \frac{-4.56}{2.72} = -1.68 \text{ s.dev. units}$$

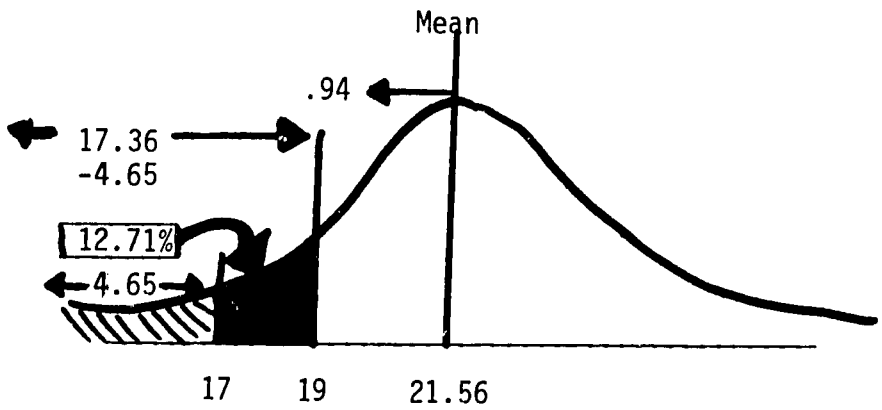
which from Table 4 (appendix) = 4.65% probability.




(b.) Between 17 and 19 q/ha

$$\frac{19 - 21.56}{2.72} = \frac{-2.56}{2.72} = -.94 \text{ s. dev. units}$$

which from Table 4 (appendix) = 17.36% probability;
Less the probability for the prior class interval of 4.65% or




(c.) Between 19 and 21 q/ha, similarly,

$$\frac{21 - 21.56}{2.72} = \frac{-.56}{2.72} = -.21 \text{ s. dev. units}$$

which from Table 4 (appendix) = 41.68% probability;
Less the probability for the prior class intervals

(4.65 + 12.71 = 17.36%)

$$\begin{array}{r} 41.68 \\ -17.36 \\ \hline 24.32\% \end{array}$$

(d.) Between 21 and 23 q/ha

Since this class interval straddles the mean,
1) the probability from 21 to 21.56 =

50 - prior class interval probabilities

= 50 - ($\Sigma$4.65 + 12.71 + 24.32)

= 50 - 41.68%

= 8.32%

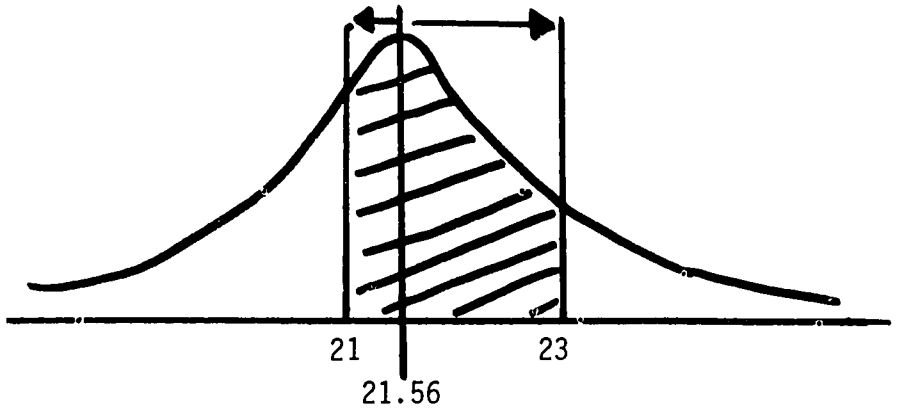
2) the probability from the mean (21.56) to 23. is

$$\frac{23 - 21.56}{2\ 72} = \frac{1.44}{2.72} = .53 \text{ s.dev. units}$$

Since this is on the right side of the mean,
Table 3 (appendix) can be used, and the probability
read directly as 20.19%

3) Thus, the probability for the class interval 21 - 23 is

8.32 + 20.19 = 28.51%




(e.) Between 23 and 25 q/ha

$$\frac{23 - 21.56}{2.72} = \frac{3.44}{2.72} = 1.27 \text{ s. dev. units}$$

which from Table 3 (appendix) is 39.8% (i.e. between the mean and 25);
thus between 23 and 25 is

39.8 - 20.19 = 19.61 %

(f.) Finally, the class interval over 25 q/ha is

50 - 39.8 = 10.2 %

As a check, the sum of probabilities of the class intervals should total 100.

4.65 + 12.71 + 24.32 + 28.51 + 19.61 + 10.2 = 100

4. Using the above data, another table should be prepared, to compare the "Observed Frequency" with the "Expected Frequency", based on the normal distribution of the populalation. Thus: -

Table 4

| Class Interval | Below 17 | 17-19 | 19-21 | 21-23 | 23-25 | Above 25 | TOTAL |
|---|---|---|---|---|---|---|---|
| Observed Freq. | 13 | 32 | 46 | 60 | 41 | 20 | 210 |
| Expected Freq. | 9.8* | 26.7 | 51.1 | 59.9 | 41.2 | 21.4 | 210.1** |

*(i.e. 4.65% of 210 = 9.77, etc.)
** 210.1 rather than 210, due to rounding off

If the null hypothesis is correct, the differences between the Observed Frequencies and the Expected Frequencies in Table 4 above should be statistically insignificant. (Let us assume .05 Level of Significance as the limit.) Thus .06 Level or less (.07, .08 etc.) would be considered insignificant, while .04 level or higher (.03, .02, etc) would be considered significant.

5. To determine the extent of variation, each class interval difference is calculated, using the following formula

$$\frac{\left(\text{Observed Frequency of Class Interval} - \text{Expected Frequency of Class Interval}\right)^2}{\text{Expected Frequency of Class Interval}}$$

The result of this calculation represents <u>the contribution of that class interval to Chi Square</u>.

Thus, for the Class Interval "below 17", the contribution to Chi Square is

$$\frac{(13 - 9.8)^2}{9.8} = \frac{(3.2)^2}{9.8} = \frac{10.24}{9.8} = 1.05$$

The results for all the class intervals are calculated similarly, as shown below.

Class Interval 17-19 $= \frac{(32 - 26.7)^2}{26.7} = \frac{5.3^2}{26.7} = \frac{28.09}{26.7} = 1.05$

Class Interval 19-21 $= \frac{(46 - 51.1)^2}{51.1} = \frac{-5.1^2}{51.1} = \frac{26.01}{51.1} = .51$

Class Interval 21-23 $= \frac{(60 - 59.9)^2}{59.9} = \frac{.1^2}{59.9} = \frac{.01}{59.9} = .00$

Class Interval 23-25 $= \frac{(41 - 41.2)^2}{41.2} = \frac{-.2^2}{41.2} = \frac{.04}{41.2} = .00$

Class Interval Above 25 $= \frac{(20 - 21.4)^2}{21.4} = \frac{-1.4^2}{21.4} = \frac{1.96}{21.4} = .09$

6. The <u>sum</u> of these contributions is the <u>Value of Chi Square</u>

Thus: 1.05 + 1.05 + .51 + .00 + .00 + .09 = <u>2.7</u>

7. Finally, this Value of Chi Square is located in the Table of Selected Critical Values of Chi Square (TABLE 6, Appendix).

Remember, from the discussion of the "T" Test, the number of Degrees of Freedom was the number of samples - 1. Similarly, for evaluating a frequency distribution of this type, the number of degrees of freedom is: -

Degrees of Freedom = Number of Class Intervals - 1

Thus, in our example, with 6 class intervals there are

6 - 1, or 5 degrees of freedom.

With 5 degrees of freedom, and a Chi Square value of 2.7 the probability of the observed frequencies being a normal distribution is quite high - off the table above .30 :--

<u>Well within the 05% limit.</u>

Thus there is no signifi difference between the Observed and Expected Frequencies, and the null hypothesis is accepted.

CHI SQUARE USE IN
NON-NORMAL DISTRIBUTIONS

The foregoing illustrated the use of the Chi Square test of significance for a normal distribution situation. However, Chi Square can also be used for any frequency distribution, whether normal or not, where the Expected Frequency can be pre-determined; or established from additional sample data by use of a "Control Group".

For Example "Soyba" is a new "nutrifood" which we want to test, to see whether or not it significantly increases the growth of malnourished children over the presently utilized foods in our Maternal Child Health Nutrition Program.

We plan to test "soyba" at one of our MCH Clinics.

Since the children attending the MCH clinic are there because presumably they were considered mal-nourished at some earlier stage, they are not a normal distribution of the children in the area. Furthermore, since the children in the clinic are already receiving a "Regular" supplement, and may have been in the program for varying lengths of time, there is no single starting point age/weight/height/skin texture, or other measure which can be used as a "Baseline".

Nevertheless, we do have some age/weight/height standards, and experience in the clinic to guide us. The following categories are established based on the clinician's judgement, and we conduct an experimental feeding program, for a certain period of time, with 146 children randomly selected to receive the new Soyba, and the remaining 179 receiving only the regular diet. Results are as follows: -

| TABLE 1 | A<br>Improved Markedly over Normal Expectation | B<br>Improved Slightly over Normal Expectation | C<br>Normal Improvement | D<br>Less than Normal Improvement | E<br>No Improvement | F<br>TOTAL |
|---|---|---|---|---|---|---|
| 1) EXPERIMENTAL SOYBA GROUP | 10 | 75 | 27 | 28 | 6 | 146 |
| 2) CONTROL REGULAR GROUP | 5 | 8 | 103 | 23 | 40 | 179 |
| 3) TOTALS | 15 | 83 | 130 | 51 | 46 | 325 |

The null hypothesis is that there is no sigficant difference between the Experimental Soyba Group and the Control Regular Group. If this hypothesis is correct, both the Experimental and Control "Samples" could have been drawn from the same Total Population with only differences due to Random Sampling Errors.

Thus since 15/325 or 4.6% of the children are in Group "A", (totally) a similar percentage should be expected for both groups (1) and (2).

For the Experimental Group this would mean 4.6% x 146 children, or 7, rounded up.

Similarly, for the Control group this would mean 4.6% x 179 = 8.

The Expected Frequencies are thus computed for every "Cell" and for convenience displayed in the format shown in Table 2, below.

TABLE 2

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| (1) | 10 / 7 | 75 / 37 | 27 / 58 | 28 / 23 | 6 / 21 | 146 |
| (2) | 5 / 8 | 8 / 46 | 103 / 72 | 23 / 28 | 40 / 25 | 179 |
| (3) | 15 | 83 | 130 | 51 | 46 | 325 |
| % | 4.6 | 25.5 | 40 | 15.7 | 14.2 | 100% |

We are now ready to compute the contributions to Chi Square, based on the differences between the Observed and Expected Frequencies in each "Cell".

If the null hypothesis is correct, the difference between O and E for any cell should be negligible; taking into consideration the relative size of the numbers. (i.e. a difference of five (5) between 100 (O) and 95 (E) is less significant than the same absolute difference of 5 between 15 (O) and 10 (E). One is 5% difference while the other is 33.3% difference.

## CONTRIBUTIONS TO CHI SQUARE

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (1)A | $\frac{(10 - 7)^2}{7}$ | = | $\frac{3^2}{7}$ | = | $\frac{9}{7}$ | = | 1.29 |
| (1)B | $\frac{(75 - 37)^2}{37}$ | = | $\frac{38^2}{37}$ | = | $\frac{1444}{37}$ | = | 39.03 |
| (1)C | $\frac{(27 - 58)^2}{58}$ | = | $\frac{(-31)^2}{58}$ | = | $\frac{961}{58}$ | = | 16.57 |
| (1)D | $\frac{(28 - 23)^2}{23}$ | = | $\frac{5^2}{23}$ | = | $\frac{25}{23}$ | = | 1.09 |
| (1)E | $\frac{(6 - 21)^2}{21}$ | = | $\frac{15^2}{21}$ | = | $\frac{225}{21}$ | = | 10.71 |
| (2)A | $\frac{(5 - 8)^2}{8}$ | = | $\frac{(-3)^2}{8}$ | = | $\frac{9}{8}$ | = | 1.13 |
| (2)B | $\frac{(8 - 46)^2}{46}$ | = | $\frac{(-38)^2}{46}$ | = | $\frac{1444}{46}$ | = | 31.39 |
| (2)C | $\frac{(103 - 72)^2}{72}$ | = | $\frac{31^2}{72}$ | = | $\frac{961}{72}$ | = | 13.35 |
| (2)D | $\frac{(23 - 28)^2}{28}$ | = | $\frac{(-5)^2}{28}$ | = | $\frac{25}{28}$ | = | .89 |
| (2)E | $\frac{(40 - 25)^2}{25}$ | = | $\frac{15^2}{25}$ | = | $\frac{225}{25}$ | = | 9 |

$\Sigma$ = Chi Square = 124.45

From table 6 (Appendix) of Critical Values of Chi Square, the probability of a null hypothesis for that value can be determined.

However, first, the Number of Degrees of Freedom must be calculated. With a working matrix (i.e. the lines & columns, excluding the totals, and the expected frequencies) the formula for degrees of freedom is calculated as follows: --

$$\text{Degrees of Freedom} = \left(\text{Number of Columns} - 1\right)\left(\text{Number of Rows} - 1\right)$$

NOTE: (When there is only 1 row, only the number of columns is calculated.)

Thus in our example, we have 5 columns (A,B,C,D & E)
and 2 rows ( 1 & 2)

Therefore the # of Degrees of Freedom to use will be

$$DF = (5 - 1)(2 - 1)$$

$$= 4 \times 1$$

$$= \text{Four } (4)$$

From Table 6, a Chi Square value of 124.45 with 4 Degrees of Freedom is not even shown. The probability is less than .001 (or 1 in 1000) that there is no significant difference between the EXPERIMENTAL and CONTROL Groups.

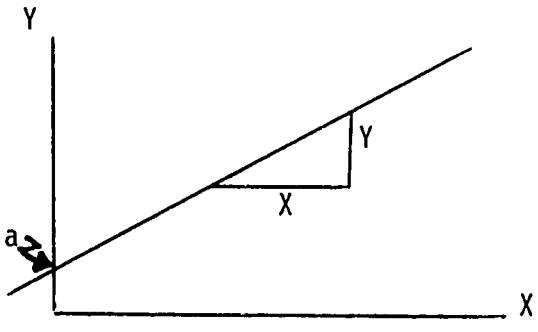
Thus the Null Hypothesis is rejected, and we conclude that while not a 100% "cure-all", Soyba does indeed have a most significant effect over the regular diet provided, in feeding most malnourished children.

# REGRESSION ANALYSIS

Frequently, management desires to make forecasts to establish realistic targets, and/or make predictions for policy analysis, based upon current trend information. This can be done by a technique known as regression analysis, which develops the "line of least squares" in the available data.

For example, continuing the previous illustration where the correlation between yields and loans was made, assuming a cause - effect relation is plausible, management might want to determine the appropriate loan size to achieve a particular level of production; assuming a linear relationship.

Essentially, the line of least squares is obtained by solving for two simultaneous equations with the data developed for the correlation analysis; then substituting the values in the formula for a straight line: $Y = a + b X$




Where

- Y = value of the Y axis data
- X = value of the X axis data
- a = the point where the line intercepts the Y axis, and the value of x is 0.
- b = the slope of the line, determined quantitatively as $\frac{\text{Y value}}{\text{X value}}$

The line of least squares is found by solving for the following equations.

(1) $\sum Y = na + b\sum x$

(2) $\sum XY = a\sum x + b\sum x^2$

Where

- $\sum Y$ = sum of Y value
- $\sum X$ = sum of X value
- $\sum XY$ = sum of XY value
- n = number of pairs of data
- $\sum X^2$ = sum of $X^2$ values

## EXAMPLE OF REGRESSION ANALYSIS

From the survey data, the following table is developed to determine the value of the elements in the formulae:

| Independent Variable "X" (000's Rupiah) | Dependent Variable "Y" (Yield q/ha) | XY | $X^2$ |
|---|---|---|---|
| 110 | 18 | 1980 | 12100 |
| 210 | 17 | 3570 | 44100 |
| 370 | 20 | 7400 | 136900 |
| 420 | 23 | 9660 | 176400 |
| 560 | 26 | 14560 | 313600 |
| 640 | 22 | 14080 | 409600 |
| 770 | 33 | 25410 | 592900 |
| 850 | 30 | 25500 | 722500 |
| 900 | 35 | 31500 | 810000 |
| $\Sigma X_i = 4830$ | $\Sigma Y_i = 224$ | $\Sigma XY_i = 133660$ | $\Sigma Xi^2 = 3218100$ |

n = Number of Pairs = 9

(1) $\Sigma Y = na + b\Sigma x$

(1) $224 = 9a + 4830b$

(2) $\Sigma XY = a\Sigma x + b\Sigma x^2$

(2) $133660 = 4830a + 3218100b$

$\Sigma X_i = 4830$

$\Sigma Y_i = 224$

$\Sigma XY_i = 133660$

$\Sigma X^2_i = 3218100$

First, we can simplify equation (2) by dividing it through by 10, thus: -

$$(3) \quad 13366 = 483a + 321810b$$

Next, we must eliminate one of the unknowns (either "a" or "b") from both equations (1) and (3). This can be done by testing for a multiplier that will set equation (1) and (3) equal for the value of "a". This is done by dividing 483 by 9, thus: -

$$\frac{483}{9} = 53.66666$$

Now multiply equation (1) by the multiplier to obtain equation (4) and round off, thus

$$(4) \quad 12021 = 483a + 259210b$$

Subtract equation (4) from equation (3):

$$\begin{array}{rcrcr} 13366 & = & 483a & + & 321810b \\ -12021 & - & 483a & + & 259210b \\ \hline 1345 & = & 0 & + & 62600b \end{array}$$

Therefore $b = \frac{1345}{62600} = .022$

Substitute this value of "b" in equation (1)

$$224 = 9a + (4830 \times .022) = 9a + 106.26$$

Transposing, $9a = 224 - 106.26$, or $117.74$

Therefore $a = \frac{117.74}{9} = 13.08$

These two values for "a" and "b" can now be substituted in the straight line equation $Y = a + bx$

$$Y = 13.08 + .022X$$

and graphically, a line of least squares can be plotted from any two data values in the table. For example,

Where $X = 110$ $\quad Y = 13.08 + (.022 \times 110) = 13.08 + 2.42 = 15.50$

and

Where $X = 900$ $\quad Y = 13.08 + (.022 \times 900) = 13.08 + 19.8 = 32.88$

By extrapolation and inspection, the values of either X or Y can be estimated for a given value of Y or X. These values can also be obtained by calculation, using either formula.

$$Y = a + bX \quad \text{or } X = \frac{Y - a}{b}$$

For example, to determine the appropriate loan size in order to obtain a harvest of 40 q/ha, from the preceding data and assuming a linear relationship

$$X = \frac{40 - 13.08}{.022} = \frac{26.92}{.022} = 1223.64$$

or approximately 1224 thousand rupiah, rounded off.

Of course, potential farm yeild is not solely a function of loan size, as there are other constraints. There is no guarantee that this size loan will produce such a yield. Nevertheless, the trend of prior experience indicates that such size loan would be an appropriate contribution towards the higher yield until diminishing returns set in from other factors, and technological ceilings were reached.

## DESIGNING THE SURVEY

The foregoing section has illustrated how to manipulate the data obtained from a survey.

This section will provide some general guidelines for gathering the data.[1]

Perhaps the most important point to remember is that despite the formidable arrays of formulae, statements of confidence, and tests of significance, the quality of data collected is the most important aspect of any survey. The "answers" provided by statistical manipulation can only be as good as the information upon which they were based. The adage "Garbage In, Garbage Out" should be borne in mind when designing surveys and gathering data, so that the processing of the data will bear useful results.

Although at the outset it is recognized that field surveys of economic and social development projects cannot produce the precision associated with carefully controlled laboratory experiments, nevertheless, a carefully designed and implemented survey can provide useful information and insights for the project manager willing to expend the effort.

Furthermore, although the cost of gathering data by survey is often used as the reason for not utilizing this method, it should be pointed out that it is often less than the hidden cost of typical 100% census type reporting from field technicians (who must divert a portion of their work-day from technical to administrative functions); and certainly more useful.

---

1 Richard Bernsten, CRIA/IRRI Cooperative Program, P.O. Box 107, Bogor, Indonesia, has prepared an excellent draft text: - Design and Management of Survey Research: A Guide for Agricultural Researchers, 1979, which covers this topic in much greater detail. I highly recommend its use by those wishing greater depth in this area.

## WHY CONDUCT A SURVEY?

1. THE FIRST STEP is to outline the purpose of the study in writing. The more precisely this is done, the better off you will be, and the sharper the study questions can be focused. Remember, that data should be gathered as a basis for learning and doing something. If nothing is to be (or can be) done with the data, then there is little purpose in collecting it.

Surveys are usually requested to provide answers for managers on problem situations which they are encountering. Sometimes, however, there is no particular "problem" but management just wants to establish a "baseline" on particular aspects of a situation against which to subsequently measure progress. Alternately, a manager may wish to be kept informed on the status of key technical areas during a project's implementation.

In any event, the first task in sample survey work is to develop a concise statement of the purpose or problem. Frequently, a manager's request is only half formulated, ambiguous, or a statement of observed symptoms about which he (she) is concerned. Often this is expressed as a question. You must get clear guidance on what to study before acting to develop a questionaire, or you will waste a lot of time and effort. Once the purpose or problem has been stated in an objective manner, the need for a study will become clearer, and detailed survey questions can be formulated.

What new knowledge do you want to obtain; what hypothesis do you wish to test; what problem do you desire to solve- what types of information should shed new light on the situation to provide guidance to management about the project, or some aspect of it?

"Brainstorming" -- is a useful technique for establishing the purpose. Accepting the suggestions of several other interested participants, uncritically, is a good way to get started. You will be surprised at the variety of ideas that can come forth from a short session. Careful follow-up discussion of each of these ideas will usually lead to formulation of one or two purposes which are far superior to anything presented in the initial go-round.

Fom the options available, select a single purpose. Many studies come to grief because they try to be too ambitious and do too much. In survey research, the "rifle" approach is generally more effective than the "shot-gun".

2. Question Development Sometimes managers have well established ideas about what they want, and can be of assistance to you in formulating the questionaire. Often, however, they don't know precisely what they want, and leave it up to you. In any event, you should not limit yourself to (or accept uncritically) the indicators that management requests. In many instances, the manager may not be totally aware of the project in terms of key indicators and may merely accept those that are familiar, or in use when he (she) became acquainted with the project, or those with which he (she) is familiar from previous specialized training. However, since you will have to persuade the manager that your survey findings are meaningful, it is well to consult with him (her) prior to establishing the data elements to survey.

A useful technique is again to "brainstorm" with some "technical experts" the possible raw data elements that could be useful in analyzing the established purpose of the survey. Don't worry about their relative importance, how feasible it would be to obtain them, their source or formatting at this stage, and don't become alarmed at the number. Just list them, for later screening.

During this brainstorming session, many "intermediate" data elements and indicators (such as rates, ratios and percentages) will be suggested. Include them, but remember that such data cannot be obtained directly. They must be calculated or derived from more basic data. Thus you will also have to identify these elements. (For example, in order to obtain the "average yield" you should obtain the total production and the total land area on which it was produced, as well as determine the most appropriate measures - kilo; bushels; hectares; acres; etc. If you cannot ask direct questions such as income, you may have to try to determine this from "proxy" questions such as "purchases" for food, luxury items etc.

A help in this stage is to screen existing files and reports to see what kinds of data elements have been reported and utilized in the past on this, or similar projects.

3. Limit the Number of Questions Management's ability to ask questions always exceeds the staff's capacity to answer them! This is also true in developing questionnaires. There are so many interesting things that would be "nice to know" once you start burrowing into a topic. Thus, refer back to the first step - the purpose. What is it you are supposed to be finding out? Don't get carried away by the enthusiasm of the brainstormers in identifying potential topics.

4. Develop a Draft Final Report An extremely useful technique is to sketch out a sample final report, complete with a set of hypothetical data. In this way, you can check whether the purpose of the study can be answered from the questions asked, or whether additional raw data questions need to be included in order to provide "intermediate" ratios, rates, etc. This is also an excellent way to purge unnecessary questions from the list -- questions which were thought important earlier but which play no part in answering the purpose of the study.

5. Review existing data sources. Often secondary data (data gathered by someone else for some other purpose) is already availabe; or at least a substantial portion of it. It is not always necessary to "re-invent the wheel". For example, most Indonesian village "statistics offices" maintain data on population by age and sex; hectares in various land use classifications (irrigated, rainfed, upland, home gardens, estate crops, etc.); number of farms by size class; hectares planted and harvested of each crop by month; etc. In addition, local officials can usually provide a lot of additional information about activities in their areas. While some of these statistics may not always be completely accurate, they can be extremely helpful in establishing basic village characteristics, for designing sampling "frames".

6. Target Your Purpose

| | |
|---|---|
| If you want to simply get a basic orientation, or "feel" for the project situation | Judgement sampling may suffice, and the sample size can be small. |
| If you want to be able to generalize about the situation or study cause/effect relationships | Scientific (random) sampling will be necessary |

7. Design the Questionnaire You are now ready to design the format of the questionnaire itself.

There is no such thing as an "ideal" questionnaire Nevertheless, there are certain useful ground rules that can facilitate their construction.

## GENERAL PRINCIPLES OF QUESTIONNAIRE DESIGN

a. Single Purpose: Whenever possible, limit the survey to a "single purpose". A poor. but frequent, practice is to try to accomodate the needs of several different management groups in one survey, rationalizing that "it doesn't take much longer to ask another question while you are there" and "it is cheaper than running a separate survey" etc. Unfortunately, a "multi-purpose shopping expedition" usually results in a cumbersome census-type document that may never be completely analyzed, but which will effectively hinder the gathering and processing of data for the primary intended purpose. Furthermore, a sample survey that is properly structured to meet a specific need is generally not a suitable vehicle for answering multi-purpose questions from the same sample base. Consequently, even if analyzed, the additional data may be invalid.

b. Limit the Number: Each question takes time (and costs money) to ask, process and analyze. Therefore be selective. Screen each proposed question carefully and decide whether such answer can be more readily obtained elsewhere. If a questionnaire becomes too long, attention and accuracy of both interviewers and respondents decreases.

c. Avoid "Leading"Questions. Many people respond to please the questioner, or to avoid embarrassment they tell what they think he/she wants to hear. Others deliberately distort their answers depending how they perceive the answer may be used. You cannot eliminate all problems in this area, but you can improve the survey considerably by being careful to phrase your questions as objectively as possible to avoid hinting at the "desirable" answer.

d. Avoid "Memory" Questions. Questions which rely on an individual's recall and cannot be verified in any meaningful way are likely to have a high degree of inaccuracy. Also, the longer the period of recall, the more inaccurate the answer is likely to be.

e. Cross-Check Questions. If there is likely to be a strong element of doubt or distortion in the answer, provide for some "probing" or objectively verifiable cross-check questions, if possible. (Note: it is not usually necessary to record the responses to probing questions.)

f. Clarity. Even though the question is clear to you, and you know precisely what you mean by it, make sure that others will interpret it in the same way. Otherwise, each interviewer will interpret it in the field in his/her own terms, and you may end up with confusing and/or useless results. If necessary, rephrase the question, and/or provide additional guidance on what it means, definitions,etc.

g. Pre-Test your questions on others before deciding on the exact wording to be used in the questionnaire. This is absolutely essential. Questions which appear clear and straight-forward to the survey designer may prove to be confusing to the respondent and elicit answers which are not relevant, because of cultural problems.

h. Language. Make sure the questions are phrased in the appropriate familiar dialect of the respondent, to ensure understanding, Not bureaucratese.

## QUESTIONNAIRE FORMAT

The following guidelines are provided to facilitate both the gathering and tabulation of the data.

a. Identification. Each question and possible response should be uniquely identified with either a number, letter, or both; so that in the processing and analytical stage they may be readily referred to without repetition or reference to the subject matter itself.

1. Question.....
.....? a. _____Yes
b. _____No

b. Multiple Choice. Structure the format so that as many questions as possible can be answered with a check mark. Spell out categories in which responses are expected.

2. Question.....
.....? a. _____Always
b. _____Sometimes
c. _____Never

c. Numbers. When numbers are required for an answer, indicate the unit that is required. Leave space for raw data to be recorded in other units. (Often in the field, responses are not in terms of the units desired, and recalculation must be done prior to tabulation.) If no space is availabe, the raw data may be inserted where the standardized unit response should go, which leads to gross errors.

3. Question.....
.....
.....? a. _____Metric Tons

d. Spacing. Leave plenty of "White Space" around each response. The answer is going to be filled in under field conditions, not small typing. Also make allowances for comments by the interviewer.

e. Block Answers. Standardize the manner for recording answers. Usually, a left hand or right hand column is easier for processing than responses scattered throughout the form, or on a single line. For multiple responses of varying length. It is easier to both record and tabulate the answers when the blank space precedes, rather than follows the item. For example: -

4. a. _____Yes
b. _____No
c. _____Don't know
d. _____Haven't made up my mind yet

Question: .....
.....
.....?

Instead of: -

4. Question: .....
.....? a. Yes_____ b. No _____ c. Don't know _____ d. Haven't made up my mind yet _____

or: -

4. Question: .....
.....
.....
.....? a. Yes_____
b. No_____
c. Don't Know _____
d. Haven't made up my mind yet

f. Think Positively. Don't phase questions negatively if it can be avoided, and never use double negatives. It only confuses people.

## PRELIMINARIES -- BEFORE CONDUCTING THE SURVEY

Before charging off and interviewing people, you need to get preliminary guidance from management on its desires and priorities; then conduct a feasibility study to plan the survey itself.

### 1. MANAGEMENT GUIDANCE

How important does management consider this survey? Although you may have been working on it full time, and it is your number 1 priority, it may be of less importance to management. Once management has established its priorities, you have a basis for determining what funding, equipment, personnel and other administrative support you may reasonably expect, to carry out the survey.

How accurate, or precise, do the results need to be in order to meet management's objectives? Remember, data collection and analysis is time consuming, and accuracy can often only be obtained at a price, and even then, perfection cannot be provided from sampling. Diminishing returns for extra effort always exist at some point, while minimizing time and cost should be an important consideration.

When does management want the results? Today, tomorrow, next week, or next month? If the results of the survey are received after the need for it has passed, the entire effort may be useless, no matter how targetted the questions, extensive the survey, sophisticated the analysis, or beautiful its presentation. Remember also, it takes time to process the data and analyze it after the field work has been completed.

What is the budget limitation for the survey? It is useless to gather a lot of information, if you cannot afford to analyze it afterwards.

When trade-offs have to be made between accuracy, timing and cost, the various options should be discussed with management before the study is undertaken; not offered up as excuses afterwards for a less than adequate job!

## 2. PLANNING THE SURVEY

When you know what the purpose of your survey is, have reviewed the possible data, have developed a questionnaire, and are ready to go to the field, the survey supervisor(s) should make a preliminary field visit and take a few judgement samples in order to become acquainted with the study environment and potential problems. This will help to make the actual study more realistic, particularly in terms of scheduling.

Administrative Coordination. Even though the actual target respondents (farmers, for example) may not be forewarned of your coming to conduct a survey, it is usually desirable that the various government administrative officials in the area are advised of your intended visit and the purpose therefore. They can often provide additional insights, maps, names, and other essential data; as well as logistical support and guides for surveyors during the conduct of the survey itself. As long as your survey is carried out on a random basis, there is little possibility that this advance notice will enable the local administration to "hide" the problems that may exist.

Data. You should look for repositories of information pertinent to your survey -- offices, addresses, phone numbers, names of key individuals, sources of possible master-lists of respondents, and other secondary data in which you are interested. Information on local market days, holidays, and key individual schedules are all useful, so that the actual survey can be timed for best effect. There is nothing more frustrating than spending several days to visit a remote province area, only to find that the key individuals whom you need to see are attending a conference in Jakarta that week! You can also gather information which will enable you to divide the survey area into approximately equal clusters, in the event that cluster surveying has to be resorted. to.

Environment. You should park your personal transportation for one day, and make a "dry run" attempt at reaching a few potential respondents (on a judgement basis) in another village, using only locally available transportation, and/or walking. Not only will this give you a much better feel for the environment than riding around in a chauffeur-driven air-conditioned jeep, but it will enable you to estimate the "worst-case" timing and interviewing expectations when your staff goes out later. Without such experience, you may place unrealistically high demands for interviews upon your field workers, which could lead to major difficulties. Either they finish on time, but with an inadequate sample; or they continue until they reach their assigned quota, delaying the processing of the data as a consquence. In either event, as the survey manager, you will face problems which you could have avoided with a little preparation.

Estimating the Sample Size. Once you have management's criteria, management's priorities, an awareness of the environment in which the survey will be conducted and estimates of travel and interview time, you should be able to estimate the appropriate sample size (as outlined earlier).

Generally, it is preferable that field staff travel in teams of two or three, rather than as individuals. In this manner, they can support each other, maintain their morale, and in the event of an emergency can usually call for assistance.

Knowing the number of teams you have available; the approximate number of interviews that can be carried out in a day, per team; and the total area to be covered; will help you to determine whether a random sample approach is feasible, or whether clustering will be required. If clustering is required, strive for the maximum number of clusters feasible within the time available for the survey. Don't forget to include travel time to, and return from, the project area, as well as non-work time due to holidays, etc.

An important point to note is that interviewing may not always be done during the work day, particularly where farmers are the respondents. They may usually be located at home during the evenings while during the day they are at their fields. Thus, interviewers should be briefed to plan their work accordingly.

The accuracy of a survey can be improved by drawing a larger sample, while obtaining less than the required minimum number of samples may severely impair the results. In the field, it is difficult to retrace one's steps to resample individuals from areas that have already been covered, and frequently individuals pre-selected may not be available. Therefore, it is a wise practice to oversample in advance, to utilize the survey team's time to the maximum extent possible. Also editing the information sheets during slack time in the day before the village (interview site) has been left is timesaving and very worthwhile. It is much easier to backtrack locally for additional data. Once you have returned to the main office, incomplete questionnaires may have to be discarded.

Team Preparation. When the samples (or clusters) have been determined, and assigned to various teams, the teams should also be furnished with a complete schedule of interviews, so that in the event of an emergency, each team knows where and when other individuals may be contacted.

Teams should also be informed of the administrative arrangements, and review the technical questionnaire again in the light of the supervisor's experience.

Scheduling A detailed work schedule for completing each major step of the survey and its analysis must be prepared at the outset, and then adhered to, in order to complete the report in time for management.

Field Supervision Field supervision and back-tracking of survey teams is an important aspect of any survey, regardless of the fact that the survey teams may be skilled, trustworthy individuals. It enables the survey supervisor to adjust his/her expectations to realistic standards that can be maintained in the field environment, and to evaluate the quality of the survey data after the interviews have been completed. The presence of a roving field supervisor enables many difficulties and misunderstandings to be resolved on-the-spot, provides additional incentive to do a thorough job, as well as giving the survey teams some reassurance that they will not get lost and left; for even in remote areas, individuals leave an "audit trail" that can be tracked. It enables the field supervisor to adjust assigned work-loads if some teams get ahead or fall behind in their interviewing for some reason or another, or encounter unforeseen obstacles.

Most important, field supervision and follow-up can provide the supervisor with an insight into the field enviornment that cannot be obtained from reading the reports in the office in the capital city. A much deeper appreciation is gained for the problems of data gathering, and in the day to day work of the respondent. Active field supervision is invaluable for team-building, both for esprit de corps and high quality work.

## CONDUCTING THE SURVEY

Some general guidelines which should be observed are as follows:

a. Brief The Interviewers. Ensure that all the interviewers have a common understanding of the purpose of the survey, definition of terms, the meaning of the questions to be asked, and a uniform way to record answers. Provide guidance on procedures to follow when they encounter difficulties. If possible provide for a "dry run" interview session to supplement the orientation process.

b. Interviewing Procedures. Differences in interviewers personalities and questioning techniques will affect the responses they obtain. The effect of this can never be eliminated, but it can be minimized. The following are general points that should be kept in mind by the interviewers.

Introduction -

Introduce yourself.
Verify who you are speaking to.
Put the individual being interviewed at ease.
Tell the reason for the survey and the use to which it will be put.
Tell the individual how he/she was selected to be interviewed.
Assure him/her of confidentiality or anonymity of results.
Tell him/her how long the interview is likely to take.
Ask if the time is convenient for an interview now.
See whether there is a suitable place to conduct the interview.
(Privacy is often desirable, especially when asking personal questions. However, in many field situations, this may be impossible to obtain as you may become the focal point of the village's "live entertainment".)

Conducting the Interview -

Use your judgment whether to follow a structured "questionnaire format" reading off each item; or an unstructured interview style using the questionnaire as a check list, but employing a lot of additional extemporaneous "probing" questions. The structured style may get a response to every answer, but you may scare or inhibit the response, especially if you record the answers in the presence of the person being interviewed. (On the other hand, some people feel more important when they see you writing down what they say, and often think if you don't write it down, you may forget and/or fail to pass on their comments.) Unstructured interviewing generally leads to a much more wide-ranging discussion, takes longer and may gather much supplementary data which may also be useful. However, even though important to note, it is not generally possible to statistically analyze such additional data. Sometimes it is critical that every respondent be given only the precisely formatted question, so that responses are standardized. Extemporaneous questioning often introduces interviewer bias.

c. Field Computations. Use local or familiar measures, and minimize computations by the respondent. Get raw data which can be converted to percentages, etc., later. Most people perform poorly in mental arithmetic, therefore record information in the terms which it is given to you. Note the conversion factor for later use in obtaining the desired unit measures.

## CAUTIONS TO OBSERVE IN CONDUCTING SURVEYS

Make sure you have the right person before you start to interview. Avoid leading questions, and verify responses for accuracy by cross checking and/or back-track repetition. Often individuals misunderstand what you are asking, or only tell you what they think you want to hear. They may be trying to impress you, gain your sympathy, or avoid discussing the topic at all for lack of knowledge or fear of embarrassment.

For instance, a farmer may understate his yield if he thinks he may be penalized (by taxes or rents), or overstate it if he is trying to compete for "farmer of the year" in the Green Revolution competition! Therefore, you may have to repeat your questions several different ways to ensure that they are understood and the person being interviewed is responding accurately to the best of his knowledge. He may also respond inaccurately if his neighbor, or the village Headman is present during the interview.

Remember - Do not promise anything (except to pass on the information) unless you have authority to take corrective action. You are usually only interviewing in the village as an observer and gatherer of facts. On the other hand the individual being interviewed may regard you as a representative of the government who can and should do something about a local problem situation. Idle promises will only result in a lack of confidence and lessen cooperation the next time around. Don't assume anything!

## ANALYZING THE DATA

After the data has been gathered and recorded on the survey forms, it must be edited, weighted, calculated and interpreted.

EDITING. Prior to use, raw data on survey forms, gathered by different enumerators, must be screened by a staff using consistent guidelines. The principal purposes of this are to review for Clarity, Internal Consistency, Correction and Mark-Up for further processing. Preliminary editing in the field should speed up this process.

Clarity. Data recorded by enumerators under field conditions is sometimes almost illegible and/or unintelligible to a staff editor. Numbers may be illegible, and many cryptic comments may have been added to the standardized responses which might qualify the answers recorded from "Yes" to "Yes, but..." When-ever possible, questionable items should be reviewed with the individual making the survey. However, this is not always possible, and even then it does not always produce success. The individual cannot always read his/her own writing, and/or does not recall the context in which the comments were made, even though they may have seemed meaningful at the time.

Where multiple choice responses have not been used, the editing staff has an extremely difficult task of developing a standardized scheme to classify "open-ended" comments received. (In fact it is often impossible at this late stage, since it is highly unlikely that all respondents would comment (or that different enumerators would solicit unstructured comments) in any systematic manner. This emphasizes the need to carefully plan and structure the survey before gathering the data, not afterwards.)

It may also develop that some things which were overlooked, or thought not to be important when designing the questionnaire, actually have great significance while other questions may no longer be pertinent. Thus some preliminary modification (or even elimination) of questions and responses may be necessary.

Internal Consistency. 1) Check marks may have been placed in more than one option of multiple choice questions even though it was originally specified that only "one of the above" was to be checked. There may be clarifying comments in the "white space" as to why, or there may be no explanation at all. 2) With "number" responses, editing is frequently required to recalculate the recorded values into the standardized units requested. Sometimes the conversion factor is provided, sometimes it has been overlooked.

Correction. Decisions have to be made on how to treat questionable data. Should the data be rejected outright as erroneous; counted at face value regardless of its apparent error; or retained but reduced in value, with an attempt to figure the "intent". This is all part of the editorial task.

Mark-up. Finally, to simplify the data processing task which follows, it may be necessary to transform the check marks in the standardized responses into "Base numbers". For example, if a series of questions were asked about rice farming which are to be analyzed in terms of hectares, the hectarage of a particular respondent's farm will be the base number to substitute for the check marks on his survey form.

To illustrate the problems of editing, a series of questions and responses on a farmer's farming practices are shown "before" and "after".

BEFORE

1. 2.3 has Area Farmed

| | a. YES | b. NO | DID YOU: - | Comments |
|---|---|---|---|---|
| 2. | x | x | use certified HYV seed? | Only for 1.5 hectares. |
| 3. | | x | use recommended amount of fertilizers? | Not enough urea available. |
| 4. | x | | use herbicides? | |
| 5. | x | x | receive credit from the bank? | Credit received too late for land preparation and transplanting. |
| 6. | x | x | receive assistance from the government technician? | Technician helped prepare farm plan and budget. Did not see him after that. |
| 7. | q/ha | | What yield did you obtain? (20 kilos/q) | 135 sacks (at 40 kilos/sack) |
| 8. | rupiah/kilo | | What selling price did you get? (000's rupiah per kilo) | Sold 80 of the above sacks for a total of 2,500,000 |

AFTER

| 1. | a. YES | b. NO | DID YOU: - |
|---|---|---|---|
| 2. | 1.5 | .8 | use certified HYV seed? |
| 3. | | 2.3 | use recommended amounts of fertilizers? |
| 4. | 2.3 | | use herbicides? |
| 5. | | 2.3 | receive credit from the bank? |
| 6. | | 2.3 | receive assistance from the government technician? |
| 7. | 117 q/ha | | What yield did you obtain? (20 kilos/q) |
| 8. | 781 rupiah/kilo | | What selling price did you get? |

$$\frac{135 \times 40}{20} = 270$$

$$\frac{270}{2.3} = 117.39 \text{ quantales}$$

$$\frac{2,500}{80 \times 40} = 781.25$$

NOTE: Question 5 & 6 could be edited in several ways. It is important therefore that a decision be reached by the "editor" and held to consistently throughout all subsequent form editings.

CALCULATING THE DATA If the data is to be hand-tabulated, a desirable practice is to have it all processed simultaneously by each team, independently.

Each team should place its own identification mark on its own questionnaires; the questionnaires from each team grouped and sequentially numbered; then the questionnaires redistributed so that no team edits its own questionnaires.

After editing preliminary calculations, and double checking, the total questionnaires can be rank ordered from high to low (or low to high) for each question, then the data transcribed onto a blackboard. From this a frequency distribution can be constructed and the mean, standard deviation, standard error, and confidence computed simultaneously by each team independently.

If these calculations are carried out simultaneously, assuming 5 separate teams, then five concurrent checks are made of each item at each step. Further processing and calculation should be delayed until agreement is reached by all teams. It is surprising how difficult it is to develop the right answer first time through -- therefore the teams should proceed slowly and carefully! It is extremely time consuming to have to recompute the data over again because of calculation errors.

When the information has been computed for one question, the entire process should be repeated for the next question; and so on, until all the questions have been calculated.

The entire data base should be reviewed, rechecked, and recalculated after a suitable time interval (the next day) to assure that it is as correct as possible.

Then, the data should be analyzed and interpreted for comparisons, trends and significance - by the use of data tables, graphs, histograms, and significance testing. This too should be done independently, simultaneously, by teams. Findings should be reviewed and discussed by team members before proceeding. In this manner, as in brainstorming, numerous insights surface which would normally be overlooked by individuals.

## PRESENTATION OF RESULTS

The final step in the survey process as far as you are concerned is to present the findings of the study. This is a very critical phase. In fact it is the point of the whole exercise. Designing questionnaires, interviewing, and statistical manipulations of various kinds were just a means to the end - providing answers to management and possibly furnishing them with some additional insights into a program for which they have responsibility. Many well-conceived, planned, and executed surveys fail miserably at this stage because they do not communicate with their intended audience. Remember management has not had the experiences that you have just had in traveling, interviewing, researching and analyzing this survey data -- so it is difficult for them to empathize with you. They will only know what you tell them, plus any impressions they may have gathered through judgement samples of their own, and other reports. It is your job to see that they get the message loud and clear.

A frequent problem is that after doing all the foregoing work, survey technicians are reluctant to summarize. They want the boss to see all the detail of everything they did so that he doesn't "miss" anything. Nothing is left out, no matter how insignificant. Unfortunately in such cases, everything is usually missed because after picking up the weighty tome and ruffling its pages, it is set aside until there is time to read it thoroughly, -- a time which rarely comes to the busy executive.

The first principle of report writing therefore is to purge -- drastically! The second principle is to simplify what is left. And then, Summarize! If you must include details because they are too precious to throw away, consider putting them in a technical appendix in which other researchers and technicians may delight to wallow but which the manager may ignore if he so chooses. Above all else -- provide the reader with a one page summary of the purpose of your study, your findings, and your conclusions. If you don't get it on one page, you haven't purged, simplified and summarized enough.

Presentation is a whole subject in itself. I will therefore limit myself to a few major points.

## MAJOR POINTS IN WRITING SURVEY REPORTS

- Avoid "technical jargon", abbreviations and acronyms unless you are sure that your intended reader is completely familiar with them.

- Round off numbers wherever possible, it won't usually distort a thing. Even though you may have been gathering data in hectares, or even tenths of hectares, when the final report is written you will probably be dealing in thousands, tens of thousands, even hundreds of thousands; so avoid data clutter and round off.

- Use graphs and charts instead of tables wherever possible -- usually it is the trend of the data that is important rather than the precise numbers. Therefore identify the point you are trying to make, then make it, simply.

- Where you do use tables - whenever possible get all the data on one page. There is nothing that will distract a reader from gleaning the message from your table, more than having to flip pages.

- Tables should be organized so that a single message is highlighted. Comprehensive matrices of basic data are only useful for researchers to analyze -- they do not communicate to management until they are interpreted. If you need the comprehensive table - the appendix is the place for it. Extract from it the point you wish to make, and then prepare a condensed version in the text at the appropriate point.

- After using a table, summarize in the narrative what the reader is supposed to learn from studying it. Some people have a mental block against numbers and only read the text -- skipping over tables.

- If you need to go into detail on a point, and it would clutter up the text, use a footnote. Remember however that a footnote is best seen at the foot of the page on which the point is raised. "Footnotes" relegated to the back of the text rarely (if ever) get read in relation to the points they are clarifying.

- Single space the narrative. This flies in the face of most research oriented training where double spaced text is required, but unless it is a draft where extensive rewrite is to be expected, no useful purpose is served by double spacing. It makes the report twice as bulky as it need be, it wastes paper, and it usually inhibits readability because the "concept density" -- the number of thoughts per page -- is halved.

## BRIEFINGS

In addition to the written report, be prepared to present an oral briefing. Used wisely, charts, slides and graphs can be much more effective in getting the message across than volumes of written documents.

If you have to present a briefing -- don't go at it alone. Consult with media specialists. In addition to giving you appropriate stimulating presentation techniques, and ideas, they will help you avoid the most common "deadly sin" of researchers -- namely transposing the pages of the written report to charts, and then reading the words to the audience!

Your job is to interpret the report's findings, not to read it. The graphics are there to help you present the message.

You must practice to speak extemporaneously, with the graphics as your notes. This increases your eye contact and rapport with the audience, keeps them awake and you alert. You shouldn't need to read the report -- after all you should be more familiar with it than anyone else at this point. Above all, in briefings speak loud and clear -- if they can't hear you or understand what you are saying -- you are not communicating, and if you are not communicating the results of your survey, then there wasn't much point in doing it in the first place!

## CONCLUSION

This booklet was written primarily as an initial introduction to, and overview of statistical survey and analysis methods for the Office of Rural Development, USAID/Indonesia, but its utility is AID-wide.

It is designed as a refresher course (in on-the-job training sessions) for those who have forgotten most, if not all of the statistics that they had in school, and for those who for one reason or another never learned. Subsequent use is intended as a ready reference, with "cook-book" examples to improve recall for the formulae when the need arises.

Obviously there is much more to the subject than is contained herein. A number of topics worthy of extensive treatment have been simplified and summarized, while others have been completely ignored. In doing this, I have tried to follow the "mini-skirt" principle - keeping it long enough to cover the subject, and at the same time, short enough to remain interesting!

Thus there should be plenty to appreciate and absorb and if applied to everyday operations where appropriate, it should result in a significant improvement in project monitoring and management.

## TABLE I

## A TABLE OF RANDOM DIGITS

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | 0 | 5 | 1 | 2 | 8 | 5 | 9 | 8 | 6 | 6 | 5 | 1 | 2 | 8 | 1 | 6 | 8 | 1 | 2 | 4 | 7 | 5 | 0 | 6 | 4 | 7 | 9 | 4 | 1 | 9 | 2 | 2 | 3 | 5 | 9 | 6 | 5 | 2 | 0 | 6 |
| 2. | 3 | 6 | 7 | 4 | 6 | 8 | 9 | 6 | 9 | 8 | 5 | 6 | 0 | 2 | 0 | 3 | 7 | 8 | 1 | 0 | 8 | 8 | 6 | 8 | 4 | 5 | 9 | 9 | 1 | 4 | 0 | 4 | 1 | 4 | 6 | 0 | 1 | 4 | 1 | 9 |
| 3. | 8 | 7 | 5 | 1 | 3 | 1 | 7 | 6 | 9 | 0 | 6 | 1 | 4 | 2 | 7 | 7 | 2 | 9 | 1 | 4 | 4 | 8 | 5 | 6 | 3 | 4 | 3 | 3 | 7 | 4 | 2 | 5 | 4 | 7 | 3 | 6 | 0 | 9 | 8 | 2 |
| 4. | 0 | 2 | 6 | 2 | 2 | 4 | 1 | 0 | 2 | 6 | 8 | 0 | 8 | 7 | 5 | 4 | 1 | 2 | 9 | 3 | 2 | 1 | 5 | 2 | 9 | 2 | 2 | 1 | 9 | 9 | 1 | 1 | 8 | 6 | 5 | 2 | 6 | 2 | 0 | 1 |
| 5. | 6 | 4 | 9 | 8 | 1 | 2 | 8 | 1 | 8 | 0 | 3 | 8 | 6 | 2 | 9 | 7 | 6 | 9 | 6 | 2 | 9 | 3 | 2 | 8 | 5 | 1 | 3 | 7 | 8 | 6 | 2 | 7 | 4 | 7 | 5 | 3 | 1 | 2 | 5 | 4 |
| 6. | 5 | 7 | 8 | 8 | 8 | 1 | 3 | 9 | 3 | 8 | 3 | 8 | 5 | 5 | 4 | 8 | 6 | 8 | 3 | 6 | 0 | 2 | 1 | 9 | 5 | 4 | 5 | 4 | 4 | 5 | 4 | 1 | 0 | 5 | 9 | 5 | 5 | 1 | 4 | 2 |
| 7. | 5 | 6 | 3 | 1 | 6 | 3 | 7 | 7 | 2 | 3 | 0 | 0 | 2 | 3 | 4 | 2 | 1 | 4 | 2 | 4 | 2 | 6 | 6 | 6 | 4 | 2 | 1 | 0 | 6 | 7 | 5 | 7 | 2 | 3 | 8 | 3 | 5 | 3 | 5 | 2 |
| 8. | 9 | 8 | 8 | 4 | 9 | 7 | 2 | 7 | 6 | 2 | 5 | 9 | 7 | 6 | 7 | 5 | 2 | 4 | 9 | 7 | 2 | 4 | 2 | 2 | 7 | 3 | 0 | 3 | 0 | 2 | 9 | 5 | 3 | 2 | 7 | 1 | 2 | 8 | 4 | 9 |
| 9. | 5 | 1 | 6 | 3 | 2 | 5 | 4 | 7 | 9 | 9 | 2 | 7 | 9 | 7 | 3 | 6 | 8 | 5 | 6 | 8 | 6 | 8 | 4 | 6 | 5 | 7 | 0 | 0 | 4 | 0 | 9 | 1 | 3 | 8 | 5 | 9 | 6 | 4 | 3 | 6 |
| 10. | 1 | 2 | 8 | 7 | 4 | 8 | 2 | 1 | 6 | 0 | 6 | 7 | 2 | 0 | 2 | 8 | 5 | 1 | 9 | 9 | 2 | 7 | 9 | 0 | 8 | 1 | 3 | 3 | 5 | 1 | 4 | 8 | 3 | 2 | 1 | 2 | 8 | 3 | 5 | 7 |
| 11. | 5 | 7 | 5 | 8 | 0 | 7 | 7 | 8 | 8 | 4 | 0 | 7 | 0 | 3 | 2 | 0 | 1 | 6 | 7 | 1 | 5 | 3 | 3 | 6 | 2 | 1 | 5 | 5 | 6 | 4 | 0 | 4 | 7 | 1 | 6 | 1 | 4 | 5 | 9 | 4 |
| 12. | 5 | 1 | 8 | 7 | 5 | 6 | 4 | 6 | 1 | 1 | 1 | 9 | 7 | 3 | 6 | 2 | 5 | 5 | 8 | 9 | 4 | 6 | 5 | 6 | 9 | 3 | 0 | 9 | 8 | 7 | 5 | 7 | 6 | 5 | 7 | 3 | 3 | 3 | 9 | 8 |
| 13. | 3 | 9 | 1 | 3 | 3 | 3 | 0 | 3 | 9 | 3 | 5 | 8 | 3 | 1 | 9 | 8 | 5 | 0 | 9 | 8 | 6 | 6 | 5 | 1 | 9 | 7 | 9 | 1 | 7 | 2 | 7 | 2 | 7 | 6 | 4 | 6 | 6 | 4 | 4 | 6 |
| 14. | 2 | 4 | 5 | 4 | 1 | 6 | 1 | 4 | 7 | 7 | 8 | 9 | 7 | 3 | 1 | 1 | 8 | 4 | 2 | 1 | 2 | 9 | 8 | 6 | 1 | 5 | 7 | 8 | 7 | 5 | 4 | 5 | 2 | 2 | 8 | 4 | 9 | 2 | 1 | 1 |
| 15. | 5 | 0 | 8 | 5 | 9 | 8 | 4 | 7 | 4 | 6 | 2 | 8 | 3 | 0 | 2 | 1 | 3 | 2 | 6 | 4 | 0 | 7 | 5 | 9 | 5 | 5 | 7 | 1 | 4 | 6 | 6 | 4 | 6 | 6 | 5 | 3 | 1 | 1 | 5 | 9 |
| 16. | 2 | 8 | 1 | 1 | 9 | 2 | 4 | 2 | 0 | 0 | 0 | 9 | 1 | 1 | 0 | 2 | 8 | 4 | 8 | 5 | 3 | 0 | 3 | 2 | 6 | 4 | 2 | 8 | 2 | 6 | 0 | 6 | 9 | 7 | 4 | 6 | 1 | 0 | 6 | 3 |
| 17. | 4 | 5 | 2 | 0 | 6 | 5 | 3 | 3 | 0 | 0 | 3 | 8 | 6 | 8 | 8 | 3 | 9 | 9 | 6 | 8 | 3 | 2 | 6 | 0 | 4 | 9 | 3 | 9 | 2 | 9 | 0 | 1 | 8 | 3 | 6 | 3 | 6 | 5 | 9 | 0 |
| 18. | 5 | 7 | 5 | 7 | 1 | 6 | 5 | 9 | 1 | 9 | 5 | 6 | 4 | 0 | 5 | 1 | 7 | 8 | 3 | 9 | 9 | 2 | 0 | 7 | 3 | 8 | 3 | 5 | 8 | 5 | 0 | 0 | 4 | 1 | 4 | 6 | 2 | 8 | 5 | 1 |
| 19. | 5 | 2 | 8 | 2 | 9 | 0 | 1 | 1 | 7 | 2 | 0 | 8 | 9 | 1 | 5 | 1 | 1 | 4 | 6 | 7 | 1 | 4 | 7 | 9 | 3 | 2 | 7 | 5 | 4 | 8 | 3 | 7 | 5 | 1 | 6 | 2 | 4 | 3 | 4 | 3 |
| 20. | 0 | 0 | 1 | 3 | 4 | 3 | 6 | 2 | 3 | 3 | 8 | 9 | 4 | 3 | 4 | 3 | 8 | 6 | 6 | 9 | 9 | 1 | 5 | 9 | 2 | 3 | 2 | 9 | 8 | 2 | 5 | 6 | 4 | 5 | 5 | 5 | 3 | 1 | 2 | 9 |
| 21. | 9 | 9 | 8 | 2 | 6 | 6 | 4 | 0 | 0 | 5 | 9 | 4 | 3 | 2 | 5 | 7 | 3 | 5 | 5 | 3 | 7 | 8 | 2 | 8 | 0 | 3 | 0 | 1 | 0 | 4 | 6 | 7 | 1 | 2 | 6 | 7 | 6 | 6 | 5 | 6 |
| 22. | 1 | 1 | 6 | 9 | 4 | 4 | 6 | 2 | 6 | 2 | 5 | 5 | 0 | 6 | 7 | 6 | 4 | 6 | 0 | 3 | 5 | 9 | 7 | 6 | 2 | 3 | 5 | 2 | 4 | 0 | 0 | 0 | 3 | 1 | 8 | 0 | 9 | 1 | 3 | 6 |
| 23. | 5 | 7 | 6 | 2 | 2 | 9 | 3 | 3 | 2 | 8 | 9 | 8 | 8 | 8 | 5 | 0 | 7 | 7 | 8 | 3 | 0 | 4 | 3 | 5 | 1 | 9 | 4 | 0 | 3 | 1 | 6 | 2 | 2 | 0 | 9 | 4 | 3 | 7 | 4 | 0 |
| 24. | 8 | 2 | 6 | 9 | 1 | 5 | 1 | 2 | 3 | 8 | 1 | 4 | 1 | 0 | 6 | 4 | 3 | 9 | 8 | 3 | 3 | 3 | 3 | 5 | 6 | 9 | 9 | 3 | 2 | 1 | 1 | 1 | 3 | 3 | 1 | 0 | 6 | 8 | 3 | 8 |
| 25. | 8 | 8 | 7 | 9 | 9 | 6 | 5 | 6 | 2 | 1 | 5 | 9 | 8 | 0 | 9 | 3 | 7 | 8 | 5 | 0 | 6 | 6 | 1 | 2 | 8 | 7 | 8 | 2 | 3 | 6 | 7 | 1 | 7 | 3 | 2 | 0 | 4 | 7 | 0 | 4 |
| 26. | 6 | 9 | 1 | 2 | 5 | 9 | 5 | 5 | 9 | 1 | 8 | 1 | 1 | 6 | 8 | 9 | 9 | 2 | 4 | 6 | 6 | 6 | 4 | 1 | 6 | 4 | 3 | 1 | 0 | 8 | 5 | 6 | 5 | 9 | 2 | 4 | 2 | 4 | 6 | 7 |
| 27. | 7 | 4 | 6 | 9 | 8 | 4 | 4 | 2 | 3 | 3 | 6 | 7 | 6 | 0 | 2 | 2 | 1 | 6 | 1 | 5 | 7 | 2 | 3 | 3 | 6 | 9 | 1 | 0 | 5 | 8 | 6 | 0 | 9 | 5 | 8 | 2 | 0 | 7 | 0 | 6 |
| 28. | 7 | 7 | 4 | 5 | 1 | 4 | 7 | 3 | 5 | 0 | 2 | 1 | 2 | 3 | 4 | 6 | 7 | 6 | 7 | 2 | 8 | 0 | 5 | 6 | 7 | 9 | 8 | 1 | 7 | 2 | 4 | 4 | 3 | 4 | 6 | 6 | 0 | 4 | 3 | 0 |
| 29. | 6 | 1 | 7 | 1 | 5 | 9 | 6 | 4 | 8 | 5 | 2 | 2 | 1 | 2 | 1 | 9 | 8 | 8 | 4 | 4 | 5 | 9 | 2 | 8 | 9 | 1 | 2 | 5 | 2 | 3 | 5 | 7 | 3 | 4 | 5 | 4 | 1 | 2 | 4 | 6 |
| 30. | 9 | 2 | 7 | 3 | 5 | 4 | 5 | 0 | 6 | 4 | 5 | 0 | 9 | 2 | 4 | 0 | 0 | 8 | 6 | 5 | 1 | 9 | 6 | 9 | 0 | 5 | 6 | 6 | 8 | 2 | 8 | 2 | 5 | 1 | 7 | 3 | 3 | 1 | 5 | 0 |
| 31. | 7 | 2 | 3 | 5 | 3 | 4 | 5 | 7 | 7 | 5 | 6 | 8 | 5 | 9 | 0 | 8 | 5 | 6 | 8 | 5 | 9 | 9 | 9 | 7 | 5 | 0 | 1 | 0 | 5 | 6 | 2 | 7 | 5 | 3 | 4 | 2 | 3 | 0 | 8 | 5 |
| 32. | 1 | 2 | 9 | 7 | 9 | 0 | 5 | 7 | 2 | 0 | 9 | 2 | 7 | 5 | 4 | 7 | 6 | 9 | 1 | 1 | 5 | 5 | 2 | 4 | 0 | 1 | 8 | 7 | 3 | 0 | 9 | 6 | 1 | 9 | 7 | 6 | 4 | 4 | 8 | 3 |
| 33. | 4 | 4 | 3 | 6 | 5 | 7 | 0 | 2 | 5 | 4 | 5 | 0 | 8 | 6 | 4 | 3 | 6 | 6 | 1 | 9 | 3 | 0 | 0 | 9 | 4 | 0 | 7 | 7 | 9 | 4 | 6 | 0 | 4 | 7 | 5 | 4 | 9 | 6 | 6 | 6 |
| 34. | 4 | 9 | 0 | 7 | 6 | 1 | 8 | 4 | 3 | 9 | 2 | 9 | 5 | 2 | 2 | 4 | 2 | 5 | 4 | 1 | 7 | 9 | 3 | 2 | 7 | 4 | 8 | 8 | 8 | 3 | 7 | 7 | 1 | 5 | 4 | 7 | 4 | 9 | 7 | 3 |
| 35. | 7 | 8 | 1 | 4 | 3 | 6 | 5 | 9 | 1 | 9 | 1 | 3 | 6 | 9 | 9 | 9 | 1 | 8 | 4 | 4 | 1 | 0 | 6 | 7 | 6 | 7 | 0 | 1 | 7 | 1 | 5 | 9 | 4 | 3 | 1 | 7 | 6 | 0 | 3 | 3 |
| 36. | 0 | 3 | 4 | 7 | 4 | 7 | 6 | 0 | 2 | 5 | 9 | 7 | 0 | 4 | 3 | 3 | 3 | 8 | 3 | 4 | 4 | 4 | 6 | 3 | 8 | 4 | 8 | 8 | 3 | 0 | 5 | 5 | 0 | 2 | 9 | 1 | 0 | 3 | 7 | 1 |
| 37. | 3 | 5 | 8 | 7 | 0 | 8 | 9 | 1 | 5 | 8 | 5 | 5 | 8 | 6 | 4 | 9 | 8 | 0 | 7 | 8 | 5 | 0 | 5 | 6 | 3 | 7 | 3 | 1 | 5 | 1 | 6 | 4 | 4 | 6 | 3 | 5 | 0 | 0 | 5 | 8 |
| 38. | 7 | 3 | 8 | 8 | 7 | 6 | 7 | 9 | 2 | 8 | 6 | 0 | 0 | 4 | 5 | 7 | 0 | 7 | 8 | 2 | 1 | 1 | 9 | 3 | 7 | 0 | 6 | 5 | 7 | 1 | 9 | 5 | 9 | 3 | 4 | 0 | 9 | 1 | 3 | 2 |
| 39. | 4 | 5 | 9 | 6 | 8 | 7 | 3 | 6 | 6 | 7 | 6 | 5 | 0 | 6 | 2 | 7 | 3 | 3 | 0 | 6 | 7 | 6 | 0 | 4 | 5 | 7 | 6 | 6 | 0 | 9 | 5 | 2 | 5 | 5 | 3 | 4 | 7 | 5 | 0 | 8 |
| 40. | 6 | 7 | 6 | 2 | 2 | 5 | 4 | 5 | 7 | 9 | 1 | 7 | 2 | 7 | 9 | 6 | 7 | 4 | 4 | 0 | 5 | 6 | 4 | 4 | 1 | 3 | 2 | 1 | 3 | 8 | 6 | 1 | 1 | 9 | 7 | 9 | 5 | 4 | 7 | 6 |
| 41. | 6 | 6 | 9 | 1 | 3 | 6 | 0 | 6 | 6 | 4 | 6 | 7 | 5 | 4 | 7 | 3 | 9 | 5 | 2 | 3 | 0 | 2 | 0 | 4 | 3 | 0 | 4 | 8 | 5 | 5 | 2 | 7 | 0 | 2 | 9 | 0 | 1 | 5 | 4 | 2 |
| 42. | 7 | 4 | 8 | 5 | 9 | 6 | 2 | 1 | 5 | 5 | 0 | 9 | 2 | 3 | 4 | 4 | 7 | 3 | 6 | 7 | 1 | 3 | 0 | 4 | 7 | 6 | 5 | 4 | 3 | 4 | 1 | 2 | 1 | 2 | 4 | 9 | 1 | 0 | 8 | 7 |
| 43. | 9 | 0 | 8 | 7 | 9 | 4 | 4 | 9 | 6 | 9 | 1 | 1 | 1 | 2 | 9 | 1 | 7 | 1 | 3 | 9 | 7 | 9 | 6 | 3 | 0 | 8 | 6 | 8 | 0 | 0 | 1 | 6 | 7 | 8 | 1 | 6 | 5 | 9 | 7 | 7 |
| 44. | 9 | 5 | 9 | 0 | 9 | 8 | 2 | 4 | 5 | 9 | 9 | 6 | 2 | 1 | 8 | 6 | 0 | 7 | 6 | 8 | 7 | 6 | 4 | 1 | 7 | 5 | 1 | 2 | 3 | 3 | 8 | 1 | 4 | 0 | 9 | 4 | 6 | 7 | 7 | 3 |
| 45. | 2 | 9 | 2 | 1 | 2 | 4 | 0 | 8 | 7 | 3 | 4 | 1 | 5 | 9 | 0 | 6 | 7 | 2 | 5 | 5 | 3 | 0 | 7 | 5 | 7 | 9 | 2 | 9 | 3 | 3 | 7 | 7 | 3 | 4 | 1 | 2 | 0 | 8 | 3 | 9 |

TABLE 2

THE NORMAL CURVE AND RELATED PROBABILITY
(Both Sides of the Mean)

Size of the Standard Error - Percentage of occurrences falling within the range

Standard Deviation - (Probability desired)
or Value of "K" - (Confidence desired)

| | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 00.00 | 00.80 | 01.60 | 02.40 | 03.20 | 03.98 | 04.78 | 05.58 | 06.38 | 07.18 |
| 0.1 | 07.96 | 08.76 | 09.56 | 10.34 | 11.14 | 11.92 | 12.72 | 13.50 | 14.28 | 15.06 |
| 0.2 | 15.86 | 16.64 | 17.42 | 18.20 | 18.96 | 19.74 | 20.52 | 21.28 | 22.06 | 22.82 |
| 0.3 | 23.58 | 24.34 | 25.10 | 25.86 | 26.62 | 27.36 | 28.12 | 28.86 | 29.60 | 30.34 |
| 0.4 | 31.08 | 31.82 | 32.56 | 33.28 | 34.00 | 34.72 | 35.44 | 36.16 | 36.88 | 37.58 |
| 0.5 | 38.30 | 39.00 | 39.70 | 40.38 | 41.08 | 41.76 | 42.46 | 43.14 | 43.80 | 44.48 |
| 0.6 | 45.14 | 45.82 | 46.48 | 47.14 | 47.78 | 48.44 | 49.08 | 49.72 | 50.34 | 50.98 |
| 0.7 | 51.60 | 52.22 | 52.34 | 53.46 | 54.06 | 54.68 | 55.28 | 55.88 | 56.46 | 57.04 |
| 0.8 | 57.62 | 58.20 | 58.78 | 59.34 | 59.90 | 60.46 | 61.02 | 61.56 | 62.12 | 62.66 |
| 0.9 | 63.18 | 63.72 | 64.24 | 64.761 | 65.28 | 65.78 | 66.30 | 66.80 | 67.30 | 67.78 |
| 1.0 | 68.26 | 68.76 | 69.22 | 69.70 | 70.16 | 70.62 | 71.08 | 71.54 | 71.98 | 72.22 |
| 1.1 | 72.86 | 73.30 | 73.72 | 74.16 | 74.58 | 74.88 | 75.40 | 75.80 | 76.20 | 76.60 |
| 1.2 | 76.98 | 77.38 | 77.76 | 78.14 | 78.50 | 78.88 | 79.24 | 79.60 | 79.94 | 80.30 |
| 1.3 | 80.64 | 80.98 | 81.32 | 81.64 | 81.98 | 82.30 | 82.62 | 82.94 | 83.24 | 83.54 |
| 1.4 | 83.84 | 84.14 | 84.44 | 84.72 | 85.02 | 85.30 | 85.58 | 85.84 | 86.12 | 86.38 |
| 1.5 | 86.64 | 86.90 | 87.14 | 87.40 | 87.64 | 87.88 | 88.12 | 88.36 | 88.58 | 88.82 |
| 1.6 | 89.04 | 89.26 | 89.48 | 89.68 | 89.90 | 90.10 | 90.30 | 90.50 | 90.70 | 90.90 |
| 1.7 | 91.08 | 91.28 | 91.46 | 91.64 | 91.82 | 91.98 | 92.16 | 92.32 | 92.50 | 92.66 |
| 1.8 | 92.82 | 92.98 | 93.12 | 93.28 | 93.42 | 93.56 | 93.72 | 93.86 | 93.98 | 94.12 |
| 1.9 | 94.26 | 94.38 | 94.52 | 94.64 | 94.76 | 94.88 | 95.00 | 95.12 | 95.22 | 95.34 |
| 2.0 | 95.44 | 95.56 | 95.66 | 95.76 | 95.86 | 95.96 | 96.06 | 96.16 | 96.24 | 96.34 |
| 2.1 | 96.42 | 96.52 | 96.60 | 96.68 | 96.76 | 96.84 | 96.92 | 97.00 | 97.08 | 97.14 |
| 2.2 | 97.22 | 97.28 | 97.36 | 97.42 | 97.50 | 97.56 | 97.62 | 97.68 | 97.74 | 97.80 |
| 2.3 | 97.86 | 97.92 | 97.96 | 98.02 | 98.08 | 98.12 | 98.18 | 98.22 | 98.26 | 98.32 |
| 2.4 | 98.28 | 98.40 | 98.44 | 98.50 | 98.54 | 98.58 | 98.62 | 98.64 | 98.68 | 98.72 |
| 2.5 | 98.76 | 98.80 | 98.82 | 98.86 | 98.89 | 98.92 | 98.96 | 98.99 | 99.02 | 99.04 |
| 2.6 | 99.06 | 99.10 | 99.12 | 99.14 | 99.18 | 99.20 | 99.22 | 99.24 | 99.26 | 99.28 |
| 2.7 | 99.30 | 99.32 | 99.34 | 99.36 | 99.38 | 99.40 | 99.42 | 99.44 | 99.46 | 99.48 |
| 2.8 | 99.48 | 99.50 | 99.52 | 99.54 | 99.54 | 99.56 | 99.58 | 99.58 | 99.60 | 99.62 |
| 2.9 | 99.62 | 99.64 | 99.64 | 99.66 | 99.68 | 99.68 | 99.70 | 99.70 | 99.72 | 99.72 |
| 3.0 | 99.74 | 99.74 | 99.74 | 99.76 | 99.76 | 99.78 | 99.78 | 99.78 | 99.80 | 99.80 |

TABLE 3



THE NORMAL DISTRIBUTION CURVE
(One Side of the Mean)

Percentage of all values included within the range formed by the mean plus, (or minus) a specified number of standard deviation (SD) units.

| SD Units | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| .0 | 00.00 | 00.40 | 00.80 | 01.20 | 01.60 | 01.99 | 02.39 | 02.79 | 03.19 | 03.59 |
| .1 | 03.98 | 04.38 | 04.78 | 05.17 | 05.57 | 05.96 | 06.36 | 06.75 | 07.14 | 07.53 |
| .2 | 07.93 | 08.32 | 08.71 | 09.10 | 09.48 | 09.87 | 10.26 | 10.64 | 11.03 | 11.41 |
| .3 | 11.79 | 12.17 | 12.55 | 12.93 | 13.31 | 13.68 | 14.06 | 14.43 | 14.80 | 15.17 |
| .4 | 15.54 | 15.91 | 16.28 | 16.64 | 17.00 | 17.36 | 17.72 | 18.08 | 18.44 | 18.79 |
| .5 | 19.15 | 19.50 | 19.85 | 20.19 | 20.54 | 20.88 | 21.23 | 21.57 | 21.90 | 22.24 |
| .6 | 22.57 | 22.91 | 23.24 | 23.57 | 23.89 | 24.22 | 24.54 | 24.86 | 25.17 | 25.49 |
| .7 | 25.80 | 26.11 | 26.42 | 26.73 | 27.03 | 27.34 | 27.64 | 27.94 | 28.23 | 28.52 |
| .8 | 28.81 | 29.10 | 29.39 | 29.67 | 29.95 | 30.23 | 30.51 | 30.78 | 31.06 | 31.33 |
| .9 | 31.59 | 31.86 | 32.12 | 32.38 | 32.64 | 32.89 | 33.15 | 33.40 | 33.65 | 33.89 |
| 1.0 | 34.13 | 34.38 | 34.61 | 34.85 | 35.08 | 35.31 | 35.54 | 35.77 | 35.99 | 36.21 |
| 1.1 | 36.43 | 36.65 | 36.85 | 37.08 | 37.29 | 37.49 | 37.70 | 37.90 | 38.10 | 38.30 |
| 1.2 | 38.49 | 38.69 | 38.88 | 39.07 | 39.25 | 39.44 | 39.62 | 39.80 | 39.97 | 40.15 |
| 1.3 | 40.32 | 40.49 | 40.66 | 40.82 | 40.99 | 41.15 | 41.31 | 41.47 | 41.62 | 41.77 |
| 1.4 | 41.92 | 42.07 | 42.22 | 42.36 | 42.51 | 42.65 | 42.79 | 42.92 | 43.06 | 43.19 |
| 1.5 | 43.32 | 43.45 | 43.57 | 43.70 | 43.82 | 43.94 | 44.06 | 44.18 | 44.29 | 44.41 |
| 1.6 | 44.52 | 44.63 | 44.74 | 44.84 | 44.95 | 45.05 | 45.15 | 45.25 | 45.35 | 45.45 |
| 1.7 | 45.54 | 45.64 | 45.73 | 45.82 | 45.91 | 45.99 | 46.08 | 46.16 | 46.25 | 46.33 |
| 1.8 | 46.41 | 46.49 | 46.56 | 46.64 | 46.71 | 46.78 | 46.86 | 46.93 | 46.99 | 47.06 |
| 1.9 | 47.13 | 47.19 | 47.26 | 47.32 | 47.38 | 47.44 | 47.50 | 47.56 | 47.61 | 47.67 |
| 2.0 | 47.72 | 47.78 | 47.83 | 47.88 | 47.93 | 47.98 | 48.03 | 48.08 | 48.12 | 48.17 |
| 2.1 | 48.21 | 48.26 | 48.30 | 48.34 | 48.38 | 48.42 | 48.46 | 48.50 | 48.54 | 48.57 |
| 2.2 | 48.61 | 48.64 | 48.68 | 48.71 | 48.75 | 48.78 | 48.81 | 48.84 | 48.87 | 48.90 |
| 2.3 | 48.93 | 48.96 | 48.98 | 49.01 | 49.04 | 49.06 | 49.09 | 49.11 | 49.13 | 49.16 |
| 2.4 | 49.18 | 49.20 | 49.22 | 49.25 | 49.27 | 49.29 | 49.31 | 49.32 | 49.34 | 49.36 |
| 2.5 | 49.38 | 49.40 | 49.41 | 49.43 | 49.45 | 49.46 | 49.48 | 49.49 | 49.51 | 49.52 |
| 2.6 | 49.53 | 49.55 | 49.56 | 49.57 | 49.59 | 49.60 | 49.61 | 49.62 | 49.63 | 49.64 |
| 2.7 | 49.65 | 49.66 | 49.67 | 49.68 | 49.69 | 49.70 | 49.71 | 49.72 | 49.73 | 49.74 |
| 2.8 | 49.74 | 49.75 | 49.76 | 49.77 | 49.77 | 49.78 | 49.79 | 49.79 | 49.80 | 49.81 |
| 2.9 | 49.81 | 49.82 | 49.82 | 49.83 | 49.84 | 49.84 | 49.85 | 49.85 | 49.86 | 49.86 |
| 3.0 | 49.87 | 49.87 | 49.87 | 49.88 | 49.88 | 49.89 | 49.89 | 49.89 | 49.90 | 49.90 |

TABLE 4



PERCENTAGE OF ONE TAIL OF THE NORMAL CURVE
AT SELECTED VALUES OF Z FROM THE MEAN

| Z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 50.00 | 49.60 | 49.20 | 48.80 | 48.40 | 48.01 | 47.61 | 47.21 | 46.81 | 46.41 |
| 0.1 | 46.02 | 45.62 | 45.22 | 44.83 | 44.43 | 44.04 | 43.64 | 43.25 | 42.86 | 42.47 |
| 0.2 | 42.07 | 41.68 | 41.29 | 40.90 | 40.52 | 40.13 | 39.74 | 39.36 | 38.97 | 38.59 |
| 0.3 | 38.21 | 37.83 | 37.45 | 37.07 | 36.69 | 36.32 | 35.94 | 35.57 | 35.20 | 34.83 |
| 0.4 | 34.46 | 34.09 | 33.72 | 33.36 | 33.00 | 32.64 | 32.28 | 31.92 | 31.56 | 31.21 |
| 0.5 | 30.85 | 30.50 | 30.15 | 29.81 | 29.46 | 29.12 | 28.77 | 28.43 | 28.10 | 27.76 |
| 0.6 | 27.43 | 27.09 | 26.76 | 26.43 | 26.11 | 25.78 | 25.46 | 25.14 | 24.83 | 24.51 |
| 0.7 | 24.20 | 23.89 | 23.58 | 23.27 | 22.96 | 22.66 | 22.36 | 22.06 | 21.77 | 21.48 |
| 0.8 | 21.19 | 20.90 | 20.61 | 20.33 | 20.05 | 19.77 | 19.49 | 19.22 | 18.94 | 18.67 |
| 0.9 | 18.41 | 18.14 | 17.88 | 17.62 | 17.36 | 17.11 | 16.85 | 16.60 | 16.35 | 16.11 |
| 1.0 | 15.87 | 15.62 | 15.39 | 15.15 | 14.92 | 14.69 | 14.46 | 14.23 | 14.01 | 13.79 |
| 1.1 | 13.57 | 13.35 | 13.14 | 12.92 | 12.71 | 12.51 | 12.30 | 12.10 | 11.90 | 11.70 |
| 1.2 | 11.51 | 11.31 | 11.12 | 10.93 | 10.75 | 10.56 | 10.38 | 10.20 | 10.03 | 09.85 |
| 1.3 | 09.68 | 09.51 | 09.34 | 09.18 | 09.01 | 08.85 | 08.69 | 08.53 | 08.38 | 08.23 |
| 1.4 | 08.08 | 07.93 | 07.78 | 07.64 | 07.49 | 07.35 | 07.21 | 07.08 | 06.94 | 06.81 |
| 1.5 | 06.68 | 06.55 | 06.43 | 06.30 | 06.18 | 06.06 | 05.94 | 05.82 | 05.71 | 05.59 |
| 1.6 | 05.48 | 05.37 | 05.26 | 05.16 | 05.05 | 04.95 | 04.85 | 04.75 | 04.65 | 04.55 |
| 1.7 | 04.46 | 04.36 | 04.27 | 04.18 | 04.09 | 04.01 | 03.92 | 03.84 | 03.75 | 03.67 |
| 1.8 | 03.59 | 03.51 | 03.44 | 03.36 | 03.29 | 03.22 | 03.14 | 03.07 | 03.01 | 02.94 |
| 1.9 | 02.87 | 02.81 | 02.74 | 02.68 | 02.62 | 02.56 | 02.50 | 02.44 | 02.39 | 02.33 |
| 2.0 | 02.28 | 02.22 | 02.17 | 02.12 | 02.07 | 02.02 | 01.97 | 01.92 | 01.88 | 01.83 |
| 2.1 | 01.79 | 01.74 | 01.70 | 01.66 | 01.62 | 01.58 | 01.54 | 01.50 | 01.46 | 01.43 |
| 2.2 | 01.39 | 01.36 | 01.32 | 01.29 | 01.25 | 01.22 | 01.19 | 01.16 | 01.13 | 01.10 |
| 2.3 | 01.07 | 01.04 | 01.02 | 00.99 | 00.96 | 00.94 | 00.91 | 00.89 | 00.87 | 00.84 |
| 2.4 | 00.82 | 00.80 | 00.78 | 00.76 | 00.73 | 00.71 | 00.70 | 00.68 | 00.66 | 00.64 |
| 2.5 | 00.62 | 00.60 | 00.59 | 00.57 | 00.55 | 00.54 | 00.52 | 00.51 | 00.49 | 00.48 |
| 2.6 | 00.47 | 00.45 | 00.44 | 00.43 | 00.42 | 00.40 | 00.39 | 00.38 | 00.37 | 00.36 |
| 2.7 | 00.35 | 00.34 | 00.33 | 00.32 | 00.31 | 00.30 | 00.29 | 00.28 | 00.27 | 00.26 |
| 2.8 | 00.26 | 00.25 | 00.24 | 00.23 | 00.23 | 00.22 | 00.21 | 00.21 | 00.20 | 00.19 |
| 2.9 | 00.19 | 00.18 | 00.18 | 00.17 | 00.16 | 00.16 | 00.15 | 00.15 | 00.14 | 00.14 |

TABLE 5

STUDENT "T" DISTRIBUTION

Value of "T" for the following Percentage Confidence Levels

| Degrees of Freedom* | 80% | 90% | 95% | 98% | 99% |
|---|---|---|---|---|---|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| ** | 20% | 10% | 5% | 2% | 1% |

* "Degrees of Freedom" is a statistical term which represents the number of independent pieces of information available about the variability of a population. There is no variability in a sample of one, one degree of freedom in a sample of two, and so forth. Each additional observation adds one additional independent piece of information about the population variance. In general, in a sample size of "n", there are "n-1" degrees of freedom. For determining correlations between two variables, in a sample size of "n" pairs, there are "n-2" degrees of freedom.

** When the table is read from the foot, the tabled values are to be prefixed with a negative sign.

TABLE 6

CRITICAL VALUES OF CHI SQUARE AT SELECTED SIGNIFICANCE LEVELS

(Probability of Null Hypothesis)

| Sig. Level | .30 | .20 | .05 | .02 | .01 | .001 |
|---|---|---|---|---|---|---|
| Degrees of Freedom | | | | | | |
| 1 | 1.07 | 1.64 | 3.84 | 5.41 | 6.64 | 10.83 |
| 2 | 2.41 | 3.22 | 5.99 | 7.82 | 9.21 | 13.82 |
| 3 | 3.66 | 4.64 | 7.82 | 9.84 | 11.34 | 16.27 |
| 4 | 4.88 | 5.99 | 9.49 | 11.67 | 13.28 | 18.46 |
| 5 | 6.06 | 7.29 | 11.07 | 13.39 | 15.09 | 20.52 |
| 6 | 7.23 | 8.56 | 12.59 | 15.03 | 16.81 | 22.46 |
| 7 | 8.38 | 9.80 | 14.07 | 16.62 | 18.48 | 24.32 |
| 8 | 9.52 | 11.03 | 15.51 | 18.17 | 20.09 | 26.12 |
| 9 | 10.66 | 12.24 | 16.92 | 19.08 | 21.67 | 27.88 |
| 10 | 11.78 | 13.44 | 18.31 | 21.16 | 23.21 | 29.59 |
| 11 | 12.90 | 14.63 | 19.68 | 22.62 | 24.72 | 31.26 |
| 12 | 14.01 | 15.81 | 21.03 | 24.05 | 26.22 | 32.91 |
| 13 | 15.12 | 16.98 | 22.36 | 25.47 | 27.69 | 34.53 |
| 14 | 16.22 | 18.15 | 23.68 | 26.87 | 29.14 | 36.12 |
| 15 | 17.32 | 19.31 | 25.00 | 28.26 | 30.58 | 37.70 |
| 16 | 18.42 | 20.46 | 26.30 | 29.63 | 32.00 | 39.29 |
| 17 | 19.51 | 21.62 | 27.59 | 31.00 | 33.41 | 40.75 |
| 18 | 20.60 | 22.76 | 28.87 | 32.35 | 34.80 | 42.31 |
| 19 | 21.69 | 23.90 | 30.14 | 33.69 | 36.19 | 43.82 |
| 20 | 22.78 | 25.04 | 31.41 | 35.02 | 37.57 | 45.32 |
| 21 | 23.86 | 26.17 | 32.67 | 36.34 | 38.93 | 46.80 |
| 22 | 24.94 | 27.30 | 33.92 | 37.66 | 40.29 | 43.27 |
| 23 | 26.02 | 28.43 | 35.17 | 38.97 | 41.64 | 49.73 |
| 24 | 27.10 | 29.55 | 36.42 | 40.27 | 42.98 | 51.18 |
| 25 | 28.17 | 30.68 | 37.65 | 41.57 | 44.31 | 52.62 |
| 26 | 29.25 | 31.80 | 38.88 | 42.86 | 45.64 | 54.05 |
| 27 | 30.32 | 32.91 | 40.11 | 44.14 | 46.96 | 55.48 |
| 28 | 31.39 | 34.03 | 41.34 | 45.42 | 48.28 | 56.89 |
| 29 | 32.46 | 35.14 | 42.56 | 46.69 | 49.59 | 58.30 |
| 30 | 33.53 | 36.25 | 43.77 | 47.96 | 50.89 | 59.70 |
| | Acceptable Range | | | Questionable/Unacceptable | | |

For Example: Given a Chi Square value of 25.00 with 15 degrees of freedom, the significance level is .05. In other words a 5 percent probability of a null hypothesis -- that there is no significant difference between the data compared. (.05 is usually an acceptable limit for management monitoring purposes.)