DOCUMENT RESUME

ED 128 411 TM 005 595

AUTHOR Smith, Kenneth F.
TITLE Statistical Survey and Analysis Handbook.
INSTITUTION Agency for International Development (Dept. of State), Manila (Philippines).
PUB DATE Mar 75
NOTE 77p.

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS *Data Analysis; *Data Collection; *Guides; Measurement; Measurement Goals; Research Design; Sampling; *Statistical Analysis; Statistical Bias; *Statistical Surveys; Statistics

ABSTRACT

The National Food and Agriculture Council of the Philippines regularly requires rapid feedback data for analysis, which will assist in monitoring programs to improve and increase the production of selected crops by small scale farmers. Since many other development programs in various subject matter areas also require similar statistical appraisals, this handbook was developed to present and explain the underlying principles and processes of scientific surveying. This includes the fundamentals of survey design, statistical sampling procedures, analytical methodologies, and presentation techniques. Often these essential steps are presented in statistical texts, which although technically complete fail to communicate with the nonmathematically oriented. This handbook has therefore been prepared as a step-by-step illustrative guidebook, with the emphasis on transmitting knowledge and creating understanding for subsequent application to typical problems. Although it can be self-studied, ideally this handbook should be used initially as the basis for intensive, practical workshop training. (Author/BW)

TM

STATISTICAL SURVEY and ANALYSIS HANDBOOK

UNITED STATES OF AMERICA

U.S. Agency for International Development
Manila, Philippines

March 1975

STATISTICAL SURVEY AND ANALYSIS HANDBOOK[1]

Kenneth F. Smith
Management Systems Advisor

U.S. Agency for International Development
Manila, Philippines

MARCH, 1975

[1] This text has been ereorganized and expanded from the initial January 1975 version based upon an intensive one week workshop seminar with NFAC/BAECON participants at the Development Academy of the Philippines, February 1975. The January 1975 text should no longer be used.

PREFACE

The National Food and Agriculture Council (NFAC) of the Philippines is involved in coordinating a number of intensive "Masagana" campaign programs to improve and increase the production of selected crops by small scale farmers. Information and Reporting Systems are in use and being further developed, to provide rapid feedback data for analysis which will assist the NFAC Management Committee in monitoring these programs. Statistical information also serves to guide decision makers in considering corrective action and/or policy changes to further the objectives of the campaigns. The Agriculture Program Evaluation Service (APES) of NFAC performs an essential role in the analysis of management information. As a regular function, they review the data reported through the formal system and analyze it for accuracy by selected follow-up surveys. The APES also conducts independent sample surveys of program implementation in the field as well as assessing the impact of natural calamities (typhoons, floods, droughts, etc.) through "quick and dirty" ad hoc studies, and occasionally more formal, longer range, in-depth analyses. Thus, the work-day of the evaluation Service staff **revolves** around surveys in some form running the full gamut from conception and design, through implementation, to final analysis and presentation.

NFAC's needs are not unique. Many other National and Provincial development programs in various subject matter areas also require similar statistical appraisals from time to time for management purposes.

It is desirable therefore, if not indeed essential for maximum utilization of resources, that all those responsible for conducting such program management appraisals be familiar with the underlying principles and procedures of scientific surveying. This includes the fundamentals of survey design, statistical sampling procedures, analytical methodologies and presentation techniques. All too often, these essential items are presented in formidable statistical texts which, although technically complete, fail to communicate with the non-mathematically oriented. Hence in many instances, scientific statistical analyses [illegible] [illegible] [illegible] obtained by management.

This Survey Handbook has therefore been prepared as [illegible] illustrative guidebook, with the emphasis on transmitting knowledge and practical understanding of [illegible] [illegible] [illegible] concerns of NFAC. However, the principles [illegible] other fields and disciplines.

No handbook can ever take the place of a good instructor. Although it can be self-studied, ideally this handbook should be used initially as the basis for intensive practical workshop training where the material is studied, discussed and worked through by the students, step by step, with supplementary illustrations and exercises. After such an experience the technician should find the handbook a useful reference in the conduct of his survey work.

The capability to conduct statistically sound surveys should complement and enhance the effort in analysis and management of any program.

March [illegible]

[illegible] F. Smith
Management Systems Advisor
USAID/Manila

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

INDEX

INTRODUCTION

> Scientific data are not taken for museum purposes;
> they are taken as a basis for doing something.
>
> If nothing is to be done with the data,
> then there is no use collecting any.
>
> W. Edwards Deming

One of the most frequent "question-statement" challenges an administrator or a technical subject-matter specialist is likely to make to the scientific approach to surveying is -

> Why should I bother to go through statistical mumbo-jumbo in order to gather and analyze data? I know my field, I have a "feel" for the situation in my area, and I know where to go to ask questions to 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 I recommend, possibly come up with findings better than mine?

Although he may not say all of the above aloud, be sure he thinks it!

There are of course several ways to make decisions without resorting to scientific statistical sample surveys:

1. Guess
2. Rely on previous experience and/or memory
3. Use logic, or "common-sense"
4. Make "spot check" and "judgement" surveys
5. Take a 100% survey

Many good decisions have been made using these approaches. Unfortunately, many bad ones have also been made. The difficulty with non-scientific approaches is that they are usually very biased, even though not intentionally so. Despite the fact that the data reported in spot checks may be accurate, there is no assurance that the conclusions drawn from it are valid and reliable. Using such information as a basis for making program management decisions is therefore a risky thing -- though again no one can say how risky.

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. At the other extreme, 100% surveys are expensive, time consuming, and often impossible to conduct.

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.

[1] Data originally gathered by someone else.

[2] New and original data.

ADVANTAGES OF SCIENTIFIC OVER NON-SCIENTIFIC SAMPLING

Unless appropriate scientific methods are used in the collection of data, statistics can be discredited in the eyes of management. Undue confidence placed in incomplete or inappropriate data may lead to wrong decisions being made.

Before we go any further then, I want to summarize the Why of scientific sampling. The rest of the booklet will emphasize How.

Principal reasons for scientific sampling	Disadvantages of judgement sampling
1. Bias and subjectivity in selecting sample units is minimized.	1. Although seemingly logical, personal biases can severely limit the 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 closely the sample can be expected to reflect 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. It is efficient, effective and economical, since the smallest size of sample 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 population itself, and cannot be projected to a larger population with any degree of confidence.

Furthermore, Sampling is generally more accurate than 100% enumeration and much more practical. This is so because there are many different sources of errors in any enumeration of mass data. For example, varying interpretations by many people of a common guideline, incompleteness of responses, errors in processing the data, delays in processing because of the volume. Such causes of error 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 program management, and policy making.

THE FIVE MAJOR STEPS IN CONDUCTING A STATISTICAL SURVEY

I CLARIFY THE PURPOSE AND DEFINE THE OBJECTIVES

II PLAN AND ORGANIZE THE SURVEY

III CONDUCT THE SURVEY

IV EVALUATE THE FINDINGS

V PRESENT THE RESULTS

Each of these steps will be discussed in more detail in the following pages.

CLARIFY THE PURPOSE AND DEFINE THE OBJECTIVES

a. Purpose/Problem Statement Surveys are usually requested to provide answers for management on problems they are encountering. Sometimes there is no particular "problem"; management just wants to be kept informed on the status of key areas of a project's implementation. In any event, your first task is to develop a concise statement of the purpose or problem. Frequently, management's request is only half formulated, ambiguous, a statement of observed symptoms or details that bother them and often it is expressed as a question. Get your guidance clear on what you are to study before you go any further, 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 becomes clearer, and the detailed survey questions can be formulated.

b. Use Why does management want the study? Often management has not thought through the use to which the answers to their questions will be put once they have been obtained. However, until you and they do understand and have defined how they intend to use it, you will be hampered in determining the kinds of questions to ask, and the manner in which the findings should be presented.

c. Importance How important does management consider the need for answers? Once this is established, you have a basis for establishing priorities, determining limitations and obtaining personnel, equipment and funding support.

d. Accuracy How accurate do the results need to be in order to meet management's objectives. Data collection and analysis is time consuming and expensive. Accuracy can only be obtained at a price, and diminishing returns for expended effort are always present at the higher levels. Minimizing time and cost aspects should be an important consideration.

e. Timing When does management want the results? Deadlines are important. If the answer is received after the need for it, the entire effort may prove useless, no matter how accurate the report, or beautiful its presentation.

f. Cost What is the budget limitation for this survey?

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

II PLAN AND ORGANIZE THE SURVEY

MAJOR ASPECTS TO CONSIDEP

a. Administrativ: What funds, staff, equipment and administrative coordination are necessary and available to conduct the survey?

b. Technical

1. Data Once the problem is understood, you should formulate a number of logical explanations (hypotheses) of what caused it. This in turn gives direction to the kind of questions that need to be asked in order to resolve which (if any) of the hypotheses are correct.

   Caution: Failure to take this step, may result in the gathering and compilation of a lot of data only to learn later that they offer no solution to your problem!

   a. What specific data are needed in order to answer the various hypotheses presented.

   b. What secondary data is already available and can be utilized -- to obviate collecting data that already exists.

   c. Source What is the most appropriate source for obtaining the required data.

   d. Method of Collection

      1. Secondary source statistics
      2. Analysis of secondary source data
      3. Personal interview
      4. Mail questionnaire
      5. Personal measurement by survey staff
      6. Personal observation by survey staff

2. Questionnaire Format Design and formatting of questionnaires is important as it improves accuracy in recording data. Wherever possible this should be pretested before actual use.

3. Master Lists If the sample is to be taken from established master lists, copies must be located.

4. Work Schedule A work schedule for completing each major step of the survey must be prepared at the outset, and then adhered to, in order to complete the work in time for management's use.

5. Sample Size and Distribution An appropriate sample size must be determined. Too large a sample will be wasteful of resources (time, money and people), while one too small, and or drawn in a biased manner may produce invalid results.

Most of the above require little or no further elaboration in a handbook of this nature. Questionnaire and Sample Size determination will be covered in more depth on the following pages.

THE QUESTIONNAIRE

There is no such thing as an "ideal" questionnaire. Questions and formats can be as varied as people. Nevertheless there are certain useful ground rules that can facilitate their construction. I will only cover the type of questionnaire that a trained interviewer would use to record information for manual tabulation, as this is the most likely form that will be utilized by NFAC in the immediate future.

QUESTIONS

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 it is analyzed, much of the additional data may be invalid.

b. Plan Ahead Work backwards, by planning the questionnaire in terms of the final report that you will be presenting to management. This will enable you to analyze whether the right questions have been included which will provide the answers requested.

c. Limit the Number Each question asked takes time (and costs money) to ask, process and analyze. Management's ability to ask questions will always exceed its staff's capacity to provide answers. Therefore be selective. Screen each proposed question carefully and decide whether the respondent is the appropriate source for the answer, or whether such answer can be more readily obtained elsewhere.

d. Avoid "Leading" Questions Many people color their answers to please the questioner. They tell him what they think he wants to hear. Others will 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.

e. 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.

f. Cross Check Questions If there is likely to be a strong element of doubt or distortion in the answer, provide for some objectively verifiable cross check questions, if possible.

g. 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 surveyer will interpret it in the field in his 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.

h. Pre-test your questions on others before deciding on the exact wording to be used in the 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 they may be readily referred to in the processing and analytical stage 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 available, 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 Question: . . . .
b. _____ No . . . .
c. _____ Don't know . . . .

Instead of --

4. Question: . . . .
. . . . ? a. Yes _____ b. No _____ c. Don't know _____

or

4. Question: . . . . a. Yes _____
. . . . b. No _____
. . . . ? c. Don't know _____

A recent survey format used by the NPAC is shown on the following page.

MASAGANA 99 MANAGEMENT INFORMATION SYSTEM
DATA VERIFICATION SURVEY

November 1974

PROVINCE ____________________

____________ 1. Mas 99 Hectares Reported PLANTED as of June 30

____________ 2. Mas 99 Hectares Reported PLANTED as of July 31

____________ 3. Mas 99 Hectares Reported HARVESTED as of October 31

____________ 4. Mas 99 Hectares HARVESTED AS A PERCENTAGE OF JUNE PLANTINGS

____________ 5. Mas 99 Hectares Reported HARVESTED AS A PERCENTAGE OF JULY PLANTINGS

6. HYPOTHESIS for APPARENT ERROR in 4 or 5 above. _____

____

7. FIELD COMMENT on accuracy of data and hypothesis, and/or reason for apparent error

____

____________ 8. Mas 99 Provincial AVERAGE YIELD reported, Çavans/Hectare

____________ 9. FIELD COMMENT on Mas 99 ESTIMATED AVERAGE YIELD

10. FIELD COMMENT on accuracy of reported yield and reason for apparent error.

____

____________ 11 Mas 99 Cumulative Hectares Reported Planted as of October 31

____________ 12. Mas 99 Cumulative Hectares Reported Harvested as of October 31 (3 above)

____________ 13 Mas 99 Estimated STANDING CROP before damage (11 minus 12)

____________ 14. Mas 99 Hectares Reported Totally Damaged

____________ 15. Mas 99 Estimated STANDING CROP AFTER DAMAGE (14 minus 13)

____________ 16. FIELD COMMENT Estimated Mas 99 STANDING CROP AFTER DAMAGE if above considered in error. (Question if reported damage is only Mas 99 or total province and Question if reported damage is Equivalent Total Damage or includes partial damage)

____________ 17. FIELD COMMENT Estimated POTENTIAL YIELD of Mas 99 Standing Crop

____________ 18. Estimated Non/Mas Cumulative Planting as of October 31

____________ 19. Estimated Non/Mas Cumulative Harvesting as of October 31

____________ 20. Estimated Non/Mas Yield C/Ha

____________ 21. Estimated Non/Mas Equivalent Total Damage Ha

____________ 22. Estimated Non/Mas Standing Crop

____________ 23. Estimated POTENTIAL YIELD of Non/Mas Standing Crop

DETERMINING SAMPLE SIZE

Statistical methods are generally useless when dealing with one, or only a few quantitative measurements.[1] It is not possible to prove a point or shed light on a problem unless a number of measurements or observations are available. At the same time, complete counts of a population are usually either impossible to obtain in most instances, or prohibitively expensive. Thus sampling is resorted to as the most expedient method for obtaining data about a population at a reasonable cost.

What size sample is appropriate for conducting a survey however? As a general rule of thumb, statistical techniques can usually be effectively applied[2] when at least 30 measurements are obtained at random.[3] This is usually insufficient however if we wish to present our findings with any quantifiable degree of confidence.

A great deal of time, money and effort can be wasted if the size of the sample is either larger or smaller than is required to meet the specified needs of management in conducting the survey. More items than required would waste resources, while fewer items than necessary would also give results with less than the required reliability.

First, we must correct two popular, but erroneous misconceptions. It is often thought that a sample should be some percentage, say 5% or 10% of the population under study. Secondly, it is often believed that 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. In fact, if the sample is less than 5 percent of the population under study, its percentage size plays no significant role in determining reliability.

Secondly, even if the sample size is thought of in terms of number of units rather than some percentage of the total population, the size of the population itself is a minor factor in determining the size of the sample.

Finally, the information derived from a survey is based on the actual units selected in the sample. The results however 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.

[1] Nevertheless many people do make such judgements -- for instance they will recommend or condemn a particular restaurant on the basis of eating one meal there, even though in the long run that may have been an unusual situation, not typical of "normal" performance.

[2] Caution. This merely enables you to generalize about a situation, but the process is not reversible. You cannot make specific inferences about a particular case. For instance, if it is found that the average amount of rainfall in Pampanga on August 1st over the past five years has been 2.13 inches, you should not use this to predict that next year it will be 2.13 inches.

[3] Randomness will be discussed in greater detail on page 29

SOME BASIC STATISTICAL CONCEPTS

Before we go any further, I want to review some basic statistical measures and concepts that are used in determining sample size.

AVERAGES

The most frequently used statistical measure for describing masses of data is the average, because it reduces the 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 group. It should be borne in mind however, that since it is a single value, it does not accurately reflect the standing of every item in the group. It merely provides a means to generalize about a mass of data.

This is sometimes misunderstood, because the variation around the average is ignored. For example, if we state that the average palay production in rainfed areas of Central Luzon is 60 ca/ha, and further assume that 60 ca/ha enables a farmer to meet expenses and make a reasonable income, it does not follow that all farmers in rainfed areas of Central Luzon 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 standard.

A further problem is that one statistical average may be used to represent groups of situations which are dissimilar. Although the resulting mathematical calculation may be correct, it may not present an accurate or useful picture of either group. For example, given that Luzon and the Visayas are experiencing heavy rainfall and flooding, while Mindanao is having a drought, it could be stated statistically that the overall rainfall level for the Philippines at that time was "Satisfactory" or "normal". A first step in calculating an average therefore is to separate the various groups to be averaged into similar groups, where known, and calculate separate averages for each group.

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 reflects the summation of the values of a group, divided by the number of items. It is often described as a mathematical "balance point", thus

$$M = \frac{\sum x}{N}$$

Where

M = mean
$\sum$ means the "sum of"
x = values of the items in the group
N = number of items in the group

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

DATA ITEM	DATA VALUE X
1	20
2	32
3	44
4	68
5	68
6	80
7	92
8	104
9	116
N = 9	x = 624

$$\text{Mean} = \frac{624}{9} = 69.33$$

Median

The median is the "mid-point" of the range of values in a data series. In the foregoing series, the 5th item, "68" is the median value. Since there is an odd number there is no problem. Otherwise we would have to take the mean of the two middle values.

The median is a useful average to employ in dealing with frequency distributions when the first and/or last grouping is open-ended and the mid-points of these groups cannot be reasonably estimated, since the values of the end groups is not required. Furthermore, when 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 is significant.

Mode

The mode is a "concentration point" - the most frequently occuring value in the data series. Again in our preceding distribution, it is "68". The mode is often used when dealing with ungrouped, 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 carabao per farm.

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

PERCENTAGES AND RANK ORDERING

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

$$\% = \frac{f}{B} \times 100$$

Where

% = percentage
f = Item frequency or value
B = Base size or value
100 = constant (100)

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 answer to the manager who wants to know the sequence standings -- 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 program data itself. Note however that rank ordering merely indicates the sequence -- it does not indicate the magnitude or the spread between each rank.

A fine point in rank ordering is that when there are "ties" for any position, the rank order should be arithmetically averaged rather than assigning the most favorable appearing rank; and subsequent ranks are unaffected. See the table below for further clarification.

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

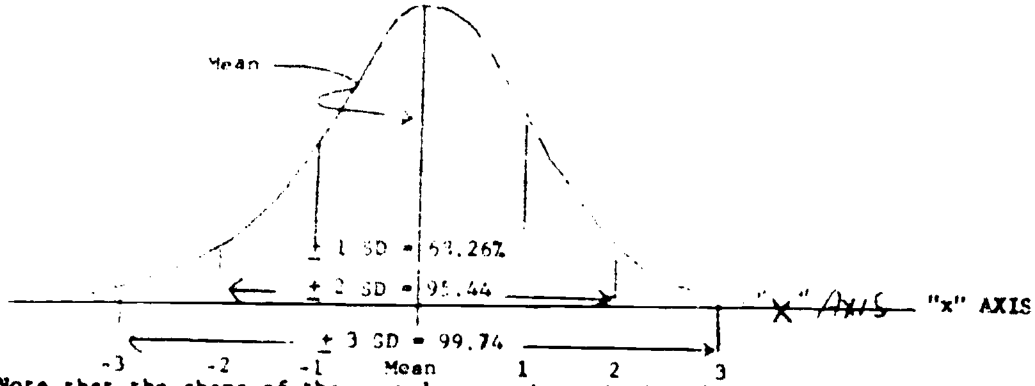
"THE "NORMAL DISTRIBUTION CURVE"

Although no two situations are ever exactly alike, statisticians have discovered that the frequency distributions of processes that can be repeated many times under similar conditions, (each occurrence of which is affected in minor ways by natural common factors and/or chance), tend to form a general symmetrical "bell-shaped" distribution pattern. This is known as the "Normal Distribution Curve". 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 the analysis of agricultural situations are symmetrical and unimodal, approximating the normal curve, and 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 range of data differs from the mean. This is done by measuring the area under the curve, from the mean to the value of the data items in question.

The normal curve has certain properties. The distance from the mean to any point is measured in terms of a unit known as 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 3 on page 72.



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.

THE STANDARD DEVIATION

Previously, we discussed the use of various averages (mean, median and mode), as "measures of central tendency". We also observed a major limitation, namely that the variation around that average was ignored, which could lead to distorted impressions of the true situation.

Averages, such as average rate of seeding per hectare, average rates of fertilization, average yields, average price per cavan, average loan, average repayment rate, etc. etc., are all familiar and useful measures in oformulating recommendations for agricultural programs, and in their management. However, we recognize that no two specific situations are exactly alike. For instance, even if both farmer Cruz and farmer Rodriguez were to follow the same guidelines to produce a rice crop, because of the many differences in their personal situations and attitudes, the natural factors which exist, and the chance occurrences which may affect either, they are both likely to obtain differing yields.

For program analysis and management purposes, the extent of the differences is extremely significant. Therefore, in addition to the foregoing averages another unit of measurement is necessary which provides a quantitative "measure of dispersion". This is the Standard Deviation, and is derived from the mean and the frequency distribution itself.

The formula for calculating the Standard Deviation from Simple-Random Samples for ungrouped data is as follows:-

$$S = \sqrt{\frac{\sum d^2}{N}}$$

Where

S = Standard Deviation
d = difference from the mean
N = number of items in the group

Let us illustrate the use of this formula with an example.

Find the Standard Deviation of this group of five numbers: 10,20,25,40, 80. By addition, the sum of the numbers is 175, and the mean is

$$\frac{175}{5} = 35$$

The difference of each value from the mean is shown in the table below. To eliminate the influence of the ± signs to obtain the sum, the difference is squared, and later the square root is taken. Thus

A Item	B Item Value	C Difference from Mean(d)	D Difference Squared(d^2)
1	10	- 25	625
2	20	- 15	225
3	25	- 10	100
4	40	+ 5	25
5	80	+ 45	2025
N = 5	$\sum$ = 175		$\sum d^2$ = 3000

By substituting in the formula,[1] the standard deviation is calculated

$$\sqrt{\frac{3000}{5}} = \sqrt{600} = 24.495 \text{ or } 24.5 \text{ rounded off}$$

Since the mean of the distribution was 35, this new measure tells us that 10.5 is one standard deviation less than the mean, (35 - 24.5) and 59.5 is one standard deviation greater than the mean (35 + 24.5). We will use such measurements later in analysing frequency distributions.

[1] This is for illustrative purposes only. Actually, "N-1" is used instead of "N" for groups of less than 30.

IMPORTANT CRITERIA FOR DETERMINING SAMPLE SIZE

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

1. Extent of variability[1] in the population under study.
2. Amount of error that will be tolerated in the findings.
3. The confidence desired when presenting the findings, that the data is accurate.
4. The amount of money, time and other resources available to obtain the data, conduct the survey and process the findings.

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

<u>For instance</u>, Management might want to know the production (ca/ha) of irrigated farmers in Iloilo during the 1973- : Wet Season.

In planning the survey, one thing you must determine is:

> How many hectares should be sampled in order
> to estimate the production (ca/ha) of irrigated
> farmers in Iloilo for the 1973 ' Wet Season?

Unfortunately management does not usually give precise directions when asking questions. It is therefore part of your task as the survey designer to acquaint management with the facts of survey life, then assist them in determining the degree of accuracy that will meet their requirements, balancing what is possible, given the time and resources available to conduct the survey. Only then can you establish an appropriate sample size. Points to stress are:

a. The final answer will be in terms of an average, or a percentage, with variability around this number.
b. No survey can be 100% accurate, therefore management must specify how accurate they need it to be.
c. Warn management that accuracy (or anything approaching it) usually costs excessively and takes time Then "bargain" with them to settle for something less than perfection.!

Practically, if management cannot or will not make these judgements, you as the designer will have to do their job for them in this situation.

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

The problem above is seeking its ultimate answer in terms of an average. We would expect our final answer to management to state "The estimated production of irrigated farmers in Iloilo for the 1973 Wet Season is YX cavans per hectare."

Let us review each of the criteria in turn, and what can be done about quantifying them for our problem.

[1] The amount of difference between individual members in the population.

VARIABILITY

Extent of variability in the population under study. How can you determine the variability in data before you have collected that data? This is a very practical question, and of course the answer is you cannot! Therefore you have to start with an educated guess. This can be based on a sample of historical data, experience in similar situations, or "expert" opinion. If this is not possible, don't make the final determination of sample size until you have taken the first 30 samples, when you can use that data to approximate the "standard deviation"[1] for the formula.

Practically, if you have any technical background in the subject you are surveying, you should be able to make "ballpark" estimates of the data, as follows:-

1 Estimate the range extremes (the lower and upper limit cases) that you expect to encounter in normal production under prevailing field conditions. Substitute in the following formula to obtain the estimated standard deviation.

$$D = \frac{b - a}{6}$$

Where:

D = Estimated Standard Deviation
b = upper limit of the range
a = lower limit of the range
6 = a constant (6) to be used
in all computations.

For example, based on your professional judgement as an agriculturalist, and prior experience in Iloilo, you might expect that the farmers in Iloilo would produce between 55 to 155 ca/ha, barring some absolute disasters or fantastically high yields.

Then

$$D = \frac{155 - 55}{6} = \frac{100}{6}$$

$$= 16.6 \text{ or } 17 \text{ rounded up}$$

If you do not have any technical background in the subject matter - consult with an "expert", and discuss your needs with him/her.

Do not become overly concerned about mathematical precision here -- use the best judgements available, round off to integers[2] and get on with the job. Thus using 17 as the estimated standard deviation is a first approximation which will suffice at this stage. Later, after you have taken the sample, judgement errors will be reflected and adjusted in the final results. The important task is to make the study and obtain those results, not to mull interminably over making a "correct" estimate of a situation before it has been studied!

[1] The standard deviation is a measure of variability in a collection of data. For a fuller discussion of the standard deviation and how to calculate it, see pages 18, 45 and 46.

[2] Whole numbers

TOLERABLE ERROR

Amount of Error that will be tolerated. Any findings developed from a sample survey, no matter how scientifically obtained, will only be approximations. This should be clearly understood at the outset. In general, the greater the desire for accuracy, the larger the sample must be. How much error will be acceptable is of course a management decision to make. However, you should be prepared to provide some additional data as a basis to help management make that decision.

First of all in our problem of Iloilo farmers what you are ultimately trying to estimate is the production rate in cavans per hectare. Try to determine how close management wants the final answer to be -- within 1 ca/ha, 5 ca/ha or what? How close is "close enough" for the purpose in this instance? What magnitude will make a difference in the use to which the findings will be put?

1. As a first step, get an idea of the size the number might be; either from historical data, prior experience, professional judgement; or more simply using the "range" data already developed to estimate the variation. Thus:-

$$M = \frac{b - a}{2} + a$$

Where:-

M = estimated average
b = upper limit of the range
a = lower limit of the range
2 = a constant (2) to be used in all computations

Following through the previous example where the upper and lower limits were estimated at 155 and 55 ca/ha, respectively, we have

$$M = \frac{155 - 55}{2} + 55$$

$$= \frac{100}{2} + 55$$

$$= 50 + 55$$

$$= 105$$

The average (or mean)[1] then, is likely to be around 105 ca/ha.

2. If this were to be so, would 100 - 110 be close enough to be of use to management?

Remember excessive accuracy is expensive, wasteful and extremely time consuming.

[1] Although "Average" is a term in common use, a more precise term is "mean" since there are several types of "average" in general statistical use. See pages 14 & 15.

CONFIDENCE

Confidence desired when presenting the findings

After you have obtained an answer, how sure do you want to be when you present it to management that the answer is correct? Of course, you'd like to be 100% correct but again in dealing with samples this is not possible and you must settle for something less. "How much less" is a decision usually made by the survey director. This decision will also have a bearing on the size of the sample to be taken.

If we took a 100% sample of a population and did everything accurately, when we calculated the "mean" of that population, we would expect our answer to be correct. When we take samples of less than 100% however we know we run the risk that our "sample mean" may not be exactly the same as the "true mean". For example, given a total population of nine numbers[1] -- 1,2,3,4,5,6,7,8,9 the true mean can be calculated as

$$M = \frac{\Sigma x}{N}$$

$$M = \frac{1+2+3+4+5+6+7+8+9}{9}$$

$$= \frac{45}{9} = 5$$

Where

M = true mean

Σ = means "the sum of"

x = values of the numbers in the population

N = population size

If we were to take random samples[2] of different sizes from this population, we might obtain results as follows:

Sample Size	Sample Data	Sample Mean
1	3	3.00
2	2,5	3.50
3	2,5,7	4.67
4	2,4,6,9	5.25
5	3,6,7,8,9	6.60
6	1,2,3,4,5,8	3.83
7	1,2,4,5,6,7,8	4.71
8	1,3,4,5,6,7,8,9	5.38

Obviously, the "means" of the various samples are not the same as the "true mean", nor, reasonably, could we expect them to be. Given such a difference though, how can we infer anything about the true mean based on any of these samples?

Statistically, there is a procedure whereby we can calculate a range of error around the "sample mean". This range (called the "standard error of the sample mean") is the range around our "sample mean" in which the "true mean" will probably fall. It is calculated as follows:-

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

Where

E = One standard error of the sample mean

S - Standard Deviation of the population from which the sample was drawn.

n = size of the sample.

Thus, it is a "standard deviation" for a special situation.

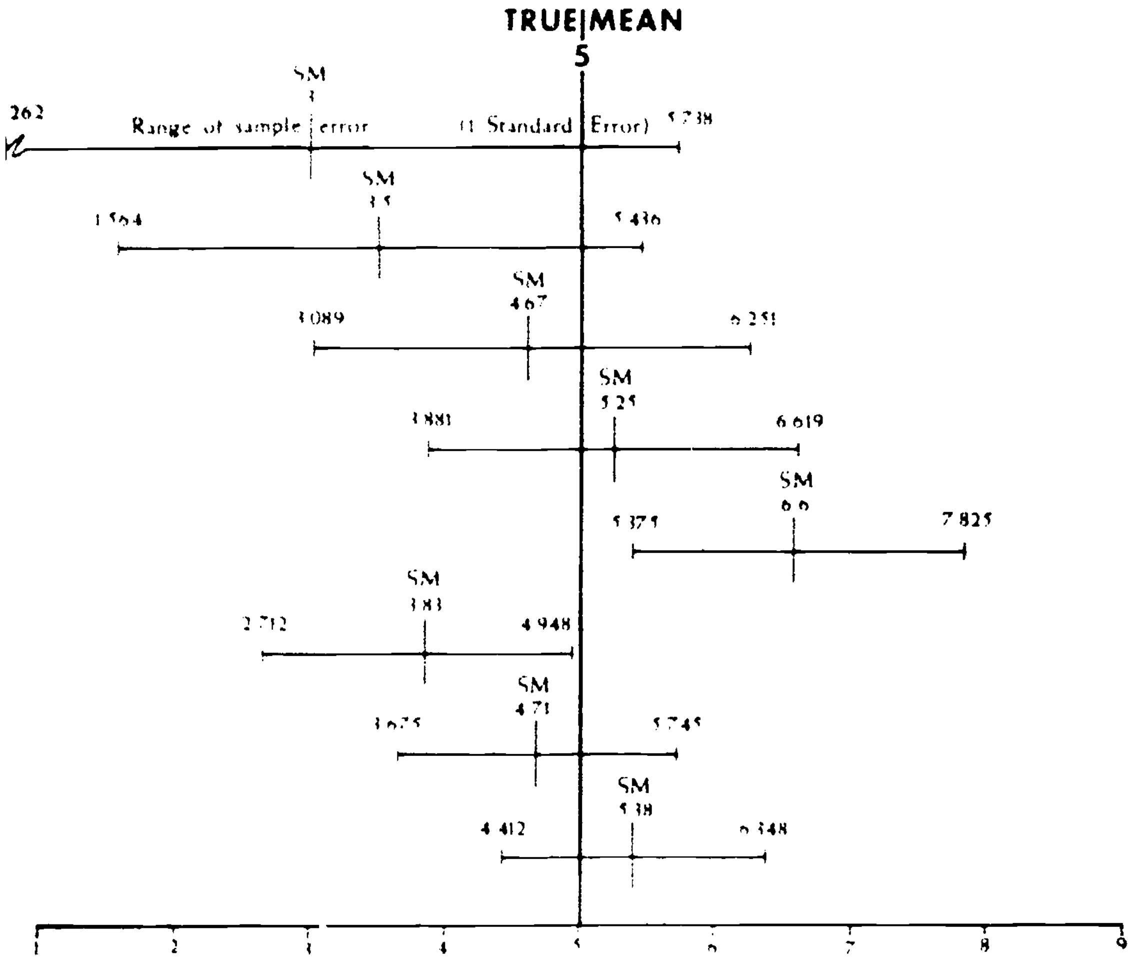
[1] For simplified illustration only a very small population and samples are used.

[2] For a discussion of randomness, see page 28.

In this example, the results can be calculated as shown in the table below -

Sample Size	Sample Data	Sample Mean	Standard Error of the Sample Mean
1	3	3.00	2.738
2	2,5	3.50	1.936
3	2,5,7	4.67	1.581
4	2,4,6,9	5.25	1.369
5	3,6,7,8,9	6.60	1.225
6	1,2,3,4,5,8	3.83	1.118
7	1,2,4,5,6,7,8	4.71	1.035
8	1,3,4,5,6,7,8,9	5.38	.968

Graphically, this can be shown as follows:



Thus in general the larger the sample, the smaller the range of "sample error", and usually (but not always) the lesser the possibility for actual numerical error in the "sample mean" due to sampling bias.

Drawing upon probability theory[1], with any sample size we can express our confidence in the "sample mean" as follows:

Number of "Standard Errors" from the Sample Mean (E)	Probability that the "True Mean" is within this range (P)	Probability that the "True Mead" is not within this range (100-P)	Chance of the "True Mean" being within this range (P/(100-P)
1	68.26%	31.74%	68.26/31.74 or 2:1
2	95.44	4.56	95.44/4.56 or 20:1
3	99.74	0.26	99.74/0.26 or 369:1

Although 1,2 & 3 "Standard Errors" are illustrated here, actually any number between 0.1 and 3.9 may be used by referral to the "Normal Curve and Related Probability" table on page 72.

Essentially, any specified sample mean will fall within a range formed by the true mean, and a given number of "standard errors" on either side of it. Thus, about 68 percent of all possible means will fall within a range ± one standard error of the mean. In other words, the probability is about 68 percent that the mean of a sample selected at random will be within this range. Conversely, the probability is 32% that it will not be. Thus the chances are 68/32 = 2:1 that it will be. As we increase the range to two standard errors, the chances are 95.5% (or about 20:1) that the true mean will be within the range of the sample mean. Generally, to increase the confidence in an estimate for a given sample size, a wider range of error must be allowed for.

When management specifies the amount of error it will tolerate, the confidence in the answer can be calculated, thus:-

$$\frac{\text{Management Tolerated Error}}{\text{1 Standard Error}} = \text{Number of Standard Errors utilized}$$

For example, continuing the foregoing illustration, with a population of 9, if management wanted to know the true mean and was willing to tolerate an error of 2.738, with a sample size of one, our confidence would be limited to 68.26%. (1 standard error).

$$\frac{2.738}{2.738} = \text{1 Standard Error}$$

Where

Sample Size = 1
E = 2.738 = 1 Standard Error
T = 2.738 = Tolerated Error

However if we were to take a sample size of eight, where 1 standard error is reduced to .968, our confidence would be increased as follows:-

$$\frac{2.738}{.968} = \text{2.83 standard errors}$$

Where

Sample Size = 8
E = .968 = 1 Standard Error
T = 2.738 = Tolerated Error

which from page 72 is equal to 99.54%.

Combining these concepts of tolerated error and confidence ahead of time, if management was willing to tolerate an error of 2.009 in our answer, and we desired to present our findings with a confidence of 89.9% probability, then from page 72, 89.9% confidence is at the 1.64 standard error point. Therefore, if an error of 2.009 is permitted, and it must fall at the 1.64 standard error limit, the size of one standard error is found as follows:

$$\frac{\text{Management Tolerated Error}}{\text{Number of Standard Errors to be utilized}} = \text{One Standard Error}$$

which in this case is $\frac{2.009}{1.64} = 1.225$

By reviewing our standard error table for the 8 different size samples illustrated, we can see that only a sample of 5 would be required in this instance. These concepts can be generalized into a formula to calculate the appropriate sample size under various conditions.

OPTIMUM SAMPLE SIZE FORMULA FOR ESTIMATING A MEAN

Having established an understanding of the elements which are involved, the following formula can now be used to determine the optimum sample size for estimating a mean.

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

Where

S = Optimum Sample Size
D = Standard deviation of data in the population
E = Size of the error in the mean that management will tolerate
K = Confidence with which we wish to present the findings

Selected Values of K	Confidence Percentage	Numerical
1	68.26	2:1
2	95.44	20:1
3	99.74	369:1

(See page 72 for more complete and precise determinations of "K".)

Let us now restate our problem of the palay production by Iloilo farmers:

Question What size sample of hectares should be used in order to estimate the palay production (ca/ha) of irrigated farmers in Iloilo for the 1973-74 Wet Season?

Management is willing to tolerate an error in the answer of as much as 3 ca/ha in either direction, and we want a 20 to 1 confidence that our answer will not exceed this degree of error. We further estimate the standard deviation in production to be approximately 17 ca/ha.

$$S = \frac{17^2}{(3/2)^2} = \frac{17^2}{(1.5)^2}$$

$$= \frac{289}{2.25}$$

$$= 128.44 \text{ or } 129 \text{ rounded up.}$$

This means that 129 samples of separate, randomly selected hectares will meet our requirements as specified in this problem, regardless of the number of hectares that are actually being harvested in Iloilo during the specified period.

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.

OPTIMUM SAMPLE SIZE FORMULA FOR ESTIMATING A PERCENTAGE

The preceding formula was useful for estimating a mean. However, it is often necessary to provide management with an answer in terms of a percentage. For example, management might have posed another question:

Question: What percentage of palay farmers in Nueva Ecija have year round irrigation on their paddies?

To determine the appropriate sample size to answer this question, the following formula is used

$$S = \frac{(100 - P) \cdot P}{\left(\frac{E}{K}\right)^2}$$

Where

S = Optimum Sample Size
100 - Constant (100) in all equations
P = Preliminary estimated percentage (The preliminary estimated answer to the question being asked)
E - Size of the error in the percentage that management will tolerate
K = Confidence with which we wish to present the findings

Selected Values of K	Confidence	
	Percentage	Numerical
1	68.26	2 to 1
2	95.44	20 to 1
3	99.74	369 to 1

See Page 22 for more complete and precise clarification of "K".

As in determining the Optimum sample size for a mean, management must specify the degree of precision it wants in its answer, as well as asking the question.

Since "E" and "K" have already been discussed at length on pages 21 through 24, that discussion will not be repeated here. We will examine "P" however.

Preliminary Estimated Percentage

Similar to the need to determine the variability of the population ("D") in the previous formula, we have a requirement in this formula to make a preliminary estimate of the answer to the question being asked. As before, if you have any technical background in the subject matter under study, you may be able to make a guesstimate. If not, you should consult with an "expert" and use his informed opinion.

The need is to select a number between 1 and 99. (0 and 100 do not compute!) As a guide to this process, you should be aware of the following general trends

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

Thus, if you have no feel for the situation, and really can get no expert opinion you can play safe by using 50 as this gives the largest possible result. Do not agonize over this preliminary answer. It is only part of a process to help determine the appropriate sample size to take. Select the number and get on with the job of finding the real answer!

Let us use this information to rephrase the question and demonstrate the use of the formula.

Question: What percentage of palay farmers in Nueva Ecija have year round irrigation on their paddies?

Management is willing to tolerate an error in the percentage of as much as 2 percent, and we want to be 99.74% sure that this degree of error will not be exceeded. We will assume that the preliminary percentage estimate is 50%.

Then, substituting in the formula-

$$S = \frac{(100 - P) \times P}{(E/K)^2}$$

Where

S = Optimum Sample Size
P = 50 = Preliminary Estimated Percentage
E = 2 = Tolerable Error
K = 3 = Confidence of 99.74%

We have

$$S = \frac{(100 - 50) \times 50}{(2/3)^2}$$

$$= 5,625$$

This is a large sample, and apart from the expense will take a long time to gather, analyse and process. Advise management of this. Perhaps, in reviewing their needs, they might relax their specifications, as follows:-

$$S = \frac{(100 - 50) \times 50}{(5/2)^2}$$

$$= 400$$

P = 50
E = 5
K = 2 (i.e. 95.44% probability)

This is a much smaller (and thus easier and less costly) study to conduct

Thus, by appropriate feedback consultation with management, the survey director can usually develop a sample size that is both feasible to conduct, within the resource constraints, and appropriate to management's needs.

As in estimating the Optimum Sample Size for a mean, it is good practice to increase the actual sample size over the optimum size, in order to protect against possible error in estimating the percentage, 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.

SCIENTIFIC SAMPLING METHODS

Once you have established "How Many" samples to draw from a population, the next important problem to be resolved is "Which ones?"

"Spot-checking" and "judgement" samples are often resorted to by people in a hurry. They tend to "play it by ear," reaching out in any or all directions to grasp for information from anyone who might be available. Such impressions may turn out to be valid, and again they may not. With experience, an individual may be able to sharpen his judgement and develop a "feel" for the situation - where to go and who to ask under varying circumstances. Nevertheless "quick and dirty" appraisals conducted in this manner are impressionistic only, and although useful to enable a policy maker to improve his mental picture of the"real world", they cannot (or should not) be used for quantitative analytical purposes, since there is no way of measuring their reliability. The "scientific way" is to use "random sampling"methods.

Contrary to popular impression, random sampling is not a process of arbitrary, haphazard selection of items from a given population. Rather it is selection in a manner which assures that each item in the population has an equal chance of being selected.

There are several approved methods for drawing samples from a population, each of which has certain advantages depending upon the circumstances. But, before you plunge in and start selecting "representative" items, you must determine the relative importance of items in the population. 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 available to conduct the survey.

Each of these will be discussed with "how to do it" illustrations.

SIMPLE RANDOM SAMPLING

Table of Random Digits

A good "scientific" method to use in simple random sampling is a table of random digits such as [illegible] 1 page 19. These tables have been carefully constructed to utilize the digits 0-9 in a completely unstructured, 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[1] under study.

Second, Use the total size of the population to determine the grouping of random digits in the table that will be used. For example, if the population is between 10 and 99, use groupings of two digits; between 100 and 999, use groupings of three digits; between 1,000 and 9,999 use groupings of four digits, and so forth.

Third, Assign sequence numbers to the population under study.

Then, Select any point in the table to start, grouping as explained above.

Finally, Proceed in any systematic manner, (i.e. down, across, etc.) selecting and recording those numbers that fall within the population range, and disregarding numbers outside the range, until the total designated sample size has been selected.

For example, let us assume we are going to select five provinces to visit from a list of forty three, using the random digit table in Table 1 page 16.

1. The population is 43 therefore use groupings of two digits.

- Assign sequence numbers to the list, thus

Sequence # & Province	Sequence # & Province	Sequence # & Province	Sequence # & Province
1 Nueva Ecija	12 Laguna	23 Quezon	34 Aklan
2 Iloilo	13 Cagayan	24 Bataan	35 Surigao del Sur
3 Pampanga	14 Ilocos Sur	25 Bohol	36 Southern Leyte
4 Pangasinan	15 Nueva Vizcaya	26 La Union	37 Antique
5 Tarlac	16 Capiz	27 Leyte	38 Misamis Occ
6 Camarines Sur	17 Mindoro Oriental	28 Davao del Sur	39 Negros Oriental
7 South Cotabato	18 Negros Occ	29 Batangas	40 Davao del Sur
8 Ilocos Norte	19 Mindoro Occ	30 Zambales	41 Bukidnon
9 Isabela	20 Albay	31 Camarines Norte	42 Zamboange Norte
10 Bulacan	21 Zamboanga Sur	32 Cavite	43 Zamboange Norte
11 North Cotabato	22 Lanao del Sur	33 Rizal	

2. Determine the groupings. In this instance, since the total population is 43, or two digits, we will use two columns for the two digit grouping.

3. Select a starting point from the random digits in this table. (Any one could be used as the starting point.) For convenience in illustration we will start with the top left pair of columns, with digits-- "05".

4. Proceed in any systematic manner, and select those numbers that fall within our population range, until five appropriate numbers have been selected. If we work down the page, the numbers are 05,86,87,02,64,57,56,98,51,12,57,51, 39,24. Those underlined fall within our range corresponding to:-

02 Iloilo; 05 Tarlac. 12 Laguna; 24 Bataan; 39 Negros Oriental

[1] Population is used in statistics to signify the total number of things from which you are drawing a sample.

RANDOM DIGITS - OPTIONAL PROCEDURE

An Optional Procedure that will speed up the selection process is to assign more than one sequence number to each item. Dividing the upper limit of the group by the population total and rounding down to the whole number will determine the appropriate amount of numbers to assign to each item. For example, in the situation above, where we have a two digit grouping (upper limit 99) and a total population of 43,

$$\frac{99}{43} = 2.3$$

two sequence numbers to each item in the population would be the appropriate allocation. What this procedure accomplishes is to lessen the number of rejected random digits since now 86 (43 times 2) of the 99 digits in the grouping are in use.

Sequence numbers would then be assigned to the list, thus

Sequence # & Province	Sequence # & Province	Sequence # & Province	Sequence # & Province
1,2 Nueva Ecija	23,24 Laguna	45,46 Quezon	67,68 Aklan
3,4 Iloilo	25,26 Cagayan	47,48 Bataan	69,70 Surigao del Sur
5,6 Pampanga	27,28 Ilocos Sur	49,50 Bohol	71,72 Southern Leyte
7,8 Pangasinan	29,30 Nueva Vizcaya	51,52 La Union	73,74 Antique
9,10 Tarlac	31,32 Capiz	53,54 Leyte	75,76 Misamis Occ
11,12 Camarines Sur	33,34 Mindoro Or	55,56 Davao del Sur	77,78 Negros Or
13,14 South Cotabato	35,36 Negros Occ	57,58 Batangas	79,80 Davao del Sur
15,16 Ilocos Norte	37,38 Mindoro Occ	59,60 Zambales	81,82 Bukidnon
17,18 Isabela	39,40 Albay	61,62 Camarines Norte	83,84 Zamboanga Norte
19,20 Bulacan	41,42 Zamboanga Sur	63,64 Cavite	85,86 Misamis Or
21,22 North Cotabato	43,44 Lanao del Sur	65,66 Rizal	

Using the same starting point and procedure as on the previous page, we would only have to run through six sequence numbers to get our quota instead of fourteen as previously, thus, 05,36,87,02,64,57, rejecting only 87. The provinces selected would then be -

05 Pampanga, 36 Misamis Oriental, 02 Nueva Ecija, 64 Cavite; 57 Batangas

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

DECK OF CARDS

A practical method for drawing random samples 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 (or disregarding and reselecting, if drawn) some cards. The deck of numbers is easily "randomized" by shuffling, cutting and drawing. As in using random digit tables, you must assign sequence numbers to the population.

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

This procedure introduces some problems as unless you are careful, it may not be as scientifically objective as a random digit table.[2] Nevertheless, it has certain practical advantages as 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 where the total population is not masterlisted, or well defined.

Psychologically, the attempt to eliminate subjectivity and the concept of chance can be more appreciated by the people you are surveying. It also serves as 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, after you have chosen their area to be surveyed by a previous sub-grouping.

For example, at the National Food and Agriculture Council (NFAC) level, although you may know in gross numbers how many farmers are enrolled in the "Masagana program" by province, you will not 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 barrios, and ultimately from the farm management technician, several farmers can be selected from his 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 symmetrical, the individual items in the population will not have an equal chance of selection.

[3] Nor should you. It is not generally necessary nor desirable to accumulate masses of detailed data at higher management levels.

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 starting point from the population at random. (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.

Example: We wish to interview 6 out of 193 technicians assigned to the Masagana program in Pangasinan. How would these he selected by systematic random sampling?

1. Assign sequence numbers from 1 to 193 to the technicians.
2. Determine the skip interval.

$$i = \frac{193}{6} = 32.16$$

Round down to the whole number, = 32.

3. Select a random starting point. Here is a working method which I could employ. (You can use your imagination to create others).

   a. Start at the upper left corner of the table. Count off the digits across the top equivalent to the skip interval. Group in three's after that (equivalent to the population size - 3 digits) and proceed from left to right, then right to left down the page, discarding until a three digit number is reached that is within our population range.

   Employing this working method, the 32nd digit would be 2, followed by the groupings "359", "652" which would be discarded, and then "069" which would be acceptable.

4. Starting with technician 69, and selecting every 32nd technician thereafter, until six technicians had been chosen, we would then have 69, 101, 133, 165, 4 and 36. (Note: 165 + 32 = 197. Since we only have 193 in our population we would have to go back to 1 and start over again. Hence, "4" would be the next selection after 165).

Caution: Sometimes, items in a population are arranged in a particular pattern or order 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 may instead all have the same characteristic.

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 3 hour skip interval, and start at 9 am -- with a sampling of activity at 9 am, 12 noon, 3 pm and 6 pm you might draw the conclusion that there is very little work going on except perhaps early in the morning, since at other times people were consistently eating lunch or merienda, 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 characteristics 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, we can usually get a more accurate picture of the total population by selecting a random sample from each group so identified. This process is known as "stratified" random sampling.

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 palay yields. The result would be much more meaningful than merely selecting farms at random without regard to such 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 South Cotabato and the province had been stratified as indicated below, the sample size for each category would 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%	200

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

CLUSTER SAMPLING

As indicated earlier, cluster sampling is often resorted to as the only practical means to gather data where time limitations and/or difficult field travel conditions make it impossible to obtain data any other way.

As its name implies, instead of selecting data from many different geographical locations, many respondents are queried at fewer locations. Whenever possible, the total appropriate population (for instance all palay farmers in a selected barrio) should be interviewed.

In practice, it may take two or more days for an interviewer to obtain responses from ten farmers by simple random sampling if they are scattered all over the province, as this may mean extensive travel from one remote barrio to another. On the other hand, by randomly selecting two barrios, and interviewing as many farmers as possible within those barrios, many more farmers may be contacted in a much shorter time period.

Because by this method the samples will be drawn from a more limited cross section of the total population, it is desirable to go beyond the minimum sample size specifications. Furthermore, as many clusters should be selected as can be accommodated by the time/budget limitations. Clusters should be approximately the same in size.

It is important to remember that the clusters themselves should still be selected on a scientific rather than a judgement basis. Furthermore, if sampling is done within the cluster rather than the entire group, it too should be done randomly.

CONDUCTING THE SURVEY

Some general guidelines which should be observed are as follows:

Brief the Interviewers A survey is rarely conducted by one individual. Therefore, 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 procedure to follow when they encounter difficulties. If possible, provide for a "dry run" interview session to supplement the orientation process.

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 was selected to be interviewed.
Assure him of confidentiality or anonymity of results.
Tell him 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 barrio's "live entertainment".)

Conducting the Interview - Use your judgement whether to follow a structured questionnaire format reading off each item, or whether to use an unstructured interview style. 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 that if you don't write it down, you may forget it, and/or fail to pass on their comment. 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, you may also miss important questions.

Field Computations Use local or familiar measures, and minimize computations by the respondent. Get raw data which you can convert to percentages, etc. later. Most people perform poorly in mental arithmetic, therefore record information in the terms which the farmer gives it to you. Note the conversion factor and do it later to obtain the desired measures.

CAUTIONS TO OBSERVE IN CONDUCTING SURVEYS

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, or gain your sympathy.

For instance, the 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, repeat your questions several different ways if necessary to ensure that they are understood and the person being interviewed is responding accurately to the best of his knowledge.

Remember - Do no promise anything, except to pass on information unless you have authority to take corrective action. You are usually only there as an observer and gatherer of facts. The individual being interviewed on the other hand usually regards you as a representative of the government who can and should do something about the situation. Idle promises will only result in a lack of confidence and lessen cooperation the next time around.

EVALUATE 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.

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 . . . " Wherever 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 own writing, and/or does not recall the context in which the comments were made, although at the time they may have seemed meaningful.

Where multiple choice responses have not been used, the editing staff has a difficult task of developing a standardized scheme to classify "open-ended" comments received. It is often impossible in fact, 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 afterwords.

It may also develop that some things which were overlooked, or thought not to be important in designing the questionnaire actually have great significance. Thus some preliminary modification or even elimination of questions and responses may be required.

Internal Consistency It may be observed on multiple choice questions that check marks have been placed in more than one option, 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. With number responses, editing is frequently required to recalculate the recorded value into the standardized units requested. Sometimes the conversion factor is provided, sometimes it has been overlooked.

Correction A whole range of important decisions therefore have to be made in the editing process on how to treat the data. Should it be rejected outright as erroneous, counted at face value regardless of its apparent error, or accepted but reduced in value, with an attempt to figure the "intent"? This is part of the editorial task.

Mark-up Finally, to simplify the data processing task which follows it may be necessary to transform all the check marks in the standardized responses into a "Base number". For example, if a series of questions have been 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 amounts of fertilizers?	Not enough area 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.	______ ca/ha		What yield did you obtain? (44 kilos/ca)	135 cavans
8.	______ pesos/ca		What selling price did you get? (50 kilos/ca)	Sold 80 of the above cavans for a total of 2,500 pesos.

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.	58.7 ca/ha		What yield did you obtain? (44 kilos/ca)	$\frac{135}{2.3} = 58.7$
8.	35.51		What selling price did you get? (50 kilos/ca)	$\frac{2,500}{80 \times 44} = 71$ ¢ per kilo $71 \times 50 = 35.51$

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.

WEIGHTING

Whenever a survey is conducted on a stratified sample basis, it is usually necessary to "weight" the raw data responses after the data has been collected This is done to avoid distortion in the evaluation process when the number of responses from each stratification differs from the original sampling scheme.

For example, we might have planned a survey of rehabilitation efforts in Central Luzon Provinces stratified according to the reported flood damage, with a sample size of 360. Because of time and distance limitations, it may not have been possible to contact many of the farmers (and hectares) as originally intended in some areas, while in other areas more hectares might have been covered. To "normalize" the data, a weighting factor is developed by dividing the original area designated to be surveyed by the area actually surveyed in each instance. Thus,

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

For example,

A	B	C	D	E	F
Province	Ha Damaged	%	Stratification (Ha to be Surveyed)	Ha Actually Surveyed	Weight D/E
Bataan	2.000	4.348	16	25	.64
Bulacan	9,000	19.565	70	40	1.75
N. Ecija	9,000	19.565	70	106	.66
Pampanga	15,000	32.609	117	98	1.19
Pangasinan	3.500	7.609	27	27	1.00
Tarlac	7,000	15.217	55	69	.80
Zambales	500	1.087	4	10	.40
Total	46,000	100%	(359)* 360	375	

Thus, from this example, an adjustment must be made to the raw numbers[1] in each survey form to reflect the normalizing effect, by multiplying the Ha actually surveyed by the weight appropriate for that province. If this were not done some areas would be overrepresented and others underrepresented in the final result.

* Due to rounding off

[1] Item E - hectares actually surveyed.

GROUPING DATA

After the survey has been completed, and the forms edited, you have a mass of "ungrouped data", usually in a disorganized state. The next task then is to organize this data into meaningful groupings. Each question to be analysed must be extracted from the individual survey form, and tabulated separately with all the other responses to that question.

For example if we were attempting to determine the average palay yield in ca/ha of rainfed farmers from a sample of 50, after weighting we might have the following responses.

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

About all we could tell from this is that the yields vary. With a little searching we might also be able to identify the range. These data could be re-grouped 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

Now a pattern is beginning to emerge. The range is readily identifiable (a span of 115, from 15 to 130) and it looks as though the mean will be in the low 70's.

We could proceed with calculations at this stage, or reduce the number of items to be manipulated by summarizing them into groups. This concentration would also have the effect of highlighting the essential pattern of the total collection. For very large collections of data, grouping into "frequency distributions" is extremely helpful to avoid a lot of tedious arithmetic. Let us follow this course of action through in this example.

Number of Groups Into how many groups should a collection of data be condensed? This is largely a judgement factor.[1] Generally, the fewer the number of items, the fewer the number of groupings. A good rule of thumb is around 15 groupings, with a range from 8 groupings for about 100 items, to 25 groupings for about 1000 items. Since the objective is to reduce the amount of arithmetical manipulation, and reveal any meaningful pattern in the data, convenience, rather than mathematical precision is the dominant consideration.

In this instance, let us select 10 as the appropriate number of groupings to use.

[1] There is a formula known as "Sturges's Rule" to determine this as follows: follows:-

Number of groups = 1 + (3.3 x logarithm of "n"), where n = number of
items in the
collection

Span of Each Grouping We want the total span of all the ten groups chosen to encompass the span of the data in our collection. As a first approximation we can determine the span of each grouping as follows:-

$$\text{Span of grouping} = \frac{\text{Range of data collection}}{\text{Number of groupings}}$$

We can find the span in our sample problem as follows:-

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

$$= \frac{115}{10}$$

$$= 11.5$$

This span is called a "class interval".

As a general rule, class intervals are established in convenient numbers, either multiples of "5", or even numbers. We should round the above up to 12. If we rounded down to 10, all the data would not be within the range.

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

Mid-Point, Limits & Range of Class Interval Because we are clustering our data (in our example, from 50 to 10 groups), for further calculations we will be using the mid-point of each class interval to represent that group. Again, to avoid cumbersome arithmetic, we should try to have an easy number to manipulate -- preferably multiples of "5" (if the class interval is set at that) or even numbers.   conjunction with setting the mid-point, we must also set the limits of the class interval. Starting with the lower end of the range of our collection of data we can establish likely candidates for the lower limit of the first class interval by calculating values of A and B.

Lower Limit of 1st class interval = Lowest Number in data collection - A or B

Where
A = Span of all class Intervals minus Span of Data Collection

B = 1/2 Class Interval

The smallest number of the above should then be selected.

Since from our example, A = 120 - 115 = 5
and B = 1/2 x 12 = 6

Therefore 5 is selected and used to establish the lower limit of the 1st: class interval.

Thus lower limit of 1st class interval = 15 - 5 = 10.

We can establish any number between 10 and 15 as the lower limit of our initial class interval, bearing in mind that we want the mid-point of that 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 opt for the middle of the class interval to be 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,   or 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.

[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, <u>any</u> 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 <u>26</u> through 38, the third <u>38</u> through 50, the fourth <u>50</u> 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:-

<u>Lower and Upper Limit</u>	<u>Mid-point</u>	<u>Frequency</u>
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.

Mean

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

A Mid-point	B Frequency	C = A x B Values
20	1	20
32	2	64
44	4	176
56	8	448
68	13	884
80	10	300
92	7	644
104	4	416
116	0	0
128	1	128
	N = 50	Σ = 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.

Median

The median is the "mid-point" of the range of values in a data series. In the foregoing frequency distribution, the value between the 25th and 26th item. Since they are both 68, there is no difficulty. Otherwise, we'd have to take the mean of those two values.

PERCENTAGE FREQUENCY DISTRIBUTIONS

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

For example, in examining the production of rice farmers under the Operation Palagad Project, we wanted to compare the cavan/hectare yield of a sampling of farmers who received government assisted credit, with those who did not. The raw data was not directly comparable however until it was converted to a percentage frequency distribution. To do this, the total number of farmers in each category (181 for borrowers, 129 for non-borrowers) was used as the base. The raw data and percentage frequency distribution derived from it are shown below:-

YIELD	NUMBERS OF		PERCENTAGE OF	
Ca/Ha	Borrowers	Non-Borrowers	Borrowers	Non-Borrowers
0 - 10	13	8	7	6
11 - 20	7	7	4	5
21 - 30	9	12	5	9
31 - 40	16	11	9	9
41 - 50	16	4	9	3
51 - 60	20	13	11	10
61 - 70	26	18	14	14
71 - 80	13	19	7	15
81 - 90	18	13	10	10
91 - 100	18	6	10	5
101 - 110	11	11	6	9
111 - 120	13	4	7	3
121 - 130	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 - 10 cavans/hectare,

$$\frac{13}{131} \times 100 = 7.1823204\ \%$$

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 entailed.

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 formula to allow for the "compaction" of varying data into clusters.

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

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

Let us recall the data from page 42 on the average palay yield of rainfed farmers in ca/ha to illustrate this. You will recall from page 43 that the mean for this distribution was 71.6. To employ this mean for calculating the difference data required in the above table would entail a lot of cumbersome arithmetic. Fortunately it is not necessary. Instead,any one of the class intervals can be selected as the "origin" and the difference from this point can be measured in <u>class intervals</u>. Thus columns D, 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		Σ 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:

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

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

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

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

$$= 12 \times 1.7464$$

$$= 20.957$$

SHEPPARD'S CORRECTION FOR GROUPED DATA

In grouped, continuous frequency distributions, because of the tendency for data to cluster around the mean, the mid-points of the class intervals to the left of the mean tend to be too small, while those to the right of the mean tend to be too large. Thus, when the differences from the mean are measured, they are too great in absolute size. Furthermore, when the values are squared, the errors are not offset, but rather are compounded. Under these circumstances, the end result is a standard deviation which is larger than would otherwise have been if the data had been left ungrouped. To compensate for this effect, an adjustment of $\sqrt{1/12}$ known as Sheppard's Correction -- is subtracted in the formula thus the Standard Deviation with Sheppard's Correction "S_{corr}" is calculated as follows:-

$$S_{corr} = i\sqrt{\frac{\Sigma f(d)^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2 - \frac{1}{12}}$$

which in the foregoing example is

$$= 12 \times \sqrt{3.14 - 0.09 - 0.0833}$$

$$= 12 \times \sqrt{2.9667}$$

$$= 12 \times 1.72095$$

= 20.65 , rather than 20.957 as calculated without the correction.

BESSEL'S CORRECTION FOR SAMPLE DATA

The foregoing formulae are employed when calculating the standard deviation for a total population. However, in most situations, the frequency distribution will represent only a sample drawn from the population, rather than the total population itself. Under these circumstances it is necessary to make a further adjustment to the standard deviation calculated for the samples to obtain a best estimate of the standard deviation for the population.

This is known as Bessel's Correction and is calculated as follows:

$$SDP = \sqrt{\left(\frac{n}{n - 1}\right) S^2}$$

Where

SDP = Best Estimate of the Standard Deviation of Population

S = Standard Deviation of the Sample

n = Size of the Sample

1 = Constant, one (1)

Thus, continuing our example where

$n = 50$ and $S = 20.65$

$$SDP = \sqrt{\left(\frac{50}{50 - 1}\right) \times 20.65^2}$$

$$= \sqrt{1.02 \times 426.4225}$$

$$= \sqrt{434.95095}$$

$$= 20.856$$

COEFFICIENT OF VARIATION

The coefficient of variation (CV) is a measurement that indicates the relative variability in the data, or process being studied. By itself, the size of the standard deviation indicates how much variability there is in the data, in absolute terms. However, in some circumstances a given number may be relatively large, while in other situations a much larger unit may be relatively small. For instance, in estimating the average seed requirements for a 1/10th hectare test bed, the standard deviation might be in grams. For the same degree of precision in estimating total seed requirements for a national production program, a standard deviation of "hundreds of cavans" might be appropriate; and cavans, although much larger than grams in absolute size would be a relatively more precise measure.

The coefficient of variation (CV) enables us to compare both of these for relative precision. The CV expresses the standard deviation as a percentage of the mean thus:-

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

Judgements about the data itself can then be made, using the following table as a guide.

CV Percentage Variation	Interpretation
Less than 20%	Highly consistent, with very small variation
20 - 39%	Fairly consistent, with moderate variation
40 - 59%	Inconsistent, with medium variation
60 - 79%	Highly erratic, with high variation
80% or more	Completely unpredictable, with extreme variation.

Thus in our example where the mean is 71.6 and the standard deviation 20.856 the coefficient of variation is

$$CV = \frac{20.856}{71.6} \times 100$$

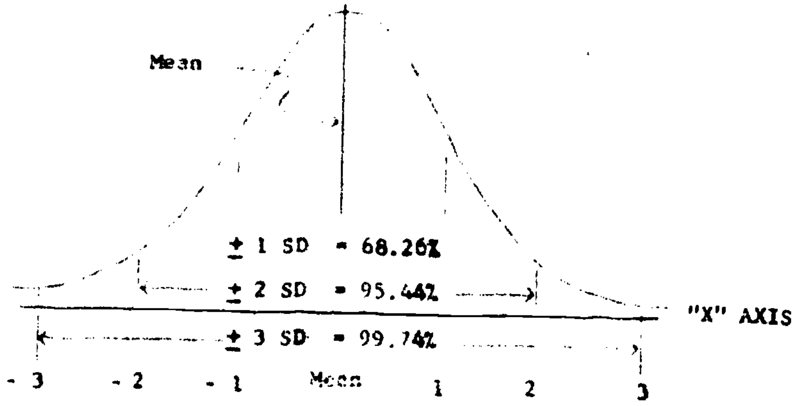
$$= .2913 \times 100$$

= 29.13% or fairly consistent, with moderate variation.

UTILIZING THE "NORMAL DISTRIBUTION CURVE"

Probability of Deviation from the Mean

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



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 our preceding problem situation where the mean of the distribution is 71.6 ca/ha and given that one standard deviation is 20.856 ca/ha.

68.26% of the farmers should obtain a harvest between

71.6 ± 20.856 = 50.744 and 92.456 ca/ha

95.44% of the farmers should obtain a harvest between

71.6 ± 41.712 = 29.888 and 113.312 ca/ha

and

99.74% of the farmers should obtain a harvest between

71.6 ± 62.568 = 9.032 and 134.168 ca/ha

Although the probabilities have been shown for ± 1,2, & 3 standard deviations, by use of the table on page 72 the probability for any range, or the range for any desired probability can be determined. This is an extremely useful feature in analyzing sample data.

Example 1 Probability for a Specified Range:

Q. Given the above mean of 71.6 and standard deviation of 20.856, what is the probability that farmers will obtain a harvest between 65 and 80 ca/ha?

To convert a data item to standard deviation units, the following formula is employed:

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

Thus 65 =

$$\text{Data Item expressed in SD Units} = \frac{65 - 71.6}{20.856}$$

$$= \frac{-6.6}{20.856}$$

$$= -0.3164556 \text{ or } {}^{-}0.32 \text{ rounded off}$$

Similarly 80 =

$$= \frac{80 - 71.6}{20.856}$$

$$= \frac{8.4}{20.856}$$

$$= +.4027617 \text{ or } +0.40 \text{ rounded off}$$

From the table 2* a standard deviation of .32 is equal to a probability of 12.55% and a standard deviation of .40 is equal to a probability of 15.54%. The specified range thus encompasses a probability of 28.09%.

Example 2 Determining the Range for a Specified Probability

Q. Given the above mean of 71.6 and Standard Deviation of 20.856 find the range within which 95% of the harvest is likely to occur.

From table 3** 95% probability occurs in the range ± 1.96 Standard Deviations from the mean.

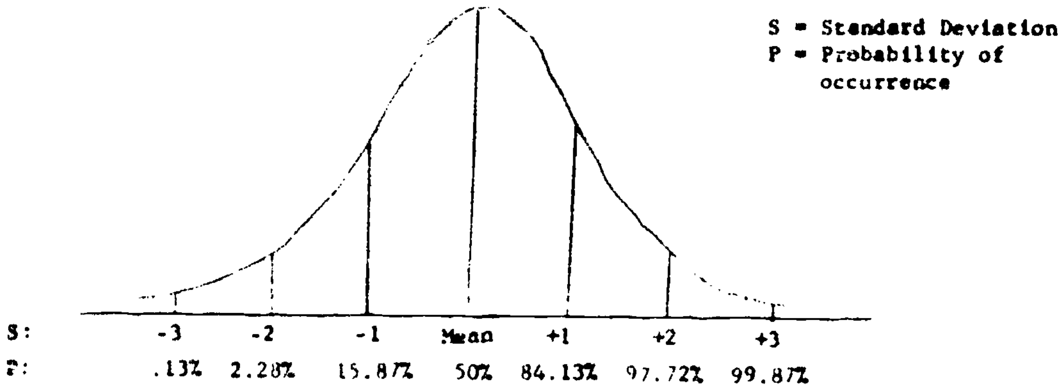
Since 20.856 ca/ha = 1 standard deviation
20.856 x 1.96 = 1.96 standard deviations
= 40.88 ca/ha

Therefore the appropriate range is
71.6 ± 40.88 = 30.72 to 112.48 ca/ha.

* page 71
** page 72

DETERMINING 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.



This is done in effect by expressing the value of the item in question in terms of its standard deviation from the mean, and then measuring the percentage of the area under the curve along the "x" axis from the extreme left of the curve to the value of the item in question.

The probabilities are shown above for several selected standard deviations, however they can be calculated for any value from - 3 standard deviations to + 3 standard deviations. See the footnote on table 2, page 71.

Thus, from our preceding problem situation, where the mean of the distribution is 71.6 ca/ha, and the standard deviation is 20.856 ca/ha, if we wished to know the probability of a farmer in this group obtaining 44 ca/ha we convert the 44 ca/ha into standard deviation units and look it up in the table, as follows:-

Data Item Expressed in SD Units $= \frac{44 - 71.6}{20.856}$

$= - \frac{27.6}{20.856}$

$= - 1.32336$ or $- 1.32$ Standard Deviations rounded off

which from table 2 is equal to 9.34 % probability. (50 - 40.66)

NON - NORMAL DISTRIBUTION

Even if a series of data is not distributed in a normal fashion, calculation 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)
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

STANDARD ERROR OF THE MEAN

Because we have been 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. Before presenting our findings to management, therefore, 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. Normal distribution theory 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,[1] 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 effect, the standard error is a standard deviation which measures the extent to which values estimated from samples differ from the true population value.

Thus in the foregoing situation, where the sample mean was 71.6, the sample size 50, and the sample standard deviation was 20.856, the standard error of the mean is thus:-

$$SEM = \sqrt{\frac{20.856^2}{50}}$$

$$= \sqrt{\frac{434.97}{50}}$$

$$= \sqrt{8.6994}$$

$$= 2.95$$

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 $\frac{2.95}{71.6} \times 100 = 4.12$ or about 4 percent

[1] "n-1" is used rather than "n" where the sample size is less than 30. If the size of the population is known, the above formula is modified as follows:

$$E = \sqrt{\frac{S^2}{N} \times \left(1 - \left(\frac{n}{N}\right)\right)}$$

Where

N = Population Size

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 ± 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 2.95, therefore we can state with a confidence of 68.26% that the true mean of the population lies between

71.6 ± 2.95, or 68.65 and 74.55 ca/ha

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" on page 72.

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

In the example, since 1 standard error of the mean = 2.95
2.81 standard errors of the mean would be 2.95 x 2.81 = ± 8.2895 ca/ha
from the sample mean of 71.6, or between 63.3105 and 79.8895

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 ca/ha, our confidence would be calculated as follows:

$\frac{1}{2.95}$ = .339 or rounded off .34 standard errors of the mean

which from the table gives us a probability of 26.62%.

STANDARD ERROR OF A PERCENTAGE

The concepts of probability are equally applicable to other measures, besides the mean. Another measure of general interest is the percentage. For instance, management might wish to know the extent to which low productivity was a problem in rainfed paddy areas.

Using the data sample on page 40 and making an assumption that 60+ ca/ha is the satisfactory cut-off point, from our sample of 50, we observe that 13 of those reported, or 13/50 = 26 percent fall in the problem area. What inference can then be drawn about the population that was sampled, from this sample information?

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[1]

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

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

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

$$= \sqrt{38.48}$$

$$= 6.2$$

To get a picture of the magnitude of the possible error, we 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%.

[1] "N-1" is used rather than "N" where the sample size is less than 30.

CONFIDENCE INTERVAL AND STANDARD ERROR OF A PERCENTAGE

Similarly, confidence associated with the sample percentage can be calculated, as it pertains to the true percentage desired by management.

Thus, 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 of the population that is unsatisfactory lies between

$$26 \pm 6.2 \quad \text{or} \quad \text{between } 19.8 \text{ and } 32.2\ \%$$

By reference to the Normal Curve and Related Probability Table on page 72 the number of standard errcrs of the percentage to utilize can be determined for any desired confidence. For example, to determine the minimum percentage unsatisfactory cases with a confidence of 99.9%, from the table 3.27 standard errors of the percentage would have to be subatracted from the sample percentage.

Since 1 standard error of a percentage = 6.2%

3.27 SEP = 6.2 x 3.27 = 20.27

or a minimum of 26 - 20.27 = 5.73 %

By the same token, it could be as much as 26 + 20.27 = 46.27 percent.

Alternately, if management wanted the answer with a range of 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, in the above situation, a range of 5 represents 2½ on each side of the sample percentage, thus

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

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

These concepts were discussed earlier on pages 10 through 27 in establishing the survey 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 merely reverse the process using the actual data drawn in the sample to determine that which we had previously guessed at.

STANDARD ERROR OF THE 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 in effect a weighted standard error, since we must take into account the fact that each of the stratified"groupings" (stratum) has its own standard error. First the mean and standard error of each stratum is calculated in the same manner as before, then the overall standard error is calculated from the following formula,

Standard Error of a Stratified Mean $= \sqrt{\dfrac{\Sigma\, SEM^2 \times P^2}{100^2}}$

Where

SEM = Standard Error of Mean of each Stratum

P = Weighted Percentage of each Stratum Population

100 = Constant 100

For example, given the following situation

A Province	B Ha Damaged	C % of Total Ha Damaged	D Stratification (Ha to be Surveyed)	E Ha Actually Surveyed	F Standard Error
Bataan	2,000	4	14	25	3.1
Bulacan	9,250	20	72	40	4.2
N. Ecija	9,250	20	72	106	3.5
Pampanga	15,000	33	119	98	2.4
Pangasinan	3,500	8	29	27	1.4
Tarlac	7,000	15	54	69	2.1
Total	46,000	100%	360	375	

$$= \sqrt{\frac{(3.1^2 \times 4^2) + (4.2^2 \times 20^2) + (3.5^2 \times 20^2) + (2.4^2 \times 33^2) + (1.4^2 \times 8^2) + (2.1^2 \times 15^2)}{100^2}}$$

$$= \sqrt{\frac{(9.61 \times 16) + (17.64 \times 400) + (12.25 \times 400) + (5.76 \times 1089) + (1.96 \times 64) + (4.41 \times 225)}{10000}}$$

$$= \sqrt{\frac{153.76 + 7056 + 4900 + 6272.64 + 125.44 + 992.25}{10000}}$$

$$= \sqrt{\frac{19500,09}{10000}} = \sqrt{1.95} = 1.396$$

or 1.4 rounded off

Note: The percentage of each stratum to be surveyed is used, not the percentage actually surveyed, otherwise some areas would be overrepresented and others underrepresented in the final result.

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 which we employ to offset the small sample size is to use "N-1" rather than "N" in the various equations, as indicated in the footnotes. A problem remains in calculating confidence estimates however. Generally, the problem with a small frequency distribution is that it 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 on page 73 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".

   Thus, 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 in the above situation, the range would be 71.6±

   2.97 x 2.145 = 6.37 or

   65.23 through 77.97.
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 ± 5, the value of T is

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

   which corresponds to a probability of 95%.

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

to adopt certain
ractices in order to improve results. For example, under the Masagana
rogram, availability and utilization of credit was seen as a major
actor which could increase farmers yields.

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

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

is is quite a complex area for analysis, and generally beyond the scope
this limited text. However, just to whet the appetite, I'd like to
ovide an example of the simplest of these correlation analysis techniques -
near relationship between two variables.

following formula can be used for this analysis:

$$\frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Where

r = coefficient of correlation
x = 1st variable values
y = 2nd variable values

above is quite a formidable looking formula, but actually it can
calculated without too much difficulty, and provides some extremely
ful guidance.

1. In effect, from a paired set of data values, a coefficient of correlation "r" is calculated. This is then compared against a scale ranging from - 1.0 to + 1.0, which is interpreted as follows:-

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

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

Percentage of Variation of Y attributable to X = $100\ r^2$

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

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.

LINEAR CORRELATION OF TWO VARIABLES

Let us illustrate the use of the above formula with an example.

Management is interested in knowing whether the availability of credit had any impact upon yields. Sample data revealed the following:

Independent Variable X Loans (Pesos) X	Dependent Variable Y Yields (ca/ha) Y	The following table is then developed to determine the values of the various elements in the formula: XY	X^2	Y^2
110	25	2750	12100	625
210	14	2940	44100	196
370	34	12580	136900	1156
420	59	24780	176400	3481
560	60	33600	313600	3600
640	43	27520	409600	1849
770	81	62370	592900	6561
850	79	67150	722500	6241
900	99	89100	810000	9801
$\Sigma X = 4830$	$\Sigma Y = 494$	$\Sigma XY = 322790$	$\Sigma X^2 = 3218100$	$\Sigma Y^2 = 33510$

N = Number of Pairs = 9

Substituting in the formula,

$$r = \frac{N \Sigma XY - \Sigma X \Sigma Y}{\sqrt{N \Sigma X^2 - (\Sigma X)^2} \sqrt{N \Sigma Y^2 - (\Sigma Y)^2}}$$

we have

$$r = \frac{(9 \times 322790) - (4830 \times 494)}{\sqrt{(9 \times 3218100) - 4830^2} \sqrt{(9 \times 33510) - 494^2}}$$

$$= \frac{2905110 - 2386030}{\sqrt{28962900 - 23328900} \times \sqrt{301590 - 244036}}$$

$$= \frac{519080}{\sqrt{5634000} \sqrt{57554}}$$

$$= \frac{519080}{2373.6 \times 239.9} = \frac{519080}{569436.5}$$

Thus $r = .912$

and $r^2 = .832$

Thus the variation in yields which can be attributed to changes in the amount of credit is 100 r^2 or

100 x $.912^2$ = 83.2 percent

and the unexplainable variation is

100 - 83.2 = 16.8 percent

Note: When "r" is based on sample data, an allowance must also be made for the fact that it is subject to sampling error.

The standard error for the correlation coefficient $= \sqrt{\frac{1 - r^2}{n - 2}}$

LINEAR RANK ORDER CORRELATION OF TWO VARIABLES

The foregoing analysis gave rise to extensive arithmetic because it compared the actual values of each data pair.

A simplified approach is to rank order each data pair and then compare the rank orders using the following modified formula[1]

$$r = 1 - \left(\frac{6\Sigma d^2}{n^3 - n}\right)$$

Where

1 = constant 1
6 = constant 6
d = difference between X and Y
n = number of pairs

Thus for the previous illustration we would have

Variable X	Rank Order X	Variable Y	Rank Order Y	Difference Between Rank Orders X and Y	Difference Squared
110	9	25	8	1	1
210	8	14	9	1	1
370	7	34	7	0	0
420	6	59	5	1	1
560	5	60	4	1	1
640	4	43	6	2	4
770	3	81	2	1	1
850	2	79	3	1	1
900	1	99	1	0	0
					$\Sigma d^2 = 10$

Substituting

$$r = 1 - \left(\frac{6 \times 10}{9^3 - 9}\right) = 1 - \left(\frac{60}{729 - 9}\right)$$

$$r = 1 - \frac{60}{720} = 1 - .083 = .917$$

and $r^2 = .841$

Thus rank ordering considerably simplifies computation. However, 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.

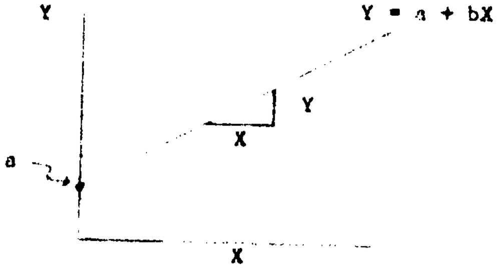
[1] Known as the Spearman Rank Order Correlation
(Note: Do not use it if you have "ties" in either of data series for example 1,2,2.5,2.5, 4 instead of 1,2,3,4)

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, management might want to determine the appropriate loan size to achieve a particular level of production, assuming a linear cause/effect relationship.

Essentially, the line of least squares is obtained by solving for two simultaneous equations with the data developed for the correlation analysis, and then substituting the values in the formula for a straight line,



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 $\frac{Y \text{ value}}{X \text{ value}}$

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

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

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

where

ΣY = sum of Y values
ΣX = sum of X values
ΣXY = sum of XY values
n = number of pairs of data
ΣX^2 = sum of X^2 values

This can be illustrated with the data from page 60, as shown on the following page.

EXAMPLE OF REGRESSION ANALYSIS

From page 60

(1) $494 = 9a + 4830\,b$ $\Sigma X = 4830$

$\Sigma Y = 494$

(2) $322790 = 4830a + 3218100b$ $\Sigma XY = 322790$

$\Sigma X^2 = 3218100$

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

(3) $32279 = 483a + 321810b$

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

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

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

(4) $26511 = 483a + 259210b$

Subtract equation (4) from equation (3)

$$\begin{array}{r} 32279 = 483a + 321810b \\ -\ \underline{26511 = 483a + 259210b} \\ 5768 = 0 \quad + 62600b \end{array}$$

Therefore $b = \dfrac{5768}{62600} = .092$

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

$494 = 9a + (4830 \times .092)$ [illegible] 444.36

transposing, $9a = 494 - 444.36$ or 49.64

therefore $a = \dfrac{49.64}{9} = 5.52$

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

$$Y = 5.52 + .092X$$

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

Where $X = 110$ $Y = 5.52 + (.092 \times 110) = 5.52 + 10.12 = 15.64$

and where $X = 900$ $Y = 5.52 + (.092 \times 900) = 5.52 + 82.8 = 88.32$

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 $Y = a + bX$ or $X = \dfrac{Y - a}{b}$

For example, to determine the appropriate loan size in order to obtain a harvest of 100 ca/ha, from the preceding data and assuming a linear relationship.

$$X = \frac{100 - 5.52}{.092} = \frac{94.48}{.092} = 1026.96$$

or approximately 1027 pesos rounded off.

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 seen as the most efficient method of gathering regular series of data - such as the Bureau of Agricultural Economics Quarterly Survey on Rice Production.

Often however, sample surveys are conducted to assist the program manager in identifying his 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 it does vary, is it worth worrying about; i.e. is it "within the ballpark"? We can improve upon the subjectivity of this question by asking "is the variation <u>statistically significant?</u>"

The size of the Standard Deviation is one 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, whereas 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, whereas carefully controlled experimental plots may produce consistently good yields; many individuals with different mental attitudes farming under varying physical conditions will 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". The statistical significance is of course based upon probability. When statistically significant data are identified in program analysis, this is an indication to management that something unusual is happening that warrants attention. If we are trying to make something unusual happen, it is good. If we are not it indicates that something is wrong for 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".

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

SIGNIFICANCE TESTING FOR A MEAN

A manager needs data to assist him in the decision making process. To meet this need, regular reports are furnished by the various operating departments, and to supplement these, sample surveys are conducted on special interest areas where it is not practical to obtain regular reporting. Periodically management should evaluate the quality of its regular reports by means of an independent sample survey. This is particularly necessary where the "operators" usually report on their own performance, but it is worth restating that rarely is "100%" reporting one hundred percent accurate, even when no vested interests are involved. There is no possibility of attaining absolute certainty even through sampling, however sampling results can be expressed in terms of probabilities. By significance testing the accuracy of the reported data can therefore be judged.

The procedure for significance testing is as follows:

1. Establish the following hypothesis, known as the "Null" hypothesis:-

   <u>There is no statistically significant difference between the sample mean and the reported mean</u>

2. Determine the criteria for significance; i.e. the minimum acceptable <u>probability</u> that the sample mean could have been drawn from a population with the reported mean.

3. Then test the Hypothesis.

   a. Calculate "Z" where

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

   b. Look up the value for "Z" in the table on page 74.

   Z indicates the probability (percentage of occurrences) that the sample mean and the reported mean could have come from the same population.

   c. **IF Z IS LOWER** than management's minimum acceptable level, **<u>THE HYPOTHESIS IS REJECTED</u>**, and we conclude **THERE <u>IS</u> A SIGNIFICANT DIFFERENCE.**

   IF Z IS EQUAL TO OR GREATER than management's minimum acceptable level, **THE HYPOTHESIS IS <u>ACCEPTED</u>** and we conclude **THERE IS <u>NO</u> SIGNIFICANT DIFFERENCE.**

   NOTE: Statistically, we cannot prove or disprove a hypothesis. We can only indicate the probability of it being as stated.

An example should clarify this.

A province reports that the average palay yield is 85 ca/ha. However, a sample survey in that province indicates that the average yield is only 78 ca/ha, and the Standard Error of the Sample Mean is calculated as 3.8.

1. Null Hypothesis - There is no statistically significant difference between 78 and 85 ca/ha.

2. Minimum acceptable probability is 5%.

   a. $Z = \frac{78 - 85}{3.8} = \frac{-7}{3.8} = -1.84$

   b. From the table on page 74

   $-1.84 = 3.29\%$

   Since Z is lower than management's minimum, the Hypothesis is rejected and we conclude there IS a significant difference.

TYPE I AND TYPE II ERRORS

By relying upon the results of significance tests in the above situation management runs the risk of making what is known as a TYPE I ERROR.

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.	TYPE I ERROR MADE

Management is too "uptight".

The risk management takes under these circumstances is to criticize the reporters unjustly, and/or look for problems in a reporting situation where none exist. The chances of making such an error can be reduced by lowering the minimum acceptable probability. For instance, in the last example there is no significant difference at the 3.29% level.

In the event that there is no significant difference indicated, and the hypothesis is accepted, management faces another risk, known as a TYPE II error.

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.	TYPE II ERROR MADE

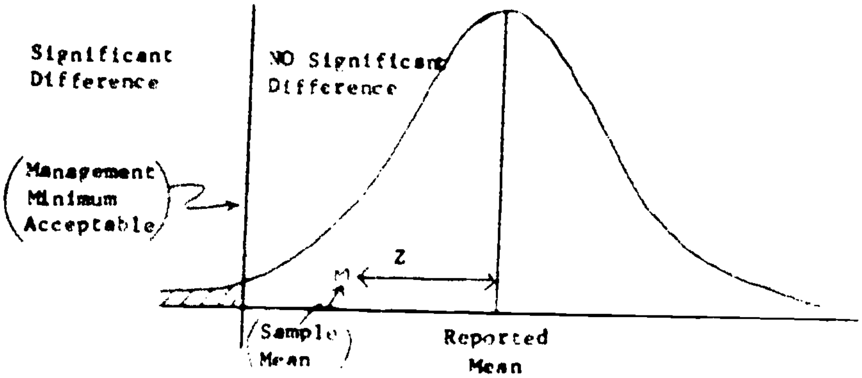
Management is "too lax".

The risk management takes under these circumstances 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 raising the minimum acceptable probability. Thus management should indicate whether it is more important to avoid Type I errors, or Type II errors, or whether both are equally as critical.

For example if management's minimum acceptable probability had been 2% in the above example, where Z = 3.29% no significant difference would have been observed.

It would not have shown up as significant until management had raised its criteria to 3.289%.

Study the sketch below to make sure you understand these concepts.



SIGNIFICANCE TESTING FOR A PERCENTAGE

Significance 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 Z test when a combination of

   $$\text{number of samples} \times \text{reported percentage}^{1} = 500 \text{ or more}$$

   For example 10 samples x 50 percent

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

2. In calculating the standard error of the sample 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. A sample survey of 25 farmers indicates however that only 60% were visited.

STEPS:

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

2. Establish the null hypothesis

   There is no statistically significant difference between the sample percentage and the reported percentage.

3. Management establishes the minimum acceptable probability 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

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

$$= \sqrt{\frac{15 \times 85}{25}} = \sqrt{\frac{1275}{25}}$$

$$= \sqrt{51} = 7.14$$

5. Calculate Z

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

   $= \frac{-25}{7.14}$

   $= -3.5$

   b. From the table on page 74

   - 3.5 = less than .139%

   Since Z is lower than management's minimum, the hypothesis is rejected and we conclude there is a significant difference.

SIGNIFICANCE TESTING -- CONCLUSION

Significance tests can be extremely useful in "quality control" of administrative program management processes, by checking regular reports against random samples. Also improvements over time can be evaluated by following up an earlier random sample and comparing the significance of the changes.

[1] Or (100 - reported percentage)

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 travelling, 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 he usually misses everything 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 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, and leave the rest to others.

MAJOR POINTS IN WRITING SURVEY REPORTS

- Avoid "technical jargon" unless you are sure that your intended reader is completely familiar with it.

- 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 instead of tables whereever 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 matrixes 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 halfed!

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 the statistical survey and analysis function for the support staff of the Philippine National Food and Agriculture Council and related agencies under the Masagana Crop Production Programs.

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 most of 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 of 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 it is all applied to everyday operations where appropriate, it should result in a significant improvement in program monitoring and management.

BLE 1

A TABLE OF RANDOM DIGITS

1.	0	5	1	2	3	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	3	9	6	9	3	5	6	0	2	0	3	7	9	1	0	3	8	6	8	4	5	9	9	1	4	0	4	1	4	6	0	1	4	1	9
3.	3	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	3	0	8	7	5	4	1	2	9	3	2	1	5	2	9	2	2	1	9	9	1	1	3	6	5	2	6	2	0	1
5.	6	4	9	3	1	2	8	1	3	0	3	3	6	2	9	7	6	9	6	2	9	3	2	8	5	1	3	7	3	6	2	7	4	7	5	3	1	2	5	4
6.	5	7	3	8	8	1	3	9	3	8	3	3	5	5	4	8	6	3	3	6	0	2	1	9	5	4	5	4	4	5	4	2	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	3	3	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	3	6	3	4	6	5	7	0	0	4	0	9	1	3	8	5	9	6	4	3	6
10.	1	2	3	7	4	3	2	1	6	9	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	3	3	5	7
11.	5	7	5	3	0	7	7	3	3	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	3	7	5	6	4	6	1	1	1	9	7	3	6	2	5	5	3	9	4	6	5	6	9	3	0	9	3	7	5	7	6	5	7	3	3	3	9	8
13.	3	9	1	3	3	3	0	3	9	3	5	3	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	3	9	7	3	1	1	3	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	3	5	9	3	4	7	4	6	2	3	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	3	1	1	9	2	4	2	0	0	0	9	1	1	0	2	3	4	3	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	3	3	3	9	9	6	3	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	3	3	7	5	1	6	2	4	3	4	3
20.	0	0	1	3	4	3	6	2	3	3	3	9	4	3	4	3	8	6	6	9	9	1	5	9	2	3	2	9	3	2	5	6	4	5	5	5	3	1	2	9
21.	9	9	3	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	0	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	3	3	3	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	3	3	3	3	3	5	6	9	9	3	2	1	1	1	3	3	1	0	6	8	3	8
25.	3	8	7	9	9	6	5	6	2	1	5	9	3	0	9	3	7	3	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	3	1	1	6	4	9	9	2	4	6	6	6	4	1	6	4	3	1	0	3	5	6	5	9	2	4	2	4	6	7
27.	7	4	6	9	4	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	3	1	7	2	4	4	3	4	6	6	0	4	3	0
29.	6	1	7	1	5	9	6	4	3	5	2	2	1	2	1	9	3	3	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	6	6	6	8	2	4	2	5	1	7	3	3	1	5	0
31.	7	2	3	5	3	4	5	7	7	5	6	3	5	9	0	3	5	6	3	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	3	7	3	0	9	6	1	0	7	6	4	4	3	3
33.	4	4	3	6	5	7	0	2	5	4	5	0	3	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	3	4	3	9	2	9	5	2	2	4	2	5	4	1	7	9	3	2	7	4	3	3	3	3	7	7	1	5	4	7	4	9	7	3
35.	7	3	1	4	3	6	5	9	1	9	1	3	6	9	9	9	1	3	4	4	1	0	6	7	6	7	0	1	7	[illegible]	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	3	3	4	4	4	6	3	3	4	3	8	3	0	5	5	0	2	9	1	0	3	7	1
37.	3	5	8	7	0	5	9	1	5	3	5	5	3	6	4	9	3	0	[illegible]	3	5	0	5	6	3	[illegible]	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	[illegible]	3	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	3	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	4	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	3	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	3	6	0	7	6	3	7	6	4	1	7	3	1	2	3	3	8	1	4	0	9	4	6	7	7	3
45.	2	9	2	1	2	4	0	3	7	3	4	1	5	3	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

Source: A Million Random Digits with 100,000 Normal Deviates, Rand Corporation; The Free Press, Glencoe, Illinois, 1956.

TABLE 2

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. To calculate cumulative probabilities see footnote below.

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	2.39	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	30.62	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	27.72	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
3.1	49.90	49.91	49.91	49.91	49.92	49.92	49.92	49.92	49.93	49.93
3.2	49.93	49.93	49.94	49.94	49.94	49.94	49.94	49.95	49.95	49.95
3.3	49.95	49.95	49.95	49.96	49.96	49.96	49.96	49.96	49.96	49.97
3.4	49.97	49.97	49.97	49.97	49.97	49.97	49.97	49.97	49.97	49.98
3.5	49.98	49.98	49.98	49.98	49.98	49.98	49.98	49.98	49 98	49.98
3.6	49.98	49.98	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99
3.7	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99
3.8	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49 99	49.99
3.9	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00

Footnote: To calculate cumulative probabilities locate the value for the standard deviation above. Then,

If the sign is + add 50. For example + 1 SD = 50 + 34.13 = 84.13%

If the sign is - subtract from 50. For example - 1 SD = 50 - 34.13 = 15.87%.

Source: Derived from Statistics for Management. B. J. Mandel, Dangary Publishing Co. Baltimore, Md. 1966, Appendix C.

TABLE 3

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 "Z" - (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	50.36	50.98
0.7	51.60	52.22	52.84	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	64.76	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	74.98	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
1.5	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	98.36	98.40	98.44	98.50	98.54	98.58	98.62	98.66	98.68	98.72
2.5	98.76	98.80	98.82	[illegible]	[illegible]	[illegible]	98.96	[illegible]	[illegible]	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
3.1	99.80	99.82	99.82	99.82	99.84	99.84	99.84	99.84	99.86	99.86
3.2	99.86	99.86	99.88	99.88	99.88	99.88	99.88	99.90	99.90	99.90
3.3	99.90	99.90	99.90	99.92	99.92	99.92	99.92	99.92	99.92	99.94
3.4	99.94	99.94	99.94	99.94	99.94	99.94	99.94	99.94	99.94	99.96
3.5	99.96	99.96	99.96	99.96	99.96	99.96	99.96	99.96	99.96	99.96
3.6	99.96	99.96	99.98	99.98	99.98	99.98	99.98	99.98	99.98	99.98
3.7	99.98	99.98	99.98	99.98	99.98	99.98	99.98	99.98	99.98	99.98
3.8	99.98	99.98	99.98	99.98	99.98	99.98	99.98	99.98	99.98	99.98
3.9	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Source. Derived from Statistics for Management, B. J. Mandel, Dangary Publishing Co. Baltimore Md, 1966, Appendix

TABLE 4

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.

Source: Derived from Fisher and Yates' Statistical Tables for Biological, Agricultural and Medical Research, Oliver and Boyd, Ltd., Edinburgh.

TABLE 5

PERCENTAGE OF ONE TAIL OF THE NORMAL CURVE AT SELECTED VALUES OF Z FROM THE ARITHMETIC 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	00.990	00.964	00.939	00.914	00.889	00.866	00.842
2.4	00.820	00.798	00.776	00.755	00.734	00.714	00.695	00.676	00.657	00.639
2.5	00.621	00.604	00.587	00.570	00.554	00.539	00.523	00.508	00.494	00.480
2.6	00.466	00.453	00.440	00.427	00.415	00.402	00.391	00.379	00.368	00.357
2.7	00.347	00.336	00.326	00.317	00.307	00.298	00.289	00.280	00.272	00.264
2.8	00.256	00.248	00.240	00.233	00.226	00.219	00.212	00.205	00.199	00.193
2.9	00.187	00.181	00.175	00.169	00.164	00.159	00.154	00.149	00.144	00.139

Source: Derived from Tables of Areas in Two Tails and in One Tail of the Normal Curve, by Frederick E. Croxton. Copyright, 1949, by Prentice Hall, Inc.

SELECTED BIBLIOGRAPHY

There is much more statistical "knowhow" than is covered by this handbook. There are also innumerable texts on the subject. In fact, judging from the quantity, one can imply that there is both an extensive felt need to disseminate and to receive statistical knowledge. Unfortunately, since mathematics is a science of concise notation, many of the experts write about statistics in the same style -- long on symbology and formulae but short on explanations. If you have a "mathematical mind" and can grasp equations and their implications readily the literature is wide open to you, and there is plenty to choose from. Otherwise you can quickly get lost -- particularly in self-study -- and become discouraged.

Three extremely useful readable books from which I personally have benefitted, and recommend to the reader who wishes to progress further, are as follows:-

B.J. Mandel, Statistics for Management, Dangary
Publishing Company, Baltimore, Maryland, 1966

M.J. Moroney, Facts from Figures, Penguin Books,
Baltimore, Maryland, 1962

D. ~~Jaff~~ Huff, How to Lie with Statistics, W. W. Norton & Co.,
New York, 1954