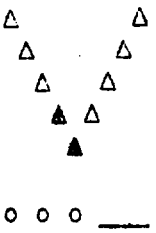
For Limited Distribution

T H E T H E O R Y O F S A M P L I N G

With Special Reference to the Collection and Interpretation
of Agricultural Statistics



Prepared by

United States Department of Agriculture
Bureau of Agricultural Economics

and

North Carolina State College of Agricultural and Engineering
Department of Experimental Statistics
Cooperating

Raleigh, North Carolina
September 22, 1942

Institute of Statistics
Mimeo. Series #1
For Limited Distribution

Table of Contents

Foreword

When textbooks in the field of "economic statistics" are subjected to a critical appraisal, it is evident that much space is given to methods of analyzing data. Methods of collecting data, on the other hand, usually receive little or no attention. This situation doubtless is the result of a misconception regarding the function of the statistician that was prevalent in the past. At one time, the main function of the statistician was to "get something out of" whatever data might be available. Collecting data was not considered a job for the statistician. In fact, the mathematical principles that make the collection of data the science it has now become, were not yet fully understood.

Scientific study of sampling techniques is a comparatively recent development in statistics. It is a subject that should now be given the prominence it deserves. The collection of the data to be used in a study is as much of a job for the statistician as the latter analysis of the data. It requires just as much thought and technical training as any of his other duties. The outline of the subject in the following pages represents an attempt to make the mathematical principles of sampling available to agricultural statisticians and economists who received their statistical education in the earlier tradition. It is intended primarily for the statistical staff of the Bureau of Agricultural Economics, U. S. Department of Agriculture. Sampling provides the basis for most of the statistical work of the Bureau, and a large part of the statistical research is directed toward the improvement of sampling techniques. During the last 4 years, this research has been augmented by a Bankhead-Jones research project designed specifically for the purpose of investigating statistical methodology in the field of agricultural statistics. The mathematical aspects of sampling, and the application of mathematical theory to practical problems, were studied under this project at the Bureau's research offices cooperating with Iowa State College and North Carolina State College. In large measure, the author of this publication has drawn upon the results of these investigations for the methods and viewpoints described in the following pages. Any merit that the publication may have is also largely due to the counsel and inspiration of W. F. Callander, formerly Head Agricultural Statistician, whose constant efforts to improve agricultural statistics are reflected in the entire statistical research program of the Bureau.

Some of the material on general statistics in the following pages may appear irrelevant to sampling work; it has been included because the development of the theory of sampling has introduced a concomitant change in the point of view from which the entire subject matter of statistics should be discussed. Although the procedures are fundamentally the same as those described in most textbooks, they are presented here with a view to relating information obtained from a sample to the population from which the sample was drawn. Textbooks generally tend to focus so much attention on the analysis of the sample that the distinction between the sample and population is sometimes overlooked.

A study of the theory of sampling should not be approached under the impression that sampling theory is a specialized branch of statistics. Sampling theory is statistics and, conversely, statistics is the theory of sampling. The only justification for choosing "Theory of Sampling" as a title

lies in the fact that most texts on so-called general statistics are not written from the viewpoint of the man whose chief concern is the collection of the primary data for an investigation. A departure from that custom seems sufficiently radical to warrant a title of its own.

It is evident, therefore, that under an all-inclusive concept of the theory of sampling as defined above, the subject matter under discussion should be no different from that usually discussed in books on general statistical theory. The only difference lies in the point of view from which that subject matter is discussed. In recent years the theory of experimental design has received considerable attention; some statisticians are under the impression that this subject is of interest only to agronomists and similar research workers, but this is not true. In its broader aspects, the subject of experimental design is a good illustration of the presentation of statistical theory from the viewpoint of the man engaged in the collection of primary data. There is no fundamental difference between the design of a well-planned experiment and the design of a well-organized sampling scheme to be used in a sample-census enumeration or similar undertaking. The particular designs that are used may differ from one another, depending upon the nature of the investigation, but the mathematical principles are identical.

As the subject of sampling theory embraces the entire field of statistical methodology, it seems evident that a well-rounded training in general statistics is a necessary prerequisite to an understanding of that theory. The present work should serve as an abridged text on general statistical methodology from the standpoint of the statistician charged with the responsibility for collecting data as efficiently, accurately, and economically as possible. His work is not spectacular and may not always be fully appreciated by those who later use the data he has assembled, but the technical training required for the adequate discharge of his duties is as extensive as that required by those who later use his results, although the latter group of workers may not always be aware of it. The picture of the collector of primary data as an unimaginative drudge engaged in the dull routine of assembling figures to be analyzed by a superior order of beings is decidedly out of date.

Walter A. Hendricks

Agricultural Statistician
Bureau of Agricultural Economics
U.S. Department of Agriculture
and
Resident Collaborator
Department of Experimental Statistics
North Carolina State College

Some Principles of Notation

Before the study of statistics is begun from any viewpoint, it is important that the student understand the symbolism or the language of the subject. It is assumed that he is already familiar with the terminology, symbolism, and procedures of elementary algebra. Statistical literature contains some symbols that are usually not mentioned in books on elementary mathematics.

At this point, it is desirable to discuss the subscript notation that is frequently encountered in statistical formulas. This system of notation can be explained most clearly by referring to a specific example. Suppose one wishes to discuss the acreages of wheat in 3 counties and wants to express these acreages by algebraic symbols. He could represent the acreage in the first county by a, that in the second by b, and that in the third by c. This would be entirely adequate for his purpose. Suppose, however, that, instead of only 3, he was dealing with 100 or more counties. The simple method of using a different letter to represent the acreage in each county can no longer be employed conveniently. There are only 26 letters in the alphabet. He could use capital letters or alphabets from foreign languages to increase the number of available symbols as required, but this would be awkward.

The subscript notation provides a solution. Instead of using a different letter for each county, he can use the same letter for each county and distinguish one county from another by attaching subscripts to that letter. For example, he can let a_1 represent the wheat acreage in the first county, a_2 that in the second, a_3 that in the third, and so on. The difficulties that might be encountered with the first system of notation are thus avoided because the possibilities for including any given number of counties are unlimited under the new system.

When this system is used by a statistician, he usually says that he is representing the wheat acreage in any given county by a_i where i assumes the values 1, 2, 3, and so on. If he has a definite number of counties in mind, such as 92, he would condense his definition into the following form,

"Let a_i; i = 1, 2, 3, ---, 92, represent
the wheat acreages in 92 counties,"

which means that he is using the letter a to represent the wheat acreage in a county and is attaching subscripts from 1 to 92 to that letter in order to distinguish one county from another. If he wishes to specify any fixed number of county wheat acreages regardless of what that number might be, he would write,

"Let a_i; i = 1, 2, 3, ---, k, represent
the wheat acreages in k counties,"

which means that he is talking about a definite number of counties but does not care, or need, to tell what that number is. This method of representation is convenient to use and simplifies the problem of algebraic notation

considerably. The symbol a_i is sometimes called the general term of the series a_1, a_2, a_3, ---, a_k because each term in the series can be derived from a_i by letting i assume a particular value.

This system can be extended to problems requiring more detailed notation. Suppose that, in addition to identifying the wheat acreage in each of a number of counties, one also wishes to distinguish the wheat acreages in particular townships. He can let the letter a represent wheat acreage, as before, and attach 2 subscripts to that letter. One subscript specifies the county and the other specifies the township. The symbol a_{32}, for example, can be used to specify the wheat acreage in the seconf township of the third county. The general term of a series of such symbols can be represented by a_{ij} where i can be assigned any number to specify a county and j can be assigned any number to specify a township to that county.

Exercise 1. - Wheat acreages were measured for 6 townships in one county, 8 townships in a second county, and 2 townships in a third county. Put the proper subscripts in the symbol a_{ij} to specify each of these 16 townships and give the meaning of each resulting expression in words.

This general method of notation should be used in complicated problems. There is no need to use it in simple problems where it is easier to use different letters of the alphabet to make necessary distinctions. When a simple system of notation will meet all requirements of a particular problem, the use of subscripts introduces unnecessary complications. It is always desirable to choose a system of notation that will present results and formulas in the clearest and least cumbersome fashion.

Summation Signs

The algebraic expression representing the sum of a series of numbers, such as the sum of the wheat acreages in 92 counties, might be represented by the expression,

$$a_1 + a_2 + a_3 + \text{---} + a_{92},$$

where a_1 is the acreage in the first county, a_2 that in the second, and so on. Mathematicians usually like to abbreviate this expression into the following form, which means the same thing and is easier to write:

$$\sum_{i=1}^{92} (a_i)$$

The Greek letter, capital sigma, is called a summation sign. The summation sign is a special case of what mathematicians call symbolic operators, because it is a symbol that stands for an operation to be performed on the

numbers that follow it. In this case the numbers that follow the summation sign are represented by the general term a_i. The individual numbers that must be added are specified by the values that the subscript i assumes. These values are indicated by the numbers written below and above the summation sign and show that i is to take the values 1 to 92 inclusive. In terms of mathematical symbols, these concepts can be expressed by the following equation that is equivalent to the definition just given:

$$\sum_{i=1}^{92}(a_i) = a_1 + a_2 + a_3 + \text{---} + a_{92} \quad \text{- - - -} \quad (1)$$

At the present time, some statisticians use the letter S instead of Σ as a summation sign. This substitution is made because the Greek alphabet is not included on most typewriters or in sets of printer's type. Since the letter S serves the same purpose, it is usually more economical for the statistician to confine himself to the English alphabet in his publications so that the purchase of additional type will not be necessary. In the present publication it seems desirable to follow this notation rather than the more classical notation that is still widely used in publications on pure mathematics. The preceding equation will thus be written in this form:

$$S_{i=1}^{92}(a_i) = a_1 + a_2 + a_3 + \text{---} + a_{92} \quad \text{- - - -} \quad (2)$$

If a statistician wishes to write an expression for the sum of a series of numbers without actually specifying how many he has in mind, he can do this by substituting a letter for the number appearing above the summation sign as follows:

$$S_{i=1}^{k}(a_i) = a_1 + a_2 + a_3 + \text{---} + a_k \quad \text{- - - -} \quad (3)$$

If, as is often the case, the text of the statistician's manuscript leaves no doubt in regard to the numbers that are to be added, it is not necessary to put so much detail into the algebraic expression representing the relationship. One might simply abbreviate equations like the preceding to the form.

$$S(a_i) = a_1 + a_2 + a_3 + \text{---} + a_k \quad \text{- - - - -} \quad (4)$$

or to the still more simple form,

$$S(a) = a_1 + a_2 + a_3 + \text{---} + a_k \quad \text{- - - - -} \quad (5)$$

Such abbreviations are common in statistical publications but should not be used if they are likely to be misunderstood. Unless the accompanying discussion is perfectly clear in regard to what is intended, it is preferable to avoid such shortcuts. The statistician should have his audience clearly in mind, so he can be sure of using a system of notation that can be followed by the persons for whose benefit he is writing. Some classes of readers will require more detailed explanation than others.

When the double-subscript notation is used, the summation extends to both subscripts. In such problems, it is customary to use two summation signs. The sum of the 16 wheat acreages, given in Exercise 1, would be written,

$$SS(a_{ij}) = a_{11} + a_{12} + a_{13} + a_{14} + a_{15} + a_{16} + a_{21} + a_{22} + a_{23} + a_{24} + a_{25} + a_{26} + a_{27} + a_{28} + a_{31} + a_{32} \qquad (6)$$

One summation sign indicates that the township acreages should be added for each county. The second indicates that the totals for each county should be added. In such problems, it is not necessary to specify the order in which the additions are performed because the grand total will always have the same value no matter which subtotals are computed as intermediate steps in the process.

Exercise 2. - Wheat acreages were measured in 5 townships for each of 6 counties. Compute the value of $SS(a_{ij})$ by the following methods:

(a) First write the expression for the 6 county totals and add the results.

(b) Write the expressions for the 5 sums obtained by adding the data for township #1, township #2, township #3, township #4, and township #5, one at a time for all counties. Add these 5 totals and show that the grand total is equal to that given by method (a).

Averages

Use of an average as a method of representing a set of numbers by a single number for purposes of summarization is perhaps one of the oldest devices of statistics. The concepts underlying the use of an average are of fundamental importance in the theory of sampling, and they involve more careful thinking on the part of the statistician than is commonly supposed, as the following discussion will indicate.

Some statisticians regard an average as a number that is "most typical" of an entire set of numbers. This most-typical-number concept does not imply that the average has to be one of the numbers in the set. The average height of a group of men might be 67 inches and this might be regarded as the typical height for the group even though no individual in that group is exactly 67 inches tall. The most-typical-number concept thus provides an early introduction to an important feature of statistics, namely, that the individuals in a particular group are of interest, not so much in themselves, but for the information they can be made to yield about the general character of groups of that kind.

As the most typical number of a set was regarded by most statisticians as some number about halfway between the smallest and the largest, averages were given the name of measures of central tendency. This terminology does not

seem so expressive as most-typical-number, but it has the advantage of directing attention to the matter of frequency distributions, which is undoubtedly the reason why it was adopted.

Experimenters in many fields of work, who were engaged in taking large numbers of measurements, soon noticed that measurements generally grouped themselves into bell-shaped frequency distributions. Extremely small or extremely large measurements occurred only rarely, but measurements near the center of the range were rather common. Therefore, it was concluded that most measurements had a tendency to concentrate about the average and, conversely, that the average should be regarded as the point about which the measurements tended to concentrate. For an illustration based on agricultural data, the student should refer to figure 1, which gives the frequency distribution of the yield per acre of cotton for the 75 counties of Arkansas in 1939. The data were computed from figures reported by the 1940 census, the yield per acre for each county being taken as an individual observation.

Thus far the general nature of an average has been discussed without describing how an average is to be computed. An average may be defined as a "most typical number" or as a "measure of central tendency" for purposes of general discussion, but, from the viewpoint of mathematical analysis, a more specific definition is required. Textbooks usually list several types of averages that may be used under different conditions. A few of these are here discussed in detail.

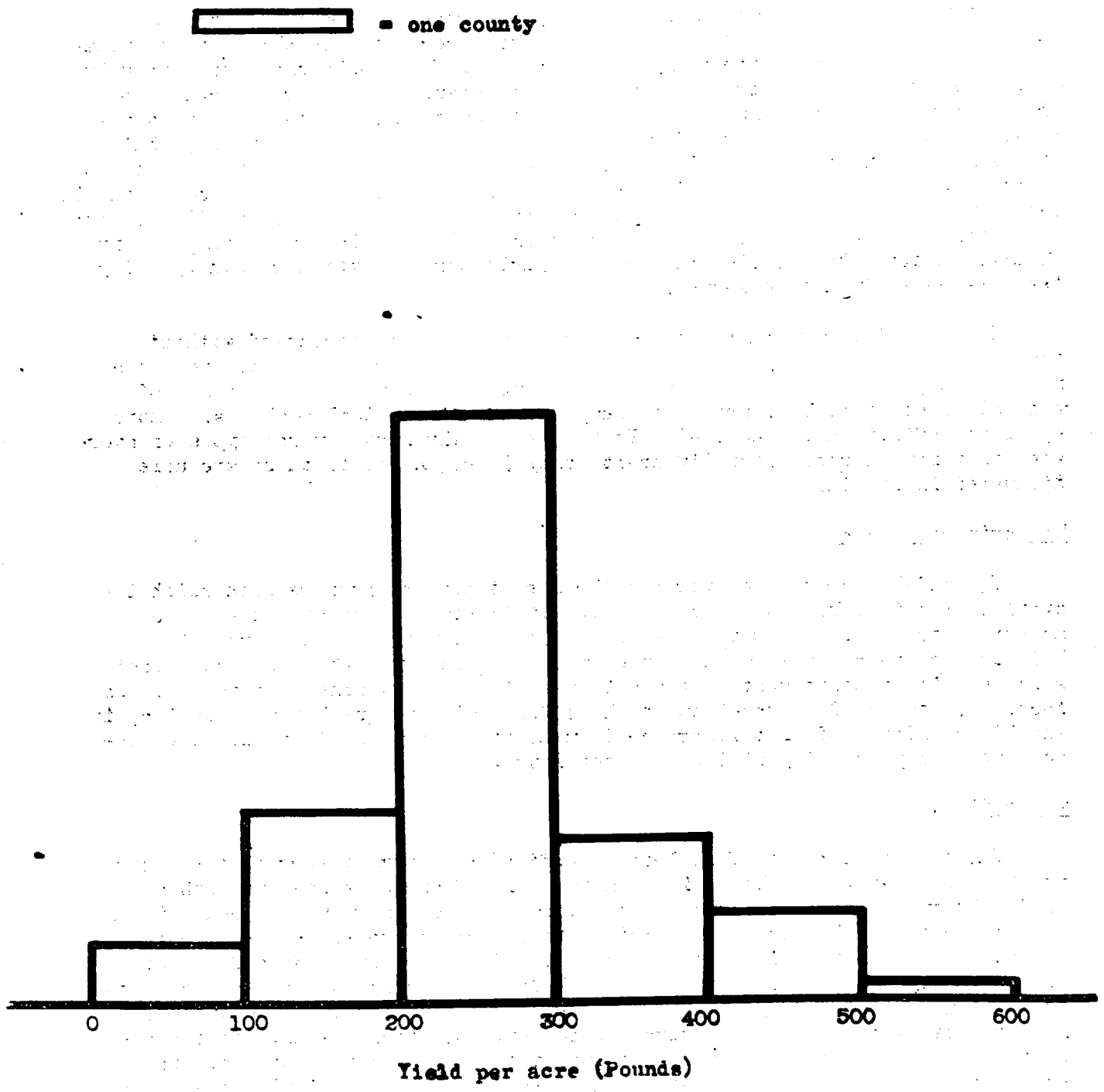
The arithmetic mean.

The arithmetic mean is perhaps the most important average with which the statistician is likely to be concerned. It is computed by adding all the measurements and dividing the result by the total number of measurements. This is the average that will be used most frequently in general statistical work and will receive most of the attention in the following discussions. In fact, when the word average appears henceforth without further elaboration, it should be understood that reference is made to the arithmetic mean. The reasons for its popularity will be evident later.

The median.

The median of a set of numbers is defined as a number such that as many numbers of the set fall below it as above it. If one were dealing with a perfectly symmetrical frequency distribution, the median would be equal to the arithmetic mean. The use of the median in preference to the arithmetic mean is usually recommended in cases where a set of numbers includes a few that differ widely from the majority. In such data the median often gives a better indication of the most typical number or central tendency than the arithmetic mean because less emphasis is given to the extreme observations. This argument undoubtedly has some merit, but its importance is sometimes exaggerated. As a matter of fact, the concept of an average as a most typical number or as a measure of central tendency is not particularly suitable for sampling work. A more useful concept is discussed later.

Figure 1. Frequency distribution of yield per acre of cotton in Arkansas
(County data based on U.S. Census, 1940)



The mode.

The mode of a set of numbers is defined as the number that occurs most frequently. If one were dealing with a perfectly symmetrical frequency distribution, the mode would coincide with the arithmetic mean and the median. It can be regarded as another measure of "central tendendy" for frequency distributions having a heavy concentration of observations near the center of the range, but it is not very satisfactory from that point of view because of its lack of stability from sample to sample.

The geometric mean.

The geometric mean of n numbers is defined as the n-th root of their product. It has been used as a measure of central tendency for frequency distributions that are not symmetrical, but it has many other applications. In modern statistics, this average appears in many mathematical operations that have no connection with measures of central tendency.

The harmonic mean.

The harmonic mean of a set of numbers is defined as the reciprocal of the arithmetic mean of the reciprocals of the numbers. This average, like the geometric mean, has other and more important applications than the measurement of central tendency. The operations performed in its computation are frequently encountered in statistics, and it is convenient to refer to the result as the harmonic mean even though nothing like the measurement of central tendency is involved.

The above-mentioned averages, and the list is by no means complete, are usually presented as measures of central tendency by many writers. The conditions under which each should be used are discussed at length in many books on statistics. Such discussions represent a viewpoint that has little bearing on sampling work and is probably out of date. The arithmetic mean is generally used in cases in which the worker is actually interested in the concepts underlying an average. This is the only statistical constant that will be called an average hereafter in the present work. In view of the importance of the arithmetic mean, the present discussion of averages will be continued with the understanding that the work "average" will henceforth apply only to the arithmetic mean.

Although the average was originally regarded as a "most typical number" or "measure of contral tendency," the trend of modern statistics has been to adopt the terminology and concepts of the theory of errors. In the study of statistics, the student is soom confronted by a diversity of concepts and nomenclature, even about such a comparatively simple topic of discussion as the average. The reason for this situation lies in the fact that the subject matter of statistics has such an extremely heterogeneous background. The subject matter of statistics, as it stands today, has evolved from the practice and researches of French social scientists and gamblers; German physicists, astronometers, and engineers; English biologists and agricultural workers; and mathematicians, economists, and philosopers of all nationalities. All have contributed more or less independently to a pool of statistical notation and technique.

This sort of background could not be expected to lead to a single universally accepted set of concepts and terminology, although the differences in fundamental concepts from one field of application to another are not so great as one might suppose. Most differences that are found are nothing more than differences in terminology and notation. Certain concepts necessarily require greater emphasis at the expense of others in any one field of application, however, and any attempt at standardization of terminology would meet some opposition. On the other hand, the concepts, terminology, and notation of the theory of errors are so general and so well adapted to all fields of application that they are coming into general use. They are particularly applicable to sampling work; in fact, sampling work can hardly be discussed satisfactorily in any other terms.

In terms of error theory, a single measurement is an estimate of the true value of the quantity that was measured. This estimate may be, and probably is, somewhat in error. Experience shows that large errors in either direction occur only rarely, whereas small errors occur often; the frequency of occurrence of small errors in either direction increases as the absolute size of those errors decreases. Furthermore, positive errors of a given size tend to occur as often as negative errors of the same size, so that the average of all errors tends to be equal to zero. For example, if a bale of cotton were weighed on the same scale by each of many men, the results would not agree exactly. Most of the errors would be fairly small, but a few errors could be expected in both directions. The errors would tend to counterbalance each other so the average of all weights would tend to be the correct weight of the bale.

As errors in the neighborhood of zero tend to occur more frequently than any other, a single measurement is more likely to be equal to the true value of the quantity measured than to any other single possible value. In other words, one can <u>expect</u> an individual measurement to equal to, or at least close to, the true value of the quantity measured. As the true value is also the average of all possible repeatedly observed measurements, this average may be called the <u>expected value</u> of the quantity measured. This concept is of utmost importance in sampling theory and is developed more fully in the following section. For the present, the reader should accustom himself to thinking of data in terms of error theory. From this point of view, for example, the various yields of cotton shown in figure 1 should be regarded as measurements of the average or expected value for the State as a whole. Deviations from the expected value should be regarded as errors of measurement. This fundamental principle must be clearly understood before the mathematical theory of sampling can assume a concrete meaning. It is important to notice that the deviation of a quantity from the arithmetic mean is regarded as an error of measurement mainly for purposes of terminology. In the theory of errors the deviation of a measurement from its expected value actually is an error of measurement. When the terminology of error theory is applied to general sampling problems, it is convenient to regard deviations from means as errors of measurement because such deviations have properties analogous to errors of measurement. For example, when the deviation of a county cotton yield from the State average is called an error of measurement, this does not imply that the cotton yield for the county was determined inaccurately. Calling such a deviation an error of measurement refers only to the accuracy with which the cotton yield for that one county represents the average yield for the entire State. In that sense, the deviation is an error of measurement even though the yield for the county was determined accurately.

Exercise 3. - The following table gives the data from which figure 1 was constructed except that each measurement within a class interval is assumed to have a value equal to the midpoint of the class interval.

Yield per acre of cotton in Arkansas by counties, 1939

Yield per acre (Pounds)	Number of counties
50	4
150	13
250	40
350	11
450	6
550	1
	75

(a) Compute the arithmetic mean.

(b) What is the approximate value of the mode? Considering the fact that you are dealing with grouped data, would you expect the value of the mode to be less than 250 or greater than 250? Why?

(c) In order to find the median, it is convenient to work with cumulative frequencies. The cumulative frequencies derived from the above table are as follows:

Yield per acre (Pounds)	Number of counties (Cumulative)
0	0
100	4
200	17
300	57
400	68
500	74
600	75

Plot this cumulative frequency curve on graph paper and find the median graphically by estimating the yield for which the cumulative frequency has the value 37.5. What kind of inaccuracies are present in locating the median by this method?

Frequency Distributions

The subject of frequency distributions was introduced in connection with the discussion of averages in the preceding section, but the importance of the various kinds of frequency distributions is so great that considerably more space must be devoted to them.

Scientific analysis of frequency distributions dates back to the earliest investigations in the theory of errors. The errors in physical measurements were observed to form symmetrical frequency distributions with an expected value of zero. In such measurements one could reasonably expect errors in the two directions to counterbalance each other. The frequency of occurrence of a measurement of a given size was found to decrease as the departure of that observed measurement from the expected value increased. Once this fact was empirically established, mathematicians began a search for a mathematical equation that would describe the relationship between the size of an error and its frequency of occurrence. The result was what is now known as the equation of the Normal Frequency Curve which is usually written in the following form:

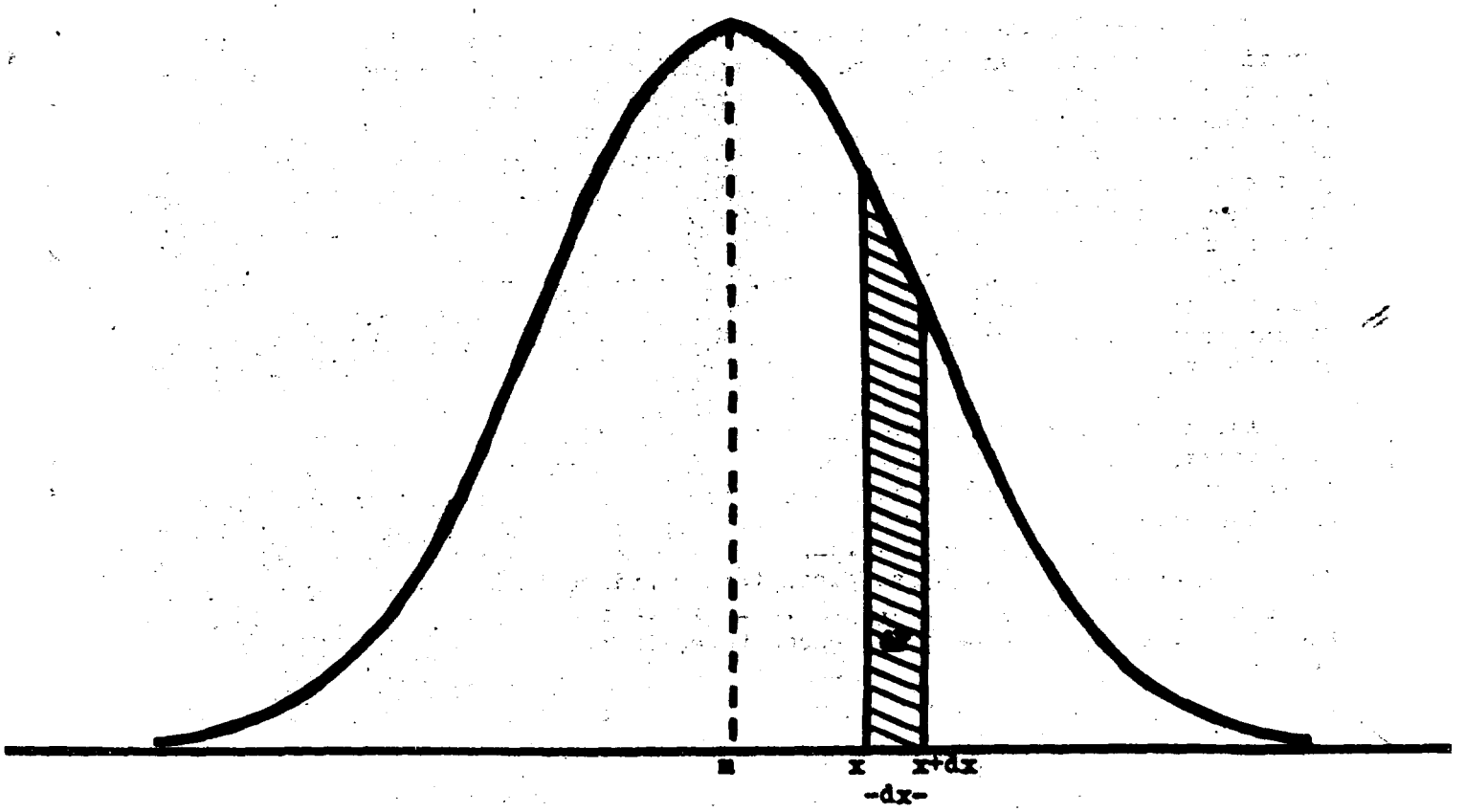
$$dF = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-m)^2} dx \qquad \text{- - - - - - (7)}$$

To understand this equation, it is necessary to develop the point of view from which the mathematical analysis of frequency distributions is conducted. It is impossible to discuss the mathematics involved without a working knowledge of the calculus, but the fundamental principles are simple and can be discussed in nonmathematical language.

The first important point to bear in mind is that mathematicians prefer to express frequencies in terms of areas. In figure 1, for example, the number of counties whose cotton yields fell within any one class interval is represented by the area of the rectangle having that class interval as a base. This is why the chart was drawn in that particular way. The vertical scale is not shown on that chart, but if it were, it would have to be graduated in such units that the sum of the areas of all six rectangles would be equal to the total number of counties in the State, which is 75. Any mathematical equation that is used to represent a frequency distribution must embody the same idea. The number of measurements in a particular class interval must be represented by an area. Equation (7) satisfies this condition and the way in which it is accomplished is shown graphically in figure 2.

The smooth bell-shaped curve in figure 2 is the graph of the expression $\frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-m)^2}$. It should be noted that the highest point on that curve is reached when x is equal to m. The shaded area under the curve represents the number of measurements in a class interval of length dx, starting at the point, x. If dx is small, that area is approximately equal to the area of a rectangle whose base is dx and whose altitude is the ordinate of the curve at the point x. Equation (7) thus represents the area of a rectangle whose altitude

Figure 2. Normal frequency distribution



is the ordinate of the curve at the point, x, and whose base is dx, where dx is a small interval. The total number of measurements is represented by the total area under the curve. This area is the limit approached by the sum of a large number of adjacent elements of area, each with a base equal to dx, as dx approaches zero.

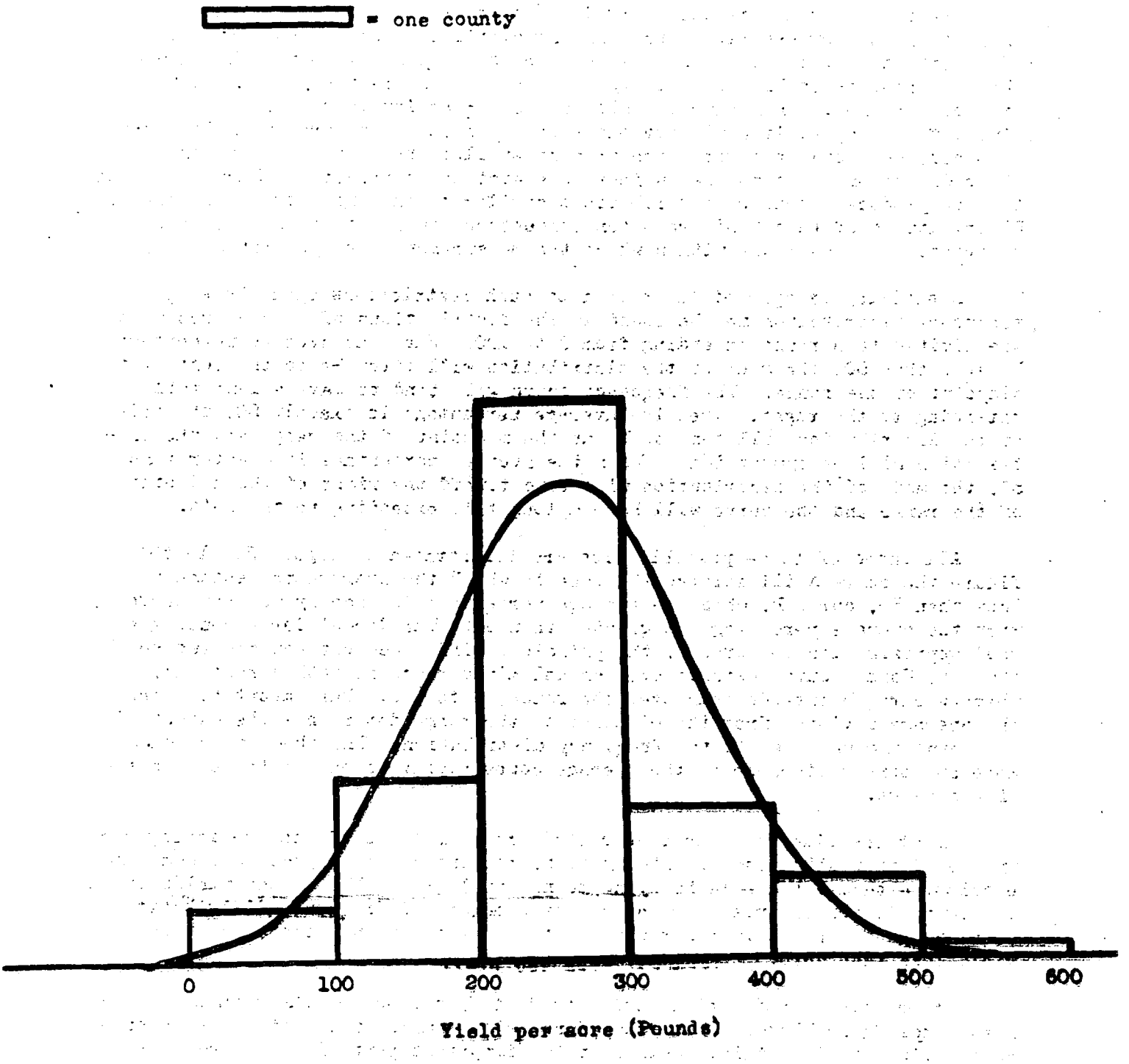
This method of representing a frequency distribution by a mathematical equation may seem awkward, but it has many advantages from the mathematician's point of view. These are not discussed here. All that is required for present purposes is an understanding of the fundamental principle underlying this kind of analysis, namely, that equations such as equation (7) represent the number of measurements, dF, that fall in the class interval bounded by x and x + dx, where dx is a number that can be made as small as desired, and x can assume any value in the range. m represents the average or expected value of the measurements, σ is a constant whose value determines how closely the measurements tend to cluster about the expected value, and N represents the total number of measurements. It should be observed that the curve is symmetrical, that is, positive deviations of measurements from the expected value occur with the same frequency as negative deviations of the same size.

The practical statistician must never forget that equation (7) is only an empirical equation that was developed to fit the kind of frequency distribution usually found by physicists and astronomers when repeated measurements were made upon a fixed quantity. There is no fundamental law from which one could deduce the fact that errors of measurement should be distributed in that particular fashion. At one time statisticians regarded the equation as a law of nature for which there must be some explanation. No such explanation was ever found, but the prominence given to the equation by the early writers on the subject is still hard to overcome. Someone once remarked that everybody believes in the Normal Law: the experimenters because they think it was proved by mathematics, and the mathematicians because they think it was established experimentally.

All that can actually be said for the equation is that it gives a good approximation to many observed frequency distributions. The esteem in which it was held by the early workers has resulted in establishing the equation as one with which everyone is now familiar, too often, unfortunately, to the exclusion of all others. Much of this popularity will doubtless be permanent. The comparative simplicity of the equation makes it peculiarly well adapted to the complex mathematical treatment used in sampling theory. In addition, the equation fits many observed frequency distributions sufficiently well to make it fairly useful in practical work. The distribution of county yields of cotton, shown in figure 1 for example, is not perfectly symmetrical, yet the Normal Curve fits it fairly well, as shown in figure 3.

For the many observed frequency distributions that the Normal Curve will not fit, a different mathematical expression must be invoked. For example, consider the frequency distribution of the 75 county cotton acreages in Arkansas for 1939, shown in figure 4. This distribution bears little resemblence to the distribution of yields (fig. 1) and the Normal Curve would not fit it. This illustration should be sufficient to convince the student that the Normal Curve is not universally applicable. It is only one of a large number of types of distributions that are encountered in practice. Mathematical

Figure 3. Normal curve fitted to distribution of county cotton yields in Arkansas



equations that fit these distributions have been available for some time, but many statisticians have failed to make much use of them. Most textbooks on elementary statistics ignore the subject and tend to create the impression that the Normal Curve represents a sort of universal law.

The reader should realize that there is no universal law governing the shape of frequency distributions. In any statistical study the particular kind of distribution which is at hand should be borne in mind. Often the nature of the measurements enables one to predict the kind of frequency distribution that will be obtained. When a very small physical quantity is measured a number of times, for example, it would obviously be impossible for large negative errors to occur; one could not get a measurement smaller than zero. But errors in the other direction would not be subject to a similar restriction. Such situations tend to produce frequency distributions similar to the one shown in figure 3. Distributions of this kind are often encountered in practice because of some restriction in the range within which the measurements are permitted to occur.

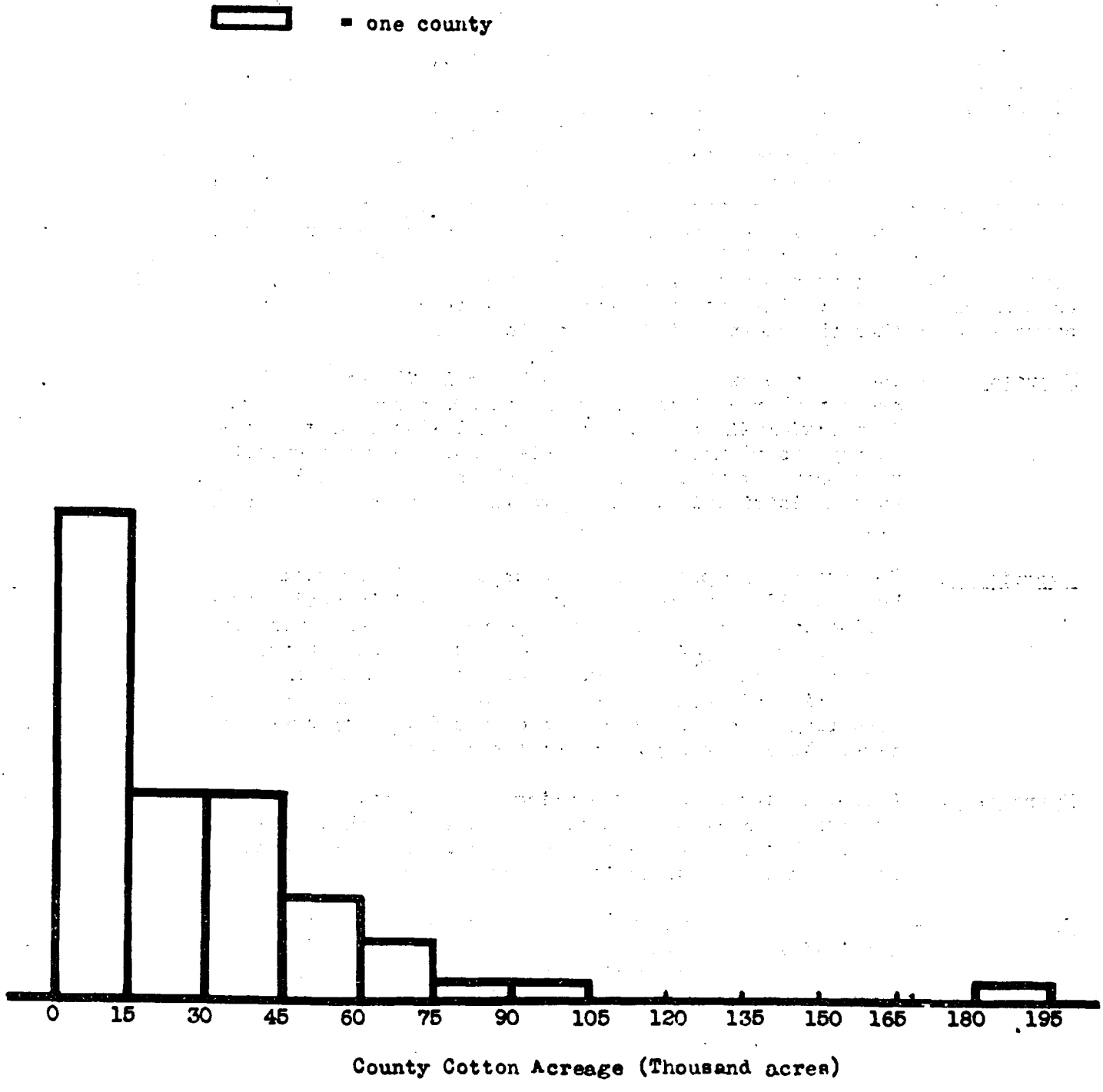
A striking example of the effect of such restrictions upon the shape of a frequency distribution may be found in the distributions of percentages which are limited to a range extending from 0 to 100. When the average percentage is less than 50, the mode of the distribution will often be to the left of the midpoint of the range. The frequency curve will tend to have a long tail extending to the right. When the average percentage is exactly 50, the mode of the distribution will tend to be at the midpoint of the range and the distribution will be symmetrical. When the average percentage is greater than 50, the mode of the distribution will tend toward the right of the midpoint of the range and the curve will have a long tail extending to the left.

All three of these possibilities are illustrated in figure 5. In this figure the curve A illustrates the case in which the average percentage is less than 50, curve B, when the average percentage is exactly 50, and curve C, when the average percentage is greater than 50. The Normal Curve would give a good approximation to curve B, the principle difference between the two being that the Normal Curve extends over an unlimited range in both directions, whereas curve B extends only over the range, 0 to 100. The amount by which the average deviates from the midpoint of the range determines the extent of the resulting skewness in the frequency distribution. The skewness becomes more and more noticeable as the average comes closer to one of the extremities of the range.

Relations like those just described are a useful guide in predicting the type of distribution that is likely to be encountered in a practical sampling problem, although the rule is by no means infallible. The cotton yields shown in figure 1, for example, form an almost symmetrical frequency distribution even though the range is restricted to the extent that no county can have a yield less than zero.

At one time statisticians tried to justify the use of the Normal Curve in most sampling problems. The Normal Curve will usually fit most types of frequency distributions fairly well over the important part of the range that includes the bulk of the observations. In recent years there has been a tendency to make greater use of the exact mathematical curves which are appropriate to the particular problem at hand. Such curves will fit the observed

Figure 4. Frequency distribution of county cotton acreages in Arkansas (U. S. Census, 1940)



distributions over the entire range. Although long available, such curves have not been widely used. The mathematical difficulties encountered in using such equations are usually greater than those encountered with the Normal Curve, but considerable progress has been made in overcoming these difficulties.

Much remains to be done in this field, however, and it is probable that statisticians will be inclined to use the Normal Curve when there is the slightest justification or excuse for doing so. For some purposes the use of the Normal Curve probably leads to no serious error. This point will be discussed later. The important thing to bear in mind at this stage is that there are many types of frequency distributions and that the Normal Curve is only one of many that are met in practice. The Normal Curve is better known than the others mainly because it has been given more publicity and more intensive study. The reader should not be misled into thinking that the prestige it enjoys represents any justification for using it in preference to all others. When it is used in preference to other possible curves, the reason is usually a matter of convenience rather than deep-seated mathematical theory.

Exercise 4. - The average size of farm in a State is 70 acres. There are a number of farms larger than 300 acres, the largest farm having 423 acres. Make a rough sketch of the frequency distribution of farm size that you would expect to obtain for that State and explain why you would expect such a distribution. Where would you expect the mode to be?

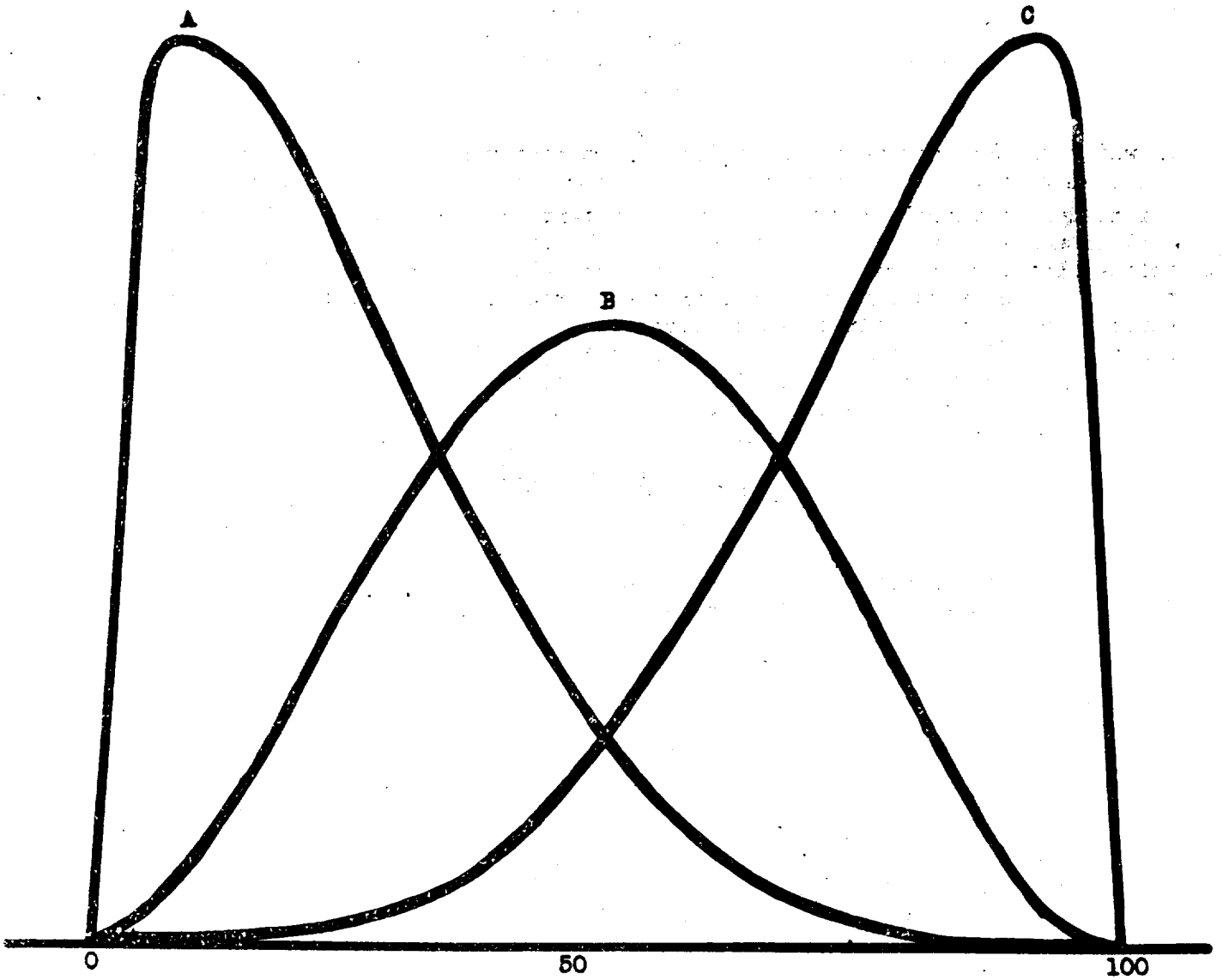
Exercise 5. - Records of egg production are kept on a large flock of hens for 10 days and the number of eggs laid during this period is recorded. The average number of eggs per hen for the 10-day period was found to be 8. Sketch the frequency distribution you would expect to get if the egg production of each hen were tabulated separately and the resulting data were used to plot the frequency distribution. Where would you expect the mode to be?

Exercise 6. - A thermometer used in measuring a large number of temperatures was tested and found to read 1.5 degrees too high. What effect would this error have on the resulting frequency distribution of temperatures?

Exercise 7. - A number of corn yields were multiplied by the same correction factor to reduce them to a moisture-free basis. What effect would this correction have on the shape of the frequency distribution and on the average yield?

Figure 5. Frequency distributions of percentages, showing effect of restricted range

A: Average less than 50 percent
B: Average equal to 50 percent
C: Average greater than 50 percent



Measurement of Variation

The preceding discussion of frequency curves has acquainted the reader with the kind of variability that is encountered in observed data. From the standpoint of sampling work, this variability gives an indication of the reliability of any one measurement as an estimate of the true or expected value of the quantity measured. One could obtain a fairly adequate opinion of the degree of reliability of any one measurement by merely looking at the frequency distribution of the measurements, but statisticians like to express it by a number. The number that is commonly used to represent the amount of variability in a set of measurements is the square root of the average of the squares of the deviations of the measurements from the arithmetic mean for the population. It is represented by the symbol σ, and its definition can be expressed by the following equation.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - m)^2} \qquad \text{- - - - - - - (8)}$$

in which the X_i represents the individual measurements, m represents the arithmetic mean of X_i, and N represents the number of measurements. This is the definition ordinarily given in elementary textbooks, although it is not general enough to fit all kinds of data. To be rigorously correct, one should define σ as the square root of the expected value of the square of the deviation of an individual measurement from the true mean. Equation (8) gives this expected value for a finite population of N independent measurements. For other kinds of data the relationship is more complicated, but the simple definition given above is adequate for the present discussion, and there is no need to confuse the reader with more complicated formulas. Equation (8) is often abbreviated into the form,

$$\sigma = \sqrt{\frac{1}{N}\sum(X - m)^2} \qquad \text{- - - - - - - (9)}$$

which is identical with equation (8) except that the subscripts are omitted. The equation is also frequently written in the form.

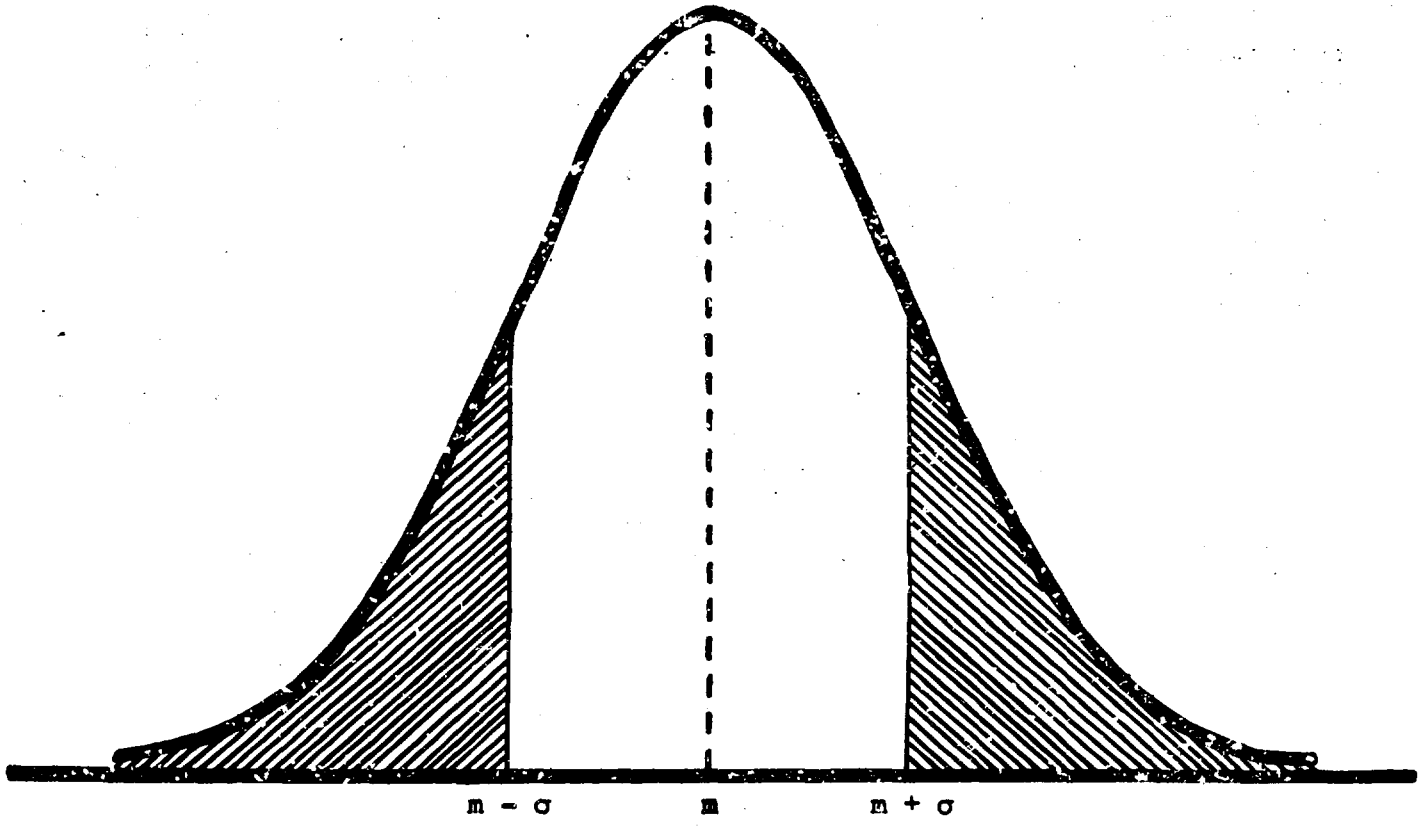
$$\sigma = \sqrt{\frac{1}{N}\sum(x^2)} \qquad \text{- - - - - - - - - (10)}$$

in which each value of x represents the deviation of a measurement from the arithmetic mean rather than the measurement itself.

This particular measure of variability was originally chosen by statisticians because it appears as an important constant in the equation of the Normal Frequency Curve. When applied to Normal distributions, it seems to be the most natural one to use. It was adopted at a time when statisticians were concerned primarily with the Normal Curve. It was given the name Standard Deviation by many statisticians, although those who were interested in error theory referred to it as the Standard Error of a measurement. The latter terminology seems preferable because it emphasizes that a measurement is an estimate

Figure 6. Relation of the standard error to the Normal Curve

The range $m \pm \sigma$ includes 68 percent of the area under the curve



of an expected value and that the deviation of a measurement from its expected value should be interpreted as an error in measurement. At present some statisticians have adopted the convention of using Standard Deviation when referring to the variability of individual measurements and Standard Error when referring to the variability of the means of several measurements. The reader who is already accustomed to this kind of terminology will doubtless wish to retain it, but such a distinction seems unnecessary. There is no fundamental difference in viewpoint when discussing the variability of means as compared with the variability of individual measurements. When different names are used, some readers may infer that the variability of means is interpreted differently than the variability of individual measurements. Nothing could be farther from the truth because the variability of means bears the same relation to the frequency distribution of such means as the variability of individual measurements bears to the frequency distribution of individual measurements. In the present work the terminology of error theory is given preference. Standard Error is used to designate the variability of individual measurements and the variability of means. When distinctions are necessary, the former will be called the Standard Error of an Individual Measurement and the latter will be called the Standard Error of a Mean.

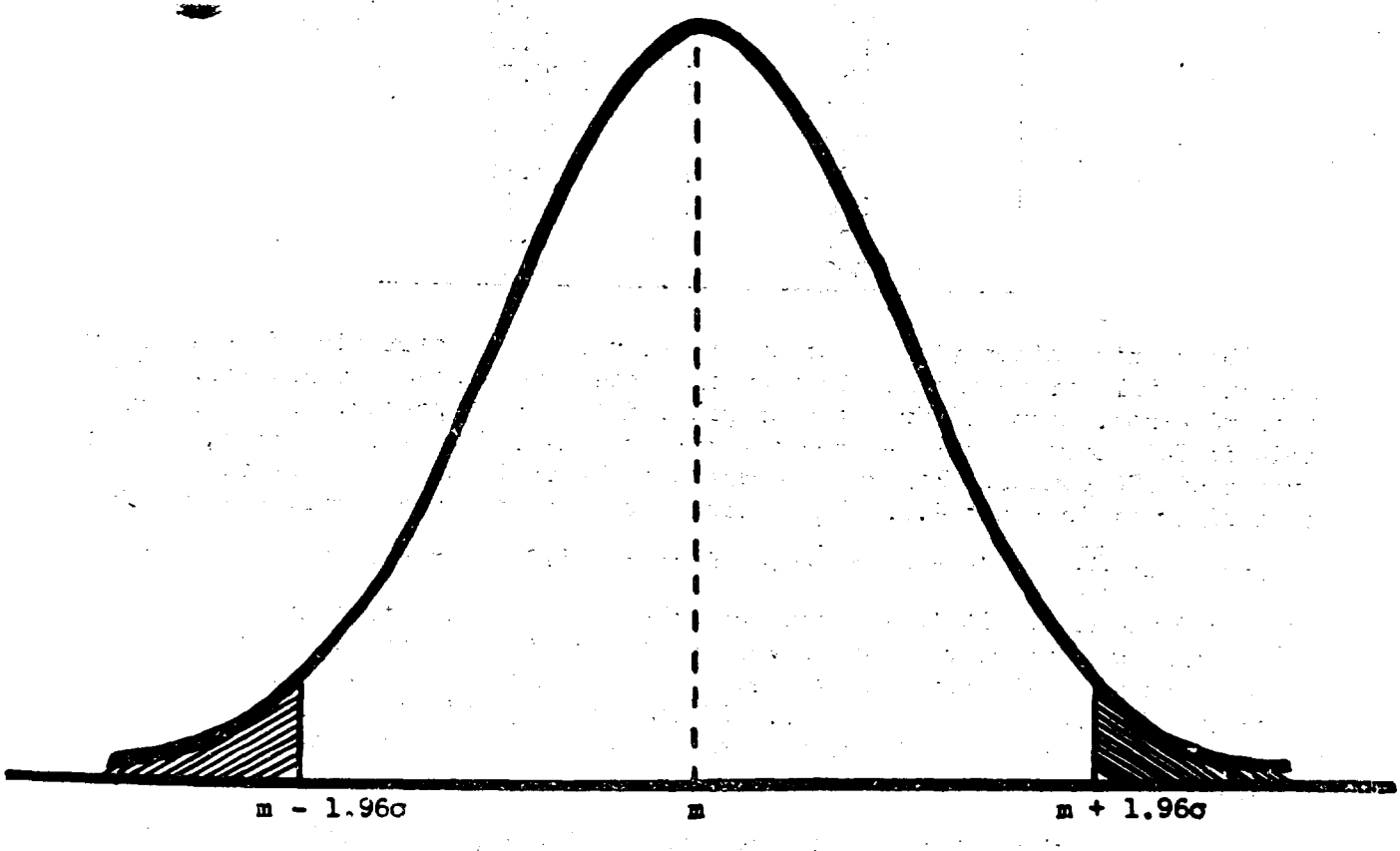
Although the standard deviation, or standard error, had its origin in connection with the Normal Curve, its field of application has been extended until it is now recognized as a general measure of variability, regardless of the shape of frequency distribution to which it is applied. The reader must remember that it has some special interpretations when it is applied to the Normal Curve, however. One of these is illustrated in figure 6, which shows that the two ordinates of the Normal Curve, erected at the values of x whose distance from the arithmetic mean is equal to σ, intersect the curve at its steepest points. This is true only of the Normal Curve. In addition, the area under the curve included between these two ordinates is about 68 percent of the total. This also is true only for the Normal Curve.

At one time, statisticians were much interested in a similar, but shorter, range that would include only 50 percent of the total area under the curve instead of 68 percent. That range was obtained by laying off a distance of about 0.6745σ on each side of the mean. The quantity, 0.6745σ, was called the Probable Error. It was widely used as a measure of variability at the height of its popularity, but the standard error is more convenient to use and serves the same purpose. For this reason, the probable error is now seldom used by statisticians and nothing of value would be lost if it were discarded entirely.

A range that has become exceedingly important in recent years is that defined by laying of a distance equal to 1.96σ on each side of the mean. That range includes 95 percent of the total area under the curve and has been generally adopted as the range within which one would expect a measurement to fall under conditions of random sampling. Theoretically, only 95 percent of the measurements are expected to fall within that range, but this figure has been arbitrarily adopted to represent the bulk of the data. The relation of this range to the Normal Curve is shown in figure 7.

A more exact relationship between the standard error and the length of the range within which measurements are expected to fall has also been developed by statisticians. It can be used to make a rough estimate of the size of the

Figure 7. Normal Frequency Curve, showing range that includes 95 percent of the area under the curve



standard error from the difference between the largest and the smallest measurement in any given set of data. The difference between the largest and smallest measurement is approximately equal to a known multiple of σ. The size of the multiplier changes as the total number of measurements changes because a worker is more likely to get both extremely large and extremely small measurements in one sample when the sample is large than when it is small. Table 1 gives the numerical value of the multiplier for samples of different sizes. This table is part of a more detailed table of the same kind given by Snedecor 1/.

Table 1. - Ratio of range to standard error for samples of different sizes

Size of sample	Ratio of Range to Standard Error
5	2.33
10	3.08
15	3.47
20	3.73
25	3.93
30	4.09
50	4.50
100	5.02
150	5.30
200	5.49
300	5.76
400	5.94
500	6.07
700	6.29
1000	6.48

The county cotton yields for Arkansas, used in constructing figure 1, may be used to illustrate the application of table 1 to a specific problem. The standard error, computed from the original data, is 89 pounds. The largest yield is 545 pounds per acre, and the smallest is 169 pounds per acre, giving a range of 376 pounds. For 75 observations the ratio of the range to the standard error is about 4.8. The estimate of the standard error, derived from the range, is $\frac{376}{4.8}$ or 78 pounds, which does not differ greatly from the exact value of 89 pounds. It is evident that the range-ratio method provides a simple and accurate check that is very useful to the practical statistician. For some purposes the approximate method of estimating the standard error from the range will be useful by itself. If only an approximate value is needed, such an estimate has much to recommend it because it can be computed so easily.

It is important to remember that the relations just described are true only for the Normal Curve. If the frequency distribution involved in a

---

Snedecor, G. W. 1940. Statistical Methods applied to Experiments in Agriculture and Biology, Ed. 3, Collegiate Press, Ames, Iowa. 422pp., illus.

particular problem is not Normal, the shape of the distribution must be considered when a specific interpretation is to be assigned to the standard error. This is often overlooked. So long as the frequency distributions are approximately Normal, no serious error is likely to be made, but if the departure from normality is marked, the error may be considerable.

Samples and Populations

By this time, the reader should be familiar with the general behavior of measurements. When many measurements are made on a physical quantity, or any other quantity or phenomenon that lends itself to measurement, three fundamental concepts should be borne in mind. First there is the concept of an expected value. This is the true value of the quantity measured, and if the measurements are properly made, it is the value that the average of the measurements will approach as the number of measurements is increased. Each individual measurement is an estimate of the expected value. The average of a number of measurements is also an estimate of the expected value, supposedly a better estimate than the result of a single measurement. The average of a number of measurements actually is a better estimate than a single measurement in the sense that it is more likely to be close to the expected value. The reader must remember, however, that the mean of a number of measurements will not always be closer to the expected value than some individual measurements. Second, there is the concept of variability in the measurements, for measurements are subject to error. The nature of the measurements has some effect on the kind of errors that are likely to be made. This brings up the third concept -- that of a frequency distribution of errors of measurement, or what amounts to the same thing, a frequency distribution of the observed measurements. Errors of measurement tend to be distributed according to a frequency curve whose general shape is not constant for all types of measurements. The approximate shape of the frequency distribution likely to be found in any particular problem can often be predicted from a knowledge of the nature of the measurement, but this is not always possible. Experience is the best guide.

All three of these concepts are important in sampling work. The first is important from the point of view of actually obtaining an estimate of the true value of the quantity measured. The second and third are involved in drawing conclusions in regard to the accuracy of the estimate.

As stated previously, a single measurement is an estimate of the true value of the quantity measured. The average of several measurements is a better estimate. The average of a larger number of measurements is a still better estimate. Usually a set of one or more measurements represents only a sample of all possible measurements that might be made on the same quantity. The set of all possible measurements that might be made is called the universe or population from which the sample is taken. In measuring the area of a wheat field, there is no limit to the number of times that area could be measured. In such cases, the universe or population of measurements is unlimited or infinite. But if one were interested in the average number of kernels of wheat per plant in that field, the situation would be different. In that case, each individual measurement would be the number of kernels of wheat

on an individual plant. The maximum number of such measurements that can be taken is limited to the number of plants in the field. Such a population is called finite. Often a finite population may itself be regarded as a sample from some infinite population. As a matter of fact, such a concept provides the foundation for the mathematical analysis of samples from finite populations that is discussed later in this work. In the case of the wheat field just mentioned, the average number of kernels per plant in the field can be treated from two viewpoints. If one is interested only in the average number of kernels per plant in that one field, the population is finite. On the other hand, one might be interested in an unlimited number of fields of that kind. From that viewpoint, the one field studied is itself a sample of an infinite population and any sample of wheat from the field could be regarded as a sample from the same infinite population.

Most of the classical theory of sampling has developed from the viewpoint of sampling from infinite populations. This point of view is entirely appropriate in a large number, perhaps the majority, of practical problems with which the statistician has to deal. On the other hand, there are special types of problems in which that theory is hardly adequate. Much of the sampling work in agricultural statistics comes under this classification. Fortunately, the modifications that must be made in the classical theory to adapt it to finite populations are not complicated and the transition can be made without difficulty.

Perhaps the most striking feature of a finite population is the fact that the true value of the quantity measured can always be obtained if one is willing to assume the labor of making all possible measurements. In the example mentioned previously, it would be possible to count the kernels of wheat on every plant in the field. The average number per plant could thus be ascertained for that field without error. Infinite populations do not have this property. The area of the wheat field could be measured as often as desired. Each additional measurement would increase the statistical precision of the estimate of that area, but one could never be certain that he had computed the true area.

The term statistical precision should be noted carefully. The fact that one sample is larger than another from the same population does not imply that the average obtained from the larger sample is necessarily closer to the true value than the average computed from the smaller sample. It implies only that the average for the larger sample has a greater chance of being close to the true value than does the average for the smaller sample.

This automatically introduces the subject of sampling errors in averages. When samples are drawn from either an infinite or a finite population, it is generally known that the average of a large sample is more likely to be close to the true value than is the average of a smaller sample. The averages for repeated samples of the same size will fluctuate from sample to sample, but this fluctuation will tend to become smaller as the size of the samples is

increased. If samples from an infinite population are taken at random, that is, in such a way that each measurement in the population has an equal chance of being included in every sample, there is a simple relationship between the standard error of the average of a number of measurements and the standard error of a single measurement.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \text{- - - - - - - - - - - -} \qquad (11)$$

In equation (11), σ represents the standard error of a single measurement, n represents the number of measurements used in computing the average, and $\sigma_{\bar{x}}$ represents the standard error of the average. The subscript, $\bar{x}$, is used instead of m to represent the average so that one may distinguish between the sample averages and the true or population value. This device is used often in the following pages of this work.

The standard error of an average bears the same relation to the frequency distribution of such averages as the standard error of a single measurement bears to the frequency distribution of individual measurements. If a large number of samples were drawn at random from an infinite population, one could compute the standard error of an average for samples of that size by means of equation (10). The interesting feature of equation (11) is the fact that it furnishes a method of computing the standard error of an average without first actually computing a number of such averages. All that is required is a knowledge of the size of the sample and the standard error of a single measurement.

If one is dealing with a finite population instead of an infinite population, the formula for computing the standard error of an average becomes

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N}} \qquad \text{- - - - - - - - - -} \qquad (12)$$

in which n is the number of measurements in the sample and N is the number of measurements in the entire population.

Equation (12) differs from equation (11) only with the respect to the factor $\sqrt{\frac{N - n}{N}}$. This factor is needed because the standard error of an average is smaller when the sample is taken from a finite population than when it is taken from an infinite population, other things being equal. If the size of the sample is small in relation to the size of the population, this correction is so small as to be unimportant. The correction becomes increasingly important as the size of the sample is increased. The limiting case is reached when the sample is so large that it includes the entire population, that is, when n = N. In this case, the factor reduces to zero and the standard error of the average, computed from equation (12), will also be equal to zero. This is perfectly logical because the true average of the entire population is estimated without error when the sample includes the entire population. The averages for repeated samples of the same size would necessarily have to be equal to each other.

Exercise 8. - Suppose that there are 1000 wheat fields with an average area of 25 acres in a county. The area of each field can be regarded as an estimate of this figure. The standard deviation of the areas of the individual fields is 15 acres. What formula should you use in computing the standard error of the average acreage that would be obtained for samples of the following sizes:

(a) 10 fields?
(b) 100 fields?
(c) 900 fields?

Compute the standard error of the average for each of these samples, first by equation (11) and then by equation (12), and explain the differences in the results. Draw a rough sketch showing how the frequency distribution of averages from repeated samples of each size should look if the distribution of individual field areas is Normal and indicate the range that would include 68 percent of the sample averages in each case.

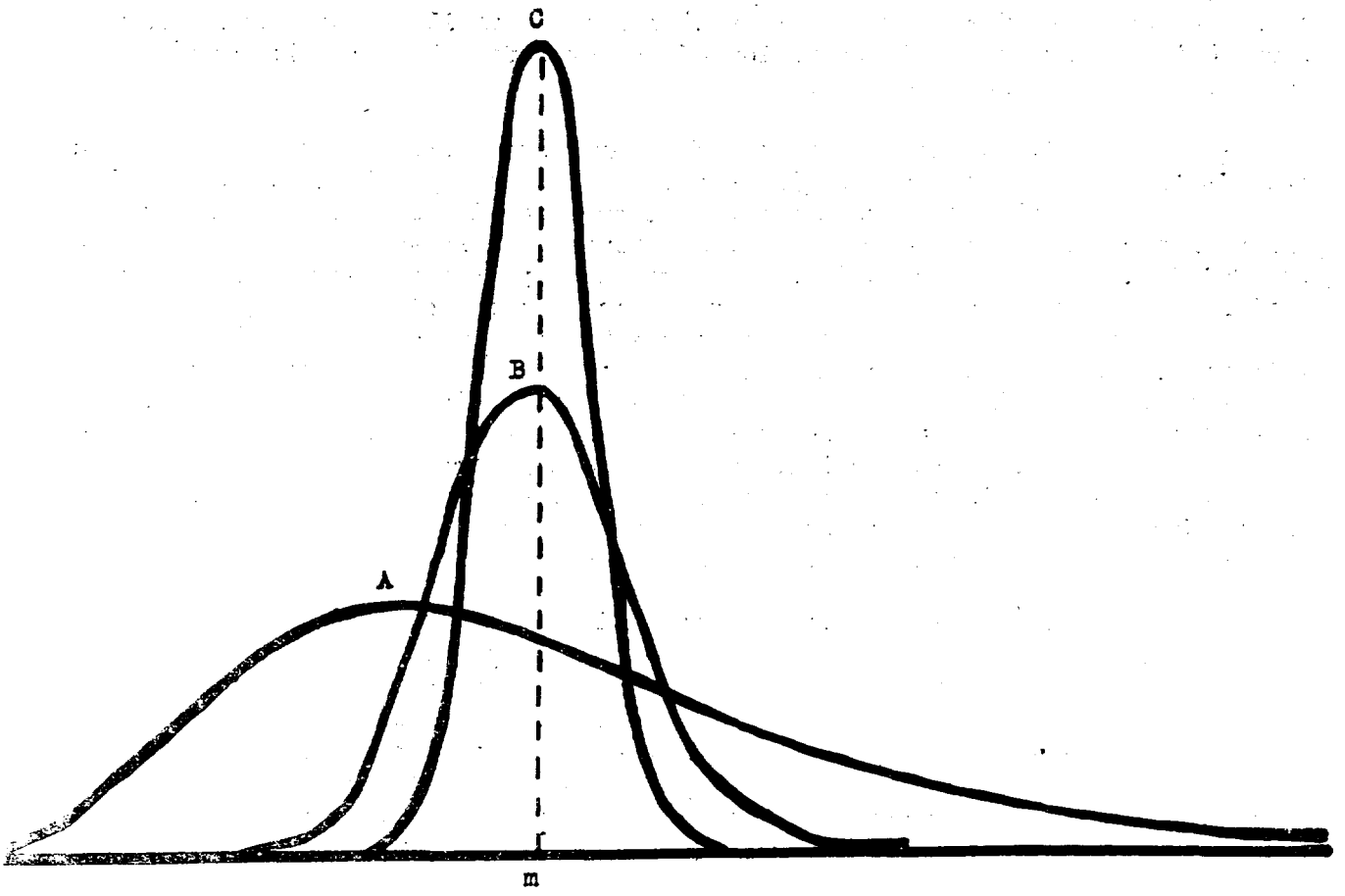
Unbiased Estimates

The theory of sampling is important because the only information that can be obtained about a population must usually be based on a study of samples from that population. If the population is infinite, there is no possible alternative. In the case of finite populations, it is theoretically possible to make a study of the entire population, but such an extensive study is seldom practicable. If one wished to learn the average number of kernels of wheat per plant in a field, there is little likelihood that he would count the kernels on every plant in the field. The logical procedure would be to take a sample of plants from that field and count the kernels of wheat on those plants only. The average number of kernels per plant in the sample would serve as an estimate of the average number of kernels per plant for the entire field. If this estimate is to serve its purpose, it must be an unbiased estimate. Bias in an estimate refers to a consistent tendency to underestimate or overestimate the quantity that is being measured. Such errors will always tend to be in the same direction from sample to sample, so that they will not counterbalance each other. Errors of this kind are the most troublesome ones with which the statistician has to deal because they will not "average out." Bias can arise from two sources: (1) an improperly drawn sample or (2) improper methods of computation. The two are distinct and must be discussed separately.

So far as the character of the sample itself is concerned, there is considerable misunderstanding in regard to what constitutes a properly drawn sample. Assuming for the present that the computations performed on the sample are all that they should be, a sample will give an unbiased estimate if it is drawn in such a way that the average of estimates based on all possible samples of that kind is equal to the true value of the quantity measured. Samples can be taken in many different ways and still satisfy this requirement.

Figure 8. Frequency distributions of averages for random samples from a non-normal population

A: Distribution of individual measurements
B: Distribution of averages from samples of 10 measurements each
C: Distribution of averages from samples of 30 measurements each



A random sample -- that is, a sample such that every measurement in the population has an equal chance of being included -- will give an unbiased estimate, but it is by no means the only kind of sample that has this property. A study of the properties of a random sample provides a good introduction to the subject, however.

The frequency distribution of arithmetic means under conditions of random sampling was mentioned in the preceding section. An arithmetic mean, computed from a random sample, is an unbiased estimate of the true population value because the average of such estimates, computed from all possible random samples, is equal to the true population value. This is the case, regardless of the kind of frequency distribution exhibited by the measurements in the population from which the samples are drawn.

Figure 8 shows an example of the relation between the frequency distribution of individual measurements and the frequency distribution of averages for random samples of different sizes when the frequency distribution of individual measurements is not Normal. Curve A is the frequency distribution of individual measurements for the entire population. Curve B is the frequency distribution of averages for all possible random samples of 10 measurements each. Curve C is the frequency distribution of averages for all possible random samples of 30 measurements each.

All three distributions have the same arithmetic mean and that arithmetic mean is the arithmetic mean of all measurements in the population. It should be observed that the standard error of the averages is smaller than the standard error of the individual measurements, as pointed out in the preceding section. In addition, the reader should notice particularly that the frequency distribution of the averages becomes more symmetrical as the size of samples increases. For all practical purposes, curve C can be regarded as a Normal distribution. This is an almost universal property of the frequency distributions of averages and explains why most practical statisticians do not show much concern about the shape of the frequency distributions of the individual measurements with which they are working. One is usually more interested in the variability of averages than in the variability of the individual measurements, and if the samples are fairly large, no serious error is introduced by assuming that the distribution of the averages is Normal.

Although a random sample from a population will yield an unbiased estimate of the population average, it will not necessarily provide an accurate estimate. If the individual measurements show a large amount of variability, the averages from random samples will also have large standard errors unless the samples are very large. Research workers have long been aware of this and have sought to overcome the difficulty by using judgement in selecting the samples. If something is known about the nature of the population from which samples are to be drawn, a worker can often control the sampling in such a way that each sample is more representative of the population than a random sample is likely to be. Such schemes can be successful, but they must be used with caution because human judgement is not infallible. The investigator may unconsciously introduce a bias into his results through an error in deciding what is representative of the population. When such control can be properly exercised, however, the results are well worth the effort. This subject is of the utmost importance to the practical statistician and is discussed later in a separate section.

The second source of bias, improper methods of computation, is harder to visualize. The reader should not conclude that "improper methods of computation" refers only to mistakes in arithmetic. The roots of the problem go much deeper. By definition, an unbiased estimate must be such that the average of estimates, based on all possible samples of the kind drawn, will be equal to the true value of the quantity measured in the population. This definition may be condensed into the statement that the expected value of the estimate must be equal to the true value of the quantity measured.

If the samples are properly drawn, the average of the arithmetic means derived from all possible samples will be equal to the true arithmetic mean for the entire population. This is not true for all statistical constants. The most familiar exception is the case of the standard deviation or standard error. Equation (8) defines the standard error of an individual observation as the square root of the arithmetic mean of the squares of the deviations of the individual measurements from the arithmetic mean for the entire population. One might suppose that the corresponding formula for computing an estimate of the standard error from a sample should be,

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2} \qquad \text{- - - - - - - -} \quad (13)$$

The symbol s is used instead of σ to distinguish the estimate from the true value just as $\bar{x}$ was substituted for m to distinguish the estimated arithmetic mean from the true value for the population. If values of s were computed for all possible random samples of size n that could be drawn from the population, the average, $\bar{s}$, of these estimates would not be equal to the population value, σ. The bias would be fairly large when n is small, although it would decrease as the size of the samples was increased. If the samples were drawn from an infinite population, n would have to be infinitely large before the bias would disappear entirely, however.

The amount of the bias can be computed by the following equation,

$$\frac{\bar{s}}{\sigma} = \frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{\sqrt{n}\,\Gamma\left(\frac{n-1}{2}\right)} \qquad \text{- - - - - - - - - -} \quad (14)$$

The numerical value of the expression on the right-hand side of equation (14) is difficult to compute, but tables that give its value for different values of n have been prepared. Table 2 presents values of the ratio $\bar{s}/\sigma$ for a few values of n and gives some idea of the amount of the bias for samples of different size. A more complete table is given by Shewhart[2/].

- -

2/ Shewhart, W. A. 1931. Economic control of Quality of Manufactured Product. Van Nostrand, New York. 501 pp., illus.

Table 2. - Values of the ratio $\bar{s}/\sigma$ for different values of n

n	$\bar{s}/\sigma$
5	0.776
10	.894
15	.931
20	.949
25	.959
30	.966
35	.971
40	.975
45	.978
50	.980
60	.983
70	.986
80	.987
90	.989
100	.990

Table 2 shows that the average of the estimated standard errors derived from the samples is consistently smaller than the true value for the population. This bias could be eliminated by multiplying each value of s obtained from a sample by a correction factor. This factor would be the reciprocal of the ratio $\bar{s}/\sigma$ corresponding to the appropriate value of n. At first glance, this would appear to be a simple method of overcoming the difficulty at hand, but when one probes more deeply into the fundamental theory of statistics, it is evident that the problem requires a more thorough study.

For example, suppose one were interested in getting an estimate of σ^2 rather than σ. The computations would be the same except that the extration of the square root would be eliminated. The estimate of σ^2 would be given by the equation,

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{x})^2 \qquad \text{- - - - - - - - - - (15)}$$

The average of all estimates of s^2 would be smaller than σ^2, just as the average of all estimates of s would be smaller than σ, but the bias would be different. The relation between σ^2 and the average of all possible values of s^2 is given in the equation,

$$\overline{s^2} = \frac{n-1}{n} \sigma^2 \qquad \text{- - - - - - - - - - - - - - (16)}$$

The bias in s^2 could be removed by multiplying each value of s^2 by $\frac{n}{n-1}$, but the resulting unbiased estimate of σ^2 would not be equal to the square of the unbiased estimate of σ discussed previously.

One would thus be confronted with the paradox of an unbiased estimate of σ and an unbiased estimate of σ^2 with the latter not equal to the square of the former. This should convince the reader that the problem of computing unbiased estimates requires careful thought. A comprehensive discussion of this subject cannot be attempted here. But it can be said that most mathematicians prefer to compute the unbiased estimate of σ^2 and to use the square root of the result as the best estimate of σ that can be obtained. Many statisticians have now adopted this concept to the extent of re-defining the standard error as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_1 - \bar{x})^2}{n - 1}} \qquad \text{- - - - - - - - (17)}$$

This estimate of σ is known as the optimum estimate and will be used exclusively in the present work. Some statisticians object to this as a definition of the standard error, but it possessos many advantages in discussions of sampling problems and has been rather generally accepted in recent years.

So far as bias is concerned, the estimate of σ given by equation (17) is a biased estimate, but it so happens that a worker is usually more interested in having an unbiased estimate of σ^2 than an unbiased estimate of σ. The square of the optimum estimate of σ will be an unbiased estimate of σ^2. The use of n - 1 as a divisor illustrates an application of what is known as the concept of degrees of freedom. The above definition of the standard error is sometimes called an estimate based on n-1 degrees of freedom. This expression is used often in the following pages and it is well to become familiar with it.

The above illustrations serve as an introduction to the problem of obtaining unbiased estimates from one or more samples taken from the population in question. The nature of the problem should be fully understood by every statistician, whatever his field of activity may be. Further discussions of this kind in these pages will deal largely with the problem of obtaining unbiased estimates of arithmetic means; in that subject one is not concerned with bias caused by improper methods of computation in the sense with which the term is used here. Much use will be made of unbiased estimates of squared standard errors. The discussion just concluded should be sufficient to prepare the reader to follow these applications of the concept of unbiased estimates without further elaboration.

Concept of Probability

To most laymen the word "probability" implies some sort of vague statement that must be made whenever the absence of specific information prevents the reaching of a definite conclusion. If the statistician were restricted to such a concept, there could be no so-called "mathematical theory" of probability. Mathematics can be applied only to a concept of probability that lends itself to measurement. Any definition that is to be of practical use to the statistician must be expressed in that kind of language.

The concept of probability that is most commonly used, at present, has proved to be useful and has been adopted by the great majority of statisticians. Under that concept, probability is defined as the relative frequency with which an event is expected to occur in a number of trials. The reader should realize at once that, under this definition, the true probability of the occurrence of an event may not be known but must be estimated from observed data. For example, if one is dealing with the areas of a number of wheat fields and finds that 10 percent of all fields measured have an area of 25 acres, the probability of getting a field with an area of 25 acres would be estimated as 0.1. If this estimate were based on a study of all wheat fields in the population, it would represent the true probability of getting a 25-acre field. If the estimate were based on only a sample of all wheat fields in the population, one would have only an estimate of the probability of getting a 25-acre field. In such a situation one would have to admit that the true probability of getting a 25-acre field was not known, but that it has been estimated to be 0.1.

The reader should not be unduly disturbed to learn that the true probability of occurrence of an event may not always be known in practice. Estimates derived from observed data are usually sufficient for practical purposes. The concept of an expected value of such an estimate is of interest mainly in academic discussions of the theory.

As probability has been defined in terms of frequency of occurrence of an event, it should be evident that probability theory is intimately associated with the theory of frequency distributions. In the previous discussion of frequency curves, it was pointed out that mathematicians customarily represent frequencies by areas under a frequency curve. In figure 2, for example, the shaded area, dF, represented the number of measurements falling in the class interval bounded by x and x + dx when the measurements are distributed according to the Normal Law. The same curve could easily be constructed on such a scale that the total area under the curve would be equal to unity instead of representing the total number of measurements. The curve would have the same general shape, but the shaded area, dF, would then represent the fraction, rather than the number, of measurements falling in the interval bounded by x and x + dx. By definition, the area, dF, would thus represent the probability that a measurement will fall within the interval bounded by x and x + dx.

In the present work, the reader will have many occasions to make use of the concept of probability as outlined above. He should always be able to interpret probability statements in terms of frequency distributions. Whenever a probability statement is encountered that does not lend itself to such an interpretation, he can be sure that the statement was not properly made.

In this publication, probability means relative frequency, and that definition should be kept in mind at all times.

Exercise 9. - In a Normal frequency distribution, 68 percent of the measurements should fall within the range bounded by $m - \sigma$ and $m + \sigma$ where m is the arithmetic mean for the entire population and σ is the standard error. Express this fact in terms of probability instead of frequency.

Exercise 10.- If measurements are distributed according to the Normal law, what is the probability that a measurement will fall within the range bounded by $m - 1.96\sigma$ and $m + 1.96\sigma$? What is the probability that a measurement will fall outside of this range?

Exercise 11.- What is the probability that a measurement will fall within the ranges bounded by the following values, assuming a Normal frequency distribution in each case:

(a) m and $m + \sigma$?
(b) $m - \sigma$ and m?
(c) $m - \sigma$ and $m + 1.96\sigma$?
(d) m and $m + \infty$?
(e) $m - \sigma$ and $m + \infty$?
(f) $m + \sigma$ and $m + \infty$?
(g) $m - \sigma$ and $m - \infty$?
(h) $m + \sigma$ and $m - \infty$?
(i) $- \infty$ and $+ \infty$?
(j) $m + 1.96\sigma$ and $m + \infty$?
(k) $m - 1.96\sigma$ and $m + \infty$?
(l) $m - 1.96\sigma$ and $m - \infty$?

The symbol ∞ is used to represent infinity in books on mathematics. In this exercise, $- \infty$ means an unlimited distance to the left. $+ \infty$ means an unlimited distance to the right. $m + \infty$ means to start at the point m and move to the right without stopping. $m - \infty$ means to start at the point m and move to the left without stopping.

Sample Means as Estimates of Population Means

A sample is usually drawn from a population for the sole purpose of obtaining some information about the population. The number of characteristics of the population about which information is desired may be large, but in almost every statistical investigation, one is interested in estimating the arithmetic mean. When the sample has been drawn and the arithmetic mean computed from the sample, the statistician must arrive at some conclusion in regard to the arithmetic mean for the population as a whole. If this could not be done, the time and effort spent on the sample would be futile. The sample is of interest only insofar as it yields information about the population from which it was drawn. In order to arrive at some conclusion about the mean for the population, it is necessary to compute an estimate of the standard error of the sample mean. But first one must obtain an estimate of the standard error of a single measurement. Abbreviating the notation in equation (17) slightly, the estimate of the standard error of a single measurement is

$$s = \sqrt{\frac{\Sigma(X - \bar{x})^2}{n - 1}} \qquad \text{- - - - - - - - - - - (18)}$$

For samples drawn from an infinite population, the estimate of the standard error of a sample mean may be written,

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \text{- - - - - - - - - - - (19)}$$

For samples drawn from a finite population this quantity would have to be multiplied by the factor, $\sqrt{\frac{N - n}{N}}$, as indicated previously.

It is important to remember that s is only an <u>estimate</u> of the standard error of a single measurement and $s_{\bar{x}}$ is, therefore, only an <u>estimate</u> of the standard error of the sample mean. The true standard error of the sample mean is $\sigma/\sqrt{n}$ where σ is the true standard error of a single measurement. The numerical value of σ is hardly ever known in practice. One must be satisfied with the estimate, s.

The estimate, $s_{\bar{x}}$, is the only statistic available for drawing conclusions about the adequacy of the sample mean as an estimate of the population mean. To show how it can be used in drawing such conclusions, it is necessary to review the subject of the frequency distributions of sample means.

If random samples are drawn from a Normal population with the arithmetic mean, m, and standard error, σ, the means from samples of n observations will be Normally distributed about m with a standard error, $\sigma_{\bar{x}}$. This is equivalent to saying that the quantity $\frac{(\bar{x} - m)}{\sigma_{\bar{x}}}$ is normally distributed about a mean of zero with unit standard error. Since the value of $\sigma_{\bar{x}}$ is usually not available and one must depend upon the estimate, $s_{\bar{x}}$, derived from the sample, one

is naturally interested in knowing how the quantity $\frac{(\bar{x} - m)}{s_{\bar{x}}}$ is distributed. This quantity is denoted by t and its frequency distribution is now well-known. The frequency distribution is symmetrical, but not Normal, and has a mean of zero. It approaches the Normal Curve as a limit when n is made large -- the difference between the two being unimportant when n is greater than about 30. For small values of n, however, the difference is fairly large.

The difference between the t distribution and the Normal is illustrated in figure 9 where the distributions of t for 4 and 9 degrees of freedom, corresponding to samples 5 and 10 measurements each, are shown in comparison with the Normal Curve drawn on the same scale. The manner in which the t distribution approaches the Normal Curve as the sample size increases can be easily distinguished. Even for samples containing only 10 measurements each, the departure of the t distribution from the Normal is not great. The most important difference between the t distribution and the Normal Curve is to be found in the areas that lie under the tails of these curves. The Normal curve approaches the base line faster than the t distribution. These areas are particularly important in making statistical tests. In such tests one is much interested in the range formed by laying off equal distances of such length on each side of the mean that the range includes 95 percent of the area under the frequency curve. The faster the frequency curve approaches the base line, the shorter this range will be. Therefore, the range will have to be longer for the t distribution than for the Normal.

For the Normal distribution of $\frac{\bar{x} - m}{\sigma_{\bar{x}}}$, the range that includes 95 percent of the area is obtained by laying off a distance equal to about 1.96 on each side of the population mean of this quantity, which is zero. For the distribution of t, which is equal to $\frac{\bar{x} - m}{s_{\bar{x}}}$, the distance that must be laid off on each side of its mean value of zero is larger than 1.96. The distance varies with the number of degrees of freedom used in estimating s and $s_{\bar{x}}$, however, and approaches 1.96 as a limit when the number of degrees of freedom becomes large. For the present, the student should think of the number of degrees of freedom as being equal to one less than the number of individual measurements used in estimating s and $s_{\bar{x}}$. Table 3 gives approximate values of this distance for a few different numbers of degrees of freedom. More detailed tabulations may be found in published tables of the t distribution that are now available in most textbooks on statistics. The differences between the t distribution and the Normal disappears as the number of degrees of freedom becomes large because s and $s_{\bar{x}}$ approach their population values, σ and $\sigma_{\bar{x}}$.

Figure 9. Distribution of t compared with Normal Curve

A ..... t distribution for 4 degrees of freedom
B _____ t distribution for 9 degrees of freedom
C _ _ _ _ _ Normal Curve

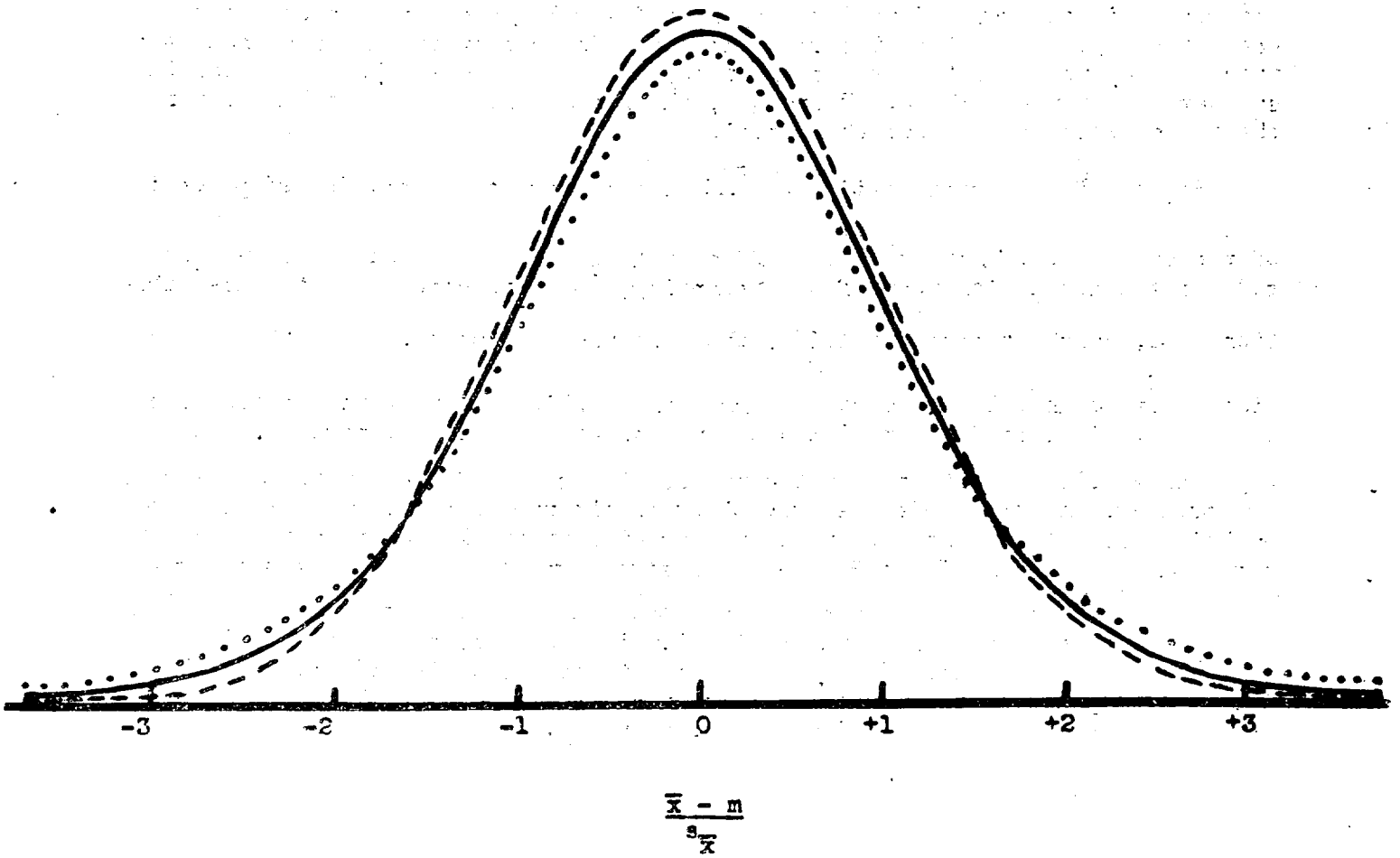


Table 3.- Distance to be laid off on each side of the mean of the t distribution to include 95 percent of all possible values.

Degrees of Freedom	Distance on each side of mean
1	12.706
2	4.303
3	3.182
4	2.776
5	2.571
10	2.228
15	2.131
20	2.086
25	2.060
30	2.042
50	2.008
100	1.984
200	1.972
300	1.968
400	1.966
500	1.965
1000	1.960
∞	1.960

The t distribution is extremely useful in drawing conclusions about the adequacy of a sample mean as an estimate of the population mean. But the way in which it must be used requires some careful thinking on the part of the statistician. The reader should notice particularly that the t distribution is formed by individual values of t, computed from separate estimates of $\bar{x}$ and $s_{\bar{x}}$ for each sample. Estimates of both will vary from sample to sample. Even if the true value of $\sigma_{\bar{x}}$ were known so that one could use the Normal Curve, one would have to be careful to avoid erroneous conclusions.

The proper way to determine how well the sample mean represents the population mean would be to compute the range extending from $\bar{x} - 1.96\sigma_{\bar{x}}$ to $\bar{x} + 1.96\sigma_{\bar{x}}$. One could then state that there is a probability of 0.95 that the range $\bar{x} \pm 1.96\sigma_{\bar{x}}$ includes the population mean, m. This is the concept of

fiducial limits or confidence intervals. It implies that, if an arithmetic mean were computed from each of all possible samples of the same size and a range were computed from each by first subtracting and then adding $1.96\sigma_{\bar{x}}$ to each sample mean, 95 percent of these ranges would include the population mean, m. This is an exact probability statement and the form in which it is given should be noted carefully because it is often misquoted. Many otherwise reputable statisticians sometimes claim that there is a probability of 0.95 that the "population mean, m will fall in the interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$" where $\bar{x}$ is a given sample mean. The fallacy in this kind of statement should be apparent at once. Probability is another word for frequency and every probability statement must be interpreted in terms of the frequency of occurrence of an event. In the

Figure 10. Frequency distribution of arithmetic means for samples of 5 measurements, drawn from a normal population with $m = 15$ and $\sigma = 6$

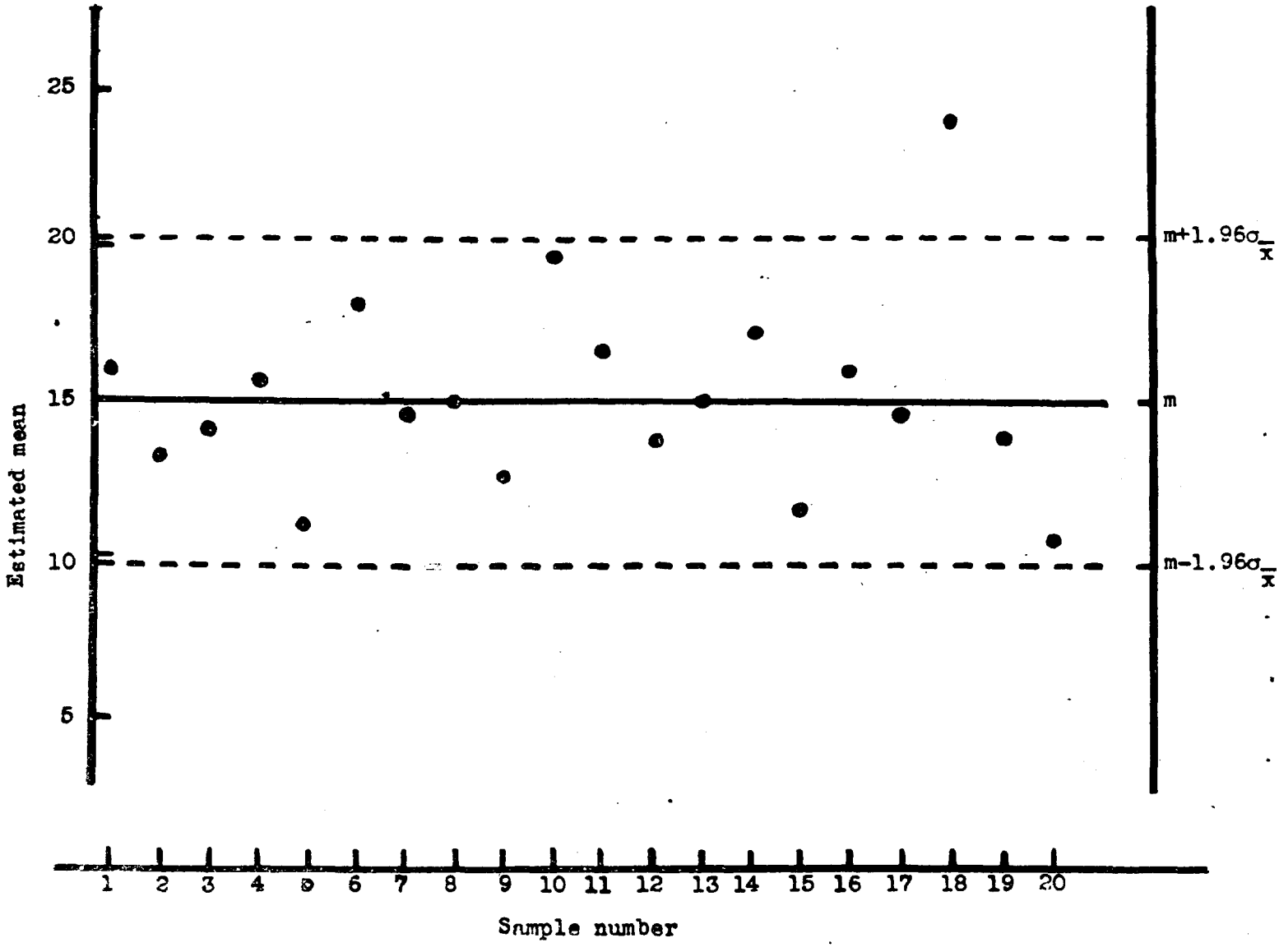
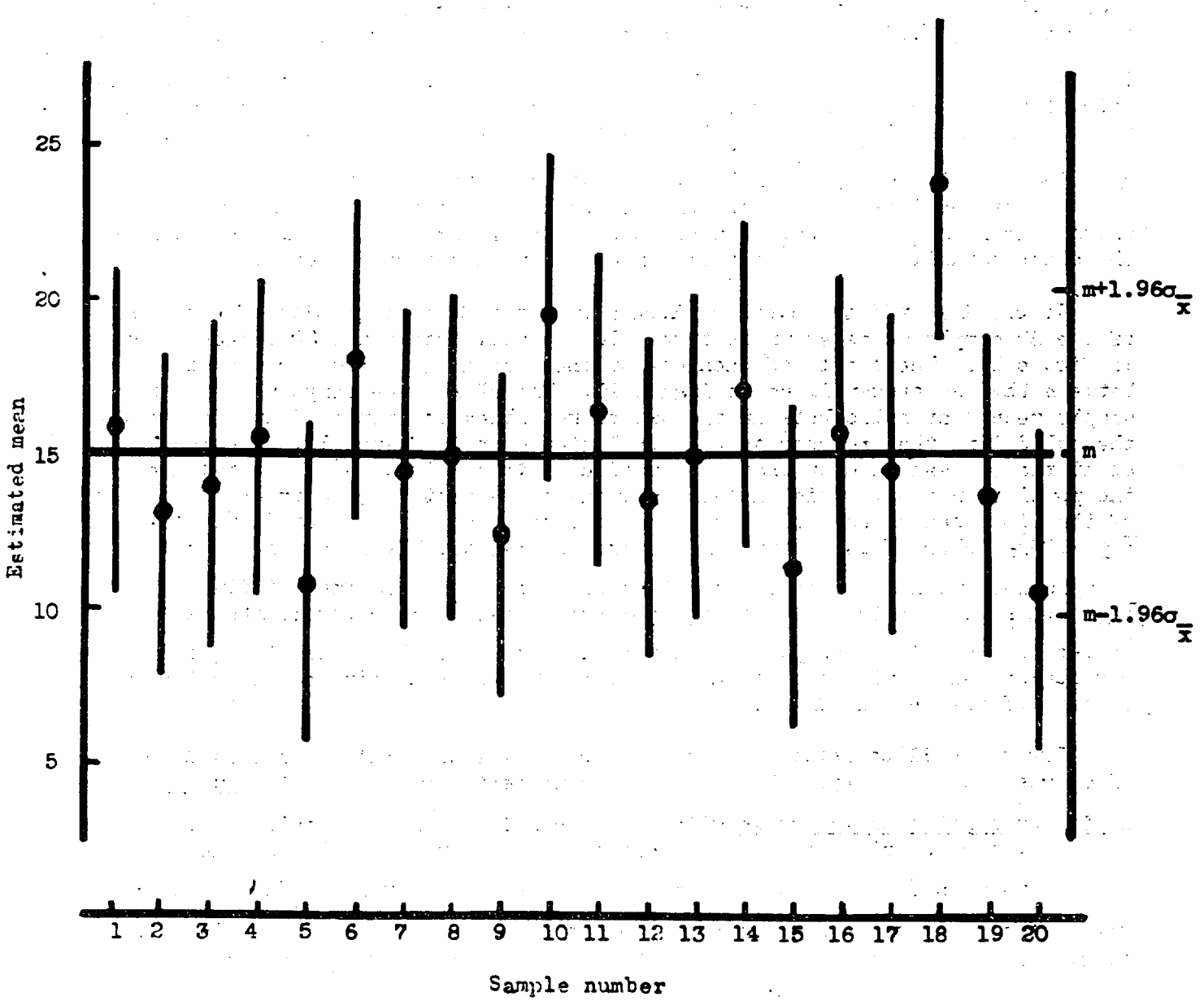


Figure 11. Fiducial limits on arithmetic means for samples of 5 measurements each, drawn from a Normal population with m = 15 and σ = 6.
(The value of σ was assumed known in this example)



present problem, the population mean, m, must be regarded as a fixed quantity. It is the computed range that varies from sample to sample. The only correct statement that can be made in terms of probability must refer to the number of times a computed range will include m.

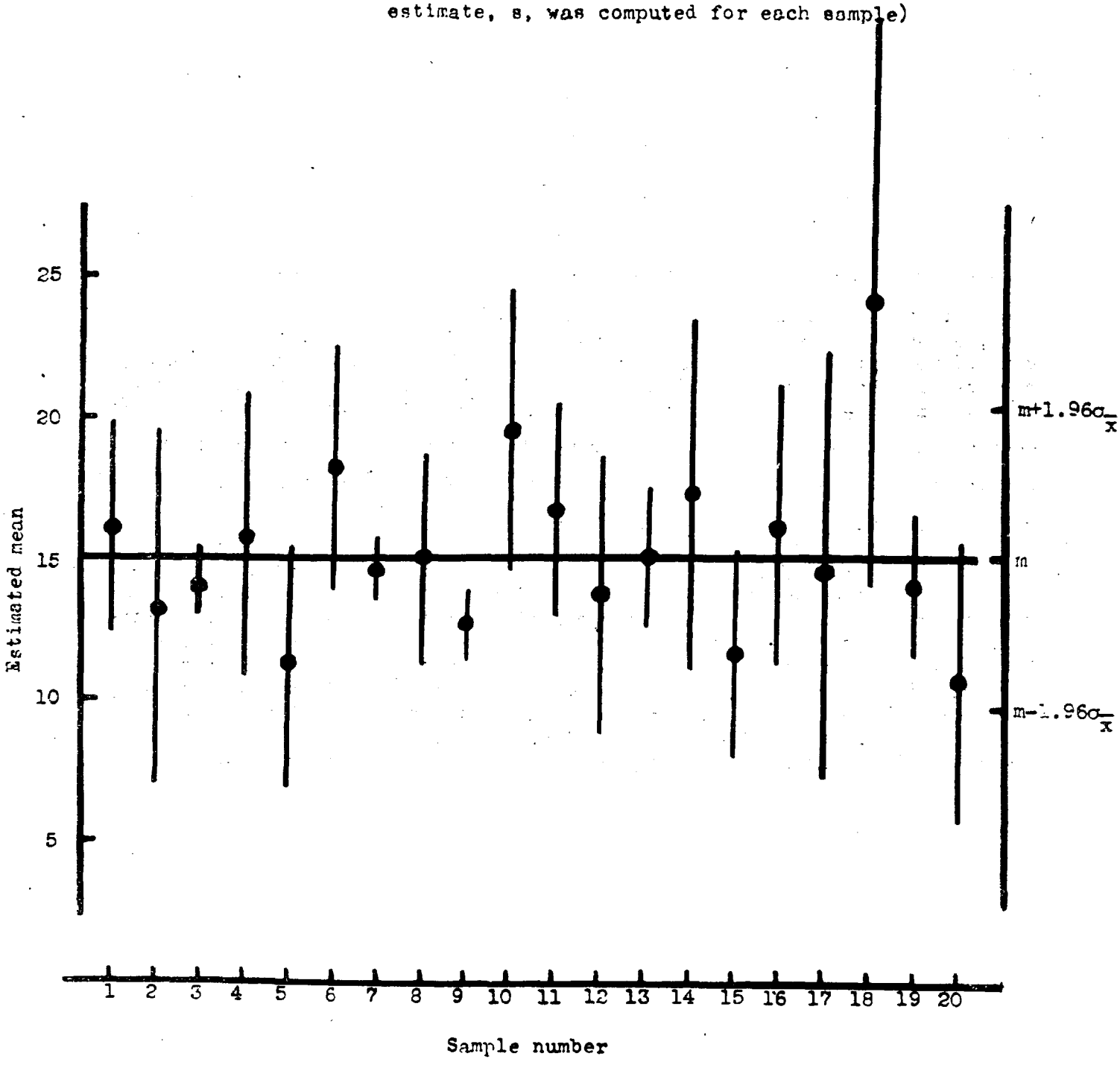
The reader may have some difficulty in seeing just why 95 percent of the ranges defined above should include the population mean. That fact follows directly from what is known about the frequency distributions of sample averages. It is known that 95 percent of the sample averages should fall within the range, $m \pm 1.96\sigma_{\bar{x}}$ where m is the population mean. The fact that the numerical value of m is unknown is immaterial. Figure 10 shows that the kind of distribution of averages one might obtain from random samples of 5 measurements, drawn from a Normal population with $m = 15$ and $\sigma = 6$. The standard error, $\sigma_{\bar{x}}$, of a sample mean is $6/\sqrt{5}$ or 2.68. The solid horizontal line shows the position of the population mean, m. The two broken lines show the range within which 95 percent of the sample means should fall. 19 of the 20 means shown in the chart actually are within this range as demanded by the theory.

If one lays off the distance $1.96 \times 2.68 = 5.25$ on each side of every one of the 20 sample means shown in figure 10, he will obtain the 20 ranges shown in figure 11. The length of each one of these ranges is equal to the distance between the two broken lines in figure 10. Therefore, the range about each sample mean that falls between the two broken lines in figure 10 must cross the solid line that represents the population mean in figure 11. The range about the one sample mean that was outside of these limits will not include the population mean. This explains why one can expect to be correct 95 percent of the time when he concludes that the population mean is included in the range $\bar{x} \pm 1.96\sigma_{\bar{x}}$, where $\bar{x}$ is a sample mean taken at random. Even though the value of m is not known, one can be sure that 95 percent of all ranges defined by $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will include m. It is important to remember, however, that $\bar{x}$ varies from sample to sample and in practice one never knows exactly which particular ranges will include m. All that one knows in advance is that 95 percent of them should do so.

The above illustration was simplified by assuming that the numerical value of $\sigma_{\bar{x}}$ was known exactly. This was done to present the concept of <u>fiducial limits</u> or <u>confidence intervals</u> with as few complications as possible. In practice, the numerical values of σ and $\sigma_{\bar{x}}$ are seldom available and one must be content with the estimates, s and $s_{\bar{x}}$, derived from samples. This naturally raises the question of how the concept just described can be applied in practice. If one were to compute the range defined by $\bar{x} \pm 1.96s_{\bar{x}}$ for each sample, these ranges would not be of the same length because the numerical value of $s_{\bar{x}}$ would fluctuate from sample to sample.

This is not the most serious feature of the situation, however. The distribution of sample means would still be as shown in figure 10, but the various ranges formed by computing $\bar{x} \pm 1.96s_{\bar{x}}$ for each sample, in addition to

Figure 13. Fiducial limits on arithmetic means for samples of 5 measurements, drawn from a Normal population with $m = 15$ and $\sigma = 6$.
(The value of σ was assumed unknown and an estimate, s, was computed for each sample)



being of unequal length, would no longer include the population mean 95 percent of the time. Too many of these ranges would be a little too short, the discrepancy increasing as the number of degrees of freedom available for estimating s and $s_{\bar{x}}$ decreases. When the factor, 1.96, is replaced by the appropriate value of t, as shown in table 3, this latter difficulty is corrected. The ranges of the type $\bar{x} \pm ts_{\bar{x}}$ will still be unequal in length, but 95 percent of them will include the population mean, m.

In the illustration previously discussed, each sample consisted of five measurements which yielded 4 degrees of freedom for estimating s and $s_{\bar{x}}$. The value of t for the 95 percent fiducial limits corresponding to 4 degrees of freedom given in table 3 is 2.776. If one were to compute the range defined by $\bar{x} \pm 2.776s_{\bar{x}}$ for each mean shown in figure 10, he would arrive at the situation depicted in figure 12. Of the ranges, 95 percent include the population mean, as was the case in figure 11 but, interestingly enough, it is not the same 95 percent that had this property before. This is immaterial, however, because all that is required is the condition that 95 percent of the ranges of the type $\bar{x} \pm ts_{\bar{x}}$ will include the population mean. This gives assurance that when a mean $\bar{x}$, and its standard error $s_{\bar{x}}$, have been computed from a given sample, there is a probability of 0.95 that the range, $\bar{x} \pm ts_{\bar{x}}$, so established will include the population mean.

The discussion just concluded indicates the kind of probability statement that can be made about the adequacy of a sample mean as an estimate of the population mean from information contained in the sample. Some statisticians have pointed out certain limitations in the utility of such a concept, but attempts at improvement have often become involved in serious controversies. These are not discussed here. So long as probability is defined in terms of the frequency of occurrence of an event, the discussion here given stands on a solid mathematical foundation. The viewpoint that it represents appears to be as satisfactory as anything that has yet been devised.

In recent years, the term variance has been used by statisticians more and more frequently. Variance is defined simply as the square of the standard error and the reader may wonder why a special name has been assigned to it. The explanation is not hard to find, however. The squared standard error is used more frequently than the standard error itself in many problems and statisticians found it cumbersome to refer continually to the "squared standard error." Variance is an easy word to say and becomes less tiresome upon repetition than "squared standard error." A few statisticians have gone a step farther and have begun to use the letter V to represent variance. This has many advantages in printing statistical formulas because the use of exponents is avoided. But the student will recall that a distinction was made previously between the true value of the standard error for the population as a whole and the estimate based upon a sample from the population. The former was denoted by σ and the latter by s. As yet, this distinction has not been made with respect to the symbol for variance and the student must be careful to determine from the context whether V is being used to represent σ^2 or s^2 in a particular formula with which he may be confronted. In practical work, it is becoming customary to associate V with the estimated variance, s^2, because it is only in theoretical discussions that one has any use for the symbol σ^2.

Any formula involving the standard error can be written in terms of variance. The estimated variance of a single measurement derived from a sample of n observations or n - 1 degrees of freedom may be written.

$$V = \frac{S(X - \bar{x})^2}{n - 1} \qquad \text{- - - - - - - - - - - - (20)}$$

The estimated variance of a mean based on n measurements from an infinite population is

$$V_{\bar{x}} = \frac{V}{n} \qquad \text{- - - - - - - - - - - - (21)}$$

The estimated variance of a mean based on n measurements from a finite population containing only N measurements is

$$V_{\bar{x}} = \frac{V}{n}\left(\frac{N - n}{N}\right) \qquad \text{- - - - - - - - - - - - (22)}$$

The student will be well-advised to become familiar with this notation because it is used often in the present work. It is difficult at first to form a mental picture of variance as a measure of variability after one has been accustomed to thinking in terms of standard errors, but this difficulty will disappear with practice. After the variance notation becomes familiar, its many advantages more than compensate for the time and effort spent in becoming acquainted with it.

One of the principal advantages of the variance notation follows from the additive property of variances. If a measurement has the variance V_1 and a second measurement has the variance V_2, the variance of the sum of those measurements is given by the relation,

$$V_s = V_1 + V_2 \qquad \text{- - - - - - - - - - - - (23)}$$

The variance of the difference between the two measurements is, surprisingly, equal to the variance of the sum and is written

$$V_d = V_1 + V_2 \qquad \text{- - - - - - - - - -} \quad (24)$$

The relationships defined by equations (23) and (24) are based on the assumption that the two measurements are independent. This means that, if all possible values of the first measurement and all possible values of the second measurement were arrayed side by side in the order in which they were taken, any numerical value of one measurement would be equally likely to be paired with a given numerical value of the other. In other words, there must be no tendency for particular values of one measurement to be paired with particular values of the other. If such a tendency exists, and this is the case more often than might be supposed, the relations given by equations (23) and (24) must be modified to make proper allowance for the effects of this tendency. But the condition of independence is satisfied in a large variety of practical problems, and the mathematical relationships based thereon are of fundamental importance in the theory of sampling.

The formula for the variance of the sum of two quantities, as given by equation (23), can be extended to the case where more than two quantities are added. The variance of the sum of k measurements is

$$V_s = V_1 + V_2 + V_3 + \text{---} + V_k \qquad \text{- - -} \quad (25)$$

in which V_1, V_2, V_3, ---, V_k are the respective variances of the individual measurements. A special case of equation (25) arises when the individual measurements have the same variance. In that case, the equation reduces to the form,

$$V_s = kV \qquad \text{- - - - - - - - - -} \quad (26)$$

in which V is the variance of each of the k measurements entering into the sum.

Equation (26) will be used often in the present work. It is needed in a large number of practical problems, particularly those arising when samples are drawn from the same population. If measurements are taken from the same population, they will necessarily have the same variance. The variance of the sum of any number of such measurements can be computed from equation (26). Many statisticians regard equation (26) as one of the most basic formulas of statistics because it enters into so many practical problems.

The formulas relating to the variances of sums and differences of individual measurements also apply to the variances of the sums and differences of arithmetic means. When the variance of the individual measurements in a set of n_1 measurements is V_1 and the variance of the individual measurements in another set of n measurements is V_2, the variances of the two means will be $\frac{V_1}{n_1}$ and $\frac{V_2}{n_2}$, respectively, provided the two samples are drawn from infinite populations. The variance of the sum of difference of the two means is $\frac{V_1}{n_1} + \frac{V_2}{n_2}$. The variance of the sum or difference of two quantities is thus given by the sum of the variances of the quantities that are added or subtracted, regardless

of whether those quantities are individual measurements or averages of several measurements. As a matter of fact, the relationship is even more general because the variance of the sum or difference of any quantities is equal to the sum of the variances of those quantities. The quantities that are added or subtracted may be any statistical constants whatever. The only necessary condition is that those quantities be independent, as previously stated, and that the variances of those quantities be known.

This discussion is closed by calling attention to another property of variances that is frequently very useful and should be impressed upon the student's memory. If a quantity, x, has a variance, V, and that quantity is multiplied by a factor, A, the variance of Ax is equal to A^2V. This is a special case of a more general property called propagation of error, which is not fully discussed at this time. But the special case just mentioned is so important that it is desirable to call attention to it. In case the student fails to see why this property is true, he will find it helpful to consider what happens to the standard error and its square when the unit of measurement is changed. When a measurement is expressed in linear feet, for example, its standard error will also be expressed in those units. When the measurement is converted into inches, the measurement itself and its standard error will both be multiplied by 12. The variance, or the square of the standard error, will then be multiplied by 12^2 or 144. In this illustration, A is equal to 12 and A^2V thus becomes 144V. This illustration should be sufficient to satisfy the student that when a measurement is multiplied by a constant factor the variance of the original measurement must be multiplied by the square of that factor to obtain the variance of the product.

This property of variances is used to derive the formula for the variance of the mean of n measurements. When one has a sample of n measurements from the same population, the variances of the individual measurements are equal and may be represented by V. By equation (26), the variance of the sum of the n measurements is

$$V_s = nV \qquad \text{- - - - - - - - -} \qquad (27)$$

The mean of the n measurements is obtained by dividing their sum by n, which is equivalent to multiplying the sum by 1/n. The variance of the result will be $(1/n)^2$ times the variance of the sum, as shown in equations (28) and (29).

$$V_{\bar{x}} = (1/n)^2 V_s \qquad \text{- - - - - - - -} \qquad (28)$$

$$V_{\bar{x}} = (1/n)^2 nV = V/n \qquad \text{- - - - - - - -} \qquad (29)$$

Equation (29) shows the variance of the mean as previously given in equation (21). This result is the one that follows from the theory of sampling for infinite populations. The corresponding formula for samples drawn from a finite population can be obtained by making use of the additive property of variances.

When a sample of n measurements is drawn from a finite population of N measurements, the variance of a single measurement, as estimated by equation (20), is actually a figure that refers to a hypothetical infinite population, of which the given finite population of N measurements is itself a sample.

This is a theoretical concept that requires the exercise of the reader's powers of imagination. When the sample of n measurements is regarded as a sample from that hypothetical infinite population, the variance of the mean is V/n, as indicated by equations (21) and (29). When the finite population of N measurements is also regarded as a sample from the same hypothetical infinite population, the mean of all N measurements has a variance equal to V/N. The variance of the mean of the n measurements, considered as a sample from the infinite population, thus consists of two parts. The first component consists of the variation of the means of samples of n, each drawn from the N given measurements. The second component consists of the variation of the means for samples of N, drawn from the hypothetical infinite population. The first component is the one to which attention must be directed because it represents the variation of the means for samples of n measurements, each drawn only from the given set of N measurements. It is obtained by subtracting the quantity, V/N, from the total variation, V/n, and one thus obtains,

$$V_{\bar{x}} = V/n - V/N = V(1/n - 1/N) = V\left(\frac{N - n}{nN}\right) = \frac{V}{n}\left(\frac{N - n}{N}\right) \qquad \text{- - - - - (30)}$$

This is the result sought, as indicated previously by equation (22).

Those relations are important in themselves, but the concepts upon which they are based are still more so. Before continuing his studies in the theory of sampling, the reader should carefully review the discussion on the variability of individual measurements and averages. He should become thoroughly familiar with the variance notation and the general properties of variances. Unless this basic material is completely understood, he will find himself in difficulties later. He should be especially careful to observe the distinction between the population and a sample from that population and to recognize the difference in point of view when samples are drawn from finite populations as contrasted with infinite populations. It is more important to understand the relationships than to memorize the formulas.

Exercise 12.-The estimated variance of the mean of 10 measurements, drawn from a finite population of 50 measurements, is 18. Compute the estimated variance of an individual measurement.

Exercise 13.-The annual egg production of each of 100 hens, chosen at random from a flock of 600, was recorded. The variance of egg production for individual birds was computed and found to equal 900. (a) To compute the variance of the mean, should one use equation (21) or (22), or could one use either? (b) If equation (22) were used, how would the interpretation of the result differ from that which would be derived by applying equation (21)?

Exercise 14.-The average of a set of 10 measurements has a variance of 24. The average of a set of 20 measurements has a variance of 12. (a) Are the variances of the individual measurements equal in both samples or are they different? (b) Compute the variance of the sum of the two means. (c) Compute the variance of the sum of all 30 measurements. (d) The unweighted mean of two averages is computed by adding the two given averages and dividing the result by 2. Compute the variance of this unweighted average. (e) Compute the variance of the average of all 30 measurements, pooled and treated as one sample.

Pooled Variance and the Significance of the Difference Between Two Averages

By this time, the student should be sufficiently well acquainted with the concept of variance to permit some additional applications of the theory to specific problems in sampling. The problem of testing the significance of the difference between two averages arises frequently in practical work. The formulas for computing the variances of differences make these tests possible.

Such tests are a special case of the more general problems involved in testing a null hypothesis. In testing the difference between two means, the null hypothesis is nothing more than an assumption that the two means are merely two different estimates of the same quantity and that the observed difference between them is only a chance fluctuation caused by vagaries of random sampling. The null hypothesis may thus be characterized as the hypothesis that there is no difference between the true values of the two means in the population, or populations, from which the two samples were drawn. The difference between the two observed sample means is then compared with the standard error of that difference to determine whether the null hypothesis should be accepted or rejected. The hypothesis will be accepted if the comparison shows that the observed difference is likely to arise by chance and rejected if the difference is so large that it would be unlikely to arise by chance in situations where the null hypothesis was true.

The test of significance thus supplies information of a rather negative type. A practical man, unfamiliar with statistics, may think it absurd to test the hypothesis that no difference exists when he may have good reason to suspect in advance that there actually is a difference. To the statistician, it represents a mathematical test that is useful because it answers a specific question even though it has some shortcomings. It tells whether an observed difference is large enough so that it would be unreasonable to conclude that no actual difference exists in the populations sampled.

So far as mathematical details are concerned, the application of the test depends upon a knowledge of the frequency distribution of the ratio of the difference of the two means to the estimated standard error of that difference, under the hypothesis that the two samples are drawn from populations having the same means. This problem, surprisingly enough, presents more complications than one might expect. The frequency distribution of the ratio of the difference between two means to the standard error of that difference is known for the case where the two samples are assumed to be drawn from the same Normal population or what amounts to the same thing, from identical Normal populations.

This kind of test is obviously somewhat specialized because it involves something more than merely testing the difference between two means. It is the test most frequently used in practice, however, and will be discussed in detail at the present time. Before proceeding with the subject, it is necessary that the reader develop a broader point of view with respect to the concept of variance than has yet been presented.

Suppose one has obtained a sample of n_1 measurements and another sample of n_2 measurements. Let $\bar{x}_1$ represent the mean computed from the first sample

and let $\bar{x}_2$ represent the mean computed from the second sample. One could estimate the variance of a single measurement from each sample by applying equation (20) to the measurements in each sample. These two estimates, which may be V_1 and V_2, will usually not be exactly equal even if the two samples were taken from the same population. They will be two independent estimates of the same quantity, however, provided that the samples were drawn from the same populations or identical populations. Since V_1 and V_2 are merely two different estimates of the variance of a single measurement, one based on $n_1 - 1$ degrees of freedom and the other on $n_2 - 1$ degrees of freedom, it is possible to obtain a single estimate, V, of this quantity by taking an average of V_1 and V_2. This average will be a better estimate than either V_1 or V_2 taken separately and will, in addition, enable the statistician to avoid the confusion of working with two separate estimates. Since V_1 and V_2 are in this case based on different numbers of degrees of freedom, the average V must be a weighted average of V_1 and V_2. The weights to be used are the respective numbers of degrees of freedom from which V_1 and V_2 were computed. One thus obtains,

$$V = \frac{(n_1 - 1)V_1 + (n_2 - 1)V_2}{n_1 + n_2 - 2} \qquad \text{- - - - -} \quad (31)$$

The estimate, V, is often called the pooled variance for the two samples because the computations indicated by equation (31) are equivalent to obtaining the sum of the squares of the deviations of the individual measurements from the sample mean separately for each sample, adding the results, and dividing the sum by the combined or pooled degrees of freedom for the two samples. The student may find it easier to think of V as an average, however. If n_1 and n_2 happen to be equal, equation (31) reduces to

$$V = \frac{V_1 + V_2}{2} \qquad \text{- - - - - - -} \quad (32)$$

The reader should verify this as an exercise.

Since V represents the variance of a single measurement, the variances of the means, $\bar{x}_1$ and $\bar{x}_2$, are V/n_1 and V/n_2, respectively. One may then compute the difference, $\bar{d}$, between the two means, $\bar{x}_1$ and $\bar{x}_2$, together with the variance and standard error of that difference.

$$\bar{d} = \bar{x}_1 - \bar{x}_2 \qquad \text{- - - - - - - - - - -} \quad (33)$$

$$V_{\bar{d}} = V\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad \text{- - - - - - - - - - -} \quad (34)$$

$$s_{\bar{d}} = \sqrt{V_{\bar{d}}} \qquad \text{- - - - - - - - - - -} \quad (35)$$

The ratio, $\bar{d}/s_{\bar{d}}$, is distributed according to the t distribution discussed previously and thus the necessary probability tables for the test of significance are available. The number of degrees of freedom to be used in reading the t

table is equal to $n_1 + n_2 - 2$. This is the number of degrees of freedom used in estimating V. The formula for computing t may be written in any of the following equivalent forms,

$$t = \bar{d}/s_{\bar{d}} \qquad \text{- - - - - - - - - - - - - - (36)}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{V\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{- - - - - - - - - (37)}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{- - - - - - - - - (38)}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad \text{- - - - - - - - - (39)}$$

The reader will find it an instructive exercise to derive equations (37), (38), and (39) from equation (36). A little algebra is all that is required.

These formulas are based on the theory of sampling from infinite populations. The necessary modifications to be made for samples drawn from finite populations are not at all complicated. For finite populations one obtains,

$$V_{\bar{x}_1} = \frac{V}{n_1}\left(\frac{N_1 - n_1}{N_1}\right) \qquad \text{- - - - - - - - - (40)}$$

$$V_{\bar{x}_2} = \frac{V}{n_2}\left(\frac{N_2 - n_2}{N_2}\right) \qquad \text{- - - - - - - - - (41)}$$

$$V_{\bar{d}} = V\left(\frac{N_1 - n_1}{N_1 n_1} + \frac{N_2 - n_2}{N_2 n_2}\right) \qquad \text{- - - - - - - (42)}$$

$$s_{\bar{d}} = \sqrt{V_{\bar{d}}} \qquad \text{- - - - - - - - - - - - (43)}$$

and, as before,

$$t = \bar{d}/s_{\bar{d}} \qquad \text{- - - - - - - - - - - - (44)}$$

The utility of the t distribution as a test of significance has been questioned by some statisticians because it does not specifically test the significance of the difference between two means. A moment's reflection will show that there is some justification for such an argument. The estimate of the standard error of the difference between the two means is subject to sampling errors. The value of t is affected by these errors as well as by the difference between the two means. An unusually large value of t could thus arise, not only when the difference between the two means is unusually large, but also when the estimate of the standard error of that difference is unusually low. Furthermore, a fairly small value of t would not necessarily mean that the difference between the two means is small. The estimate of the standard error might be considerably too large. A significantly large value of t thus cannot be ascribed entirely to a large difference between the means nor can a small value of t be used as an indication of a small difference between those means. But the test is a fairly sound indication as to whether the two samples are drawn from the same population or from identical populations, and the reader should focus his attention on this aspect of the problem.

Attempts have been made to derive similar tests for problems in which the variance of an individual measurement is not the same in the two populations. The results to date have not been very satisfactory. If the variance of an individual measurement in the first population were V_1 and the variance of an individual measurement in the second population were V_2, the variance of the difference between the two means would be $\frac{V_1}{n_1} + \frac{V_2}{n_2}$. The standard error of the difference would be the square root of this quantity. The ratio of the difference between the two means to the above-mentioned estimate of the standard error of that difference does not follow any simple law of distribution. The best that can be done in such cases is to work with samples sufficiently large so that V_1 and V_2 can be regarded as reasonably accurate estimates of σ_1^2 and σ_2^2. The ratio of the difference to its standard error may then be assumed to follow the Normal frequency distribution and the significance of the difference between the two means can be tested by using tables based on the Normal Curve. At present, there seems to be no way of adapting the exact t test to problems of this kind except in a few special cases which are not of much interest to the practical statistician.

Analysis of Variance

The t test provides a method of testing the significance of the difference between two means, but it cannot be used to test the significance of the differences among three or more. For this purpose a more general test is required. The general solution of this problem resolves itself into what is known as Analysis of Variance. For the special case involving a comparison of only two means, the method of analysis of variance yields results identical with those given by the t test. It is becoming common practice to use the more general test in all such problems so that the same method is used throughout, regardless of the number of means that are to be compared.

The general principle underlying analysis of variance is fairly simple. The variability of the individual measurements in two or more samples is used to predict the variance of the means of those samples. The actual variance of those means is then computed by making use of equation (20), or a modification of it if the samples are not of the same size. The individual sample means are used as the values of x in this equation and the general mean for all samples is used as the value of $\bar{x}$. The number of samples, or the number of individual sample means to be compared, is used as the value of n in the formula. This procedure yields two estimates of the variance of the sample means, one predicted from the variance of the individual measurements and the other obtained by actually measuring the variability of those means. If there is no significant difference between the means, that is, if the differences between the means can be explained by fluctuations of random sampling alone, the two estimates will be approximately equal. On the other hand, when the means are significantly different, the observed variance of those means will be greater than the variance that is predicted from the variance of the individual measurements. This concept should be clear to the student before he attempts to use analysis of variance in practical work. It is comparatively simple, but it is so fundamental to work in this field that its importance can hardly be overestimated.

So far as the details of the test are concerned, the mechanics of computation differ slightly from the procedure indicated above. Instead of actually computing the predicted and observed variances of the means, it is more convenient to express these variances in terms of the variances of individual measurements. The observed variance of the individual measurements in the samples and the observed variance of the means of those samples is computed. From the observed variances of the means, one works back to compute the variance of individual measurements that would be required to account for the observed variances of the means. This second estimate of the variance of individual measurements is compared with the observed variance of the individual measurements, computed directly from the individual measurements in the samples. It is immaterial, from the standpoint of mathematical theory, whether this kind of comparison is made or whether predicted and observed estimates of the variance of the means are compared. The results will be identical. But there are certain advantages in making the comparison on the basis of individual measurements as just described, and it is customary to apply the test in that fashion. The various computations that must be made are fairly simple.

First, the variance of the individual measurements must be estimated by computing a pooled variance of individual measurements from all samples whose means are to be compared. Let this estimate be denoted by V_2. If there are n samples and the numbers of measurements in the various samples are k_1, k_2, ---, k_n, V_2 is given by the equation,

$$V_2 = \frac{SS\left[(X - \bar{x}_i)^2\right]}{S(k_i) - n} \qquad \text{- - - - - - - -} \qquad (45)$$

In this equation the double summation sign is used to indicate the pooled sum of squares of the deviations of the measurements from the sample means. In other words, the sum of the squares of the deviations of individual measurements from the sample mean is computed separately for each sample and the results are added. The quantity, $S(k_i) - n$, represents the pooled degrees of

freedom. The first sample contributes $k_1 - 1$ degrees of freedom, the second contributes $k_2 - 1$, and so on. Since there are n samples, the sum of all these degrees of freedom is $S(k_i) - n$. This is the total number of measurements in all samples minus the number of samples. The estimate V_2 may be regarded as the average variance of the individual measurements within the samples and involves nothing more than a simple extension of a principle already discussed in connection with equation (31) in the preceding section.

After obtaining V_2, which is the estimate of the variance of the individual measurements obtained from those measurements themselves, the other estimate must be computed from the observed sample means. This estimate is denoted by V_1 and may be computed from the equation,

$$V_1 = \frac{S\left[k_i(\bar{x}_i - \bar{\bar{x}})^2\right]}{n - 1} \qquad \text{- - - - - - - - -} \quad (46)$$

In this equation the $\bar{x}_i$ represent the n sample means and $\bar{\bar{x}}$ represents the mean of all measurements in the n samples. V_1 is the value that the variance of the individual measurements would have to assume in order to account exactly for the observed variance of the n sample means. If there were no significant difference between the means, V_1 would not differ significantly from V_2. But if the means were significantly different, V_1 would be significantly greater than V_2.

To decide whether V_1 is significantly greater than V_2, it is more convenient to work with the ratio, V_1/V_2, than with the difference, $V_1 - V_2$. This ratio is denoted by F. Its frequency distribution has been worked out, so all the necessary machinery for applying the test of significance is available. When V_1 and V_2 are approximately equal, F will be approximately equal to unity. If the sample means are significantly different, F will be larger than unity. The frequency distribution of F depends upon the number of degrees of freedom used in estimating V_1 and V_2. Comprehensive tables of the distribution have never been published, but the values of F which must be equaled or exceeded for significance have been tabulated for a large number of degrees of freedom. Tables of these critical values now are given in most textbooks. In practical problems one is thus able to determine whether an observed value of F is sufficiently large to indicate a significant difference between the sample means.

When only two means are compared, there is 1 degree of freedom for estimating V_1. In this case the F test will yield exactly the same results as the t test and many statisticians prefer to use it in such cases. If a number of tests of significance must be made, it is usually more convenient to use the F test throughout than to use the t test for comparisons involving two means and applying the F test to the others. As the reader gains some facility in the application of these tests, he will be likely to feel more and more inclined to use the F test in preference to the t test in problems where he has a choice. The computations are really easier to perform and their general nature is such that they appear as natural steps to more extensive analysis of the data.

This will be more apparent after the reader becomes more familiar with modern statistical techniques. For the present, the reader should be content to understand the meaning of the F test. As applied in practice, it involves only a comparison of the variance of the individual measurements in the samples, as computed directly from those measurements, with a hypothetical value which that variance would have to assume in order to account for the observed differences between the sample means.

It may be difficult for the student to understand why the estimate V_1, computed from equation (46), represents an estimate of the variance of individual measurements. The relationship is easier to visualize for the special case in which each of the samples contains the same number of measurements. In that case equation (46) reduces to

$$V_1 = \frac{k\, S\left[(\bar{x}_i - \bar{x})^2\right]}{n - 1} \qquad \text{- - - - - - - - -} \qquad (47)$$

where k represents the number of measurements in each sample. The quantity $\frac{S\left[(\bar{x}_i - \bar{x})^2\right]}{n - 1}$ represents the variance of the means of the n samples, or what amounts to the same thing, the variance of means for samples of k measurements each. The fundamental relation between the variance of means and the variance of individual measurements is indicated by equation (21). It shows that the variance of means for samples of k measurements is obtained by dividing the variance of the individual measurements by k. Conversely, given the variance of the means, the variance of the individual measurements may be computed by multiplying the variance of the means by k. Equation (47) thus involves nothing more than computing the variance of the means and multiplying that variance by the number of measurements in each sample to derive an estimate of the variance of the individual measurements.

Equation (46) could be derived on the same basis except that the problem would appear more complicated because the numbers of measurements in the various samples are unequal. It would be necessary to introduce the concept of weighting into the discussion and, for the present, it is better to hold that subject in abeyance. All that is required at this stage is a good understanding of the principles underlying equation (47) and a realization that equation (46) should be given a similar interpretation.

V_1 is often called the variance between means or variance between samples, but most statisticians like to refer to this quantity as the mean square between samples. There has frequently been a tendency for inexperienced workers to confuse V_1 with the variance of the means. The use of the term mean square helps to emphasize that V_1 is a quantity computed from the variance of the means but does not represent the variance of the means. It should not be forgotten that V_1 is actually an estimate of the variance of individual measurements that is computed from the observed variance of the means.

The estimate V_2 which represents the variance of the individual measurements computed directly from those measurements, is generally called the

mean square within samples to keep the terminology consistent. If the mean square between samples is significantly greater than the mean square within samples, it may be concluded that the means of the samples are significantly different.

This discussion closes with a short reference to an interesting property that is characteristic of analysis of variance, namely, the additive property of sums of squares. The quantity $S\left[k_i(\bar{x}_i - \bar{x})^2\right]$ in equation (46) is called the sum of squares between samples. Similarly, the quantity $SS\left[(X - \bar{x}_i)^2\right]$ in equation (45) is called the sum of squares within samples. The sum of these two sums of squares is equal to the sum of the squares of the deviations of the measurements for all n samples from the mean of all those measurements. This sum is called the total sum of squares. The relationship may be expressed algebraically by the equation,

$$S\left[k_i(\bar{x}_i - \bar{x})^2\right] + SS\left[(X - \bar{x}_i)^2\right] = S\left[(X - \bar{x})^2\right] \quad -- (48)$$

The corresponding degrees of freedom have a similar property. The degrees of freedom between samples is n - 1. The degrees of freedom within samples is $S(k_i) - n$. The sum of these is equal to $S(k_i) - 1$, which is the total number of degrees of freedom or one less than the total number of measurements. These relationships are summarized in table 4 in much the same way that the results of an analysis of variance on actual data are presented.

Table 4. - Structure of an analysis of variance table.

Source of variability	Degrees of freedom	Sum of Squares	Mean square
Between samples	$n - 1$	$S\left[k_i(\bar{x}_i - \bar{x})^2\right]$	$\dfrac{S\left[k_i(\bar{x}_i - \bar{x})^2\right]}{n - 1}$
Within samples	$S(k_i) - n$	$SS\left[(X - x_i)^2\right]$	$\dfrac{SS\left[(X - x_i)^2\right]}{S(k_i) - n}$
Total	$S(k_i) - 1$	$S\left[(X - \bar{x})^2\right]$	$\dfrac{S\left[(X - \bar{x})^2\right]}{S(k_i) - 1}$

So far as the degrees of freedom are concerned, it is easy to see why the additive property holds true. It is more difficult to verify the corresponding relationship for the sums of squares, but this can be done by algebraic manipulation without much effort. If the reader is interested in mathematics, the verification of equation (48) will provide him with an instructive exercise.

The total mean square shown in table 4 is not required for testing the significance of the difference between the sample means. It has been inserted in the table at this time merely to complete the picture. But it will be required in some applications of analysis of variance that will be discussed later.

Exercise 15.-From a sample of 15 measurements, the variance of a single measurement was estimated to be 8.26. A similar estimate from a second sample of 10 measurements was 9.04. Under the assumption that the two samples were drawn from identical populations, compute (a) the pooled or average variance for the two samples; (b) the variance of the mean for each sample; (c) the variance of the difference of the means for the two samples.

Exercise 16.-Suppose that the arithmetic mean for the first sample in Exercise 15 is 12.45 while that for the second sample is 7.18. Compute the value of t by equation (39) or its equivalent.

Exercise 17.-From the data given in Exercise 15 and 16 construct an analysis of variance table, as indicated by table 4, by computing the various degrees of freedom, sums of squares, and mean squares.

Exercise 18.-Compute the value of F for the analysis of variance prepared in Exercise 17. Compute $\sqrt{F}$ and compare the result with the value of t obtained in Exercise 16. How do they compare? This relationship is always true when there is only 1 degree of freedom between samples and shows why either the F test or the t test can be used in problems of this kind.

Short Cuts in Computation

The arithmetical work required in most statistical investigations is one of the necessary evils with which the statistician must contend. Any device that can be invented to reduce this overhead should be put to use immediately. Ingenious computers soon learn to eliminate unnecessary steps in an analysis wherever possible and this kind of activity should be encouraged. Many such devices have been discovered and are in general use at the present time. One of these, which relates to the computation of standard errors and the various sums of squares involved in an analysis of variance, may be discussed to good advantage at this point.

As a step in the computation of the variance of individual measurements, it is necessary to compute the sum of squares, defined by $S\left[(X - \bar{x})^2\right]$, which represents the sum of the squares of the deviations of the individual measurements from their arithmetic mean. The numerical value of this expression could be computed by subtracting the mean from each individual measurement, squaring each of these deviations, and adding the results. But the same result could be achieved more easily by making use of the relationship,

$$S\left[(X - \bar{x})^2\right] = S(X^2) - \left[S(X)\right]^2/k \quad \text{- - - - - - - - - - - - - -} \quad (49)$$

In this equation, $S(X^2)$ represents the sum of the squares of the individual measurements, $S(X)$ represents the sum of those measurements, and k represents the number of measurements. The mathematically minded student should derive this equation as an exercise. The derivation is fairly simple and anyone familiar with ordinary algebra should have no difficulty in verifying the relationship.

Equation (49) shows that the sum of the squares of the deviations of the individual measurements from their arithmetic mean can be computed without actually obtaining the deviations of the individual measurements from the mean. All that is necessary is to square the measurements themselves and to subtract from the sum of the measurements the correction term obtained by squaring the sum of the measurements and dividing by the number of measurements. In addition to eliminating the need for computing the individual deviations, this procedure is actually more accurate because errors involved in rounding the mean to a given number of decimal places are avoided.

When the individual measurements are expressed in terms of large numbers, the squares of the measurements may be awkwardly large. In such cases, it is often desirable to code the data by subtracting a constant from each measurement before applying equation (49). Such an adjustment has no effect whatever on the numerical value of $S(X - \bar{x})^2$ because the value of $\bar{x}$ changes with the adjustment in the same manner as the values of X change. Thus a new set of data is provided which can be manipulated more easily without any loss of precision.

Equation (49) can be adapted to the computation of the various sums of squares in an analysis of variance without much difficulty. If there are n samples and the numbers of measurements in those samples are k_1, k_2, ---, k_n, the total sum of squares is given by the equation,

$$S\left[(X - \bar{x})^2\right] = S(X^2) - \left[S(X)\right]^2/S(k_i) \quad - - - \quad (50)$$

In this equation, $S(X^2)$ represents the sum of the squares of the measurements in all samples combined, $S(X)$ represents the sum of those measurements, and $S(k_i)$ represents the total number of measurements.

The sum of squares within samples is obtained by applying equation (49) to each sample separately and adding the results. The final result can be expressed by the equation,

$$SS\left[(X - \bar{x}_i)^2\right] = S(X^2) - S\left\{\left[S(X_i)\right]^2/k_i\right\} \quad - - - \quad (51)$$

Equation (51) looks somewhat complicated, but the reader should not let the algebraic notation frighten him unduly. As stated previously, equation (51) represents the final result of applying equation (49) separately to each sample and then combining the n sums of squares into a total. This total can be expressed as the difference between the sum of the squares of the individual measurements in all samples and the sum of the n separate correction terms of the form, $S(X_i)^2/k_i$, as shown in equation (51).

The sum of squares between samples is given by the equation,

$$S\left[k_i(\bar{x}_i - \bar{x})^2\right] = S\left\{\left[S(X_i)\right]^2/k_i\right\} - \left[S(X)\right]^2/S(k_i) \quad - - - - \quad (52)$$

It should be observed that the first quantity in the right-hand member of this equation is identical with the quantity used as the correction term in equation (51). The correction term in equation (52) is identical with the correction term used in equation (50).

It is evident that much time can be saved by this scheme of computation because computations performed for one step in the analysis can also be used in succeeding operations. Furthermore, this method of computing can be organized on a systematic basis that makes the various operations easy to remember. The preferences of computers vary with respect to the order in which the various operations are performed, but the student should learn to follow a definite routine of some kind in performing these computations. One satisfactory scheme is to compute the total sum of squares first of all. This involves computing $S(X^2)$ and $\left[S(X)\right]^2/S(k_i)$. The sum of squares between samples is computed next. This involves the computation of $S\left\{\left[S(X_i)\right]^2/k_i\right\}$ and the use of the correction term used in the preceding step. The sum of squares within samples can then be obtained merely by substracting the quantity, $S\left\{\left[S(X_i)\right]^2/k_i\right\}$, computed in the preceding operation, from $S(X^2)$ which has also been computed previously. As a check, the sum of squares between samples and the sum of squares within samples should be added to make sure that the sum agrees with the total sum of squares that was computed previously as the first step in the analysis. This check should always be performed, but it should be noted that it does not provide a complete check on the accuracy of the work. An error in the quantity, $S\left\{\left[S(X_i)\right]^2/k\right\}$, will introduce compensating errors into the sum of squares between samples and the sum of squares within samples so that the sum of the two will agree with the total sum of squares computed previously. The check <u>will</u> verify that the scheme of analysis was properly followed.

The allocation of the various degrees of freedom should also be performed systematically. The total number of degrees of freedom is one less than the total number of measurements in all samples. The degrees of freedom between samples is one less than the number of samples. To compute the number of degrees of freedom within samples it is necessary to remember that the degrees of freedom are computed separately for each sample and that these separate numbers are then combined into a pooled value. The number of degrees of freedom contributed by each sample is one less than the number of measurements in that sample. These degrees of freedom are merely added to arrive at the degrees of freedom within samples. When the various degrees of freedom have been computed, a useful check is provided by adding the degrees of freedom contributed by all sources and comparing the result with the total which was computed first. This kind of check enables the statistician to verify that he has kept his thinking straight.

Most experienced workers prefer to compute the degrees of freedom for an analysis of variance before computing the sums of squares. The process of assigning these degrees of freedom provides a convenient method of thinking the problem through before the heavy work is begun. The steps involved in the process form an outline for the succeeding computations that the statistician will find extremely helpful. This fact may not be so apparent for the simple analysis of variance described in the preceding section, but when the reader begins to work with more complicated problems it will be evident without further argument.

Exercise 19.-The following data are cotton yields (pounds per acre) reported by 30 farmers in three Arkansas counties on December 1, 1939.

Cleburne	Fulton	Izard
200	125	210
208	300	190
160	250	235
214	133	225
228	225	175
240	225	240
100	160	200
135	250	250
255	150	
175	210	
195	166	

Compute the analysis of variance outlined in table 4, using the short-cut methods described in this section. The data for each county can be regarded as a sample of measurements but it is more appropriate to speak of variation between counties and within counties rather than of variation between samples and within samples. Such terminology is more descriptive of the problem at hand. Are the mean yields for the 3 counties significantly different?

Application of Analysis of Variance to Sampling Problems

Analysis of Variance was originally developed for testing the significance of differences in experimental data, particularly in agronomy work. This method of analysis proved so useful in that field that it was applied to data in other fields at an early date. But its application to a comparison of the relative efficiencies of different sampling schemes is fairly recent. The current interest of statisticians in sample census methods has stimulated such applications considerably, and their value is now fully appreciated. In such applications, one is primarily interested in segregating and measuring the component parts of the total variability in data obtained with a particular sampling scheme. The analysis furnishes information that shows whether a better sampling scheme can be devised, and if so, what form it should take. There is more emphasis on measuring the magnitude of the variation than on testing the significance of differences in the data. Applications of analysis of variance in this field are thus conducted from a slightly different point of view than those ordinarily encountered in other types of research work.

To illustrate the utility of analysis of variance in sampling work, the following practical problem may be considered. A county in North Carolina contains 2,238 farms distributed over nine townships. An estimate of the acres of cropland per farm is required for the county, and this estimate is to be derived from an enumeration of a sample of farms from the county. The question to be answered is this: Should the sample be a random sample of farms from the county, or should the sampling be controlled so that some farms will be taken from every township? The answer to this question depends upon the way in which the cropland for individual farms varies from one township to another and from

farm to farm within each township. If the acreage of cropland per farm varies considerably from township to township, but is fairly constant from farm to farm within any one township, the controlled sample would yield better results than a random sample of the same size. On the other hand, if there is no greater difference between farms in different townships than between farms in the same township, no precision could be gained by using a controlled sample.

A preliminary sample can be taken as a means of obtaining sufficient information about the population to serve as a guide in planning the main survey. Suppose five of the nine townships are chosen at random and 20 farms, selected at random in each of these townships, are enumerated with the results shown in table 5.

Table 5. - Acres of cropland on 100 farms from five townships in a North Carolina county containing 2,238 farms distributed over nine townships, as indicated by 1939 State Farm Census.

Farm Number	Cropland				
	Township 1	Township 2	Township 3	Township 4	Township 5
	acres	acres	acres	acres	acres
1	18	34	10	17	12
2	2	21	110	169	92
3	33	24	16	30	18
4	37	26	16	25	74
5	25	13	24	23	7
6	66	30	8	52	17
7	17	21	38	13	3
8	11	36	32	41	19
9	100	20	68	45	6
10	14	26	70	63	17
11	28	21	32	24	42
12	19	39	19	54	12
13	44	40	4	35	11
14	20	55	26	73	37
15	29	35	35	36	73
16	24	6	14	22	10
17	1	42	9	19	24
18	12	5	21	38	24
19	24	13	24	23	36
20	3	84	27	30	22
Total	527	591	603	832	556

These data yield the analysis of variance shown in table 6. The reader should verify the computations as an exercise.

Table 6. - Analysis of variance of cropland on 100 individual farms equally apportioned among five townships in a North Carolina County.

Source of variation	Degrees of Freedom	Sum of Squares	Mean Square
Between townships	4	2,939	735
Within townships	95	62,377	657
Total	99	65,316	660

If the differences between farms in different townships were no greater than those between farms in any one township, the mean square between townships would be equal to the mean square within townships except for ordinary sampling fluctuations. Table 6 shows that the mean square between townships is slightly larger than the mean square within townships. This indicates that there are some consistent differences from township to township with respect to acreage of cropland on individual farms. The choice between controlled sampling and random sampling as a method of conducting a more extensive survey depends upon the relationship between the "within township" mean square and the "total" mean square. If the sampling were controlled so that observations would be taken at random only within townships, the sampling error of the final result would depend upon the mean square within the townships. If a random sample were taken from the county as a whole without regard to the particular townships in which farms were selected, the sampling error of the final result would depend upon the total mean square in the population.

The mean square within townships shown in table 6 was estimated from only five of the nine townships in the county, but it seems reasonable to suppose that this is a fair estimate of the average variability within all nine townships. The total mean square shown in table 6 does not represent an exact estimate of the total mean square for the entire population, but it serves as a good approximation in this case and in similar problems. Using 657 as an estimate of the mean square within townships and 660 as an estimate of the total mean square, the relative efficiency of a controlled sample, as compared with a random sample from the county as a whole, is 660/657 or 100.5 percent. In other words, a random sample would be 0.5 percent less efficient than a controlled sample of the same size. This difference is hardly large enough to justify the use of a controlled sample on statistical grounds. Unless there are some administrative advantages to be gained by regionalizing the sampling work, one might as well take a random sample of farms from the county as a whole.

The relative efficiency of a controlled sample can be estimated more accurately by actually estimating the total mean square for all farms in the county. As stated previously, this estimate will not differ greatly from the figure given in table 6, but it is of considerable theoretical interest. To derive such an estimate, a table like table 6 must be constructed for the county as a whole. The data in table 6 can be used to construct such a table, but first it is necessary to understand exactly what the various means squares in this table represent.

The mean square between townships, 735, was obtained by computing the variance of the five township means and multiplying that variance by the number of farms from each township, or 20. In other words, the observed variance of the township means is 735/20. This quantity may be regarded as consisting of two components. The first component, which may be designated by V_t, represents the actual variance of the true township means. The second component represents the sampling error introduced by the fact that each observed township mean is only an estimate of the true township mean derived from 20 of the farms in the township. The mean square within townships, which was estimated as 657 in table 6, represents the variance for individual farms in the same township and may be regarded as a measure of the sampling variance for an individual farm. The sampling variance of a mean, computed from data for 20

farms, is, therefore, 657/20. The value of V_t can thus be estimated by subtracting 657/20 from the variance of the observed township means, 735/20. The result is 78/20 or 3.9. These relations may be summarized by saying that the mean square between townships is an estimate of the quantity, $kV_t + V$, in which k is the number of farms from each township, V_t the variance of the true township means, and V the sampling variance of observed cropland for an individual farm.

The reader may wonder how this procedure could have been applied if the numbers of farms from the five townships had been unequal. In that case, the numbers of farms from the five townships would be represented by k_1, k_2, k_3, k_4, and k_5. The mean square between townships would then be an estimate of $k_0V_t + V$, where the value of k_0 would be given by the equation,

$$k_0 = \frac{1}{n-1}\left[S(k_i) - S(k_i^2)/S(k_i) \right] \qquad \text{- - - - - - -} \qquad (53)$$

in which n represents the number of townships in the sample. The value of k_0 may be regarded as a sort of average of k_i, but it will always be somewhat less than the arithmetic mean of the k_i unless all of the k_i are equal. If all the k_i are equal, the value of k_0 given by equation (53) will simply be equal to the number of farms in a single township.

In constructing the analysis of variance for the entire population, it is convenient to set up the skeleton of a table like table 6 and to record all of the data so far available. First it is necessary to enter the various degrees of freedom as shown in table 7.

Table 7. - Analysis of variance of cropland for all farms in a North Carolina county predicted from analysis of a sample.

Source of variation	Degrees of Freedom	Sum of Squares	Mean Square
Between townships	8	12,848	1,606
Within townships	2,229	1,464,453	657
Total	2,237	1,477,301	660

As there are nine townships in the county, the number of degrees of freedom between townships is equal to eight. As there are 2,238 farms in the county and a degree of freedom is deducted for each of the nine townships, the number of degrees of freedom within townships is 2,229. The total number of degrees of freedom is one less than the total number of farms, or 2,237.

The mean square within townships shown in table 6 can be accepted as an estimate of the average mean square within all nine townships in the county and may be recorded in the appropriate space in table 7. The mean square between townships will differ from the value given in table 6. The mean

square given in table 6 is an estimate of $kV_t + V$ in which k is equal to 20. The mean square to be entered in table 7 must be an estimate of $K_0V_t + V$ in which K_0 depends upon the number of farms actually present in the individual townships. The number of farms in each township, indicated by the 1939 State Farm Census, is shown in table 8.

Table 8. - Distribution of farms used in constructing table 7.

Township	Number of farms
1	205
2	77
3	497
4	214
5	227
6	255
7	220
8	276
9	267
Total	2,238

As the numbers of farms in the nine townships are unequal, the value of K_0 must be computed from the equation,

$$K_0 = \frac{1}{N - 1} S(K_i) - S(K_i^2)/S(K_i) \qquad - - - - - (54)$$

which is identical with equation (53) except that the various quantities entering into the equation are population, rather than sample, data. The numerical value of K_0 for the data at hand is 1/8 (2238 - 653178/2238) or 243.27.

The values of V_t and V were computed previously and found to be 3.9 and 657, respectively. The mean square between townships for the entire population, which is equal to $K_0V_t + V$, can now be computed. Its numerical value is (243.27)(3.9) + 657 or 1606.

The sum of squares between townships and within townships is computed from the mean squares and degrees of freedom by multiplication. Adding these two sums of squares yields an estimate of 1,477,301 for the total sum of squares for the entire population.

The total mean square for the entire population is obtained by dividing this last figure by the total degrees of freedom, or 2,237. The result is 660 which agrees perfectly with the corresponding value in table 6 to three significant figures.

This kind of result is frequently encountered in practice. The estimate of the total mean square for the entire population usually differs little from the total mean square obtained in the analysis of variance of a sample. It is generally advisable to compute the population value, however, because the bias in the sample value is sometimes large enough to justify the small amount of additional work involved.

The analysis of variance shown in table 7 is the analysis that would be expected if the individual farm data for the entire county were used in the computations. In this case the results were unusually good. An actual analysis performed on the 2,238 individual farm records yielded the results shown in table 9.

Table 9. - Analysis of variance of cropland for all farms in a North Carolina county computed from data for all farms in the county.

Source of variation	Degrees of Freedom	Sum of Squares	Mean square
Between townships	8	12,341	1,543
Within townships	2,229	1,486,427	667
Total	2,237	1,498768	670

The analysis in table 7 represents an inflation of the analysis of variance obtained for the sample and is consequently no more accurate than that analysis. It should not be used to test the significance of differences, its only purpose is to estimate the effects of the various sources of variability in the population as a whole. These usually have different weights in the population than in the sample. On the basis of the analysis in table 9, the relative efficiency of a controlled sample, as compared with a random sample of the same size, is 670/667 or 100.4 percent which agrees closely with the results obtained previously from the sample data. Ordinarily, one will find that the relative efficiency of a controlled sample, obtained from a predicted analysis of variance for an entire population, will not differ much from the value that would be obtained from an actual analysis of the entire population. Values like the estimated mean square between townships shown in table 7 are subject to large standard errors because the degrees of freedom associated with this mean square are usually small in relation to the total. The mean square between townships shown in table 7 agrees more closely with the corresponding true value in table 9 than would ordinarily be expected under conditions of random sampling. This is not a serious matter in estimating the relative efficiencies of different methods of sampling because the total mean square in tables like table 7 is of more interest than the mean square between townships. The mean square between townships could fluctuate over a fairly wide range without much effect on the estimate of the total mean square.

The preceding discussion should give the reader an indication of the value of analysis of variance in sampling work. An analysis of a small preliminary sample supplies the necessary information for the designing of a sampling scheme that is best adapted to a particular problem. Students are often surprised to learn that one kind of sample enables the prediction of the behavior of a different kind of sample, but as indicated above, there is nothing mysterious about the process. Possible sources of variability in the preliminary data are identified and measured. Once this has been done, it is fairly easy to compute the effects of these sources of variability upon a different kind of sample. It is important to remember that such results are only estimates, but they are exceedingly useful. Some caution must be exercised in applying them, particularly when some of the sources of variability do not appear to be statistically significant. Sometimes this means that such sources of variability should be ignored, but it might also mean that they should be measured more accurately.

Exercise 20.4 Convert table 5 into a table with unequal numbers of observations for the five townships by deleting the last two observations for township 1, the last observation for township 3, the last 5 observations for township 4, and the last 4 observations for township 5. Compute (a) the analysis of variance corresponding to table 6, (b) the values of k_o and V_t, and (c) the predicted analysis of variance for the entire county corresponding to table 7.

Some General Principles of Sampling

Before proceeding with the mathematical analysis of sample data, it is desirable to summarize the essential features of several sampling schemes that have been used by statisticians. As stated previously, a sample is drawn from a population and studied to obtain information about the population. The sample is usually of little interest in itself. Consequently, research workers strive to draw the sample in such a way that its characteristics will resemble the characteristics of the population as much as possible. Research workers often refer to a "random sample," a "representative sample," an "adequate sample," or a "fair sample," in an attempt to embody this fundamental concept into a single descriptive term, but at times there appears to be some lack of understanding of the mathematical definitions that should be kept in mind.

Apparently, a good deal of misunderstanding is prevalent in regard to the mathematical definition of a random sample. Some workers confuse the word "random" with "representative." A random sample is defined as a sample taken in such a way that every individual in the population has an equal chance of being included. Nothing in this definition gives assurance that a particular random sample will be representative; in fact, a random sample may sometimes be far from representative. The noteworthy feature of a random sample is that it is likely to be representative, which is different from saying that it is always representative. In taking a random sample of 100 farms from a county, for example, it is possible to get either the 100 largest or the 100 smallest in any one sample; but some of each can usually be expected.

When random samples are drawn repeatedly from the same population, the aggregate of a large number of samples is more likely to be representative of the population than any one of the individual samples. Thus the statement that, in the long run, random samples will tend to be representative is justified. Confidence in the result of the sampling would increase as the size of such samples, or the number of samples, increased.

It should be noted that, in random sampling, individuals are taken from a population without any attempt to force the sample to be representative. The tendency of random samples to be representative is inherent in the method of sampling itself.

The sampling errors caused by the failure of some individual random samples to be representative become troublesome when it is necessary to draw conclusions about a population from a single small sample or a few such samples. Research workers learned at an early date that such sampling errors could be reduced if samples were drawn in such a way as to enforce some similarity

between the characteristics of the sample and the population. Such a procedure presupposes some advance information about the population so that the worker has some knowledge in regard to what would constitute a representative sample.

The general procedure in stratified sampling is to divide the population into subregions or "strata" in such a way that the differences between individuals in the same stratum are as small as possible while the differences between the strata are as great as possible. The sampling can then be controlled in such a way that a predetermined number of individuals is taken at random from every stratum. If the differences between individuals from different strata are greater than those between individuals from the same stratum, such a sample can be made more representative of the population than a random sample of the same size. The different types of individuals segregated by the process of stratification can be included in the sample in their proper proportions.

In taking a sample of farms from a county, for example, the county may be stratified by townships to good advantage if the differences between farms in different townships are greater than the differences between farms in the same township. This is often the case because differences in type of farming are generally associated with differences in location. Such regionalization of the sampling work often possesses administrative advantages in addition to the gain in accuracy. From the standpoint of statistical precision, the sampling errors of results derived from stratified samples are smaller than corresponding sampling errors for random samples of the same size. Such errors depend only on the variability within strata which should be less than the variability for the population as a whole because the strata are chosen in such a way that a large portion of the total variability has been removed from the estimate of error.

The degree to which this kind of control can be exercised is limited only by the extent of information about the population that is available before the sample is drawn. If information regarding type of farm is available, for example, all the farms in a county could be grouped by type within every township and thus a double stratification of the population would be provided. The process could be continued almost indefinitely.

This method of sampling can be extremely useful, but there is an element of danger in using it because the research worker may be mistaken in regard to some of his preliminary ideas about the population. Such misconceptions may lead to a bias in the final result which is constant and cannot be reduced by increasing the size or number of the samples as random sampling errors are reduced. A bias would arise if each part of the population was not sampled in its proper proportion. In some types of work it seems preferable to risk the occurrence of a possible bias when it appears that this bias will be smaller than the sampling error that could be expected if the sampling were random. Such a point of view does not have much justification in experimental work where the interest lies in testing the significance of the differences between samples, but it is a logical position to take when the interest is in using a sample to obtain descriptive information about a population that is of interest in itself.

Subsampling may be described as a special case of stratified sampling. The number of classes or strata to be sampled may be so large that available facilities do not enable the taking of samples from all strata as in ordinary stratified sampling. If it is found necessary to reduce the scope of the sampling without sacrificing all of the benefits of stratified sampling, it is often desirable to choose some strata at random and to take random samples from those strata only. This method of sampling may also be substituted for random sampling for administrative reasons. The restriction of sampling to a limited number of strata often produces economics such as a reduction in the travel and supervision required for assembling the data.

The data reported in table 5 illustrate a problem in subsampling. The nine townships in the county provided a geographic stratification. A sample of farms from the five townships chosen at random could always be enumerated with less effort than a stratified sample involving all nine townships. Random samples of farms from the county as a whole could be expected to cover more than five townships most of the time and would also involve more work than the subsampling scheme.

When the variation between strata is large in relation to the variation within strata, subsampling may not give as accurate results as a complete stratified sample or a random sample. Subsampling is used in preference to other methods of sampling mainly for administrative reasons. This kind of sampling is recommended whenever the variation of the individuals within the larger units sampled is not small in relation to the differences between the larger units. This should be borne in mind when a sampling scheme is to be chosen.

If it is desired to attach a standard error to an estimate of a population characteristic computed from a sample, it is necessary to retain an element of randomness in the sampling scheme. In other words, the sampling should not be completely controlled. In stratified sampling this element of randomness is achieved by taking individual observations at random within the classifications established by the stratification. This principle is violated in sampling schemes that follow a systematic design so that each observation in the sample is selected according to some fixed rule.

An example of systematic sampling may be found in surveys of farms in which every tenth farm along a road is enumerated. Such samples will usually yield unbiased estimates of arithmetic means, so long as the starting point is chosen at random, but no accurate estimates of standard errors can be obtained from individual samples taken in this way. Some recent research indicates that standard errors may be estimated from special kinds of systematic samples drawn from some populations, but as a general rule, it is at present feasible to estimate standard errors for systematic samples by empirical methods only. If repeated systematic samples are drawn from the same population, the observed variation in the arithmetic mean from sample to sample can be used to estimate the precision of such samples for the particular kind of population sampled. Such an estimate obviously cannot be obtained from a single sample.

Systematic sampling has a strong appeal for many research workers because this kind of sample insures a good spacing between the individual sampling units. The possibility of many individual sampling units being taken from a

particular part of the population, with a corresponding lack of coverage of other parts of the population, is thus avoided. For this reason, systematic samples are frequently more representative of the population than random samples. If one is primarily interested in estimating arithmetic means, there is no reason why systematic sampling should not be used so long as the starting point is taken at random. But the difficulty of estimating standard errors from such samples should not be forgotten.

A practical sampling scheme known as double sampling has been studied rather thoroughly and seems to be useful in some types of work. If information is wanted about a population characteristic, which may be represented by y, and that characteristic is difficult to measure, it may be preferable to measure a characteristic, x, that is correlated with y and is easier to measure. The relationship between y and x can be determined from a small sample in which values of both are obtained. Then, if a large sample is used to get an accurate estimate of x for the population, a corresponding estimate of y can be computed from the relation between y and x determined previously from a small sample.

This method of sampling has many applications to economic surveys. For example, it might be difficult to get information regarding a farmer's income from the sale of hogs by direct questioning because of a natural tendency on the part of some farmers to be reticent about their income. Furthermore, some farmers might not remember the exact amounts received. Information regarding the number of hogs sold could be obtained more easily and accurately. If the relationship between income from the sale of hogs and the number sold were known, the desired income data could be computed. This relationship could be established by getting information on both items from a few farmers who were willing and able to supply it.

Double sampling is also useful when it is necessary to learn the relative numbers of individuals in various strata for weighting purposes. In such cases a large sample is taken to determine those numbers, but only a small fraction of it need be studied in detail to derive the other information sought. The entire sample is used only to determine the weights that should be applied to derive an unbiased average. For example, we might take a large sample of farms to measure the relative numbers of livestock, dairy, and field-crop farms in a State and then investigate labor requirements for a smaller sample of farms. The larger sample gives information on the relative numbers of farms of each type so the types can be properly weighted in the sample to give an unbiased average of the labor requirements per farm in the State.

Random Sampling

The process of taking a random sample from a particular population is more difficult than one might suppose. The use of tables of random numbers is a helpful device that is being employed by most statisticians at the present time. The individual sampling units in the population are numbered consecutively and reference to a table of random numbers provides one with a selection of sampling units that is free from bias. If a random sample of 100 from a county containing 2,238 farms was wanted, such a table of random numbers

would enable the selection to be made without difficulty, provided a list of all farms in the county, such as a tax assessor's list, were available. When no such list is available, the problem is more complicated. In such cases, the tables of random numbers may be used for the selection of points on a map at random. The farms located at these random points will constitute a good approximation to a random sample of farms, provided the density of the points available for selection is proportional to the number of farms in each part of the county.

Random samples are usually recommended when information about the population is insufficient to permit stratification or when it is known in advance that stratification by various criteria would not reduce the sampling errors in the final results. Random samples will give unbiased estimates of arithmetic means and will permit the obtaining of estimates of the variances of those means. The main disadvantage of random sampling is the comparatively large sampling errors that are usually found in results obtained by this method when there is much variability within the population.

The variance of a mean for a random sample can be estimated very easily. When the number of individuals in the sample is small in relation to the number in the population, the formula, $V_{\bar{x}} = V/n$, gives a good approximation. For larger samples use should be made of the more exact formula, $V_{\bar{x}} = \frac{V}{n}\left(\frac{N - n}{N}\right)$. Figure 13 presents a comparison of the results given by these two formulas in estimating the variance of the average cropland per farm in a North Carolina county for samples of different sizes. The data required for the construction of these charts were obtained from table 9. V is equal to 670 and N is equal to 2,238. It is evident that the estimate of the variance of an average is too large when the population is assumed to be infinite. This error becomes relatively more important as the size of the sample approaches 100 percent of the population. The exact formula is so easy to use that it seems desirable to avoid such errors whenever sufficient information about the population is available to do so.

Stratified Sampling

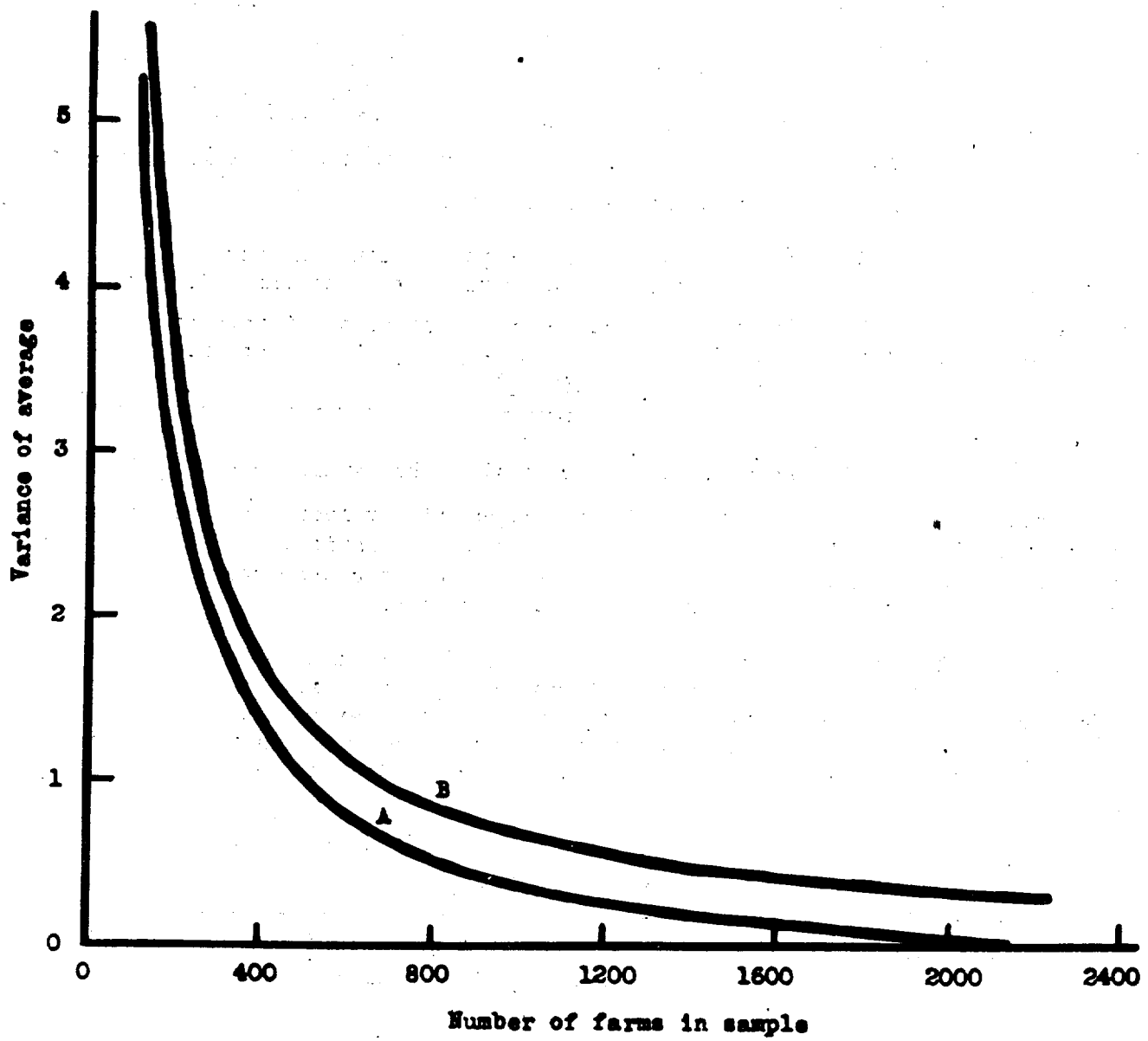
In practice it is usually possible to make use of some form of stratification. In drawing a sample of farms from a county, for example, the civil divisions of the county provide a convenient basis for stratification that can nearly always be used to good advantage. Such stratification is desirable for administrative reasons and will also provide more accurate estimates than random samples in many cases. Farms in the same location tend to be more nearly alike than farms in different locations because type of farming generally varies from one geographic area to another.

In making use of stratified samples, it should be remembered that each stratum is sampled at random so that a separate random sample of observations is obtained from each stratum. When an arithmetic mean is computed from the combined data for all strata, it is necessary to give each stratum its proper weight in order to arrive at an unbiased estimate of the true population mean. The true mean of the individual strata may not be equal to each other.

Figure 13. Variance of average acres of crop land per farm for random samples of farms from a county containing 2238 farms.

Curve A: Correct value

Curve B: Assuming infinite population



Therefore, the true mean for all strata combined is a weighted mean of the individual stratum means. This weighted mean can be represented by the equation,

$$m = \frac{S(K_i m_i)}{S(K_i)} \qquad \text{- - - - - - - - -} \qquad (55)$$

in which the K_i represent the numbers of sampling units in the various individual strata, the m_i represent the true values of the corresponding stratum means, and m represents the weighted mean of the individual stratum means. Equation (55) shows that m may be computed by multiplying each stratum mean by the number of sampling units in the stratum, adding the products, and dividing the result by the total number of sampling units in all strata. This procedure would be equivalent to computing the arithmetic mean of all observations in the population because each product of the type, $K_i m_i$, merely represents the sum of all observations in a single stratum and the expression, $S(K_i m_i)$, represents the grand total of the observations in all strata.

When a stratified sample is taken from a population, the numerical values of the K_i should be known in advance. The sample data from each stratum provide estimates of the individual stratum means. The best estimate of the population mean that can be made from the data is given by the equation,

$$\bar{x} = \frac{S(K_i \bar{x}_i)}{S(K_i)} \qquad \text{- - - - - - - - -} \qquad (56)$$

which is identical with equation (55) except that the sample means, $\bar{x}_i$, for the various strata are substituted for the population values. $\bar{x}$ represents the best estimate of the true mean for the entire population. $\bar{x}$ is an unbiased estimate of the weighted population mean, m, because the observed mean for each stratum is given its proper weight in the computations.

As an illustration of this procedure, consider the data in table 5. The best estimate of the average acreage of cropland per farm for the five townships would be a weighted average of the five township averages. The weight to be applied to each township average would be given by the number of farms in each township. The five township averages and their weights are shown in table 10.

Table 10. - Township weights and average acres of cropland per farm, based on samples of 20 farms per township.

Township	Farms in Township	Cropland per farm indicated by sample of 20 farms
	number	acres
1	205	26.35
2	77	29.55
3	497	30.15
4	214	41.60
5	227	27.80
Total	1,220	

Applying equation (56) to these data, the average acres of cropland per farm for each township is multiplied by the number of farms in the township. The sum of the five products is 37,875. Dividing this figure by the total number of farms gives an estimate of 37875/1220 or 31.05 for the weighted mean, $\bar{x}$.

If the sample were drawn in such a way that the number of sampling units taken from each stratum were proportional to the number present, the sample would be self-weighting. In such cases the arithmetic mean of all observations in the sample would be an unbiased estimate of the population mean. This can be proved very easily. If the number of sampling units taken from each stratum were proportional to the number present, the number taken from each stratum would be represented by k_i in the equation,

$$k_i = aK_i \qquad \text{- - - - - - - - - - - - - -} \qquad (57)$$

In this equation K_i represents the number of sampling units present in a stratum and a represents the fraction taken in the sample. Solving equation (57) for K_i, one obtains

$$K_i = \frac{1}{a} k_i \qquad \text{- - - - - - - - - - - - - -} \qquad (58)$$

When the quantity, $\frac{1}{a} k_i$, is substituted for K_i in equation (56) the equation can be written in the form,

$$\bar{x} = \frac{S(\frac{1}{a} k_i \bar{x}_i)}{S(\frac{1}{a} k_i)} \qquad \text{- - - - - - - - - - - - - -} \qquad (59)$$

or

$$x = \frac{\frac{1}{a} S(k_i \bar{x}_i)}{\frac{1}{a} S(k_i)}$$

Dividing numerator and denominator of the right-hand member of this equation by 1/a gives the required result,

$$\bar{x} = \frac{S(k_i \bar{x}_i)}{S(k_i)} \qquad \text{- - - - - - - - - - - - - -} \qquad (60)$$

The variance of the weighted mean computed by equation (56) is given by the equation,

$$V_{\bar{x}} = \frac{S(k_i^2 V_{\bar{x}_i})}{S(K_i)^2} \qquad \text{- - - - - - - - - - - - - -} \qquad (61)$$

in which $V_{\bar{x}_i}$ represents the variance of the i-th stratum mean. If the variance of the individual observations is the same for all strata, the variance of an individual stratum mean is given by the equation,

$$V_{\bar{x}_i} = \frac{V}{k_i}\left(\frac{K_i - k_i}{K_i}\right) \quad \text{- - - - - - - - - - - - (62)}$$

in which V is the variance of the individual observations within strata, K_i is the number of sampling units in the i-th stratum, k_i is the number of sampling units taken from the i-th stratum to estimate the stratum mean, $\bar{x}_i$.

For the data in table 10, an estimate of V is provided by the mean square within townships in table 6. The variances of the five individual township means in table 10 are

$$V_{\bar{x}_1} = \frac{657}{20}\left(\frac{205 - 20}{205}\right) = 29.64$$

$$V_{\bar{x}_2} = \frac{657}{20}\left(\frac{77 - 20}{77}\right) = 24.32$$

$$V_{\bar{x}_3} = \frac{657}{20}\left(\frac{497 - 20}{497}\right) = 31.53$$

$$V_{\bar{x}_4} = \frac{657}{20}\left(\frac{214 - 20}{214}\right) = 29.78$$

$$V_{\bar{x}_5} = \frac{657}{20}\left(\frac{227 - 20}{227}\right) = 29.96$$

The variance of the weighted mean, $\bar{x}$, estimated by equation (61) from these data, is

$$V_{\bar{x}} = \frac{(205)^2(29.64)+(77)^2(24.32)+(497)^2(31.53)+(214)^2(29.78)+(227)^2(29.96)}{(1220)^2} = 8.12$$

It is important to note that this estimate is only a measure of the accuracy with which the weighted mean of the five township averages was computed. As there are more than five townships in the county, this variance cannot be interpreted as a measure of the accuracy with which the weighted mean of the five township averages represents the mean for the entire county. Such an estimate would have to include a component introduced by the variation of the true township averages because the five townships were only a sample of all townships in the county.

The reader should note that the accuracy with which the weighted mean for the five townships has been estimated depends, not only upon the total number of farms in the sample, but also upon the way those farms were apportioned among the townships. The sample used in the preceding computations consisted of 20 farms from each of the five townships. A different number of farms from each township would also yield an unbiased estimate of the mean for the five townships, but the variance of that estimate would be different, even though the total number of farms from all townships remained equal to 100.

In practice, it is usually desirable to apportion the sample in such a way that the variance of the weighted mean, $\bar{x}$, will be as small as possible. If the variance of the individual observations within strata is the same for all strata, the most efficient sample is obtained by making the number of sampling units from each stratum proportional to the number present in that stratum. For the example discussed previously, the 100-farm sample constitutes 100/1220 or 8.197 percent of all farms in the five townships. Under the assumption that the variance of cropland for individual farms within each township is equal to 657, the most efficient sample would be obtained by taking 8.197 percent of the farms in each township. To the nearest whole number, the farms in the sample should be allocated according to the scheme,

$$k_1 = (0.08197)(205) = 17$$

$$k_2 = (0.08197)(77) = 6$$

$$k_3 = (0.08197)(497) = 41$$

$$k_4 = (0.08197)(214) = 17$$

$$k_5 = (0.08197)(227) = 19$$

$$S(k_i) = (0.08197)(1220) = 100$$

If this allocation of farms had been used instead of the one given in table 5, the variances of the five township means would have been

$$V_{\bar{x}_1} = \frac{657}{17}\left(\frac{205 - 17}{205}\right) = 35.45$$

$$V_{\bar{x}_2} = \frac{657}{6}\left(\frac{77 - 6}{77}\right) = 100.97$$

$$V_{\bar{x}_3} = \frac{657}{41}\left(\frac{497 - 41}{497}\right) = 14.70$$

$$V_{\bar{x}_4} = \frac{657}{17}\left(\frac{214 - 17}{214}\right) = 35.58$$

$$V_{\bar{x}_5} = \frac{657}{19}\left(\frac{227 - 19}{227}\right) = 31.69$$

The variance of the weighted mean would have been

$$V_{\bar{x}} = \frac{(205)^2(35.45)+(77)^2(100.97)+(497)^2(14.70)+(214)^2(35.58)+(227)^2(31.69)}{(1220)^2} = 6.03$$

Using the most efficient allocation of farms, instead of taking 20 farms from each township, would have resulted in an estimate of $\bar{x}$ with a variance of 6.03 instead of 8.12. In problems of this kind, proportional sampling thus results in a more accurate estimate of the weighted average in addition to simplifying the computation of that average as indicated previously.

The mathematical analysis of stratified samples has so far been discussed under the assumption that the variance of the individual observations is the same in all strata. Whenever this assumption is not justified, the analysis of the data should be modified accordingly. When the variances within the strata do not differ much, the error introduced by using the methods previously described is so small that it may be neglected. When such differences are large, it is desirable to use methods of analysis that take those differences into account. No satisfactory method has yet been developed for adapting analysis of variance to such data. The other computations can be made to conform to the requirements of the data without difficulty.

For illustrative purposes, consider the data in table 5. These data have been treated as though the variance of cropland for individual farms was the same within all townships. But when a separate estimate for each township is actually computed, the following results are obtained:

$$V_1 = 532$$

$$V_2 = 323$$

$$V_3 = 657$$

$$V_4 = 1153$$

$$V_5 = 618$$

These variances differ sufficiently to warrant the conclusion that cropland for individual farms is more variable in some townships than in others. Under this hypothesis, the most efficient allocation of a total sample of 100 farms would be obtained by making the number from each township proportional to the product of the number of farms in the township and the standard error of cropland for individual farms in that township.

The total number of farms in each township is given in table 10. The corresponding standard error of cropland for individual farms in each township can be obtained by extracting the square roots of the 5 variances given above.

$$s_1 = \sqrt{532} = 23.07$$

$$s_2 = \sqrt{323} = 17.97$$

$$s_3 = \sqrt{657} = 25.63$$

$$s_4 = \sqrt{1153} = 33.96$$

$$s_5 = \sqrt{618} = 24.86$$

The number of farms to be taken from each stratum should be proportional to the quantity $K_i s_i$. The most convenient way to compute this number of farms is to compute a product of the type, $K_i s_i$, for each township. This product is divided by the sum of all products of that type and the result is multiplied by the total number of farms to be taken from all townships. In mathematical

language, the number of farms to be taken from each township is given by the equation,

$$k_i = n \frac{K_i s_i}{S(K_i s_i)} \qquad \text{- - - - - - - - -} \quad (63)$$

where n is the total number of farms to be taken from all townships. The necessary computations for apportioning a sample of 100 farms among the five townships at hand are given in table 11.

Table 11. - Allocation of a sample of 100 farms among five townships on the basis of the number present and standard error of cropland for individual farms

Township	Farms in Township K_i	Standard error of cropland s_i	$K_i s_i$	$\frac{K_i s_i}{S(K_i s_i)}$	Farms in sample k_i
	number	acres			number
1	205	23.07	4729	0.1489	15
2	77	17.97	1384	.0436	4
3	497	25.63	12738	.4011	40
4	214	33.96	7267	.2288	23
5	227	24.86	5643	.1777	18
Total	1220		31761	1.0001	100

The allocation of farms given in the last column differs slightly from the result obtained by making the number of farms from each township proportional to the number present in the township. The precision of the weighted average should be computed from equation (61) because the number of farms present in each township is still used to compute the estimate of the weighted average from the individual township averages. The variances of the five township averages are

$$V_{\bar{x}_1} = \frac{532}{15}\left(\frac{205 - 15}{205}\right) = 32.87$$

$$V_{\bar{x}_2} = \frac{323}{4}\left(\frac{77 - 4}{77}\right) = 76.56$$

$$V_{\bar{x}_3} = \frac{657}{40}\left(\frac{497 - 40}{497}\right) = 15.11$$

$$V_{\bar{x}_4} = \frac{1153}{23}\left(\frac{214 - 23}{214}\right) = 44.74$$

$$V_{\bar{x}_5} = \frac{618}{18}\left(\frac{227 - 18}{227}\right) = 31.61$$

The variance of the weighted average is

$$V_{\bar{x}} = \frac{(205)^2(32.87)+(77)^2(76.56)+(497)^2(15.11)+(214)^2(44.74)+(227)^2(31.61)}{(1220)^2} = 6.21$$

This estimate is slightly larger than the value 6.03 obtained for the most efficient allocation under the assumption of equal variances within townships. But the difference is not large and shows that the assumption of equal variances used previously did not introduce any serious error. It is only under rather extreme conditions that much concern need be felt about differences in variability within the strata. In most practical problems use of an average value is justified. It is fortunate that this is the case because the assumption of equal variances usually simplifies the statistical analysis.

The above discussion covers the essential mathematical principles underlying stratified sampling. The student should notice particularly that something must be known about the nature of the variability in a population before these principles can be applied. After a sample has once been taken, an analysis of that sample will yield the information required to design an efficient sampling scheme to be used in future work. The examples given in the preceding discussion illustrate the general nature of the process. The methods described can be extended to more complicated problems without difficulty. A more detailed study of the variability in the population is all that is required to investigate the advantages of more complicated stratifications. These will not be described at present, lest the details of computation divert attention from the fundamental principles now under discussion.

Exercise 21.-In a given population, the variance of individual observations is the same in all strata and the number of observations taken from each stratum is proportional to the number present. Then $k_i = aK_i$ where a is the fraction taken from each stratum. Under these conditions, show that equation (61) can be reduced to the simple form,

$$V_{\bar{x}} = \left(\frac{1 - a}{a}\right) \frac{V}{S(K_i)}$$

where V is the variance within strata.

Exercise 22.-In the text, the variance of a weighted mean for a sample like that described in Exercise 21 was computed from equation (61). There were 1,220 farms in the five townships and the fraction taken from each was 0.08197. The variance within townships was 657. The variance of the weighted mean was found to be 6.03. Show that the equation derived in Exercise 21 gives the same result. Which do you think is easier to use?

Exercise 23.-In the text, the variance of a weighted mean was computed under the assumption that the variance of individual observations was different in each of five townships. The most efficient allocation of the sample was used in the example. Compute the variance of the weighted mean for the case where the k_i are all equal to 20 and for the case where $k_i = 0.08197K_i$. Compare the results with the value of 6.21 given in the text and explain the differences.

Subsampling

The examples of stratified sampling given in the preceding section were concerned with the problem of estimating the average cropland per farm for a population of all farms in five townships. If it were desired to use that average as an estimate for the entire county, the problem would be one of subsampling instead of stratified sampling. This distinction should be self-evident. The average for the five townships could be ascertained without error by enumerating every farm in those townships, but such an average would not necessarily be equal to the average for the county. It would be an estimate of the county average, but would be subject to error.

If one wishes to interpret the average for the 5 townships as an estimate of the county average, the formulas given in the preceding section for computing the variance of that average no longer apply. Those formulas apply to a county average only when every township in the county is sampled. The variance of the five-township average as an estimate of the county average must include an additional term to allow for the variation between townships. This involves an extension of the mathematical methods described in the preceding section. The same data can be used to illustrate the procedure.

Twenty farms were taken at random from each of five townships which were themselves a random sample of the nine townships in the county. The cropland on the 100 farms in this sample is given in table 5. The problem at hand is to derive an estimate of the average cropland per farm for the entire county, together with an estimate of the variance of that average. The formulas given in the preceding section for computing the weighted average from the five-township sample still apply. The weighted average computed from the sample by those methods serves as an estimate of the average for the county as a whole. The only difference lies in the variance of that estimate.

The formula used to compute the variance of the weighted average from the subsampling point of view is developed in two steps. First, assume that the five townships are a sample of an unlimited number of townships and that each township contains an unlimited number of farms. Under these conditions the variance of an observed township mean would be

$$V_{\bar{x}_i} = V_t + \frac{V}{k_i} \qquad \text{- - - - - - - - - -} \quad (64)$$

In this equation, V_t is the variance of the true township means, V is the variance within townships, and k_i is the number of farms taken from the township. The variance of the weighted mean for a sample of n townships would be,

$$V_{\bar{x}} = \frac{\sum_{i=1}^{n} \left[K_i^2 \left(V_t + \frac{V}{k_i} \right) \right]}{\left[\sum_{i=1}^{n} (K_i) \right]^2} \qquad \text{- - - - - - -} \quad (65)$$

As the county is not an infinite population but a finite population of N townships with K_i farms in individual townships, equation (65) will overestimate

the sampling variance of the weighted average. The estimate given by that equation must be reduced by deducting the quantity that represents the sampling variance of the mean of the entire finite population, when that population is itself considered as a sample from the hypothetical infinite population. This is the same reasoning that was used in deriving equation (22) from equation (21). The quantity that must be subtracted is

$$\frac{\sum_{i=1}^{N}\left[K_i^2\left(V_t + \frac{V}{K_i}\right)\right]}{\left[\sum_{i=1}^{N}(K_i)\right]^2}$$

It should be observed that each of the summations in this correction term includes N items, one for each township in the county. The corresponding summations in equation (65) include only n items, one for each township in the sample. The equation to be used in estimating the variance of the weighted average, when that average is considered as an estimate for the county, is thus of the form,

$$V_{\bar{x}} = \frac{\sum_{i=1}^{n}\left[K_i^2\left(V_t + \frac{V}{k_i}\right)\right]}{\left[\sum_{i=1}^{n}(K_i)\right]^2} - \frac{\sum_{i=1}^{N}\left[K_i^2\left(V_t + \frac{V}{K_i}\right)\right]}{\left[\sum_{i=1}^{N}(K_i)\right]^2} \quad \text{- - - - - - (66)}$$

All of the quantities entering into this equation have been defined previously. N is the number of townships in the county. n is the number of townships in the sample. K_i is the number of farms in the i-th township and k_i is the number taken from that township. V_t is the variance of the true township means, estimated from the analysis of variance for the sample, and V is the variance within townships. For the data at hand

N = 9		n = 5
$K_1 = 205$	Townships included in sample	$k_1 = 20$
$K_2 = 77$		$k_2 = 20$
$K_3 = 497$		$k_3 = 20$
$K_4 = 214$		$k_4 = 20$
$K_5 = 227$		$k_5 = 20$
$K_6 = 255$		
$K_7 = 220$		
$K_8 = 276$		
$K_9 = 267$		

From the analysis of variance given in table 6 is obtained,

$$V_t = 3.90 \qquad V = 657$$

The weighted average for the five townships in the sample was computed previously from the data in table 10. The average cropland per farm, estimated from the five-township sample with 20 farms taken from each of those townships, is 31.05 acres. If this estimate is to be used as a measure of the cropland per farm for the entire county, the variance of that average from such a point of view would have to be computed from equation (66).

The quantity $V_t + \frac{V_i}{k_i}$ has the value $3.90 + \frac{657}{20} = 36.75$ for all five townships in the sample because exactly 20 farms were taken from each township.

The quantity $V_t + \frac{V}{K_i}$ will differ for each of the nine townships in the county because the number of farms present in each township is not constant. The values of this quantity for the nine townships are:

$$3.90 + \frac{657}{205} = 7.10$$

$$3.90 + \frac{657}{77} = 12.43$$

$$3.90 + \frac{657}{497} = 5.22$$

$$3.90 + \frac{657}{214} = 6.97$$

$$3.90 + \frac{657}{227} = 6.79$$

$$3.90 + \frac{657}{255} = 6.48$$

$$3.90 + \frac{657}{220} = 6.89$$

$$3.90 + \frac{657}{276} = 6.28$$

$$3.90 + \frac{657}{267} = 6.36$$

The variance of the weighted average is:

$$V_{\bar{x}} = \frac{1}{(1220)^2}\,(205)^2(36.75)+(77)^2(36.75)+(497)^2(36.75)+(214)^2(36.75)+(227)^2(36.75) -$$

$$\frac{1}{(2238)^2}\,(205)^2(710)+(77)^2(12.43)+(497)^2(5.22)+(214)^2(6.97)+(227)^2(6.79)+(255)^2(6.48)+(220)^2(6.89)+(276)^2(6.28)+(267)^2(6.36) =$$

$$9.69 - 0.80 = 8.89$$

The variance of the weighted average as an estimate of the county average is thus 8.89. The variance of the same average, considered only as an estimate of the true average for the five townships in the sample, was 8.12. When the weighted average is considered as an estimate for the county, the sampling error is larger than when the average only is considered as estimate for the townships included in the sample. This is the kind of result that could be expected from the difference in viewpoint. It serves to emphasize that the particular population to which the results of a statistical analysis apply must be borne in mind when the analysis is made. A sampling error attached to an average is meaningless when the population to which the average applies is not specified. The sampling error of the same average can have many different numerical values as the interpretation of that average changes. The one that is used depends upon the particular population average of which the sample average is supposed to be an estimate.

Exercise 24.-Suppose that the n values of k_i are equal to each other and are represented by k. Also suppose that the N values of K_i are equal to each other and are represented by K. Under these conditions, show that equation (66) reduces to the form,

$$V_{\bar{x}} = V_t\left(\frac{1}{n} - \frac{1}{N}\right) + V\left(\frac{1}{nk} - \frac{1}{NK}\right)$$

Exercise 25.-When n = N equation (66) reduces to equation (61), where the $V_{\bar{x}_i}$ are as defined in equation (62). Prove that this is true and explain why one could expect such a result on the basis of the difference between subsampling and stratified sampling. In working this exercise you should notice that

$$\frac{V}{k_i}\left(\frac{K_i - k_i}{K_i}\right) = V\left(\frac{1}{k_i} - \frac{1}{K_i}\right)$$

Exercise 26.-The example given in the text to illustrate how the variance of a weighted average is computed in subsampling was based on a sample of 20 farms per township for each of five townships. Compute the variance when the number of farms taken from each township is proportional to the number present. That is, let

$$\begin{aligned} k_1 &= 17 \\ k_2 &= 6 \\ k_3 &= 41 \\ k_4 &= 17 \\ k_5 &= \underline{19} \\ &\quad 100 \end{aligned}$$

Fiducial Limits for Means From Stratified Samples and Subsamples

The problem of establishing fiducial limits or confidence intervals for means estimated from random samples was discussed in an earlier section. Similar methods can be applied to means estimated from stratified samples and subsamples.

For a mean from a stratified sample in which the variance of individual observations is the same in all strata, such limits are obtained by computing $\bar{x} \pm ts_{\bar{x}}$. The value of t to be used depends upon the number of degrees of freedom from which the variance within strata was estimated. If the data in table 5 are considered as a stratified sample from this kind of population, the variance within townships would be estimated from 95 degrees of freedom as shown in table 6. The weighted mean for the five townships was 31.05 and the variance of that mean was 8.12. The standard error of the mean would be $\sqrt{8.12} = 2.85$. The value of t for 95 degrees of freedom would be about 2, as indicated by table 3. The fiducial limits on the observed weighted average would thus be $31.05 \pm (2)(2.85)$ or 25.35 and 36.75. One would thus have 95 chances out of 100 to be correct if he concluded that the range 25.35 to 36.75 included the true mean for the five townships.

When equal variances within the townships are not assumed, the t-table should not be used to compute such ranges. The above procedure is rigorously correct only when such variances are equal. When the variance of the average is computed from separate estimates of the variances within individual townships, only approximate results can be obtained. Small-sample theory to fit this case has not yet been developed and some approximation must be used. It is usually safe to assume that the frequency distribution of averages is Normal when fairly large samples are used. The fiducial limits would then be approximately $\bar{x} \pm 1.96s_{\bar{x}}$ as demanded by the Normal Curve. The factor 1.96 is so close to 2 that most statisticians prefer to use 2 instead of the exact value. Usually other approximations are involved in the analysis, such as the assumption of Normality itself, so that the error introduced by using 2 as a factor instead of 1.96 is relatively unimportant.

The t distribution should also not be used to compute fiducial limits for averages obtained by subsampling. The variance of such an average involves the quantity V_t, which can only be estimated approximately from the sample There is no justification for attempting the refinement represented by the t distribution when other approximations are involved in the analysis. When the weighted mean, 31.05, is interpreted as an estimate of the mean for the entire county, the variance of that mean is 8.89. The value of $s_{\bar{x}}$ is thus $\sqrt{8.89}$ or 2.98. The fiducial limits from this point of view are $31.05 \pm (1.96)(2.98)$ or 25.21 and 36.89.

The fact that the t distribution cannot be adapted to problems in subsampling, or to problems in stratified sampling when the variances within strata are unequal, leads to no serious difficulty in most practical problems. The samples with which the agricultural statistician or economist has to work

are usually sufficiently large so that the refinements represented by small-sample theory are not particularly important. The t distribution approaches the Normal Curve so rapidly as the sample size increases that the t distribution can usually be dispensed with. In general, the expression $\bar{x} \pm 2s_{\bar{x}}$ can be used to represent the 95 percent fiducial limits on an observed average with sufficient accuracy for all samples likely to be encountered in practice. Unless the samples are very small, there is no need to be unduly concerned about refinements like those represented by the t distribution or the difference between the factors 2 and 1.96. Other details are usually more worthy of attention in the operation of a sampling study. But when approximations are used, it is well to be aware of their nature.

Exercise 27.-In exercise 23 the variance of the weighted mean for five townships was computed for the sample shown in table 5 under the assumption that the variance of cropland for individual farms was different for each township. The weighted mean was 31.05. Compute the 95 percent fiducial limits from this variance and compare the results with those given above in the text, under the assumption that the variances were equal. Do you think that the extra work involved in computing separate variances within townships makes enough difference in the final result to be worth the effort?

Sampling Units and Expansion Factors

In estimating a quantity like the average cropland per farm in a county, it might be supposed that the individual farm would have to be taken as the sampling unit. This is not necessarily true. If individual farms were taken as the sampling units, the sample would consist of individual farms taken at random from the county as a whole or from various strata in the county. If the farms were to be enumerated by mail, this kind of arrangement would be as satisfactory as any other. On the other hand, if the information were to be obtained by actually visiting each farm in the sample, the amount of travel required could be excessive. As a practical matter, the travel could be reduced by visiting groups of contiguous farms chosen at random. Under this scheme, each group of farms would constitute a sampling unit. A sample of farms taken in this way will usually yield less accurate estimates than a sample of the same size in which individual farms are the sampling units. Neighboring farms tend to be more nearly alike than farms at a distance from each other. The loss in accuracy caused by grouping depends upon the degree of similarity between neighboring farms. The more nearly such farms are alike, the more information will be lost by grouping. But the lower cost of enumerating such a sample often enables one to increase the total number of farms enumerated. This increase in the total size of the sample tends to compensate for the loss of precision introduced by grouping. At times the compensating effect is so great that the grouped sample will give more accurate results than any sample by individual farms that could be enumerated at the same cost.

The scientific study of sampling units of different kinds and sizes for surveys of different types is a field that has not yet been fully explored by statisticians. Many of the mathematical principles have been worked out, but the data required to make use of them are still incomplete. In designing a survey to estimate the cropland on farms, for example, the degree of similarity between contiguous farms must be known before it is possible to judge the relative merits of individual farms versus groups of contiguous farms as sampling units. Problems of this kind are made still more difficult by a lack of consistency in the behavior of different items that usually must be estimated from the same enumeration. Neighboring farms are often similar with respect to some characteristics and dissimilar with respect to others. A sampling scheme that would be efficient for estimating some items could be inefficient for others.

For any one item, the variance of a per farm average, estimated from a random sample consisting of n groups of farms with k farms in a group is

$$V_{\bar{x}} = \left(\frac{V_g}{n} + \frac{V}{nk}\right)\left(\frac{N - n}{N}\right) \quad - - - - - - - - - \quad (67)$$

In this equation, N represents the number of such groups present in the population, V_g represents the variance of the true group means, and V represents the variance within groups. V_g can be estimated from an analysis of variance in the same way that the quantity V_t was estimated previously. Once the numerical values of V_g and V are known, equation (67) can be used to estimate the precision of an average obtained by grouping neighboring farms into aggregates of different sizes. Such computations involve the assumption that the variance between farms within a group is constant for groups of different sizes. This assumption is not strictly accurate because more variability could logically be expected between farms in a large group than between farms in a small group. So long as the group sizes under consideration do not cover too wide a range, no serious error is likely to be introduced. Ordinarily, one would only be interested in comparing groupings within a fairly narrow range. If large groups were to be considered, it would be more practicable to use the method of subsampling than to enumerate all farms in every group.

The principles underlying this kind of sampling have many applications. A few years ago, officials of the Agricultural Adjustment Administration were interested in estimating the yields of individual corn fields by taking blocks of four hills each at random from every field. The ears of corn on these hills were weighed and the average weight used to derive an estimate of the yield for the entire field. In this study, each four-hill block was a sampling unit. The corn was weighed separately for each pair of two hills in every sampling unit so that a measure of the variation within sampling units could be obtained. The analysis of variance of the weights of the two-hill samples is given in table 12.

Table 12. - Analysis of variance of corn weights for two-hill samples taken from individual fields in blocks of four hills each.

Source of variability	Degrees of Freedom	Sum of Squares	Mean Square
Between 4-hill blocks	491	363.920	0.74118
Between 2-hill samples within blocks	496	213.399	.43024
Total	987	577.319	

The analysis of variance shown in table 12 was based on data from five fields. A separate analysis was made for each field, after which the degrees of freedom and sums of squares for the five fields were combined to give the analysis shown in the table. The mean squares shown in the table are, therefore, average values for the five fields used in the analysis. These mean squares can be used to show how the precision of the final result would be affected by changing the number of hills included in a block. The mean square between blocks in table 12 is an estimate of $2V_g + V$ in which V_g is the variance of the true block averages and V is the variance of the 2-hill sample averages within blocks. V is equal to 0.43024, as given in the table, and V_g thus has the value,

$$\frac{0.74118 - 0.43024}{2} = 0.15547$$

As the number of four-hill blocks taken from each field was small in relation to the total size of the field, the factor $\frac{N - n}{N}$ in equation (67) may be neglected. The variance of the average corn weight per two-hill sample can be written

$$V_{\bar{x}} = \frac{V_g}{n} + \frac{V}{nk} \qquad \text{- - - - - - - - - - - - (68)}$$

in which $V_g = 0.15547$, $V = 0.43024$, n is the number of blocks and k is the number of two-hill samples per block.

The precision of an average, based on any given number of blocks with an assigned number of hills per block, can be computed from equation (68) without difficulty. As the average is expressed in terms of a two-hill average, the variance of that average is on the same basis regardless of the number of hills actually present in a block. For example, the quantity $\frac{0.15547}{50} + \frac{0.43024}{(50)(1.5)}$ represents the variance of the average weight per two-hill sample, when that average is computed from 50 blocks with 1.5 two-hill samples, or three hills, per block. The effect of changing the size of block can be observed in figure 14 which gives the variances of averages based on 100 two-hill samples taken in blocks of different sizes. The number of blocks decreases as the number of hills per block increases.

Figure 14. Variance of mean corn weight per 2-hill sample for 100 2-hill samples taken from a field in blocks of different sizes (all weights in pounds).

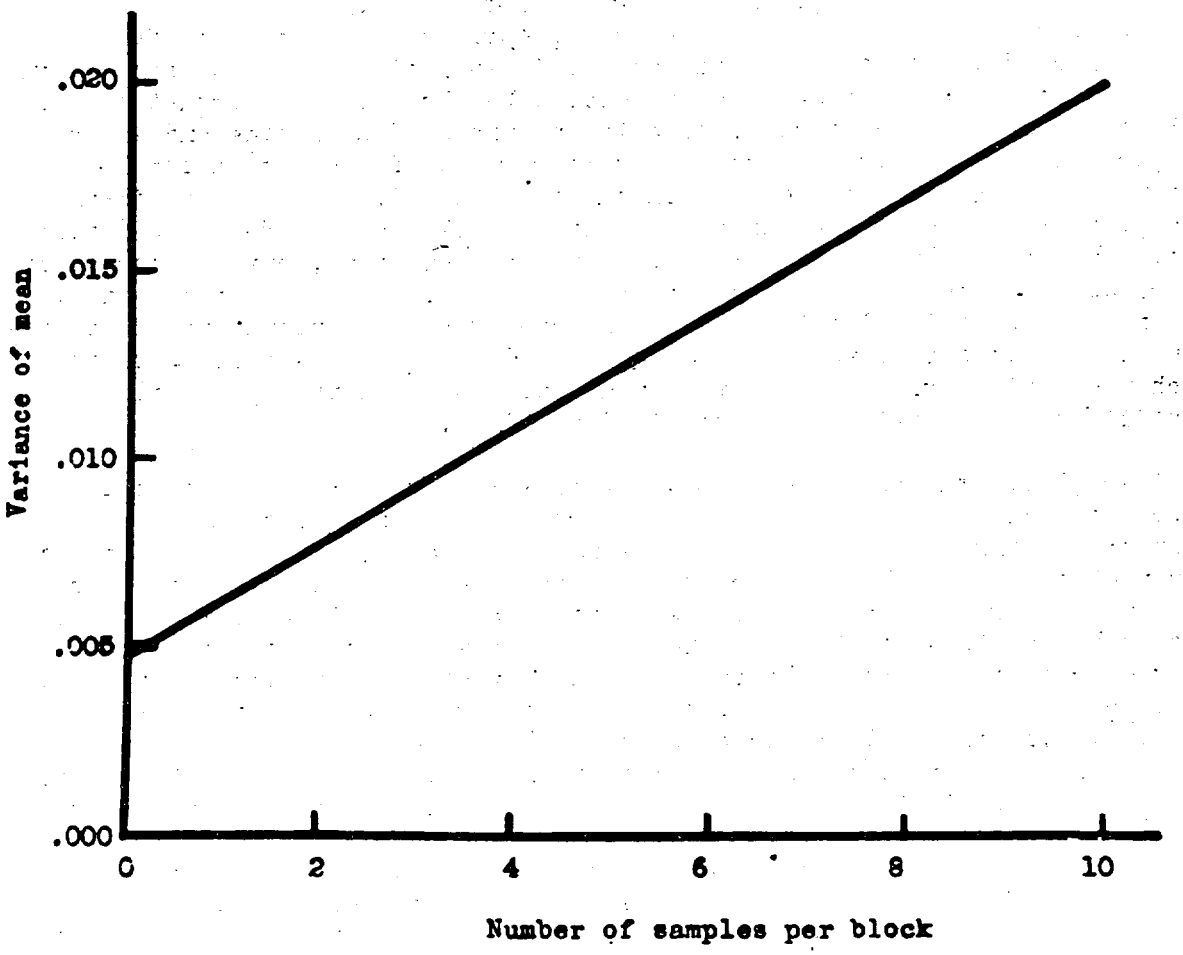


Figure 14 shows that the variance of the mean increases as the samples are grouped into larger aggregates. This effect is characteristic of grouping individuals into larger sampling units. The relative numerical values of V_g and V determine how marked the effect will be in any particular example. If V_g is small in relation to V, little precision will be lost by grouping. On the other hand, if V_g is relatively large, the loss of precision will be comparatively large. The size of sampling unit that is best suited to a particular problem depends upon the relative sizes of V_g and V, together with the relative costs involved in using sampling units of different sizes. In some practical problems, the saving of time and expense brought about by using the larger sampling units is so great that the number of such units can be increased more than enough to compensate for the detrimental effects of grouping. Even when such is not the case, this kind of grouping is often necessary to keep the cost of a survey within reasonable bounds.

The discussion of sampling units given above indicates the general nature of the problems encountered in choosing appropriate sampling units for a particular study. As stated previously, this subject is one that has not yet been investigated as thoroughly as it should be, but it is important in all applications of sampling theory, both in the field of economics and in the biological sciences. The subject has so far received most attention in agronomic research in connection with the design of field-plot experiments. Its application to other sampling problems has only begun. The mathematical principles are fairly well understood at the present time, but the necessary data required to make use of them can be obtained only by experimentation. Some progress in this direction has been made by a few agricultural statisticians and economists, but much remains to be done. This is a field of research that offers rich rewards to anyone interested in the application of scientific sampling methods to practical problems.

A sampling unit should generally be chosen in such a way that it is possible to expand sample indications to population estimates. For example, the interest might be in learning how many people are living on all farms in a given State rather than in the per farm average. If a per farm average has been computed from a sample of farms, a State estimate of the farm population can only be obtained if the total number of farms in the State is known with a fair degree of accuracy. If the total number of farms in the State is not known, the per farm average is of no use, insofar as an estimation of the farm population in the State is concerned. Sometimes it is possible to make use of additional information to derive an expansion factor when the number of sampling units in the population is not known accurately. When the total farm land or total cropland in the State is known, a "per acre of farm land" average or a "per acre of cropland" average can be computed for the farms in the sample. The former can be multiplied by the total acreage of farm land in the State to derive a State estimate, whereas the latter can be multiplied by the total acreage of cropland.

So long as there is no bias in the sample of farms, all of these expansions will give essentially the same results. But the estimates will have different sampling errors. When a choice of expansion factors is permissible, it is desirable to use the one that will yield the estimate with

the smallest variance. In general, the number of farms applied to the per farm average will give the most accurate results for items that are not correlated with size of farm. For items that tend to increase with size of farm, an expansion based on farm land or cropland will be more accurate than one based on number of farms.

It should be noted that these expansion factors yield unbiased State estimates only when they are known accurately and when the sample is free from bias. When there is a preponderance of large farms in the sample, allowance must be made for this bias in expanding the sample average to a State estimate. This is generally possible when the expansion factors are free from error. For items that are independent of size of farm, the estimate obtained from the per farm average of the sample and the number of farms in the State is the one to use. For items that tend to be multiples of farm land or cropland, the farm land or cropland expansion will give better results. For items that are correlated with farm land or cropland without being simple multiples of them, neither of these expansions will give satisfactory results. Under such conditions a method of expansion that corresponds to the particular relationship at hand should be used. Before such methods can be discussed, the student should have some understanding of the principles of regression. This subject will be taken up in the next section.

Linear Regression and Correlation

In working with experimental or sample data, it is often found that measurements on one variable are related to those on another. Such measurements are said to be correlated with each other. Relationships of this kind are of utmost importance in all statistical work. An illustration from agricultural sample data is given in figure 15.

Figure 15 is an example of a two-way frequency tabulation that is often used to get some preliminary information about the degree of relationship between two variables. Grouping the data into class intervals takes less time than plotting the individual data on graph paper. In this case the tabulation shows that the number of sharecroppers on a farm tends to increase as the acreage of cropland in the farm increases. A similar tabulation for the same farms was made to investigate the possibility of a relationship between cropland and number of workers belonging to the operator's family. The results are given in figure 16. This tabulation indicates that there is little relationship between the acreage of cropland on a farm and the number of workers belonging to the operator's family.

These indicated relationships conform to what one would expect. Small farms have no need for sharecroppers. Only the larger farms are likely to be subdivided into sharecropper units, and the number of such units on a farm should be roughly proportional to the size of the farm. The situation with respect to family workers is entirely different. In most regions perhaps a few more family workers could be expected on large farms than on small farms because members of an operator's family can find employment at home when the farm is large. When the farm is small, the adult members of the operator's family usually seek employment elsewhere. This cannot be

Figure 15. Sample of 171 farms from a North Carolina Crop-Reporting District classified according to crop land and number of share croppers per farm (March 1942 mailed Farm Employment survey)

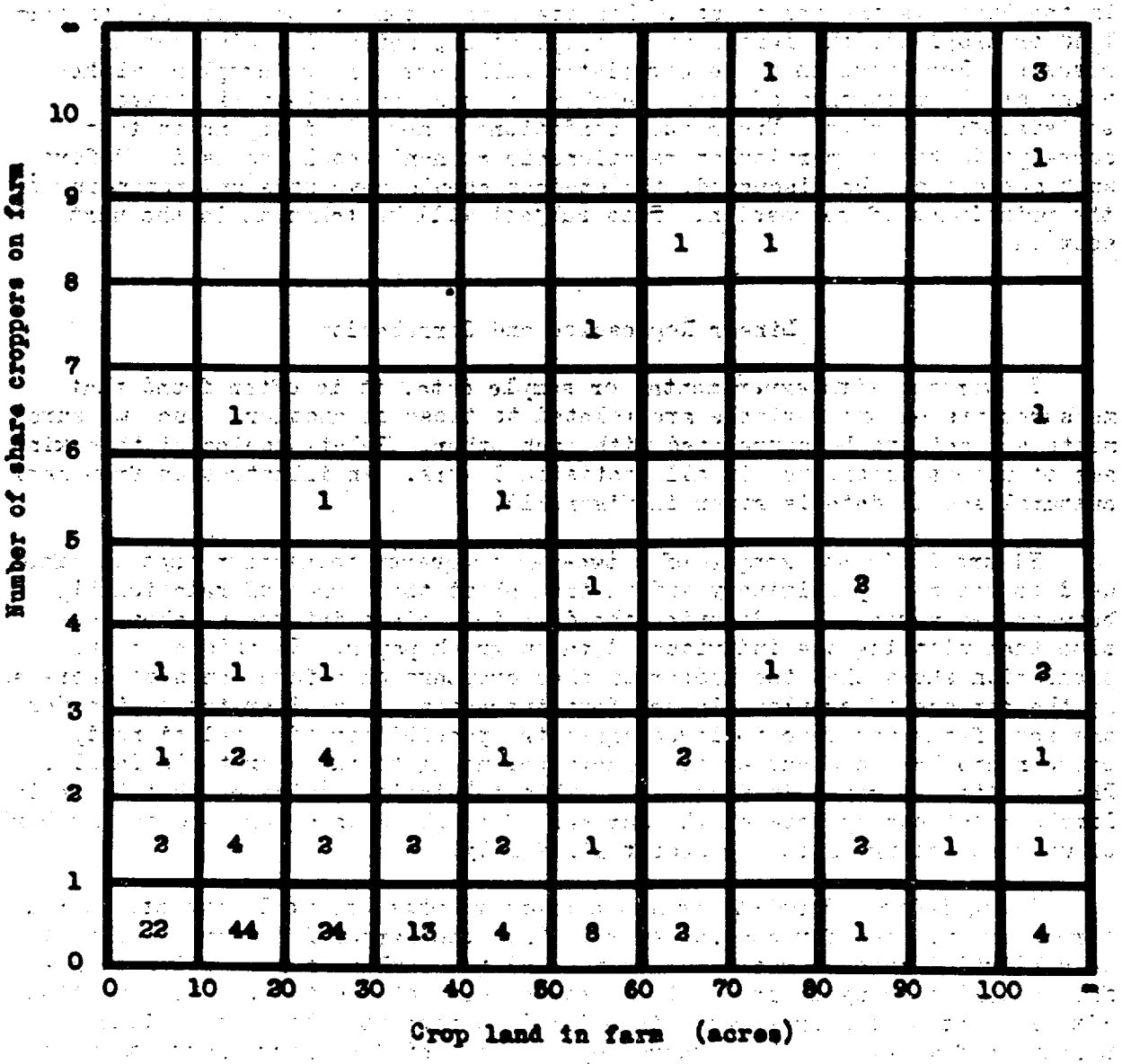
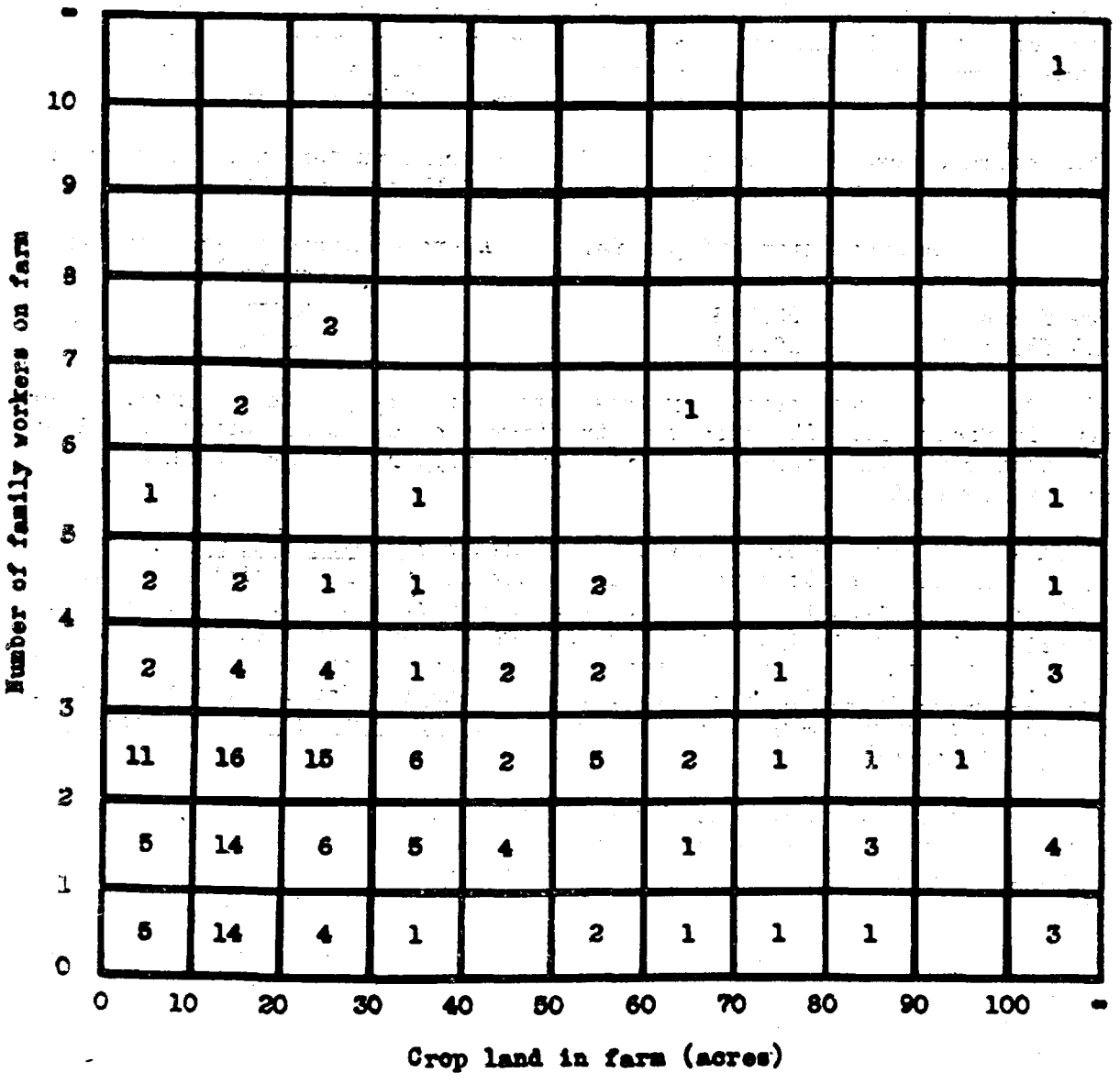


Figure 16. Sample of 171 farms from a North Carolina Crop-Reporting District classified according to crop land and number of family workers per farm (March 1942 mailed Farm Employment survey)



regarded as a hard and fast rule because the poorer families on small farms often have more children than the more prosperous families living on the larger farms. In such areas a small decrease in the number of family workers could be expected with an increase in the size of the farm. This is actually the case for some areas in North Carolina and is probably characteristic of many of the poorer farming communities throughout the country.

An analysis of the data used in constructing figures 15 and 16 was made to determine the average numbers of sharecroppers and family workers that could be expected on farms with any given acreage of cropland. The easiest way to do this is to separate the 171 farms into two groups on the basis of the acreage of cropland. The farms with an acreage of cropland below the average were placed in one group and the farms with an acreage of cropland above the average were placed in another. This provides a separation of the 171 farms into a group of small farms and a group of large farms.

The average acreage of cropland and the average number of sharecroppers per farm were computed for each group with the following results:

	Average cropland per farm (acres)	Average sharecroppers per farm (number)
Large farms	93.723	3.170
Small farms	16.911	.581

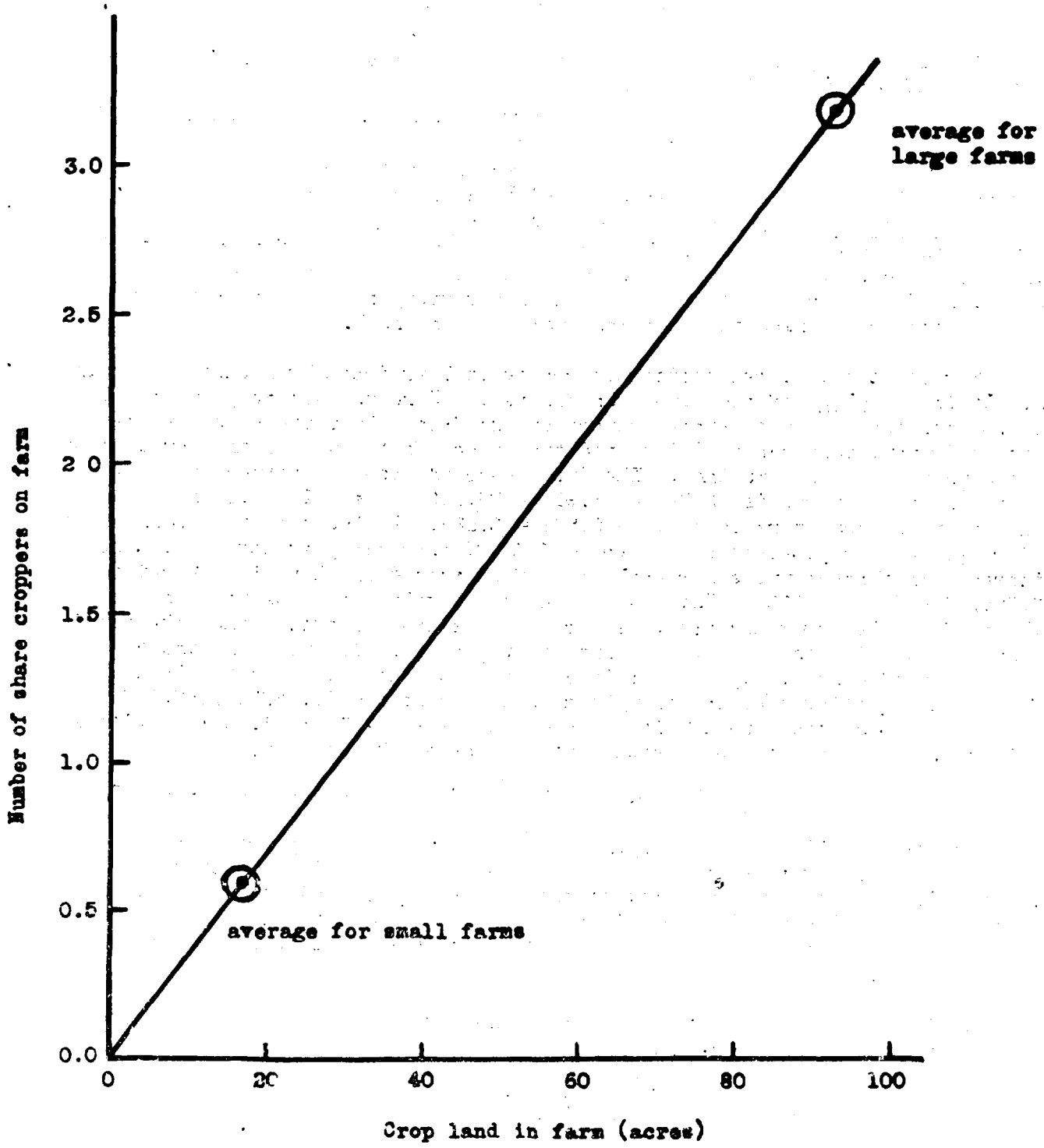
The two points represented by these averages were plotted on graph paper as shown in figure 17. The straight line drawn through these points provides a chart that gives the average number of sharecroppers on farms with any given acreage of cropland.

A line like the one in figure 17 is called a regression line. In this particular example, the line passes almost through the zero point where the vertical and horizontal scales intersect each other. The zero point is called the origin. The fact that the regression line passes almost through the origin indicates that the average number of sharecroppers on a farm tends to be a simple multiple of the acreage of cropland in the farm. In other words, the number of sharecroppers on a farm is roughly proportional to the acreage of cropland in the farm.

The relationship between cropland and the average number of family workers on a farm is somewhat different. For the two groups into which the 171 farms in the sample were divided, the average cropland and numbers of family workers are as follows:

	Average cropland per farm (acres)	Average family workers per farm (number)
Large farms	93.723	2.000
Small farms	16.911	1.758

Figure 17. Relation between crop land and number of share croppers for farms in a North Carolina Crop-Reporting District (March 1942 mailed Farm Employment survey)



These two points, and the regression line established by them, are shown in figure 18. It is evident that the average number of family workers on a farm is about the same for farms of different sizes. A small increase is associated with increasing acres of cropland, but the relationship is not nearly so marked as in the case of numbers of sharecroppers.

A relationship of the kind illustrated in figures 17 and 18 can be represented by an equation of the form,

$$Y = a + bX \qquad \text{- - - - - - - - - - - - - -} \qquad (69)$$

When the numerical values of a and b are given, a value of Y can be computed for any assigned value of X. Equation (69) is sometimes called a linear regression equation because the values of Y computed for different values of X will lie on a straight line when they are plotted against the values of X on graph paper. Any straight line on a chart can thus be expressed by an equation of that form. To find the equation corresponding to a given line, it is necessary to find the numerical values of a and b.

The constant, b, represents the change in Y produced by a unit change in X. It is the slope of the line. In figure 17, X represents cropland and Y represents the number of sharecroppers on the farm. From the group averages computed previously can be determined the change in the number of sharecroppers for a unit change in cropland. The difference in cropland between the large farms and small farms is 93.723 - 16.911 = 76.812 acres. The difference in number of sharecroppers is 3.170 - 0.581 = 2.589. An increase of 76.812 acres of cropland thus produced an increase of 2.589 sharecroppers per farm. The increase in sharecroppers for each acre of increased cropland is, therefore, 2.589/76.812 = 0.033706. This is the numerical value of b for the line in figure 17. The numerical value of a can be obtained by noting that the final equation must fit the two fixed points in the chart, although only one of these is needed here. Applying this condition to the point for the larger farms, the equation must satisfy the condition that a + (0.033706)(93.723) = 3.170. Solving this equation for a, a = 3.170 - (0.033706)(93.723) = 0.011 is obtained. The complete regression equation may now be written,

$$Y = 0.011 + 0.033706X \qquad \text{- - - - - - - - - -} \qquad (70)$$

This equation enables one to compute the number of sharecroppers expected on a farm with any given acreage of cropland instead of reading the chart in figure 17. The reader should notice that the quantity 0.011 is the value of Y obtained when X = 0. It thus gives the point at which the line in figure 17 intersects the vertical axis. This constant is the Y intercept of the regression line.

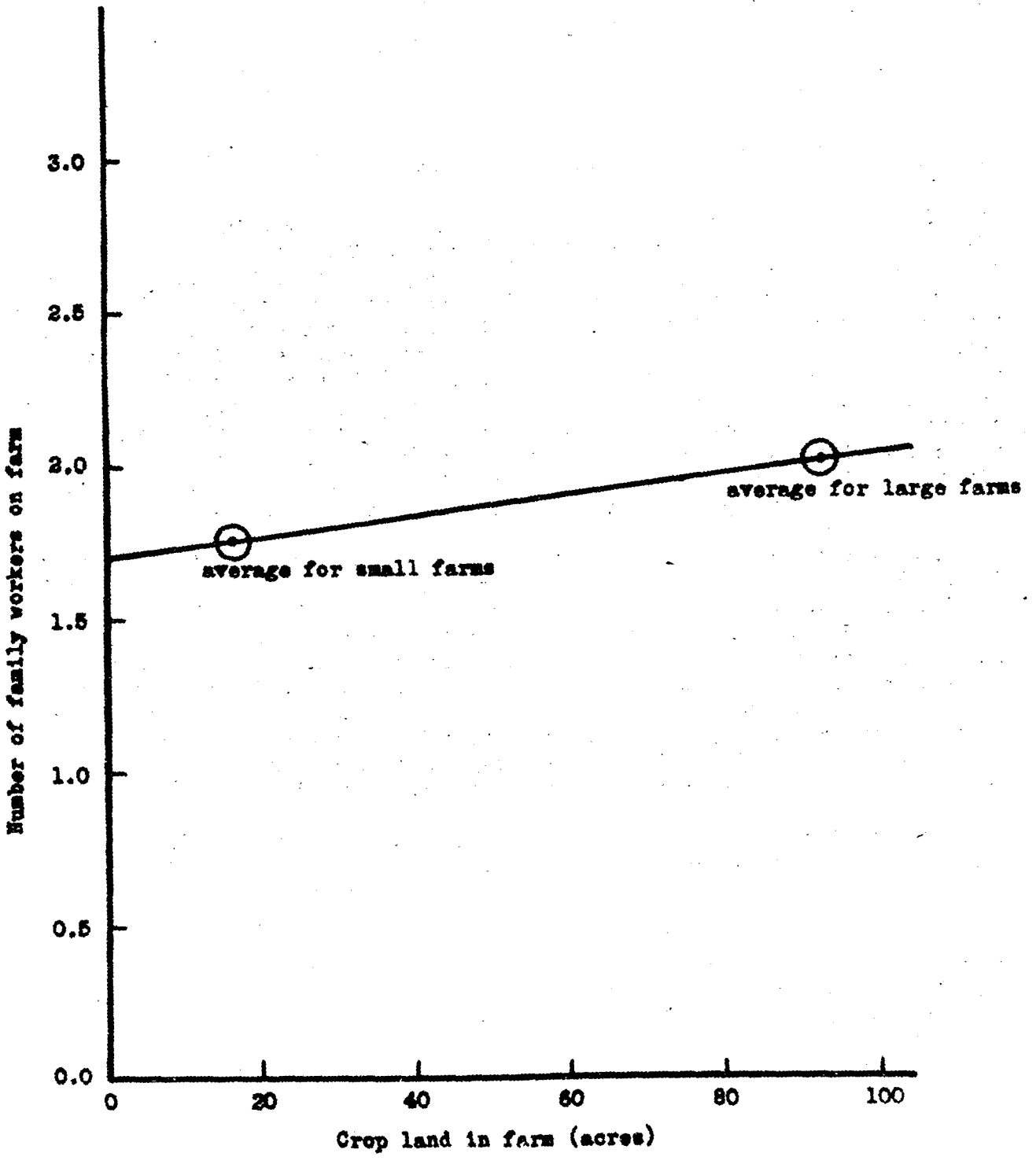
A similar analysis of the data on family workers gives the equation of the straight line in figure 18.

$$b = \frac{2.000 - 1.758}{93.723 - 16.911} = \frac{0.242}{76.812} = 0.003151$$

$$a = 2.000 - (0.003151)(93.723) = 1.705$$

$$Y = 1.705 + 0.003151X \qquad \text{- - - - - - - -} \qquad (71)$$

Figure 18. Relation between crop land and number of family workers for farms in a North Carolina Crop-Reporting District (March 1942 mailed Farm Employment survey)



The values of Y obtained by assigning different values of X would fall on the straight line shown in figure 18 if they were plotted in the chart. This equation differs from equation (70) just as one would expect from a comparison of the straight lines in figueres 17 and 18. The slope is only 0.1 as great and the Y intercept is considerably larger than zero.

These relationships are important in estimating the total numbers of sharecroppers and family workers on all farms in the district from which the sample of 171 farms was taken. There are 23,142 farms, with a total cropland of 328,171 acres in that district. The average acreage of cropland per farm for the entire district is 328171/23142 = 14.181 acres. There are 171 farms with 6,502 acres of cropland in the sample. The average acreage of cropland per farm in the sample is thus 6502/171 = 38.023 acres. The sample evidently contains too many of the larger farms in the district.

The number of sharecroppers reported for the 171 farms in the sample is 221. The average number per farm is 221/171 = 1.2924. If the total number of sharecroppers in the district were to be estimated from the number of farms in the district and the average number of sharecroppers per farm in the sample, that estimate would be (23142)(1.2924) = 29,909. This figure would be a poor estimate. The number of sharecroppers on a farm increases rapidly with the acreage of cropland. The farms in the sample contain more cropland, on the average, than the farms for the district as a whole. Therefore, the average number of sharecroppers per farm in the sample is larger than the average for the entire district. An estimate of the total number of sharecroppers in the district, based only on the number of farms in the district and the average number of sharecroppers per farm in the sample, would be too high.

Now consider an expansion based on cropland rather than on the number of farms. The average number of sharecroppers per acre of cropland in the sample is 221/6502 = 0.033990. If this figure were multiplied by 328,171, (328171)(0.033990) = 11,155 would be obtained as an estimate of the total number of sharecroppers in the district. This figure is only about one-third as large as the estimate based on the number of farms. As figure 17 shows that the expected number of sharecroppers on a farm is almost exactly proportional to the cropland in the farm, this second estimate is much nearer to the truth than the other. It would be slightly in error, however, because the regression line does not pass exactly through the origin.

Now consider the case of the family workers. The 171 farms in the sample reported a total of 312 family workers. This gives an average of 312/171 = 1,8246 per farm and 312/6502 = 0.047985 per acre of cropland. The corresponding district estimates are (23142)(1.8246) = 42225 for the estimate based on the number of farms in the District and (328171)(0.047985) = 15747 for the estimate based on the cropland in the district. These estimates bear the same relation to each other as the corresponding estimates of the number of sharecroppers in the district. But in this case, the larger of the two estimates is the better because figure 18 shows that the number of family workers on a farm is almost independent of the acreage of cropland in the farm. But it would be slightly too high, because there is some increase in number of family workers with an increase in cropland.

This kind of situation is frequently met in practice. It should be noted that an estimate based on the number of farms differs from an estimate based on cropland only when the average cropland for farms in the sample is too high or too low. If there were no such bias in the sample, the two estimates would be essentially the same. The matter of statistical precision is the only point it would then be necessary to consider. In most sample data, and especially those obtained from mailed questionnaires, there is a consistent tendency for the larger farms to be overrepresented. Situations like those just described are the general rule rather than the exception.

In view of this fact, it is desirable to use a method of expansion that is theoretically correct under all conditions. The regression equations corresponding to the lines in figures 17 and 18 provide the basis for such a method of expanding sample indications to population estimates. As the relationship between number of sharecroppers and cropland for individual farms is given by equation (70), the average number of sharecroppers per farm in any sample is given by the equation,

$$\bar{y}_s = 0.011 + 0.033706\bar{x} \qquad (72)$$

In this equation $\bar{x}$ represents the average acreage of cropland per farm in the sample and $\bar{y}_s$ represents the average number of sharecroppers per farm in the sample. It is easy to see that the number of sharecroppers per farm in the sample will vary with the average acreage of cropland per farm in the sample. To estimate the average number of sharecroppers per farm in the population, it is necessary to know the average acreage of cropland per farm in the population. If m represents the average acreage of cropland per farm in the population, the average number of sharecroppers per farm in the population is given by the equation,

$$\bar{y}_p = 0.011 + 0.033706m \qquad (73)$$

This equation enables the computation of an adjusted estimate of the average number of sharecroppers per farm for the sample data. It is an estimate of the average that would have been obtained directly from the original data if there had been no bias in the sample. Since $\bar{y}_p$ represents the average number of sharecroppers per farm in the population, the total number in the population can be estimated by multiplying this value by the number of farms in the population. This estimate is given by the equation,

$$E = 0.011N + 0.033706Nm \qquad (74)$$

In this equation N represents the number of farms in the population and E represents the estimated number of sharecroppers in the population.

The reader should notice that the product Nm represents the total cropland in the population because N represents the number of farms and m represents the average cropland per farm. The estimate of the number of sharecroppers in the district obtained by this method thus consists of two parts. The constant 0.011 is multiplied by the number of farms in the district and the constant 0.033706 is multiplied by the cropland in the district. These two components are added to derive the district estimate.

Carrying out these operations yields (0.011)(23142)+(0.033706)(328171) = 11316 as the estimated number of sharecroppers in the district. This estimate does not differ much from the figure 11,155 obtained from the simple cropland expansion. But it is more accurate because it is based on the exact relationship between number of sharecroppers and cropland for individual farms. In this case the regression line passes almost through the origin. That is why the estimate based on the simple cropland expansion comes so close to the correct value.

Applying the same procedure to the data on family workers, the best estimate of the total number of family workers in the District is (1.705)(23142)+(0.00315)(328171) = 40491. This estimate is fairly close to the figure 42,225 based on the number of farms. This is a direct consequency of the fact that the number of family workers on a farm does not vary much with the acreage of cropland. But there is a small difference between the two estimates, because farms with a large acreage of cropland tend to have a few more family workers than the smaller farms.

Population estimates based on regression equations like those described above are automatically built up from a part that is independent of farm size and a part that varies with farm size. Each part exerts its effect in proper proportion. When the slope of the regression line is equal to zero, the estimate derived from the regression equation is identical with the estimate obtained from the number of farms. When the regression line passes through the origin, the estimate from the regression equation is equal to the one based on cropland. When the regression line has a slope different from zero, but does not pass through the origin, the estimates based on number of farms and cropland will both be in error. The estimate based on the regression line is the only one that will be correct. As a general procedure, the use of the regression equation will yield correct results, regardless of the position of the regression line. Special cases like a = 0 or b = 0 take care of themselves automatically when they occur. The constants a and b always contribute to the final result in proper proportion. The amount contributed by each depends upon its numerical value in any particular sample.

This sort of estimating procedure has many applications. It can be used for estimating crop acreages, livestock numbers, and other items as well as farm employment. Total land in farm may be substituted for cropland in the equations whenever it is necessary or desirable to do so. Cropland was used to estimate farm-employment items in North Carolina because the non-cropland appears to be uncorrelated with such items. The use of cropland instead of farm land thus provides estimates of greater precision.

The discussion given above indicates one application of the principle of regression. Regression equations are also useful for predicting values of one quantity from measurements of another. For example, consider the data in table 13. This table gives the North Carolina cotton yields for a 15-year period, together with data on acreage, August condition of the crop as reported by farmers, and reported weevil infestation.