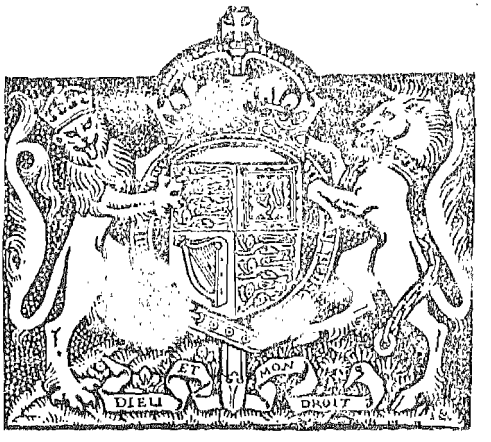


MG1PC—84—III-1-93—22-8-45—5,000.

SAMPLING METHODS IN FORESTRY AND RANGE MANAGEMENT

IMPERIAL AGRICULTURAL
RESEARCH INSTITUTE, NEW DELHI.

MGIPC—84—III-1-93—22-8-45—5,000.

DUKE UNIVERSITY
SCHOOL OF FORESTRY
BULLETIN 7

SAMPLING METHODS IN FORESTRY AND RANGE MANAGEMENT

BY

F. X. SCHUMACHER

Professor of Forestry, School of Forestry
Duke University

AND

R. A. CHAPMAN

Associate Silviculturist, Southern Forest Experiment Station, Forest Service,
United States Department of Agriculture

DURHAM, NORTH CAROLINA
JANUARY, 1942

PRINTED IN THE UNITED STATES OF AMERICA BY
THE SEEMAN PRINTERY, DURHAM, NORTH CAROLINA

PREFACE

The concept of sampling error is essentially simple. It implies that the discrepancy—real, but unknown— between a true magnitude, which is the subject of inquiry, and the sampling estimate thereof, may be evaluated precisely.

The practice of forestry is replete with problems of sampling. In many of them, however, as in timber cruises, the essential simplicity of the concept of sampling error is obscured by failure on the part of foresters to recognize that the body of data gathered from a systematic pattern of strips or line-plots, upon which estimates of timber volumes and values are commonly based—and which they have been taught in their college courses in forest mensuration—does not contain information on sampling error.[1] Unquestioning acceptance of the systematic pattern as the only kind worthy of consideration has resulted in attempts to extract sampling error that are more akin to the art of the conjurer than to scientific assay.

The development of mathematical statistics, particularly of that part concerning the theory of small samples, is exerting remarkable influence upon the scientific endeavor of research foresters and range ecologists, by making available experimental methods of logical structure which are at once capable of yielding efficient estimates of effects, and valid tests of hypotheses pertaining thereto.

Less apparent, perhaps, but nonetheless genuine, is the growing influence of mathematical statistics upon the everyday work of practicing foresters and range examiners. Administrative decisions pertaining to management of a forest or range business commonly rest upon *estimates* of the amount, or condition, of forest or range values. Thus the maximum number of cattle a range can support without deterioration; or the volume of a given class of timber which may be removed from a forest compartment without harm to the residue; these are deduced from estimates of existing magnitudes of forest or range values, arrived at by means of some planned sampling procedures.

While each such estimate is obviously encumbered with a *real* error, it has not been universally recognized that it is the job of practicing foresters, or range technicians, to acquire the art of planning—and executing—suitable sampling procedures, such that (1) the real error may be assessed unambiguously; and (2) the best estimate is obtainable (and,

[1] One of us (F. X. S.) takes this occasion to indict himself as co-author of a text on forest mensuration in which systematic cruise patterns are the only kinds discussed.

consequently, the real error is least) consistent with the time and funds available for the sampling work.

It is the purpose of this treatise to discuss this twofold aspect of the problem of sampling, of the kind encountered in the practice of forestry.

Such use as is made of mathematics in the following pages presupposes no special training in the subject beyond the modest requirements of a forestry curriculum. Occasionally, when a needed demonstration seemed to become heavy, or to distract attention from the main theme, it has been relegated to the Appendix.

We are indebted to Professor E. S. Pearson, of University College, London, for permission to reproduce a page of Tippett's Random Sampling Numbers; and to R. A. Fisher, and his publishers, Messrs. Oliver and Boyd, for permission to reproduce the table of t. But we cannot adequately express our appreciation of the work of those mathematicians and scientists—particularly of Professor Fisher and his associates—to whose vision and insight the development of small-sample theory is due. Without the foundation of their labors the present work would not have been attempted.

We are also deeply indebted to James G. Osborne, Chief of Forest Measurements, Division of Forest Management Research, United States Forest Service, for a critical reading of the manuscript and many valuable suggestions.

DURHAM, NORTH CAROLINA
January, 1942

F. X. SCHUMACHER
R. A. CHAPMAN

TABLE OF CONTENTS

PART 1. STATISTICAL BACKGROUND

PART 2. DIRECT ESTIMATES BY SAMPLING

APPENDIX

FIGURES

PART I

STATISTICAL BACKGROUND

CHAPTER I

INTRODUCTION

1.1 The Art of Sampling. A sample is a part or portion of anything presented as evidence of the quality of the larger whole from which it has been drawn. Thus if the timber volume on 2 acres of a 10-acre woodlot is 8 M feet board measure, it is a sample, and from it *something* is known about the volume of the whole woodlot.

But how much? Even if the volume on 2 acres *is* 8 M feet b.m., questions immediately arise concerning the sample. Is 8 M feet b.m. the volume of a single 2-acre area of, perhaps, the best timber? Or the poorest? Or is it, perhaps, the aggregate volume of 20 square chains of area scattered throughout the 10-acre woodlot?

Questions such as these are essential features of every inference concerning the *population* as may be derived from the sample; for the volume on the 10-acre woodlot may be considered as the aggregate, or *population*, of volumes according to 2-acre subdivisions; or, again, as a population of volumes according to the 100 square chains of area into which the woodlot may be subdivided.

Were other samples presented, volume would vary among them. Now the only means of quantitatively appraising variation is by the use of statistical methods, which is the process of extracting from one or more samples all the information they contain concerning the population they represent. Furthermore, when combined with professional experience in populations such as are met with in forestry and range management practice, statistical methods give rise to the art of sampling them.

The art of sampling consists in making the most efficient use of available resources so as to afford the best possible estimate concerning the quality of a population under consideration as is consistent with the ever-present limitation in time and funds.

It is therefore apropos that a measure of statistical background be acquired by way of introduction to practical problems in sampling.

1.2 The Mean and the Standard Deviation of the Sample. A line is drawn from pith to cambium on the surfaced cross-section of a tree stem. The width of a particular annual ring is then measured along the line by an experienced observer using a microscope-caliper, the least count of which is 0.01 mm. Following are four measurements:

227, 226, 227, 230.

As the conditions for precise work are favorable, and the observer is experienced and careful, it is to be assumed that the discrepancies among the several observations are beyond his control.

When a set of discordant observations, which have been taken on some physical magnitude, are all supposed equally good, their arithmetic mean is generally accepted as the best single value characteristic of the set. The mean of the four observations on annual ring width is

$$\frac{1}{4}(227+226+227+230)=227.5$$

in units of 0.01 mm.

Conventionally, the arithmetic mean, $\bar{y}$, of a set of n values of y is expressed

$$\bar{y}=\frac{1}{n}\overset{n}{S}(y)$$

where $\overset{n}{S}$ denotes summation over the n values of the enclosed quantities following it.

But the mean alone is not enough; for the degree of confidence it invites depends not only upon its weight in number of observations but upon the variation among individual observations as well. It is necessary therefore to give some special attention to variation.

The difference between the observed values of the sample and their arithmetic mean, that is, quantities expressed individually in the form

$$(y-\bar{y})$$

are called *residuals*. Thus the observations which exceed the mean supply positive residuals, and those which fall short of the mean supply negative residuals.

Two important properties of residuals are the following:

1. *The algebraic sum of residuals is zero.* This follows at once, for

$$\overset{n}{S}(y-\bar{y})=\overset{n}{S}(y)-n\bar{y}=0$$

since the product of the mean and the number of observations upon which it is based is equal to the sum. In the case of the sample of four annual ring measurements the sum of the residuals may be expressed

$$(227-227.5)+(226-227.5)+(227-227.5)+(230-227.5);$$

and this may be written

$$(-0.5)+(-1.5)+(-0.5)+(2.5)=0,$$

or, in alternative form, as

$$227+226+227+230-4(227.5)=0.$$

2. *The sum of squares of residuals is minimum.* If a set of n measurements of y is to be characterized by some unknown constant, say a, the sum of squares

$$\overset{n}{S}\left[(y-a)^2\right]$$

is a minimum when a is the mean value of the set. For upon differentiating the above expression with respect to the (as yet) unknown a, and equating the first derivative to zero, it follows that

$$-2\overset{n}{S}(y)+2na=0$$

whence, after dividing by 2

$$a=\frac{1}{n}\overset{n}{S}(y)=\bar{y}$$

and this is the mean value of the set. Therefore

$$\overset{n}{S}\left[(y-\bar{y})^2\right]$$

is the minimum sum of squares which can be derived from the sample.

The average value of the squared residuals of a sample of n observations, that is

$$\frac{1}{n}\overset{n}{S}\left[(y-\bar{y})^2\right]$$

is known as the *variance of the sample.* Its square root, taken positively, is called the *standard deviation of the sample.*[2]

The standard deviation of the sample of four tree ring measurements, is, accordingly, the square root of

$$\frac{1}{4}\left[(-0.5)^2+(-1.5)^2+(-0.5)^2+(2.5)^2\right]=2.25.$$

This form, which follows directly from

$$\frac{1}{n}\overset{n}{S}\left[(y-\bar{y})^2\right]$$

[2] This definition of the variance (or standard deviation) of the sample should be kept clearly in mind. Later it is to be distinguished from estimates of the *population* variance (or standard deviation) as derived from the sample.

is simplest, for purposes of calculation, only when the sample is small, and when the arithmetic mean does not contain continuing decimals. With larger samples, particularly if a calculator or table of squares is at hand, the preferred method of calculation is indicated by the expansion of the sum of squares of residuals; that is,

$$\overset{n}{S}\left[(y-\bar{y})^2\right]=\overset{n}{S}(y^2)-2\bar{y}\overset{n}{S}(y)+n\bar{y}^2$$

which may be written

$$\overset{n}{S}\left[(y-\bar{y})^2\right]=\overset{n}{S}(y^2)-\bar{y}\overset{n}{S}(y) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

after noting that

$$n\bar{y}^2=\bar{y}(n\bar{y})=\bar{y}\overset{n}{S}(y).$$

Upon applying the right-hand member of equation (1) to the four annual ring observations, we have

$$(227)^2+(226)^2+(227)^2+(230)^2-(227.5)(910)$$

and this is equal to

$$207{,}034-207{,}025=9.00.$$

The variance of the sample is $(1/n)$th of the sum of squares and the standard deviation of this sample is therefore the square root of 2.25, as before.[3]

The standard deviation is, accordingly, a *measure of dispersion* among the observations. Its range is from zero, in which case the observations are all identical, through small values if they are fairly consistent, to high values as they become discordant. It is through the standard deviation that one arrives at the accuracy of the observations or the degree of confidence one is entitled to place in conclusions drawn from them.

1.3 The Sample and the Population. The four measurements on annual ring width used in the previous section comprise a sample of observations drawn from a hypothetical *infinite population* of such measurements of the same physical magnitude, as might occur under essentially the same conditions. In this case the population is wholly the outcome of accidental errors of hypothetical measurements. In this sense, the numerical value of the *population mean* cannot be known exactly. It

[3] Certain shorter methods of calculating the mean and standard deviation of the sample have been found useful when samples are large. *See*, for example, Bruce, D, and F. X. Schumacher, Forest Mensuration. McGraw-Hill Book Co., New York, 1935, Chapter VI.

does not follow, however, that the population mean is the *true* magnitude, unless *bias*, or systematic errors which tend to affect all observations alike, have been completely eliminated. Sources of bias and their elimination will be discussed in Sec. 12.2.

The following discussion is not concerned with systematic errors. It should be stressed, however, that it is the part of any worth-while observational program to eliminate systematic errors insofar as possible.

An accidental error, accordingly, may be regarded as an observed value, say y, as a deviation from the mean, μ, of the population from which it is drawn; hence, symbolically,

$$y-\mu$$

is an accidental error. An accidental error may be considered as the effect of a multiplicity of causes, each of which contributes independently either a positive or a negative portion, the error itself being the sum of the contributed portions.

In forestry, and other biological work, however, one is not usually concerned with populations of accidental errors due to measurements taken on the same physical magnitude. One deals most commonly with populations of measurements taken on different magnitudes of the same class, as for instance, the population of individual tree diameters which occur in a forest. Such populations are the outcome of biological variation, the causes of which are not entirely independent of one another.

Any population may be considered as characterized by certain numerical constants, or *parameters*—such as its mean or its standard deviation—the exact values of which, in the case of infinite populations, cannot be known, except perhaps with certain games of chance. From a sample, however, one may calculate exact numerical constants, or *statistics*, as estimates of corresponding parameters of the population. An illustration will clarify the distinction.

A population is chosen the parameters of which are known a priori. Suppose a pack of 10 playing cards is made up of an ace, 2, 3, 4, 5, 6, 7, 8, 9, and one other to represent zero. If the pack be shuffled so that a card (say the top one) to be drawn therefrom has exactly the same chance of selection as any other of the ten, it can be said that the card selected has been drawn *at random* from the pack; hence, the probability that it represents any particular digit of the supply

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9$$

is exactly 1/10. If the card be replaced, the pack reshuffled, and a second draw made, the value of this draw is quite independent of that of the first, and the probability is, again, exactly 1/10 that the new card

represents a particular digit. It can be said that the two cards were drawn *independently and at random* from the *infinite* population of digits represented; for the drawing of a card with replacement from a finite population is tantamount to the drawing of a card without replacement from an infinite population.

If the game now be carried on; that is, after shuffling the pack, let the top card be withdrawn, its value noted, and replaced in the pack. In n such draws the *expected frequency* of occurrence of each digit is, of course, $\frac{1}{10}n$, and the distribution in the population is said to be rectilinear. Since each card is drawn independently and at random, the n observations together make up a *random sample* from the unlimited supply, or infinite population of such digits.

In this game of chance the exact values of the parameters are known. The population mean, μ, is the arithmetic average[4] of the digits

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9,$$

and therefore

$$\mu = 4.5 \text{ exactly;}$$

while the variance of the population—commonly symbolized as σ^2—is the average of the squares of the 10 quantities $(0-4.5)$, $(1-4.5)\ldots.$ $(9-4.5)$, or,

$$\sigma^2 = 8.25 \text{ exactly.}$$

Sampling the digits by means of the card game becomes tedious. Recourse will therefore be had to Tippett's Random Sampling Numbers, a collection of over 40,000 digits which have been taken at random from census tracts and reports.[5] Tippett's numbers are particularly suited to the great variety of problems designed to test statistical theorems by means of artificial random samples. Figure 1 is a reproduction of the first page. The tract consists of 26 such pages. Until these numbers were available, artificial sampling was based upon drawing varicolored balls from an urn, cards from a shuffled pack, or the tossing of coins or dice. Such methods are not always free from bias and they are usually time-comsuming. The labor of drawing random samples by means of Tippett's numbers is trifling by comparison.

Another set of random numbers is that of Fisher and Yates (1938, Table XXXIII). They constructed the set from the 15th to 19th digits of a 20-figure logarithm table.

[4] The authors will usually adhere to the convention of denoting parameters of the infinite population by Greek symbols, and sample statistics by Roman symbols.

[5] Tippett, L. H. C., Random Sampling Numbers. Tracts for Computers XV. Cambridge University Press, 1927.

(1) RANDOM SAMPLE NUMBERS

1–4	5–8	9–12	13–16	17–20	21–24	25–28	29–32
2952	6641	3992	9792	7979	5911	3170	5624
4167	9524	1545	1396	7203	5366	1300	2693
2730	7483	3408	2762	3563	1089	6913	7691
0560	5246	1112	6107	6008	8126	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446
5870	2859	4988	1658	2922	6166	6069	2763
9263	2466	3398	5440	8738	6028	5048	2683
2002	7840	1690	7505	0423	8430	8759	7108
9568	2835	9427	3668	2696	8820	1955	6515
8243	1579	1930	5026	3426	7088	3991	7151
5667	3513	9270	6298	6396	7306	7898	7842
1018	6891	1212	6563	2201	5013	0730	2405
6841	5111	5688	3777	7354	3434	8356	6424
2041	2207	4889	7346	2865	1550	5960	5479
5565	4764	2617	5281	1870	6497	5744	9576
4508	1808	3289	3993	9485	4240	2835	9955
2152	6473	5692	9309	7661	1668	5431	7658
6917	4113	7340	6853	1172	7229	1279	5085
8241	4124	4131	9500	5657	3932	5942	3317
7913	3709	5944	9763	2755	4211	4996	8657
9385	7125	3230	0737	2957	1013	6369	4494
3436	6293	6025	9384	3343	1071	1468	4801
9094	1634	5070	0664	6510	0918	4601	4294
9226	9296	2796	7097	4057	2074	6297	2587
7781	3760	2895	7653	0091	7012	1308	1946
9742	9694	7347	0017	9572	1850	0116	1899
9420	9210	8787	9375	4663	0396	6717	5562
1179	3571	5992	3059	9015	5608	2348	8144
0708	4011	4057	1550	1674	1376	5243	4427
6350	3996	3795	2176	8182	4514	6349	3483
1414	7152	3658	1636	0638	3443	4440	3086
7041	8985	7011	5676	7570	6685	1776	3154
3243	2783	0840	9054	8862	5173	8433	9117
7922	4931	5753	6160	6566	8602	3423	9074
8769	3513	8976	0780	6382	0029	2619	6982
2510	7274	8743	0000	1850	2408	3602	5179
0224	2404	9811	6641	9732	1662	9158	1404
3009	8516	7245	9409	2844	0717	1072	3137
7489	0221	7921	2351	2696	4906	2484	3868
5188	1825	2220	9382	0532	1915	1790	2081
1198	2545	2482	9607	0067	3744	9866	5096
3908	4676	7816	6517	9121	3171	4119	3615
1094	2223	1675	2282	3712	8191	1330	1454
1817	7723	5582	7153	9518	0231	7782	5742
6208	9598	9623	2114	7747	2096	5027	0561
4752	4519	2749	8020	4642	1190	7302	8350
0486	6993	3115	5025	4887	1571	9819	6804
4942	3004	1442	2810	1479	0970	7302	3775
4930	9785	7460	3996	2864	0559	3985	8092
2349	1594	7152	0257	4041	4105	3180	9806

FIG. 1. A page of Tippett's Random Sampling Numbers.

Using Tippett's Random Sampling Numbers by way of illustration the sampling of the population is easily performed by noting the digits as they occur, commencing, for instance, with the first column of page 1 (Fig. 1). The actual frequency distribution of the 100 digits in the first two columns is shown graphically in Figure 2, and is tabulated in Table 1. If this is, in fact, a random sample, the deviations of the observed frequencies from the expected frequency of 10 of each digit is entirely fortuitous.

Denoting the individual observations by y, and the mean of the sample by $\bar{y}$, we have

$$\bar{y} = \frac{1}{n} \overset{n}{S}(y) = \frac{1}{100}(435) = 4.35$$

as the estimate of the population mean, a statistic of 4.35 as an estimate of the parameter 4.50.

The variance of the sample, that is,

$$\frac{1}{n} \overset{n}{S}\left[(y-\bar{y})^2\right] = \frac{1}{n}\left[\overset{n}{S}(y^2) - \bar{y} \cdot \overset{n}{S}(y)\right],$$

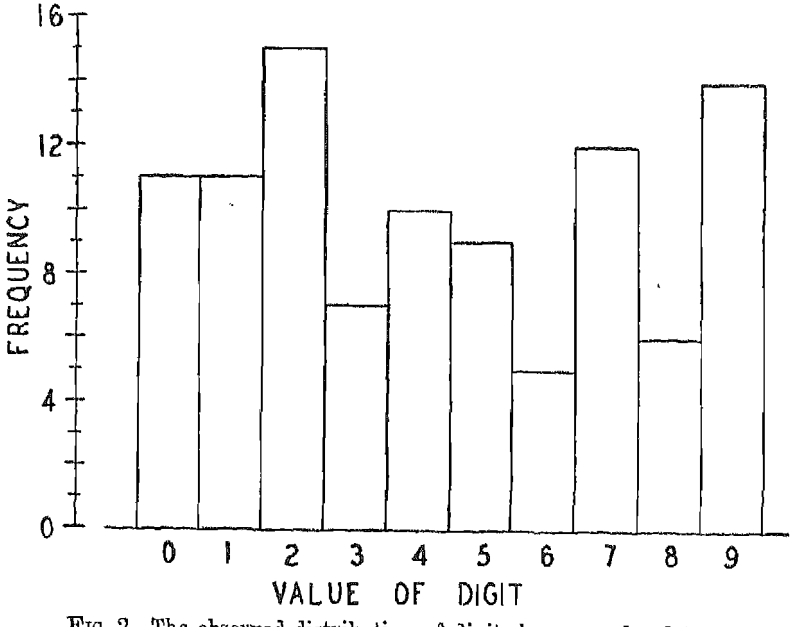


FIG. 2. The observed distribution of digits in a sample of 100, drawn from Tippett's Random Sampling Numbers.

is, numerically,

$$\frac{1}{100}(2{,}805 - 1{,}892.25) = 9.1275;$$

whereas the corresponding population parameter is 8.2500. Calculations leading up to these values are given in Table 1.

TABLE 1. Frequency Distribution of a Random Sample of 100 Digits, and Calculation of Mean and Sum of Squares of Residuals

Digit Value y	Frequency f	fy	$y \cdot fy$
0	11	0	0
1	11	11	11
2	15	30	60
3	7	21	63
4	10	40	160
5	9	45	225
6	5	30	180
7	12	84	588
8	6	48	384
9	14	126	1,134
Sum	100	435	2,805

Calculation of

Mean	Sum of Squares of Residuals		
$\bar{y} = \frac{1}{100}(435)$	$\overset{100}{S}(y^2)$	=	2,805
$= 4.35$	$\bar{y} \cdot \overset{100}{S}(y)$	=	1,892.25
	$\overset{100}{S}\left[(y-\bar{y})^2\right]$	=	912.75

1.4 The Distribution of Means of Independent Observations and the Normal Curve of Error. As each of the digits in the population just used occurs with equal frequency, the distribution of digits is rectilinear. But the distribution of means, of two or more digits, takes on a different form, as we may observe by direct sampling.

In Figure 3 the frequency distribution of the means of 550 samples of five digits each, taken from Tippett's Random Sampling Numbers, is presented; and in Figure 4 the distribution of the means of 550 samples of 10 digits each, from the same source.

From these distributions it is apparent that the sample means tend to cluster around the population mean of 4.5, the larger of the two sample sizes (Fig. 4) with noticeably less dispersion.

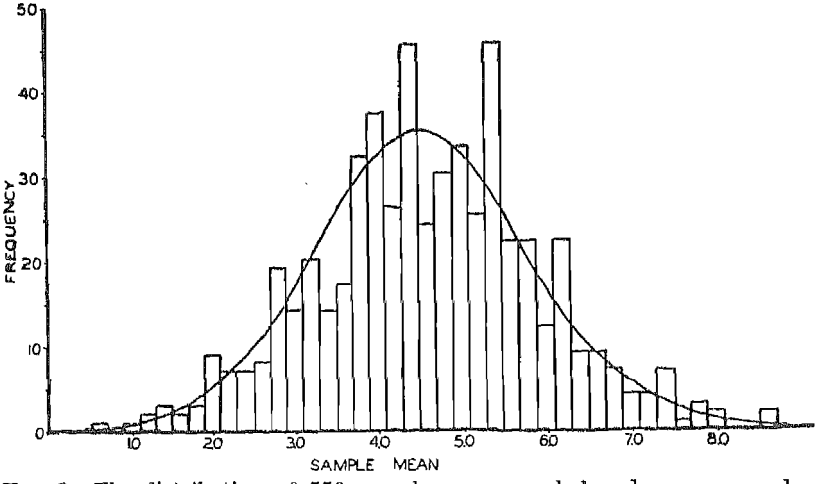


FIG. 3. The distribution of 550 sample means, each based upon a random sample of five observations from a rectilinear population.

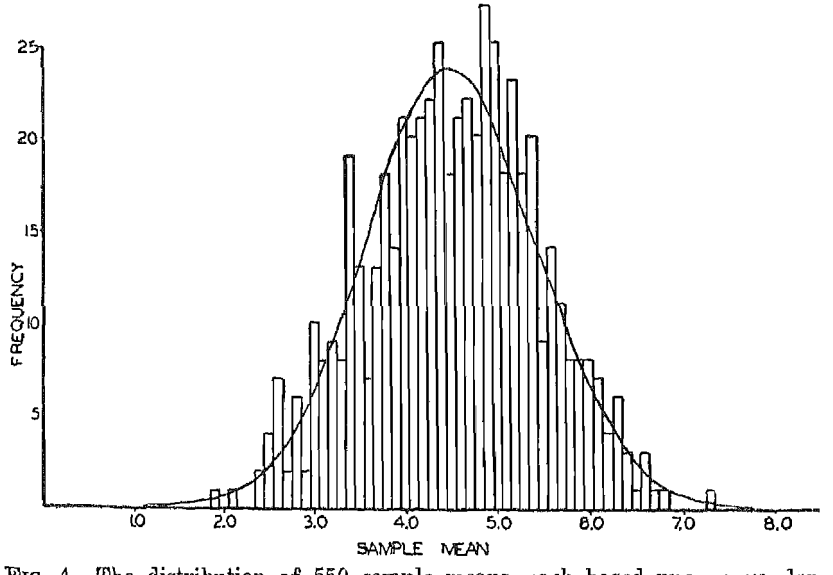


FIG. 4. The distribution of 550 sample means, each based upon a random sample of ten observations from a rectilinear population.

These observations conform with experience. A great number of investigations of a wide variety of kind has demonstrated that the distribution of sample means, when each is based upon a given number of

independent observations, tends to a definite form, in common with the distribution of accidental errors of measurements taken on a given physical magnitude. Certain general features of such distributions are the following:

1. Positive and negative errors are equally likely to occur.
2. Small errors are considerably more likely than large errors.
3. Errors beyond some undefined magnitude do not occur.

The distribution of accidental errors has led to what is known as the normal curve of error, or the normal distribution.[6] The curve representing it is symmetrical about zero, relatively high in the center, and falls off to exceedingly small values at any considerable distance from the center. Its equation is

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

where Y, the ordinate of the curve, is the relative frequency, or probability, of an error in the infinitesimal range dy; $(y-\mu)$ is the error; and σ, called the *standard deviation*, is a measure of the dispersion of the individual errors. The numerical equivalents of μ and σ are the only characteristics of a normal population that are needed to define its distribution completely. These will be further discussed shortly.

In Figures 3 and 4 the distributions of observed sample means are compared with the normal curve of error fitted thereto.[7]

The great utility of the normal curve of error lies in the fact that the distribution of many statistics—such as means, or sums, of random samples—tends to the normal form as the size of sample is increased, even though the distribution of single observations—or single variates, as they are called—be of radically different form.

Insofar, then, as a single sample, of size n, supplies a satisfactory estimate of the standard deviation of *means* of n single variates, the sample statistics can be made to afford the information concerning the probable discrepancy between the true, but unknown, population mean and the estimate thereof as derived from the sample.

1.5 Variance of the Sample and of the Population. The variance of the population is the numerical value towards which the variance of the sample tends as the size of the sample approaches that of

[6] The development of the normal curve of error may be found in any complete text on least squares, such as Brunt, 1931.

[7] Methods of fitting the normal curve of error to observational data need not be given here. For details, *see*, for example, Bruce, D., and F. X. Schumacher, Forest Mensuration. McGraw-Hill Book Co., New York, 1935.

the population from which it is drawn. In unlimited natural populations—such as occur in forestry, or biology generally—the exact value of the population variance, σ^2, is not known, just as the exact value of the population mean, μ, is not known. It is required, however, to make unbiased estimates of these parameters through the device of drawing one or more random samples.

The sample mean is in itself an unbiased estimate of the population mean. The sample variance, on the other hand, is not an unbiased estimate of the corresponding population parameter, σ^2. In order to clarify the nature of the bias—which is particularly acute when samples are small—we need to distinguish, again, between a residual, the deviation of an observed value, y, from the mean of the sample, $\bar{y}$, that is,

$$y-\bar{y},$$

and an error, which is the deviation of the observed value from the mean of the population, or

$$y-\mu.$$

This distinction may be illustrated by means of a sample from the hypothetical supply of digits 0, 1, 2,. . . ., 9, for it is known, in this game of chance, that $\mu=4.5$ exactly. Turning to page 1 of Tippett's Random Sampling Numbers (reproduced in Figure 1), the first five digits of column 1 are found to be

$$2, 4, 2, 0, 2.$$

Calculations based upon this sample are shown in Table 2, the first column of which lists the numbers in the order of draw. In the second column are the squares of the residuals, and in the third column the squares of the errors.

The mean square of the errors, 7.85 in this case, is entirely independent of the sample mean and is, therefore, an unbiased estimate of 8.25, the population variance.

The mean square of the residuals, 1.60, which is the minimum mean square to be derived from these numbers (Sec. 1.2) is immediately recognized as the variance of the sample. It cannot be greater than the mean square of the errors; and it is less than the latter whenever the sample mean differs from the population mean, regardless of whether in positive or negative direction. This bias in the sample variance—correction of which will be treated shortly—becomes of little practical importance with sufficient increase in sample size; for in large samples, residuals tend toward errors by the fact that the difference between sample and population means tends toward zero. The difference between the variance of the sample and the mean square of the errors is precisely

the square of the difference between sample and population means. This may be checked in the table.

TABLE 2. A Random Sample of Five Digits from a Population of Known Mean

Sample digits y	Squared residuals $(y-2.0)^2$	Squared errors $(y-4.5)^2$
2	0.00	6.25
4	4.00	0.25
2	0.00	6.25
0	4.00	20.25
2	0.00	6.25
Mean 2.0	1.60	7.85

For the sake of simplification in later work, the population variance will be expressed in slightly different symbolic form. Let ϵ be a deviation from the population mean, that is, an error; let n be the sample size in number of observations of ϵ; let N be the total (indefinitely large) number of hypothetical samples of size n in the population. Then the population of errors consists of N sets of n values of ϵ, and the population variance, σ^2, may be written

$$\sigma^2 = \frac{1}{Nn} \overset{N}{S} \overset{n}{S}(\epsilon^2)$$

where the double summation, $\overset{N}{S} \overset{n}{S}$, denotes summation over the N samples, of the sums over the n observations of ϵ^2 in each sample. The above may also be expressed

$$\sigma^2 = \frac{1}{N} \overset{N}{S} \left[\frac{1}{n} \overset{n}{S}(\epsilon^2) \right],$$

and this is the average value of the squares of all errors in the population. An unbiased estimate of the population variance, as afforded by a random sample of n observations of ϵ, may therefore be expressed

$$s_\epsilon^2 = \frac{1}{n} \overset{n}{S}(\epsilon^2) \rightarrow \sigma^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where s_ϵ^2 is the variance of the errors of a sample and an unbiased estimate[8] of the population variance, σ^2, since the latter is the average of s_ϵ^2 over all samples.

[8] The symbol $\rightarrow$ is read "is an estimate of."

Expression (2) is not a practical estimate of a population variance because no sample contains, in itself, the errors ϵ. The expression is, however, a logical step in the elucidation of estimates of population variance, useful in practice. As such, it will be used in the next two sections.

1.6 Variance of Sums and of Means of Independent Observations. It has been noted that if every single variate of a population is regarded as having an equal and independent chance of being drawn, one that is actually taken may be said to have been drawn independently and at random. This particular one may, of course, have a positive or negative error, high or low. But whatever its error, it indicates nothing concerning the error of a second—or any succeeding—observation drawn under the same conditions. A sample of such observations is a random sample.

The implication contained in a random sample may be readily illustrated. Suppose, for instance, all the samples of a very large population are available, each consisting of just two observations drawn independently and at random. Suppose, further, that the first observation of each sample is plotted upon the second in a system of rectangular coordinates, the ordinate of the graph representing the scale of error of the first observation, and the abscissa that of the second. It is not at all necessary that the errors be normally distributed.

Before a sample is drawn, then, each of the four quadrants has precisely the same chance of receiving it. Consequently, after all sample points have been plotted the graph exhibits a circular cluster with center at the zero origin of coordinates. If, now, one calculates the product of the two errors representing each sample, that is, the product of ordinate and abscissa of each point, those which fall within the first and third quadrants are positive, while those within the second and fourth quadrants are negative. And the sum of products over the four quadrants is zero because of the symmetry of the cluster.

From the above discussion it follows that in random samples of two errors, ϵ, the average value of the square of their sum, over the entire population, is equal to the average value of the sum of their squares, since the average value of the product of the two errors is zero. Extending this line of reasoning to random samples of any size n, each of the $n(n-1)$ products of errors of different order of draw in the same sample, totals to zero exactly, over all samples in the population. Consequently, the average value of

$$(\epsilon_1+\epsilon_2+\ldots.+\epsilon_n)^2$$

where the subscripts represent the order of draw, is, over all random samples of size n in the population, equal to the average value of

$$\overset{n}{S}(\epsilon^2).$$

The importance of these deductions lies in the fact that the variance of the sample sum, or sample mean, is immediately expressible in terms of the variance of single variates. Given a population of single variates y, each of which has a true error, ϵ, such that

$$\epsilon = (y - \mu)$$

then the variance of y is the variance of ϵ. And the *sampling variance*—that is, the estimate of the average variance over all samples of the population—of the sum of n random values of y, which we may symbolize

$$V(y_1 + y_2 + \ldots + y_n) = V\left[\overset{n}{S}(y)\right]$$

where V denotes the sampling variance of the enclosed terms following it, may be written (expression (2), Sec. 1.5)

$$V\left[\overset{n}{S}(y)\right] = \overset{n}{S}(\epsilon^2) \rightarrow n\sigma^2 \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The sampling variance of a sum of n single variates is therefore n times the variance of single variates.

The sampling variance of a mean follows at once. By definition, the mean of n values of y is

$$\bar{y} = \frac{1}{n}\overset{n}{S}(y),$$

and the variance of this mean is the average over all samples of

$$\left[\frac{1}{n}(\epsilon_1 + \epsilon_2 + \ldots + \epsilon_n)\right]^2 = \frac{1}{n^2}\overset{n}{S}(\epsilon^2).$$

Comparing this with equation (3), it is evident that the sampling variance of mean y may be expressed

$$V(\bar{y}) \rightarrow \frac{1}{n}\sigma^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4),$$

that is, the sampling variance of mean y is the variance of y divided by the number of observations upon which the mean is based.

1.7 Estimate of Population Variance from a Sample. While dealing with estimates of the population variance, and the sampling variance of sums and means, it has been supposed that the popula-

tion mean was a known parameter; hence, we were enabled to make unbiased estimates of population variances directly from the known errors. In practice, however, the individual errors are not known because the population mean is unknown. The sample itself merely supplies residuals as estimates of corresponding errors.

The problem now pertains to the estimation of the mean square of errors—the population variance—from the mean square of residuals—the sample variance.

Suppose one has at hand a random sample of n values of y drawn from a population whose mean value, μ, is unknown. The sample mean, $\bar{y}$, is an observable statistic and an unbiased estimate of μ. Let each

$$\epsilon = y - \mu$$

be the unknown error of an individual y. And let

$$\bar{\epsilon} = \bar{y} - \mu$$

be the unknown error of mean y, that is, of $\bar{y}$. Then, of course, each

$$y - \bar{y} = \epsilon - \bar{\epsilon},$$

the right-hand member being the expression of an error in terms of corresponding residual. The variance of the sample of y is then

$$\frac{1}{n} S^{n}\left[(y-\bar{y})^2\right] = \frac{1}{n} S^{n}\left[(\epsilon-\bar{\epsilon})^2\right]$$

Upon comparing the right-hand member with equation (1), Sec. 1.2, the above identity may be written such that

$$\frac{1}{n} S^{n}\left[(y-\bar{y})^2\right] = \frac{1}{n}\left[S^{n}(\epsilon^2) - n\,\bar{\epsilon}^2\right]$$

$$= \frac{1}{n} S^{n}(\epsilon^2) - \bar{\epsilon}^2$$

$$= s_{\epsilon}^2 - \bar{\epsilon}^2$$

where s_{ϵ}^2 is the estimate of the population variance σ^2, it being the mean square of the n errors. And $\bar{\epsilon}^2$ is the square of the error of the sample mean.

Now the average of s_{ϵ}^2 over all samples of the population is σ^2 (Sec. 1.5), and the average value of $\bar{\epsilon}^2$ is the variance of the sample means, which from equation (4), Sec. 1.6, is $\frac{1}{n}\sigma^2$. Hence, over all samples of size n in the population the average variance of the samples is the average of

$$\frac{1}{n} S^{n}\left[(y-\bar{y})^2\right]$$

over all samples; and this average may be expressed

$$\sigma^2-\frac{\sigma^2}{n}=\sigma^2\left(\frac{n-1}{n}\right)$$

It follows, then, that the variance of a sample, as calculated from the n observations of a single sample, is

$$\frac{1}{n} S^{n}\left[(y-\bar{y})^2\right] \rightarrow \sigma^2\left(\frac{n-1}{n}\right).$$

Upon multiplying both sides by $\left(\frac{n}{n-1}\right)$ we find that

$$\frac{1}{n-1} S^{n}\left[(y-\bar{y})^2\right]=s^2 \rightarrow \sigma^2.$$

Hence an unbiased estimate of the population variance, σ^2, is obtained by dividing the sum of squares of residuals of a single sample by one less than the number of observations. And it is said that this estimate of σ^2 is based upon $(n-1)$ *degrees of freedom.* The equivalent of one observation has been sacrificed since the sample does not directly supply the sum of squares of the errors.

It is helpful to bear in mind that one degree of freedom is sacrificed because the sample mean is taken as the estimate of the population mean and, consequently, that the estimate of the population variance is based upon the squares of residuals, that is, of deviations about the sample mean. If the deviations are measured from any locus, the choice of which is quite independent of information contained in the sample, the degrees of freedom and number of observations are identical.

Consider, by way of illustration, a sample of original observations

3, 4, 5.

Each of these is, by definition, a deviation from zero. Their sum of squares

$$9+16+25=50$$

rests upon three degrees of freedom. Should the sum of squares about zero be adjusted to the sum of squares about the sample mean of 4, by deducting the product of mean and sum, that is

$$4(3+4+5)=48,$$

this correction term is itself based upon one degree of freedom, so that the sum of squares of residuals

$$50-48=2$$

rests upon

$$3-1=2$$

degrees of freedom.

From a slightly different point of view it may be construed that since the estimate of the population variance must rest upon the squared residuals, which depend in turn upon the sample mean, there are but $(n-1)$ independent comparisons in a sample of n observations. In particular, there are two independent comparisons in the sample

$$3,\ 4,\ 5,$$

for two of these observations can take on any value whatever but the third must thereby be fixed in order that the mean be 4.

CHAPTER II

OBSERVATION AND EXPECTATION

2.1 A Few Points about the Normal Curve of Error. It was brought out in the preceding chapter that if the variance of the population of single observations is σ^2, then the means of random samples of n observations tend to be distributed normally with variance $\frac{1}{n}\sigma^2$, even when the original observations are not so distributed.

The great utility of the normal curve of error in the biological sciences follows directly from this fact.

The equation of the normal curve was given in Sec. 1.4 as follows:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

This form may be appreciably simplified if the error $(y-\mu)$ is measured in units of the standard deviation, σ; that is, let

$$t = \frac{y-\mu}{\sigma}$$

be an error expressed in *standard units*, or units of σ. In these units, distributions of errors of entirely different order of absolute magnitude are comparable. If the frequencies are expressed as relative parts of the total, the area under the normal curve is unity, and the curve may be expressed

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

This equation is shown graphically in Figure 5, the abscissal units being identical with units of t.

Interest lies more commonly in the area under certain sections of the curve than in its ordinates; for the area bounded by a segment of the base line—that is, between two values of t—and corresponding ordinates is proportional to the expected frequency of observations between the same limits.

Areas beneath the normal curve of error are listed in Table 3 according to selected values of t. As the curve is symmetrical about $t=0$,

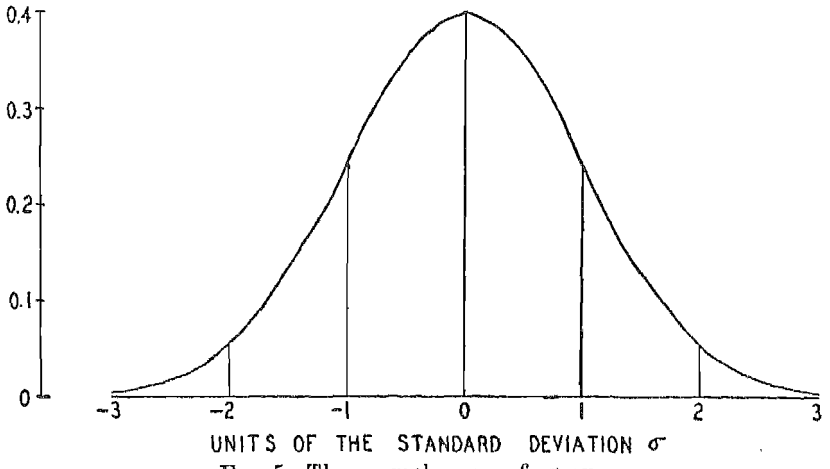


FIG. 5. The normal curve of error.

the positive half alone is given, but the area as listed is the accumulation from the left extreme of the curve up to the given positive t. Subtraction of a listed area from total area, unity, leaves the area which lies to the right of the selected positive value of t, or to the left of the corresponding negative value.

It is useful to remember that 32 percent of the area lies outside the limits bounded by plus and minus one standard unit; hence, the odds are 68 to 32, or about 2 to 1, that an observed value drawn at random from a normally distributed population is within these limits.

Three standard units mark the practical range of the curve for most purposes, since only 0.3 percent lie in the tails beyond $t=3$ on both sides of the zero origin.

TABLE 3. Area Under the Normal Curve of Error from the Left Extreme to Given Positive Values of t

Number of standard units t	Area	Number of standard units t	Area
0	0.5000	1.6	0.9452
0.2	0.5793	1.8	0.9641
0.4	0.6554	2.0	0.9772
0.6	0.7257	2.2	0.9861
0.8	0.7881	2.4	0.9918
1.0	0.8413	2.6	0.9953
1.2	0.8849	2.8	0.9974
1.4	0.9192	3.0	0.9987

Another form of tabulation of the area under the normal curve of error is according to Table 4. In the use of this table one starts with a selected relative area in both tails of the normal curve and reads off the number of standard units, t, which divides the area into this, and the remaining, proportion.

The use of these table will next be illustrated.

TABLE 4. Values of t, Outside the Range of Which, in Both Tails, Lie Selected Proportions of the Area Under the Normal Curve of Error*

Relative area in both tails	t	Relative area in both tails	t
1.0	0	.4	0.8416
.9	0.1257	.3	1.0364
.8	0.2533	.2	1.2816
.7	0.3853	.1	1.6449
.6	0.5244	.05	1.9600
.5	0.6745	.01	2.5758

*This table is taken from the bottom line of Fisher's *Table of t*. See Table 7.

2.2 Calculation of Expected Frequencies of Normally Distributed Variates. By way of illustration of the use of Table 3, consider the following problem: On the supposition that for practical purposes the means of samples of five digits taken from Tippett's Random Sampling Numbers are distributed normally, in what proportion should the mean be 6.0 or less?

The observed distribution of 550 means of five are listed in Table 5.

TABLE 5. Distribution of 550 Means of Five Digits from Tippett's Random Sampling Numbers

Mean	Frequency	Mean	Frequency	Mean	Frequency	Mean	Frequency
.6	1	2.8	19	5.0	33	7.2	4
.8		3.0	14	5.2	25	7.4	7
1.0	1	3.2	20	5.4	45	7.6	1
1.2	2	3.4	14	5.6	22	7.8	3
1.4	3	3.6	17	5.8	22	8.0	2
1.6	2	3.8	32	6.0	12	8.2	
1.8	3	4.0	37	6.2	22	8.4	
2.0	9	4.2	26	6.4	9	8.6	2
2.2	7	4.4	45	6.6	9		
2.4	7	4.6	24	6.8	7		
2.6	8	4.8	30	7.0	4		

The mean of the population is 4.5 (Sec. 1.3). The population of single digits has a variance of 8.25 (Sec. 1.3). Hence the standard devi-

ation of means of five is the square root of $\frac{8.25}{5}$ or 1.285 (Sec. 1.6). These parameters, 4.5 and 1.285, are all that are needed to define a normal curve of error completely. But the actual distribution is not continuous. It is a series of discrete classes, 0.2 units wide. In order to compare it with the continuous curve, we note that 6.1 would be the upper limit of the class 6.0, and the lower limit of the class of 6.2. Hence the number of standard units which divides the distribution into classes 6.0 and less, on the one hand, and 6.2 and greater, on the other, is

$$t=\frac{6.1-4.5}{1.285}=1.25 \text{ approximately.}$$

Referring this positive value of t to Table 3, the area to the left of it is found to be 0.89 approximately. Hence 89 percent of the means of random samples of five digits should in the long run be less than 6.1, and 11 percent should be greater. In 550 such samples, these percentages correspond to frequencies of 490 and 60 respectively. The observed frequencies (Table 5) are 480 and 70.

As another illustration, the expected proportion whose means are between 3.0 and 6.0 inclusive, might be calculated.

The lower limit of a continuous variate grouped into classes of 0.2 units interval whose mid-point is 3.0 is, of course, 2.9. The corresponding standard unit is

$$t=\frac{2.9-4.5}{1.285}=-1.25 \text{ approximately.}$$

In this case we need the relative area under the normal curve between the limits $t=-1.25$ and $t=+1.25$. Evidently $89-50=39$ percent of the area is between the zero origin and $t=+1.25$. Because of the symmetry of the curve this proportion also lies between $t=-1.25$ and the origin. Hence 78 percent lies between the standard units of plus and minus 1.25. In a total of 550 samples, then, about 429 should have mean values between 3.0 and 6.0 inclusive. By direct observation in Table 5, we have 418.

As an illustration of the use of Table 4, let it be required to calculate the range within which 95 percent of the means of five digits should fall. One needs therefore a value of t which encloses, between its positive and negative values, just 95 percent of the area under the normal curve of error, and outside of which in both tails of the normal curve lies 5 percent of its area. The proper value of t from Table 4 is 1.9600.

Since the standard deviation of the distribution under discussion is 1.285, the range we seek is from

$$4.5-1.9600(1.285)=1.98$$

to

$$4.5+1.9600(1.285)=7.02$$

or approximately from 2 to 7 inclusive. Hence in 550 samples we should expect—on the supposition of a continuous distribution—95 percent of 550 or about 522 to have means between the classes 2.0 and 7.0 inclusive.

The corresponding observed frequency, from Table 5, is 519.

2.3 Sample Size and the Normality of Distribution of Sample Means. The above illustrations are concerned with problems of distribution. We have supposed that means of random samples of five single digits from a rectilinear population of digits is distributed according to the normal curve of error. Such an hypothesis in not untenable unless attention is focused upon comparisons between observation and expectation near the extremes of the distributions. In these regions the hypothesis that the observations are normally distributed is incompatible with fact; for the actual distribution is limited between 0 and 9, whereas the normal curve is unlimited.

The criterion as to whether the normal curve of error is a satisfactory description of the distribution of means of random samples, is a practical one. It depends upon the number of standard units between the population mean and the limit of its range that is considered to be a sufficient approach to infinity. This number may be conveniently set at 4 for most purposes, for the area in both tails beyond $t=4$ is only about 64 parts in a million.

How a knowledge of the distance between known limits of a range is useful may be illustrated by means of a concrete example.

Suppose an estimate of the number of 1-year-old seedlings on a forest floor is needed. It would be convenient to conceive the area as subdivided into many small quadrats, each of which contains one of the numbers, 0, 1, 2, etc., of seedlings. If, now, the population mean be 1.0 seedlings to the quadrat, and the standard deviation of quadrats be also 1.0 seedlings, what should be the minimum number, n, of quadrats in a random sample such that the sample mean be normally distributed?

The standard deviation of the sample mean, based upon n quadrats, will be $\frac{1}{\sqrt{n}}$, (Sec. 1.6) since $\sigma=1$ seedling. This is one standard unit

of such sample means. The distance from the mean, 1.0, to the zero limit, is next equated to 4 of these, that is,

$$1.0 = \frac{4}{\sqrt{n}}$$

whence

$$n = 16.$$

It should be kept in mind that this problem is not concerned with precision of sample means, but only with the estimate of minimum sample size such that the mean is distributed in a known way, that is, according to the normal curve of error.

Precision is to be gained by increasing n to a size such that $\frac{\sigma}{\sqrt{n}}$ is sufficiently small for the job at hand. Should an estimate of the average number of seedlings to the quadrat be required with a precision such that the chances are 2 to 1 that it be correct within $\frac{1}{10}$-seedling, this is tantamount to the requirement that the standard error of the mean of an unknown number of quadrats be 0.10 seedlings; that is, that

$$\frac{\sigma}{\sqrt{n}} = 0.10$$

and since $\sigma = 1$, in the problem under discussion, we find that $n = 100$.

2.4 Estimate of the Mean of an Infinite Population from a Large Sample. In the applications of the normal curve of error in Sec. 2.2 we started from a population of known parameters, μ and σ, and inquired about the distribution of the means of random samples drawn therefrom. Our object was merely to show that the distribution of such means is sufficiently normal for the practical purpose at hand.

The deductive procedure from population to sample is, however, of only trifling value except, perhaps, in the use of gambling devices. Seldom can we specify biological populations with sufficient exactitude to deduce the distribution of random samples therefrom. The practical object in the sampling of populations is the application of the reverse process—that of specifying unknown population parameters, as nearly as may be done, from known statistics as derived from random samples of the population.

Let us now try this latter process. Suppose one is given the data of Table 5 and all that is known is that they are a single random sample of 550 observations from *some* population. The problem is to specify, as nearly as one can, the mean of the population represented.

The data are presented again in Table 6, together with a shortcut scheme, leading to the calculation of the mean and variance, in *coded* units, x. Upon comparing the first and third columns of Table 6, the *code* is found to be

$$x = 5\ (y - 4.4)$$

TABLE 6. Calculation of the Mean and Standard Deviation of Coded Observations

y	f	x	fx	$x.fx$
.6	1	−19	− 19	361
.8				
1.0	1	−17	− 17	289
1.2	2	−16	− 32	512
1.4	3	−15	− 45	675
1.6	2	−14	− 28	392
1.8	3	−13	− 39	507
2.0	9	−12	−108	1,296
2.2	7	−11	− 77	847
2.4	7	−10	− 70	700
2.6	8	− 9	− 72	648
2.8	19	− 8	−152	1,216
3.0	14	− 7	− 98	686
3.2	20	− 6	−120	720
3.4	14	− 5	− 70	350
3.6	17	− 4	− 68	272
3.8	32	− 3	− 96	288
4.0	37	− 2	− 74	148
4.2	26	− 1	− 26	26
4.4	45	0	0	0
4.6	24	+ 1	24	24
4.8	30	+ 2	60	120
5.0	33	+ 3	99	297
5.2	25	+ 4	100	400
5.4	45	+ 5	225	1,125
5.6	22	+ 6	132	792
5.8	22	+ 7	154	1,078
6.0	12	+ 8	96	768
6.2	22	+ 9	198	1,782
6.4	9	+10	90	900
6.6	9	+11	99	1,089
6.8	7	+12	84	1,008
7.0	4	+13	52	676
7.2	4	+14	56	784
7.4	7	+15	105	1,575
7.6	1	+16	16	256
7.8	3	+17	51	867
8.0	2	+18	36	648
8.2				
8.4				
8.6	2	+21	42	882
Sum	550		+508	25,004

In Units of x:

$\bar{x} = \dfrac{+508}{550} = 0.9236$

$S(fx^2) = 25{,}004.$

$\bar{x}.S(fx) = 469.2$

$S\left[f(x-\bar{x})^2\right] = 24{,}534.8$

$\dfrac{1}{549}\,S\left[f(x-\bar{x})^2\right] = 44.69$

$s_x = 6.685$

and the decoding equation is

$$y = 4.4 + 0.2x$$

The value of a code, in cases such as this, lies in the simplification of the arithmetic involved in the calculation of the mean and standard deviation in the coded units x. These statistics are conveniently transcribed to the original units of y at the end.

As worked out in the table, mean x is

$$\bar{x} = 0.9236;$$

hence, mean y is

$$\begin{aligned}\bar{y} &= 4.4 + 0.2(0.9236) \\ &= 4.585\end{aligned}$$

Next is needed the precision of this estimate of the population mean. From Table 6,

$$s_x = 6.685$$

and upon multiplying by the class interval,

$$\begin{aligned}s_y &= 0.2\ (6.685) \\ &= 1.337;\end{aligned}$$

hence, the sampling error, or *standard error*, of $\bar{y}$, that is, of 4.585,

$$\begin{aligned}SE(4.585) &= \frac{1.337}{\sqrt{550}} \\ &= 0.0570\end{aligned}$$

This is the estimate of the standard deviation—or one standard unit—of the distribution of means of 550 observations each. Combining it with the mean of 550 observations, one may now make exact probability statements concerning the range within which the population mean, μ, must lie. For instance, the probability is 0.68 that

$$\mu = 4.585 \pm 0.0570.$$

This means that the probability is 0.68 that the true population mean lies between $4.585 - 0.0570$, and $4.585 + 0.0570$, because 0.0570 is the value of one standard unit of the distribution of means of 550 random observations of the population, and the area under the normal curve between the positive and negative standard unit is 68 percent of the entire area under the curve.

2.5 The Probability of Discrepancy. An estimate of the population mean based upon a large sample—of the order, say, of hundreds of observations—is made with considerable confidence of precision, because the sampling variance of the means of samples of size n is always equal to

$\frac{1}{n}$ of the variance of the individual variates; and a large sample supplies an exact, or nearly exact, value of the true variance, σ^2. In such cases the distribution of t, where

$$t=\frac{\bar{y}-\mu}{\sigma/\sqrt{n}},$$

follows the normal curve of error with unit variance. This expression is of utmost importance in sampling work; for the numerator represents the *real, but unknown*, error of the sample mean. It is, of course, as likely to be negative as positive, for the normal curve is symmetrical. One may therefore write

$$t=\frac{|\bar{y}-\mu|}{\sigma/\sqrt{n}};$$

the numerator being enclosed between bars to indicate that it is taken without regard to sign. It is thus the *real discrepancy* between sample and population means.

The use of this expression with large samples may be illustrated by means of the sample of 550 observations of the previous section, for which

$$\bar{y}=4.585, \qquad SE\ (4.585)=0.0570.$$

Inserting these into the above, we have

$$t=\frac{|4.585-\mu|}{0.0570}$$

and, by transposition,

$$|4.585-\mu|=0.0570(t)$$

Now the numerical equivalent to t depends only upon the degree of confidence we wish to express. Suppose, for example, we set the chance at 1 in 20 that

$$|4.585-\mu|>0.0570(t).$$

This[9] corresponds to a probability of 0.05, and from Table 4 the probability is 0.05 that $t > 1.9600$. Consequently, with probability of 0.05

$$|4.585-\mu|>0.0570(1.9600)$$

that is, that the real, but unknown, discrepancy exceeds 0.112. Another way of stating this result is that the probability is 0.95 that

$$\mu=4.585\pm0.112,$$

for this is the range which encloses μ with the given probability.

2.6 Small Samples and the Probability of Discrepancy. With small samples, on the other hand, the estimate s, of σ, defined by

[9] The symbol > is read "is greater than."

$$s^2 = \frac{1}{n-1} \overset{n}{S} \left[(y - \bar{y})^2 \right]$$

while entirely satisfactory, will, nevertheless, differ more or less from the true value, σ. Furthermore, the distribution of

$$t = \frac{(\bar{y} - \mu)}{s/\sqrt{n}}$$

for small samples does not follow the normal law of error, although it approaches it rapidly as the number of degrees of freedom upon which s is based exceeds 30 – 50. The exact distribution of t depends upon the

TABLE 7. Table of t. Values of t, Outside the Range of Which in Both Tails Lie Selected Proportions of the Total Area*

Degrees of freedom	Relative Area in Both Tails											
	.9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01
	Values of t											
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	.12566	.25335	.38532	.52440	.67449	.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582

*This table is taken by consent from Statistical Methods for Research Workers by Professor R. A. Fisher, published at 15/– by Oliver and Boyd, Edinburgh. Attention is drawn to the larger collection in Statistical Tables by Professor R. A. Fisher and F. Yates, published by Oliver and Boyd, Edinburgh.

number of degrees of freedom available. It was first investigated by Student (1908) and has been tabulated by R. A. Fisher. It is given in Table 7, and also in the Appendix.

The relative areas listed correspond to those of Table 4. The bottom line of Table 7, in fact, contains the same entries as Table 4 (except that the latter have been rounded off to four decimals) for the degrees of freedom upon which s is based are here taken as infinity, and, consequently, s, for these values of t, is σ exactly.

Figure 6 shows a graphic comparison of the distribution of t corresponding to four degrees of freedom with that of the normal distribution.

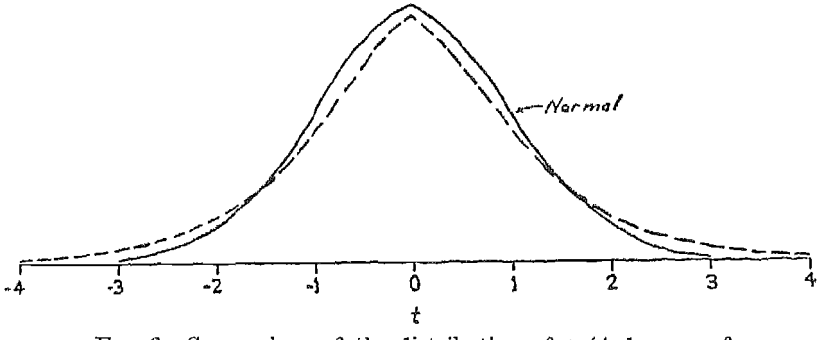


FIG. 6. Comparison of the distribution of t (4 degrees of freedom) with the normal curve of error.

Suppose, now, one is given the observations

$$64,\ 42,\ 49,\ 39,\ 49;$$

and the only other pertinent information is that they are a random sample from an infinite population, hypothetically of normal distribution. Let it be required to make an exact probability statement, consistent with the hypothesis, concerning the population mean.

Denoting the separate numbers by y, the sample mean becomes

$$\bar{y} = \frac{1}{5}\overset{5}{S}(y) = \frac{1}{5}(64+42+49+39+49) = \frac{1}{5}(243)$$
$$= 48.6$$

The sum of squares of the deviations from the sample mean,

$$\overset{5}{S}\left[(y-\bar{y})^2\right] = \overset{5}{S}(y^2) - \bar{y}\overset{5}{S}(y)$$

is numerically

$$(64)^2+(42)^2+(49)^2+(39)^2+(49)^2-(48.6)\ (243)$$

or

$$12{,}183 - 11{,}809.8 = 373.2$$

As there are four degrees of freedom among the five observations, the estimate of the population variance (Sec. 1.7) is

$$s^2 = \frac{1}{4}(373.2) = 93.3$$

whence, the estimate of the variance of the sample mean

$$V(48.6) = \frac{s^2}{5} = 18.66,$$

where V denotes the sampling variance of the enclosed term following it. The standard error of the sample mean is the square root of this, or 4.32. Hence the mean with its standard error may be expressed

$$48.6 \pm 4.32$$

and since the number of degrees of freedom upon which s is based is known, one needs only to choose the numerical equivalent of the confidence to be placed in the statement in order to complete it. Should the chance of error be fixed at 5 out of 100, then the probability with which Table 7 is entered is 0.05, and the value of t corresponding to this probability and four degrees of freedom is 2.776. Hence with probability of 0.95

$$\mu = 48.6 \pm 4.32(2.776)$$
$$= 48.6 \pm 12.0$$

Suppose, as an alternative, one is willing to take a 50-50 chance of error. This corresponds to a probability of 0.5, and the value of t for this probability and four degrees of freedom is, from Table 7, 0.741. Hence the chance is even that the true mean of the population

$$\mu = 48.6 \pm 4.32\ (0.741)$$
$$= 48.6 \pm 3.2$$

Each observation of the sample is, in fact, the sum of 10 digits from Tippett's Random Sampling Numbers whose population mean is 45.0. The estimates thereof are reasonable.

PART 2

DIRECT ESTIMATES BY SAMPLING

CHAPTER III

SIMPLER CASES OF SAMPLING FINITE POPULATIONS

3.1 Infinite and Finite Populations. In the previous chapters it was supposed that the populations sampled are made up of an infinite number of variates. This is the usual conception and a very common-sense one. Hypothetically there is an infinite number of measurements an observer may make on the same physical magnitude under a given set of conditions; his sample, however large, representing only an infinitesimal part of the whole. In like manner there is, hypothetically, an infinite number of digits, ranging from 0 to 9, represented by a random sample of them.

In games of chance the distinction in conception between an infinite and a finite population is easily made. Imagine an urn containing 100 marbles of the same size and consistency, indistinguishable to touch. If a number is painted upon each, the urn may be said to contain a population of such numbers. Now suppose a random sample of size n be drawn from this population. Each draw must, of course, be made such that each marble in the urn has exactly the same chance of being drawn as any other. If after drawing a marble and recording its number that marble is replaced before the next draw, the n draws supply a random sample from the hypothetical *infinite* population. But if the sample marbles are not replaced during the course of the n drawings, the random sample is from the *finite* population of 100 numbers. In the latter case n is, of course, less than 100.

Finite populations are the rule in most of the sampling problems with which forestry is concerned. Timber cruising is a sampling job on a finite area of timber stand or forest. The estimate of natural reproduction on logged-over areas, and the evaluation of forage in the meadow, are everyday problems of the forester in sampling finite populations.

It often happens that the distinction between a population known to be finite and the hypothetical infinite population is of no practical consequence. If N, the population size, is considered a sufficient approach to infinity, and if $\frac{n}{N}$ is sufficiently close to zero, where n is sample size, the distinction, as will be seen in later sections, is inconsequential. Fortunately, the practical consequences of neglecting the distinction where it should be recognized, become readily apparent in every case.

3.2 Sampling Units. Suppose one needs to know with fair precision the number of pine seedlings on a sample plot, one square chain in area. If the plot is covered with herbaceous vegetation so that it is difficult to distinguish the pine therefrom without diligent attention to detail of observation, a sampling job is indicated; unless, perforce, the time and expense involved in obtaining a complete tally of the entire population of pine seedlings is not a consideration.

In sampling an area, the constituent parts of the sample are to be located independently and at random. But as these constituent parts may be visualized in a variety of ways, the sampling units to be used hereafter are defined and illustrated as follows:

Ultimate Unit. The small plot or area that is not subdivided. For the square-chain population of pine reproduction it is the smallest practicable unit of area upon which counts are made, as for instance, the quarter milacre square.

Random Sampling Unit. A constituent part of the sample, which is drawn independently and at random. It consists of one or more ultimate units, as for example a strip, $\frac{1}{20}$-chain wide and one chain long, across the square chain of pine reproduction referred to above; this strip containing 20 ultimate units of a quarter milacre each.

Sample. The set of random sampling units. The sample must contain a minimum of two random sampling units, for a single random sampling unit does not contain the information on sampling error.

3.3 Sampling a Small Rectangular Area. The use of the above terms is easily demonstrated by means of a simple experiment. Let it be required to estimate the sum of the 100 numbers in Figure 7 from just 20 of them.

The ultimate unit in this case is the cell, a particular one of which will be referred to by its column and row number. Thus cell 50 is that of column 5 and row 0, and its observed value is 47; whereas the observed value of cell 05 is 61.

Even a hasty perusal of the figure discloses that the numbers which make up this population are not scattered wholly at random among the cells, for there is greater variation among cells of the same column than among cells of the same row. The middle rows run to higher numbers than the top and bottom rows. This effect of layers is called *stratification.* Stratification is a common characteristic of populations in nature. In the forest, for example, the better sites are usually found along the lower slopes, while the poorer sites are commonly along the

ROW NUMBERS	0	1	2	3	4	5	6	7	8	9
9	32	40	40	37	44	45	40	32	39	39
8	42	38	40	46	47	50	45	37	48	44
7	51	55	51	59	55	55	53	53	52	57
6	51	55	60	58	64	56	58	56	57	62
5	61	62	69	65	65	61	66	66	67	66
4	65	62	72	75	73	73	64	66	73	68
3	62	62	67	64	71	66	59	61	63	61
2	59	52	57	64	62	59	55	58	57	58
1	51	55	51	57	56	58	54	49	59	54
0	45	49	47	50	52	47	49	43	48	50

COLUMN NUMBERS

FIG. 7. The population of 100 numbers.

ridges. When observed, even in its broader effects, stratification may be made to enhance the efficiency of sampling. In practice the timber cruiser acts upon his recognition of stratification by conducting strips, or lines of plots, at right angles to the direction of general drainage, that is, across the strata.

Of the great many methods which may be devised for sampling the numbers of Figure 7, two are chosen for certain distinctive features, although both are forms of unrestricted random sampling.

Strip Method. If the column is taken as the random sampling unit it follows that the population of 100 numbers is to be visualized as 10 numbers, each a column sum. The population of columns is presented diagrammatically in Figure 8. As a sample of 20 percent is required, the process of sampling these 10 column sums is analogous to drawing two marbles—without replacement—from an urn containing 10 of them, each representing a particular column. In practice two numbers from 0

to 9 are drawn from Tippett's Random Sampling Numbers in some preassigned order, such as the first two digits of column 1 of a given page. Should the second number be a duplicate of the first, it is, of course, rejected and the next one taken, as a *finite* population of 10 separate strips is to be sampled.

The numbers 4 and 8 happen to be drawn. These indicate the columns whose sums (of 10 ultimate units each) are taken as random sampling observations. They are 589 and 563, which total to 1,152, or one-fifth the estimate of the population sum. The latter, then, is 5,760.

Single Plot Method. The ultimate unit, the cell, is also taken as the random sampling unit. The population is visualized as made up of 100 of these. It is presented diagrammatically in Figure 9. Twenty of the numbers from 0 to 99 are drawn by the aid of Tippett's Random Sampling Numbers, the first digit indicating the column and the second the

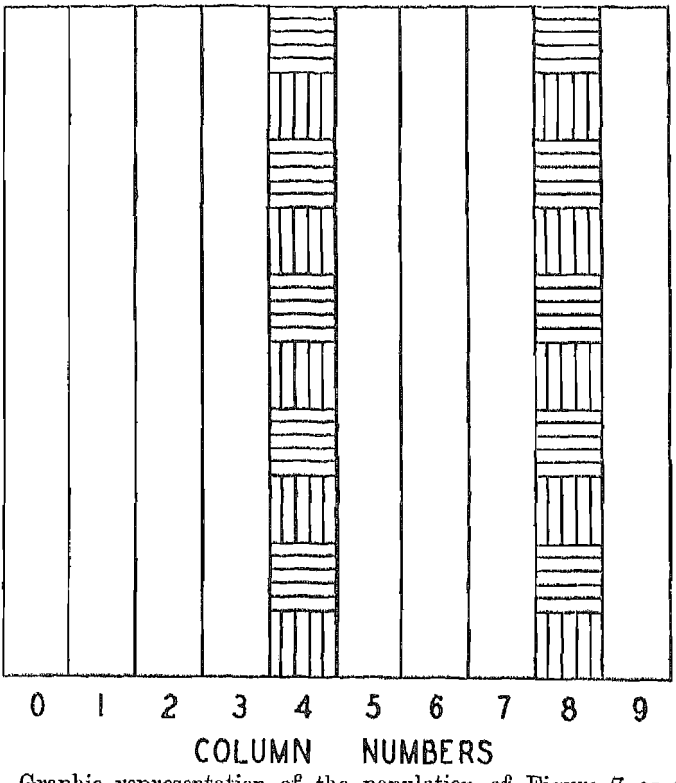


FIG. 8. Graphic representation of the population of Figure 7 as that of 10 strips. The shaded strips comprise two random sampling units of a sample.

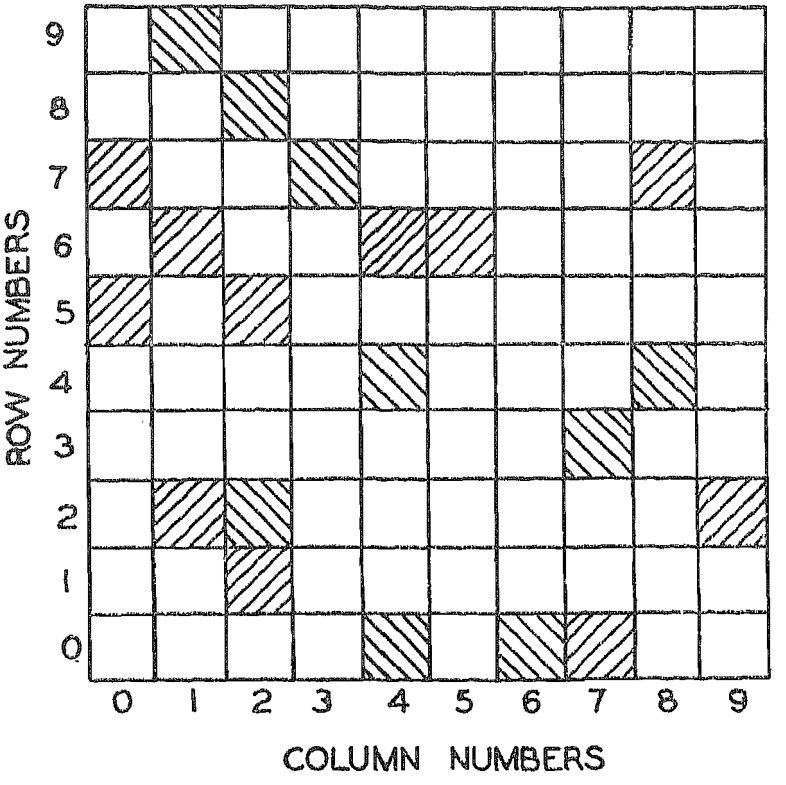


FIG. 9. Graphic representation of the population of Figure 7 as that of 100 cells. The shaded cells comprise 20 random sampling units of a sample.

row. A possible second draw of any cell is rejected, and another number taken in its place.

Following are the cell numbers and observations in order of draw:

Cell number	Observation	Cell number	Observation
12	52	56	56
05	61	87	52
92	58	73	61
07	51	40	52
28	40	70	43
21	51	25	69
16	55	84	73
44	73	19	40
60	49	22	57
37	59	46	64
		Total.....	1,116

From the total of 1,116 the estimate of the population sum is 5,580.

Other sampling designs, of only slightly greater complexity, might be applied to the sampling of this population. Those just illustrated are among the simpler ones.

3.4 The Variance of the Mean of a Random Sample from a Finite Population. When a random sample of n observations of y is drawn from an infinite population, the estimate of the variance of y was given in Sec. 1.7 as

$$s^2 = \frac{1}{n-1} \overset{n}{S} \left[(y-\bar{y})^2 \right],$$

where $\bar{y}$ is the mean of the sample. The sampling variance of mean y as given in Sec. 1.6, is

$$V(\bar{y}) = \frac{s^2}{n}.$$

But when the sample consists of n random sampling units of y, from a finite population of just N such values of y, we require to adjust the sampling variance of mean y as follows:[10]

$$V(\bar{y}) = \frac{s^2}{n} \left(\frac{N-n}{N} \right)$$

It is at once evident that as the population size, N, approaches infinity the quantity $\left(\frac{N-n}{N} \right)$ approaches unity and the variance of the mean is identical to that from an infinite population. On the other extreme, as sample size, n, approaches and becomes population size, N, the variance of the mean approaches and becomes zero, since there can be no sampling error if the entire population is enumerated.

The tools are now available for estimating the limits of the probable discrepancy between the estimate of the population aggregate, according to each of the sampling designs of the previous section, and the corresponding true value.

According to the strip method, the mean of the two random sampling units, 589 and 563, is 576. Hence the estimate of the population variance of random sampling units is

$$(589-576)^2 + (563-576)^2 = 338$$

on one degree of freedom. Were the sample from an infinite population, the estimate of the variance of the mean of the two random sampling

[10] The derivation of the sampling variance of the mean of a random sample from a finite population is somewhat cumbersome to be included here. It is given in detail in the Appendix, Sec. 3.4.

units would be half of this, or 169. But since in this case $n=2$, and $N=10$, we have

$$V(576)=169\left(\frac{10-2}{10}\right)$$
$$=135.2$$

on one degree of freedom. The standard error of the mean is the square root of this or 11.6. Hence, on the random sampling unit basis

$$\bar{y}=576\pm 11.6,$$

and the estimate of the population aggregate is ten times this quantity or

$$5{,}760\pm 116$$

on one degree of freedom. Hence, as discussed in Sec. 2.6, the probability equation appropriate to the estimate of the population aggregate is

$$5{,}760\pm 116(t)$$

where the numerical value of t, to be taken from Table 7, corresponds to the probability chosen. If the probability of error is set at 0.05, t is 12.706 on one degree of freedom. Hence with probability of 0.95 the population aggregate is

$$5{,}760\pm 116(12.706)$$

or

$$5{,}760\pm 1{,}474$$

In other words, with probability of 0.95 this estimate is subject to an error not to exceed $\left(\frac{1474}{5760}\right)$ or 26 percent of the estimated value.

This would not seem to be very satisfactory estimating, and we shall compare it with the single plot design used in sampling the same population. The estimate of the mean and variance of random sampling units according to this design is given in Table 8. The mean of the 20 random sampling units is 55.8, while the estimate of the variance of random sampling units is 89.642 on 19 degrees of freedom. Since $N=100$, and $n=20$, the estimate of the variance of the sample mean is

$$\frac{89.642}{20}\left(\frac{100-20}{100}\right)=3.586$$

and its standard error is the square root of this, or 1.89. Hence on the random sampling unit basis

$$\bar{y}=55.8\pm 1.89$$

and the estimate of the population aggregate is 100 times this number or

$$5{,}580\pm 189$$

TABLE 8. Calculation of the Mean and Variance of 20 Random Sampling Units (Cells). Data from Figure 7

Observation	
y	y^2
52	2,704
61	3,721
58	3,364
51	2,601
40	1,600
51	2,601
55	3,025
73	5,329
49	2,401
59	3,481
56	3,136
52	2,704
61	3,721
52	2,704
43	1,849
69	4,761
73	5,329
40	1,600
57	3,249
64	4,096
1,116	63,976

Calculation of mean:

$\bar{y} = \frac{1}{20}(1{,}116) = 55.8$

Calculation of variance:

$\overset{20}{S}(y^2) = 63{,}976.$

$\bar{y} \cdot \overset{20}{S}(y) = 62{,}272.8$

$\overset{20}{S}\left[(y-\bar{y})^2\right] = 1{,}703.2$

$s^2 = \frac{1}{19}\overset{20}{S}\left[(y-\bar{y})^2\right] = 89.642$

on 19 degrees of freedom, and the probability equation appropriate to this estimate is

$$5{,}580 \pm 189(t).$$

If the probability of error is set at 0.05, the corresponding value of t based upon 19 degrees of freedom is 2.093. Therefore with probability of 0.95, the population aggregate is within the limits

$$5{,}580 \pm 189(2.093)$$

or

$$5{,}580 \pm 396.$$

This is an estimate of considerably more precision than that of the strip method, for the probability is 0.95 that the error does not exceed $\left(\frac{396}{5{,}580}\right)$ or 7.1 percent of the estimated aggregate.

The estimate of the limits of discrepancy between sample and population is equally valid in both of these sampling methods. Their magnitudes, however—1474 and 396 on a probability of 0.95 that the real discrepancy is not exceeded by these estimates—are very discordant, reflecting as they do the efficiency of the sampling designs used. The seemingly

abnormal 1474 of the strip method is due to the paucity of degrees of freedom available for its estimate. This becomes evident upon a glance at the table of t (Table 7). For a single degree of freedom the value of t (the ratio of discrepancy to its standard error) is 12.706 at the 5 percent level. It drops abruptly to 4.303 for two degrees of freedom, and for 19 it is 2.093. For an indefinitely great number of degrees of freedom, t approaches its limiting value of 1.960. Consequently, if only very few degrees of freedom are available for the estimate of sampling error, the discrepancy between the sample mean and corresponding population parameter is likely to be too large for practical sampling work.

3.5 Sampling a Small Area of Irregular Boundaries. The two sampling designs used above may be applied very widely. They are not at all confined to populations which are distributed over a rectangular area, although their application is most advantageous in populations of simple geometrical outline.

A small population of irregular outline is presented in Figure 10. In this case, the strip (for example, column) method would be somewhat less simple, for the different strips are of variable lengths and, consequently, the strip means are of variable precisions. Special cases having to do with variable precision involving weighted observations will be treated later (Chapter VII).

The single plot sampling design, as illustrated in Sec. 3.3, might also be somewhat troublesome to apply to this population, for it would require advance information—involving, perhaps, a map—concerning the exact number and location of the plots in the population *before the draw*, in order to assure equal chance for every plot to make the sample.

But a class of sampling design, analogous to the strip method used previously, is easily applied to populations of irregular outline.

Returning for the moment to the strip method of Sec. 3.3, the reader will remember that the random sampling unit consisted of the 10 ultimate units of the same designation (for example, column 4) one taken *from each row*.

By analogy, a population of irregular outline, as in Figure 10, may be considered as containing N—not very large—random sampling units, where each random sampling unit consists of an (as yet) unknown number of ultimate units of cells. For instance, if the population of Figure 10 is traversed by moving up the first column on the left, down the second, up the third, and so on, until N ultimate units have been encountered, numbered in order,

$$1, 2, \ldots, N;$$

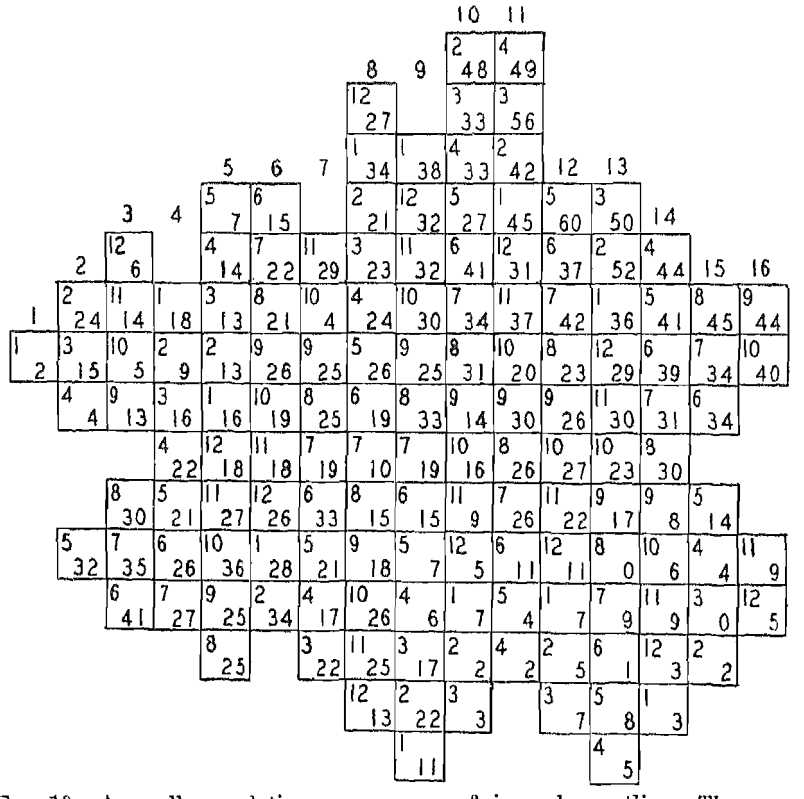


FIG. 10. A small population on an area of irregular outline. The upper figure of each cell represents the ultimate unit number in each set of 12 units of strip.

then upon continuing in this fashion, the population is found to be made up of a number of *sets* of N ultimate units. Suppose, now, that all ultimate units numbered 1 be taken as the first random sampling unit, all ultimate units numbered 2 as the second, and so on to the Nth random sampling unit. Thus the population is conceived as made up of N random sampling units, of an unknown number—until the observational work is complete—of ultimate units each.

As a concrete illustration, let the random sample to be drawn from Figure 10 contain three random sampling units and 25 percent of the population. Thus N is fixed at 12. Accordingly, three random numbers, from 1 to 12, are drawn, to represent the ordinal number of ultimate unit *in each and every set of 12 units of strip*. The entire population is then traversed by moving up the first column on the left, down the sec-

ond, up the third, and so on; and record is made of the indicated ultimate unit observations as encountered.

The ultimate unit numbers within each set have been selected by running down two adjacent columns of Tippett's Random Sampling Numbers (columns 7 and 8 of page 1 in this case) and listing the first three different numbers of value 12 or less. They happen to be 11, 7, and 8, as may be checked by referring to Figure 1. Accordingly, the sum of every 7th, every 8th, and every 11th cell value in the sets of 12 units of strip make up the three random sampling units of the sample. These are recorded in Table 9, while Figure 11 gives a good idea of how well the sample represents the population.

From Table 9, then, the three random sampling observations are

$$308,\ 304,\ \text{and}\ 261$$

out of a possible 12 such numbers in the population. The mean of the sample, $\bar{y}$, is 291. As deviations therefrom, the observations $(y-\bar{y})$ are

$$17,\ 13,\ \text{and}\ -30,$$

whence the estimate of the population variance among random sampling units is

$$s^2=\frac{1}{2}(289+169+900)=679$$

on two degrees of freedom. In accordance with the sampling design used, the population size, N, is 12, and the sample size, n, is 3. The estimate of the variance of the mean from this limited population

$$V(\bar{y})=\frac{s^2}{n}\left(\frac{N-n}{N}\right)=\frac{679}{3}\left(\frac{12-3}{12}\right)=169.75$$

The mean of the the three sampling units with its standard error is, accordingly,

$$291\pm 13.0$$

and as this is one twelth of the estimate of the population aggregate, the latter is

$$3{,}492\pm 156(t)$$

where t, on two degrees of freedom, corresponds to the probability selected. With the probability of greater error fixed at 0.05, t, is 4.303 from Table 7. With probability of 0.95, then, the estimate of the population aggregate is

$$3{,}492\pm 156(4.303)$$

or

$$3{,}492\pm 671$$

TABLE 9. Ultimate Unit Observations whose Sums Supply Three Random Sampling Units from the Population of Figure 10

Ordinal number of ultimate unit in each 12 units of strip		
7	8	11
Observations on ultimate units		
35	30	14
27	25	27
22	21	18
19	25	29
10	15	25
19	33	32
34	31	9
26	26	37
42	23	22
9	0	30
31	30	9
34	45	9
Sum 308	304	261

As indicated above, the sampling scheme used here is distinct from that of the strip method as applied to the rectangular area in Sec. 3.3, only because the present population is distributed over an area of irregular boundaries.

3.6 Systematic Versus Random Sampling. The question may be raised as to why a systematically chosen sample does not have all the virtues of a random sample as a method of estimating population means or aggregates.

In the early application of the theory of errors to problems of sampling populations in confined areas—such as the timber volume of a forest—major emphasis was placed upon the necessity of selecting a sample free from personal bias. So simple an expedient as the mere mechanical selection of plots or strips at equidistant intervals solved this difficulty entirely. Unfortunately, however, it was not at once discovered that the removal of the personal equation does not entirely fulfill the condition of sufficiency.

The mathematical requirements for the solution of sampling problems imply that the constituent parts upon which sampling error is based —the random sampling units as defined above—be located independently and at random.

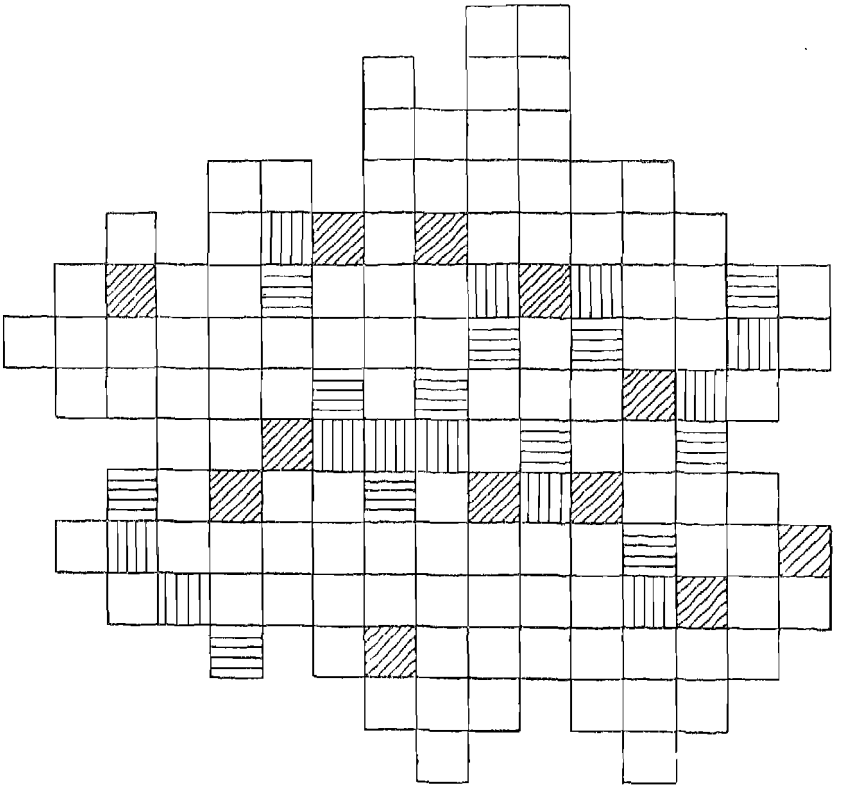


FIG. 11. Each aggregate of the cells designated alike, represents a random sampling unit of a sample of three observations of the population of Figure 10.

The failure of the systematic pattern of plots or strips to provide information concerning the probable discrepancy between the estimated and the true characteristic may be clarified in the light of the simple experiments used in this chapter. Should the numbers which make up the populations illustrated, have been assigned to the cells strictly at random, any systematic sample would clearly have contained all the information of a random sample of equal weight. The systematic sample would, indeed, have been a random sample as well, just as a systematically chosen set of digits from a page of random numbers is a random sample of digits.

It follows, therefore, that when the calculator derives what he considers the standard error from a body of data taken systematically in

natural populations, such as the forest, he assumes that Nature has been obliging enough to have randomized for him. But foresters, like naturalists, know that the components of any characteristic of forest populations they may wish to examine, are perhaps never *arranged* independently and at random within a forest. Their only alternative, if they are at all concerned about the probable discrepancy between the true values sought and their estimates thereof, is to base the estimates upon samples *drawn* independently and at random from the population.

CHAPTER IV

REPRESENTATIVE OR STRATIFIED RANDOM SAMPLING

4.1 The Principle of Representative Sampling. The sampling designs illustrated in the preceding chapter are forms of unrestricted random sampling of populations distributed over confined areas. A population was conceived as made up of numbers, or magnitudes of a certain characteristic, which occurred on the N subdivisions into which the whole area was partitioned. Each of these subdivisions was then a possible random sampling unit. The sample consisted of n of them, drawn independently and at random, from the entire number, N.

Now stratification is a well-known property of practically all forest and field populations. Yields of different parts of the same subdivision of land tend to be more uniform than yields of different subdivisions. Under these conditions, the precision of an estimate of a population mean may be appreciably enhanced by recognizing stratification and modifying the sampling design accordingly.

The area to be sampled is subdivided into strata, or *blocks,* and a random sample of the characteristic to be estimated is drawn from each block. No great care need be taken to have block boundaries coincide with visual limits of soil fertility gradients, or density changes in vegetation. The exigencies of the problem usually demand that a balance be struck between the theoretically desirable and the practically feasible. Precision is gained by dividing the population into as many blocks as expedient, even though the number of random sampling units taken from each block is the minimum of two.

Representative sampling, then, is the process of drawing a *representative set of samples,* consisting of a random sample from each block, or stratum, of the population sampled. It may be illustrated by means of the population of Figure 12, which contains 200 numbers, divided into 10 blocks of 20 numbers each. There is obvious stratification here as the top tier of blocks runs to lower numbers than the bottom, and the left-hand block of each tier runs to lower numbers than the right-hand one.

By way of illustration, let a representative set of samples be drawn from the population of Figure 12, by direct observation of just 10 percent of the entire population. Taking the cell to be the random sampling unit as well as the ultimate unit, two numbers are drawn between 1 and 20, *independently and at random for each block,* by means of Tippett's

Random Sampling Numbers. These and the observations they supply are listed in Table 10. The general mean of the 10 samples is 95.2. Complete analyses of these will be treated shortly.

In the meantime we shall show the comparison of results between representative and unrestricted random sampling of the same population.

BLOCK 1	48	23	64	59	38	87	54	65	58	45
	34	48	81	67	60	60	71	91	108	86
	78	74	46	62	52	52	94	82	94	76
	36	40	35	61	82	48	83	87	92	82
3	44	51	56	65	72	73	84	77	90	95
	60	60	107	87	84	80	111	91	82	104
	46	72	76	52	78	78	90	74	124	114
	62	78	43	77	78	108	115	99	122	116
5	44	97	62	97	90	103	94	123	108	117
	44	66	65	75	98	86	103	99	136	144
	66	78	80	80	90	116	112	130	108	146
	52	64	119	109	96	118	125	115	130	120
7	46	85	86	85	98	117	140	137	124	151
	84	92	91	89	84	98	123	147	140	150
	74	72	72	96	128	114	130	122	148	184
	94	76	101	89	126	94	125	153	154	178
9	68	67	94	103	146	135	148	157	170	179
	60	100	113	101	136	136	145	147	170	186
	76	76	122	120	122	142	160	174	182	184
	104	106	129	133	146	130	149	165	192	150

BLOCK 2

BLOCK 4

BLOCK 6

BLOCK 8

BLOCK 10

LEGEND OF SAMPLING NUMBERS

1	5	9	13	17
2	6	10	14	18
3	7	11	15	19
4	8	12	16	20

FIG. 12. A population of 200 cells, divided into 10 blocks of 20 cells each.

4.2 Comparison of Representative with Unrestricted Random Sampling. The efficiency of representative sampling as compared to unrestricted random sampling of stratified populations may, perhaps, be most convincingly demonstrated by graphic comparison of estimates of the population mean as based upon repeated sampling trials.

In Figure 13A are plotted 11 estimates of the mean of the population of the 200 numbers of Figure 12. Each estimate is the mean of 20 cell

TABLE 10. A Representative Set of Samples from the Population of Figure 12

Block	Random sampling unit members		Observations	
1	4	18	36	60
2	10	11	91	82
3	2	14	60	87
4	7	11	90	74
5	20	10	96	65
6	20	5	120	94
7	8	17	76	98
8	8	9	125	137
9	10	14	113	101
10	8	20	149	150
Sum			1,904	
Mean			95.2	

values, drawn at random from the population as a whole according to the single plot method used in Sec. 3.3.

Figure 13B shows 11 estimates of the mean of the same population according to the representative sampling design of the preceding section. Each mean is again based upon 20 cell values, but the drawing was made with the restriction that two cells be taken independently and at random from each of the 10 blocks.

Clearly the means of samples drawn by unrestricted random sampling as used here, are dispersed more widely around the true population mean of 98.015 (also shown in Fig. 13) than are the means of the representative sampling trials.

The variance of the 11 means of Figure 13A is 129.65 by direct calculation, while that of the 11 means of Figure 13B is 15.44. Since

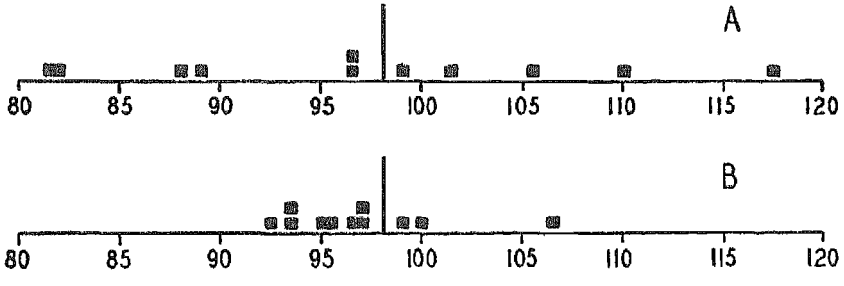


FIG. 13. Distribution of eleven estimates of a population mean based upon 20 random sampling units. In A, by unrestricted random sampling; in B, by representative random sampling of two units in each block.

$$\frac{129.65}{15.44} = 8 \text{ approximately,}$$

one estimate according to representative sampling is worth about eight estimates of unrestricted sampling of this population.

4.3 The Variance of the Mean of a Representative Set of Samples. Detailed observations of representative sampling of the population of Figure 12 are given in Table 10. The general mean is 95.2, and we require the variance of this estimate of the population mean.

Consider, first, block 1 alone. The observations, 36 and 60, are a random sample of the population of this block, and the estimate of the block mean is, therefore, 48.0. The sum of squares of deviations is

$$(36-48.0)^2+(60-48.0)^2=288$$

and since it is based upon a single degree of freedom, this is also the estimate of variance among the random sampling units of block 1, though not yet adjusted for the finite population sampled.

When an estimate of variance is based upon just two random sampling unit observations, a short-cut method of calculation is to be preferred. Let x_1 and x_2 be two such observations from an infinite population of x. Their mean is

$$\frac{1}{2}(x_1+x_2),$$

and the sum of the squares of deviations from this mean is

$$\left[x_1-\frac{1}{2}(x_1+x_2)\right]^2+\left[x_2-\frac{1}{2}(x_1+x_2)\right]^2,$$

which may be written

$$\left[\frac{1}{2}(x_1-x_2)\right]^2+\left[\frac{1}{2}(x_2-x_1)\right]^2$$

or, since these two terms are identical, as

$$\frac{1}{2}(x_1-x_2)^2$$

on one degree of freedom. The estimate of the variance of the mean of the two observations is (Sec. 1.6) ½ of this, that is,

$$V\left[\frac{1}{2}(x_1+x_2)\right]=\frac{1}{4}(x_1-x_2)^2,$$

and the variance of the sum of the two is twice the variance of single observations, or

$$V(x_1+x_2)=(x_1-x_2)^2.$$

For block 1, then, the variance of the mean, 48.0, is

$$\frac{1}{4}(36-60)^2=144$$

whereas the variance of the sum of the two observations, that is,

$$V(36+60)=(36-60)^2=576$$

not yet corrected for the finite population of the block.

It is somewhat simpler to calculate the variance of the individual block sums, as shown here for the first block, rather than of the block means. Having a separate and independent sample from each block in the population, the variance of the sum over all blocks is the sum of the separate variances.

The calculations are given in Table 11. The sum over all 10 samples is 1,904, and the variance of this sum is 4,052, not yet corrected. Since in each block $N=20$ and $n=2$, we have

$$V(1,904)=4,052\left(\frac{20-2}{20}\right)$$
$$=3,646.8$$

on 10 degrees of freedom, as each one of the blocks supplies a single degree of freedom. The standard error of the observed sum is the square root of this variance, or 60.4; whence the observed sum is

$$1,904\pm 60.4$$

on 10 degrees of freedom. As the samples compose 10 percent of the population, the estimate of the population aggregate is

$$19,040\pm 604(t)$$

TABLE 11. Analysis of a Representative Set of Samples

Block	Random Sampling observations		Block sum	Variance of sum
	x_1	x_2	(x_1+x_2)	$(x_1-x_2)^2$
1	36	60	96	576
2	91	82	173	81
3	60	87	147	729
4	90	74	164	256
5	96	65	161	961
6	120	94	214	676
7	76	98	174	484
8	125	137	262	144
9	113	101	214	144
10	149	150	299	1
Sum			1,904	4,052

where t, taken from Table 7, is for 10 degrees of freedom on the probability chosen.

If one prefers to work through the mean of the random sampling units rather than their sum, one should note that 1,904 is the sum of the 20 observations whose mean is 95.2. Furthermore, if σ^2 is the variance of single observations, the variance of the sum of n observations (Sec. 1.6) is

$$n\sigma^2$$

whereas the variance of the mean of n observations is

$$\frac{1}{n}\sigma^2.$$

Consequently, the variance of the mean of n may be derived from the variance of the sum of n by dividing the latter variance by n^2.

In block 1, by way of illustration, the block mean is $\left(\frac{96}{2}\right)$; and as the estimate of the variance of the block sum is

$$V(96) = 576$$

the variance of the mean may be written

$$V\left(\frac{96}{2}\right) = \frac{576}{(2)^2}.$$

Similar calculations performed on the other blocks and added together provide the sum of the 10 block means and the variance of this sum; that is,

$$V\left[\frac{96}{2} + \frac{173}{2} + \ldots + \frac{299}{2}\right] = \left[\frac{576}{(2)^2} + \frac{81}{(2)^2} + \ldots + \frac{1}{(2)^2}\right]$$

whence the mean of these, that is, the general mean, and its variance may be written

$$V\left[\frac{1,904}{(2)(10)}\right] = \frac{4,052}{(2)^2(10)^2}$$

not yet corrected to the finite population sampled. Upon applying the correction, we have

$$V(95.2) = \frac{4,052}{400}\left(\frac{20-2}{20}\right)$$
$$= 9.117$$

on 10 degrees of freedom; and its square root is 3.02. The estimate of the mean of the 200 numbers is then

$$95.2 \pm 3.02$$

Upon multiplying by 200, the same estimate of the population aggregate as given above is obtained.

4.4 Disproportional Sampling by the Representative Method. In the representative sampling treated above, the blocks were of exactly the same size, and the same proportion (10 percent) of the block populations was sampled in each. It was then necessary merely to sum the observations over all blocks and to multiply the grand sum by 10, the product being the estimate of the population aggregate.

The necessary variances were calculated almost as easily.

It may happen, however, that the practical requirements of sampling a given population preclude the direct observation of the same proportion of all block areas. There may be greater interest in certain blocks than in others, or, perhaps, irregular boundaries of the population may not readily permit its division into equal parts. As an outcome, sampling may be more intensive in some blocks, and the several samples may have different precisions.

The populations of irregular outline used previously (Fig. 10) will serve to illustrate the case. It is reproduced in Figure 14, and is subdivided into five blocks of different numbers of ultimate units or cells. Necessary information concerning the population, as well as the observations following from disproportional random sampling of the several blocks, are given in Table 12. The variation in the number, n, of observations to the block is introduced only for illustrative purposes.

TABLE 12. Disproportional but Representative Sampling of the Population of Figure 14

Block number	N	n	Random sampling observations	Sum	Factor $\frac{N}{n}$	Estimate of total
1.....	26	2	38, 50	88	13	1,144
2.....	28	4	26, 31, 8, 26	91	7	637
3.....	30	3	4, 5, 5	14	10	140
4.....	28	2	21, 21	42	14	588
5.....	32	4	9, 14, 25, 23	71	8	568
Estimate of population aggregate.....						3,077

The second column is the listing of the area of each block in number of cells, or ultimate units, which are again taken as random sampling units; while the third column gives the number of these which make up the samples. The random sampling observations follow, and then their sums. In the next to last column is the factor $\left(\frac{N}{n}\right)$ by which the sam-

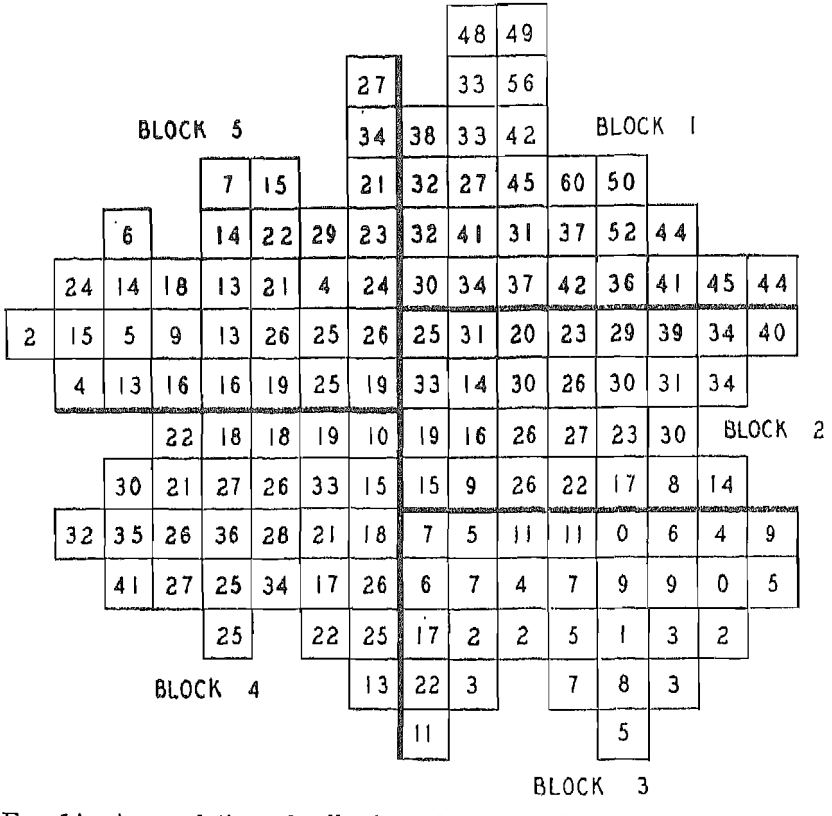


FIG. 14. A population of cells, irregular in outline, and subdivided into blocks of different areas.

ple sums are multiplied, the products affording the estimates of the block totals, in the last column. The grand sum of these, 3,077, is, of course, the estimate of the population aggregate. And it is the precision of this estimate that is now required.

Since 3,077 is the sum of independent estimates of each of the five block sums, its variance is the sum of the five separate variances of the same block sums.

For block 1, the variance of the sample sum, 88, is

$$(38-50)^2=144$$

its standard error being the square root of this, or 12. Since

$$13(88)=1{,}144$$

is the estimate of the block population, the standard error of this estimate is

$$(13)(12)=156,$$

and its variance, being the square of the standard error, is

$$(13)^2(144)=24{,}336$$

based upon one degree of freedom, though not yet corrected for the limited population of the block. The correction factor, based upon the sample number, $n=2$, out of the population number, $N=26$, is

$$\frac{N-n}{N}=\frac{24}{26}$$

for this particular block. Applied to the uncorrected estimate, 24,336, the estimate of the variance of the limited population of block 1 is 22,464.

These values are listed in the first line of Table 13. The results of corresponding operations upon the observations of the other blocks are given in the succeeding lines of the table. For block 2, as another example, the estimate of the variance of 91, is, from Table 13, four times the variance of the individual observations, or, since this represents three degrees of freedom, it is 4/3 times the sum of squares of deviations from the block mean. Numerically, this is 4/3 of

$$(26)^2+(31)^2+(8)^2+(26)^2-\frac{1}{4}(91)^2$$

or 409, based upon three degrees of freedom. As the sample consists of one seventh of the block population, the estimate of the latter is

$$7(91)=637$$

and its variance is

$$(7)^2(409)=20{,}041$$

TABLE 13. Estimate of the Variance of a Finite Population Aggregate, from Disproportional, but Representative Sampling

Block number	Sample sum	Estimate of variance of sample sum	$\frac{N}{n}$	Estimate of block populations	Estimate of variance of block populations (uncorrected)	Correction factor $\frac{N-n}{N}$	Estimate of variance of finite block populations
1....	88	144	13	1,144	24,336	24/26	22,464
2....	91	409	7	637	20,041	24/28	17,178
3....	14	1	10	140	100	27/30	90
4....	42	0	14	588	0	26/28	0
5....	71	228	8	568	14,592	28/32	12,768
Estimate of population sum and its variance.....				3,077	.....	.....	52,500

on three degrees of freedom. The correction factor for this block being $\left(\frac{24}{28}\right)$, the corrected estimate of the variance of the block population is 17,178.

The sum of the variances of the block populations is the variance of the sum of the block populations; hence, from the bottom line of Table 13, the variance of our estimate of the population aggregate, that is, of 3,077, is 52,500, based upon 10 degrees of freedom; the degrees of freedom in the total over all blocks being the sum of those in the individual blocks. Out of the 15 random sampling units of the five blocks, one degree of freedom was used in the estimate of each of the five block means.

The estimate of the population sum with its standard error is

$$3{,}077 \pm 229$$

or, with probability of 0.95, the population aggregate is

$$\begin{aligned} &3{,}077 \pm 229(2.228) \\ =&3{,}077 \pm 510, \end{aligned}$$

for which 2.228 is the value of t on 10 degrees of freedom.

CHAPTER V

SIMULTANEOUS SAMPLING OF MORE THAN ONE POPULATION

5.1 The Problem and an Illustration. In the previous chapter it was shown that an efficient estimate of a population mean (or sum) is the outcome of suitable sampling design. As the populations of forest and field are characteristically—one is tempted to say universally—heterogeneous and stratified, the area to be sampled is divided into sub-areas, or blocks, such that the variation among random sampling units of the same block is less than variation among random sampling units of different blocks. Then each block is sampled so as to provide the necessary and sufficient conditions for exact evaluation of the limits of the probable discrepancy between the magnitude of a true, but unknown, mean (or sum) of the population sampled—such as the timber volume of a forest property—and the estimate of it as derived from the samples.

The problem may be extended to more than one population of the forest or range. If, for example, the timber volume of a forest property is distributed over a number of timber-species groups, it may be required to estimate the volume of each group separately, as well as the combined volumes of any two or more groups, together with exact evaluation of the limits of probable discrepancy between the true, but unknown, volume of each group, or any combination of two or more, and the corresponding estimate of their magnitudes as derived from the samples.

The problem may be illustrated for the area represented diagrammatically in Figure 15. The cells contain two populations. These may be construed as the volume, on the cell area, of each of two species groups, made up, respectively, of the upper and lower numbers. Let it be required to estimate the total volume of each group separately, and of both groups combined from direct observation of 20 percent of the populations.

Four blocks are delineated. The random sampling unit of each population will also be the ultimate unit, that is, the cell. Four of the 20 cells of each block are drawn, independently and at random, by means of a set of random sampling numbers, and the observations are listed, according to group x (upper cell number) and group y (lower cell num-

ber) in Table 14. The sample totals, which are 20 percent of the estimate of the population aggregates, are the following:

Group x: 474
Group y: 302
Both Groups: 776

BLOCK 1 BLOCK 2

23 30	20 24	25 20	20 27	30 13	27 19	35 2	47 1
29 26	38 25	32 18	45 18	40 7	27 17	34 13	43 --
29 16	26 20	35 16	39 8	31 13	30 8	36 6	31 --
21 25	47 13	42 13	24 28	27 12	32 12	48 --	42 6
13 33	38 22	37 11	37 6	39 13	44 1	52 --	54 --
6 54	15 41	19 28	35 21	23 11	26 19	39 --	47 --
16 35	13 28	4 37	31 12	33 14	37 19	33 11	38 --
8 49	21 29	35 14	42 24	28 10	23 16	30 --	45 --
20 39	25 38	32 21	35 18	35 13	28 12	41 --	25 --
9 32	23 29	31 16	37 10	35 16	24 3	42 --	30 --

BLOCK 3 BLOCK 4

FIG. 15. Two populations—upper and lower numbers within the cells—which are not known to be distributed independently of one another.

Only the variances of these estimates are now needed in order to complete the solution.

TABLE 14. Representative Sampling Observation of the Populations of Figure 15

Block number	Group		Total	Block number	Group		Total
	x	y	$x+y$		x	y	$x+y$
1	47	13	60	3	15	41	56
	29	26	55		9	32	41
	38	25	63		25	38	63
	24	28	52		35	14	49
2	35	2	37	4	23	11	34
	42	6	48		30	0	30
	27	17	44		33	14	47
	27	19	46		35	16	51
				Totals. . . .	474	302	776

5.2 Variances and Covariances Involved. The variance of each group is calculated in the usual way. Given n random sampling unit observations of x in one of the blocks, the estimate of the variance of

$$\overset{n}{S}(x),$$

uncorrected for the finite population of the block, is

$$\frac{n}{n-1}\overset{n}{S}\left[(x-\bar{x})^2\right]$$

where $\bar{x}$ is the mean of the observed random sampling units of x in the block. In like manner, the estimate of variance of

$$\overset{n}{S}(y)$$

of the same block is

$$\frac{n}{n-1}\overset{n}{S}\left[(y-\bar{y})^2\right].$$

The corresponding estimate of the variance of both groups combined, that is of $(x+y)$, where x and y are observed values of the two groups on the same random sampling unit, is based upon the sum of squares of deviations of $(x+y)$ around the block mean of the n random sampling observations of $(x+y)$. Since mean $(x+y)$ is mean x plus mean y, the variance needed is

$$\frac{n}{n-1}\overset{n}{S}\left[\left\{(x+y)-(\bar{x}+\bar{y})\right\}^2\right]$$

which, for convenience, may be written

$$\frac{n}{n-1}\overset{n}{S}\left[\left\{(x-\bar{x})+(y-\bar{y})\right\}^2\right].$$

Upon expanding, this becomes

$$\frac{n}{n-1}\overset{n}{S}\left[(x-\bar{x})^2\right]+\frac{n}{n-1}\overset{n}{S}\left[(y-\bar{y})^2\right]+\frac{2n}{n-1}\overset{n}{S}\left[(x-\bar{x})(y-\bar{y})\right]$$

The first term of this expanded form is immediately recognized as the variance of the sample sum of x; and the second term as the variance of the sample sum of y. The third term—disregarding the factor 2—is known as the *covariance* of the sample sums of x and y. Obviously it may be positive or negative. In general, then, the *variance of* $(x+y)$ *is the variance of* x, *plus the variance of* y, *plus twice the covariance of* x *and* y.

The term *covariance* designates a mean product in the same sense that variance designates a mean square.

If two variables, such as x and y, are distributed independently over the area sampled, their covariance in the population is zero. Now while the random sampling units in each block are drawn independently, there is no assurance whatever that the magnitude of x on any random sampling unit is independent of the magnitude of y *on the same random sampling unit.* In variables such as timber volume according to species groups, it is, indeed, to be expected that the greater the volume of one of the groups, on plots or strips of given area, the less will be the volume of some, or all, of the remaining groups. Covariances among groups, consequently, ought usually to be negative.

In the calculation of sum of products upon which covariances are based, short-cut methods are available, quite analogous to computation schemes already used in arriving at the sum of squares of deviations about the mean of a sample. In the latter case it was noted that

$$\begin{aligned}\overset{n}{S}\left[(y-\bar{y})^2\right] &= \overset{n}{S}(y^2)-\bar{y}\overset{n}{S}(y)\\ &= \overset{n}{S}(y^2)-\frac{1}{n}\left[\overset{n}{S}(y)\right]^2\end{aligned}$$

where either of the two forms

$$\bar{y}\overset{n}{S}(y) \text{ or } \frac{1}{n}\left[\overset{n}{S}(y)\right]^2$$

which are known as correction terms, are deducted from the sum of

squares of the original values of y, the residue being the sum of squares of deviations about the mean.

The appropriate correction term to be applied to the sum of products becomes apparent upon expanding the expression which represents the sum of products; that is,

$$\overset{n}{S}\left[(x-\bar{x})(y-\bar{y})\right] = \overset{n}{S}(xy) - \bar{y}\overset{n}{S}(x)$$

$$= \overset{n}{S}(xy) - \bar{x}\overset{n}{S}(y)$$

$$= \overset{n}{S}(xy) - \frac{1}{n}\left[\overset{n}{S}(x)\right]\left[\overset{n}{S}(y)\right]$$

The three forms of the correction term—the second term of each right-hand member—are identical. In computational work one should choose that form which is handiest.

Calculations leading to the variances of the three sums at the bottom of Table 14 are given in Table 15. Taking the data of block 1, by way of illustration, it is found that the sums of x^2, y^2, and xy, are

5,070, 2,254, and 2,987,

respectively. The corrections to these, in turn, are

$$\frac{1}{4}(138)^2,\ \frac{1}{4}(92)^2,\ \text{and}\ \frac{1}{4}(138)(92),$$

or

4,761, 2,116, and 3,174;

hence, the corresponding sums of squares and products of deviations about the sample means are

309, 138, and −187,

as given in the table. Each of these is based upon three degrees of freedom among the four random sampling unit observations involved. Similar operations upon the observations of the remaining blocks are also given in Table 15; and these are combined in Table 16 in the second line from the bottom. In order to convert these sums of squares and products into variances and the covariance of the sample totals which are given in the bottom line of Table 14—listed again in Table 16—they are to be multiplied by the factor 4/3; that is, the factor $\left(\frac{n}{n-1}\right)$. Finally, the correction factor for the finite populations sampled is

$$\frac{N-n}{N} = \frac{16}{20},$$

the product of the two factors being

$$\left(\frac{4}{3}\right)\left(\frac{16}{20}\right) = \frac{16}{15}.$$

TABLE 15. Calculation of Sums of Squares and Products among the Random Sampling Unit Observations of Table 14

	x	y	x^2	y^2	xy
BLOCK 1					
	47	13	2,209	169	611
	29	26	841	676	754
	38	25	1,444	625	950
	24	28	576	784	672
Sums	138	92	5,070	2,254	2,987
Corrections			4,761	2,116	3,174
Deviations	..	..	309	138	−187
BLOCK 2					
	35	2	1,225	4	70
	42	6	1,764	36	252
	27	17	729	289	459
	27	19	729	361	513
Sums	131	44	4,447	690	1,294
Corrections			4,290.25	484	1,441
Deviations	..	..	156.75	206	−147
BLOCK 3					
	15	41	225	1,681	615
	9	32	81	1,024	288
	25	38	625	1,444	950
	35	14	1,225	196	490
Sums	84	125	2,156	4,345	2,343
Corrections			1,764	3,906.25	2,625
Deviations	..	..	392	438.75	−282
BLOCK 4					
	23	11	529	121	253
	30	0	900	0	0
	33	14	1,089	196	462
	35	16	1,225	256	560
Sums	121	41	3,743	573	1,275
Corrections			3,660.25	420.25	1,240.25
Deviations	..	..	82.75	152.75	34.75

TABLE 16. Assembly of Sums of Squares and Products, and Calculation therefrom of Variances and Covariances of the Totals of Table 15

Block number	Observed sums			Sums of squares and products			Degrees of freedom
	x	y	$x+y$	x^2	y^2	xy	
1.....	138	92	230	309	138	−187	3
2.....	131	44	175	156.75	206	−147	3
3.....	84	125	209	392	438.75	−282	3
4.....	121	41	162	82.75	152.75	34.75	3
Total.....	474	302	776	940.50	935.50	−581.25	12
Variances and covariances of sample sums (16/15 of above).....				1,003.2	997.9	−620.0	

Since the samples make up 20 percent of the population, the estimates of the population aggregates, with their standard errors, are five times the following:

For x: $474 \pm \sqrt{1{,}003.2}$; or 474 ± 31.67

For y: $302 \pm \sqrt{997.9}$; or 302 ± 31.59

For $(x+y)$: $776 \pm \sqrt{(1{,}003.2)+(997.9)+2(-620.0)}$; or 776 ± 27.59.

Thus the estimates of the population aggregates are

For x: $2{,}370 \pm 158.4$

For y: $1{,}510 \pm 157.9$

For $(x+y)$: $3{,}880 \pm 137.9$

5.3 Simultaneous Sampling of More than Two Populations. Although practical considerations may prescribe otherwise, the most efficient class of sampling designs appropriate to stratified populations on confined areas of land, secure representativeness by subdividing the whole area into as many blocks as a minimum of two random sampling units to the block will permit. The computational operations to be performed upon samples of two submit to systematic and compact tabulation as well.

The necessity for systematic and self-checking computational work can hardly be overemphasized, particularly when the characteristic sampled is distributed over several populations, and it is required to estimate it according to each population separately, as well as according to combinations of any two or more of them.

Let it be required to sample the type areas of Figure 16 by direct observation of 5 percent of the population within the large rectangle.

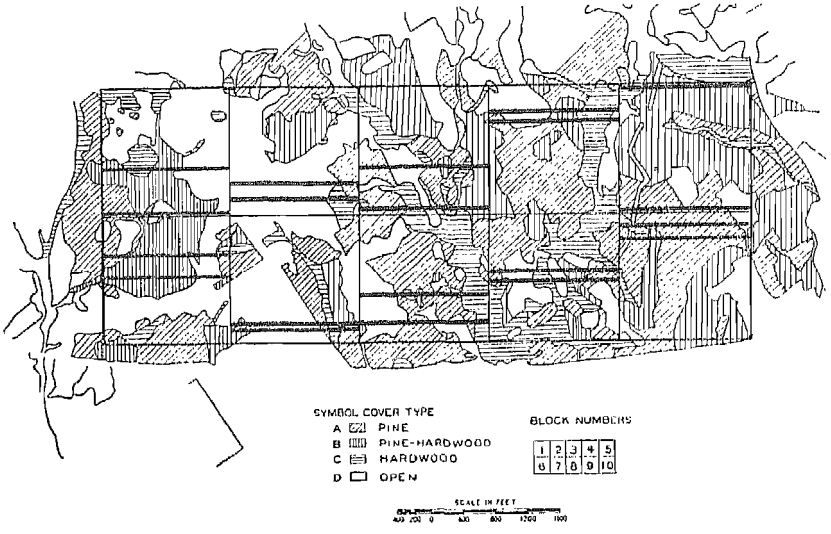


FIG. 16. The type map, subdivided into blocks and showing the two random sampling units of each block.

The objective is to estimate the area of each of four types, and of combinations of any two or three of them, so as to provide exact measures of the probable discrepancy between the true value and the sampling estimate thereof.

The map from which Figure 16 was made is on a scale of 1 inch to 800 feet. The first problem is the division of the area into blocks, each of which is to furnish a separate sample of two random sampling units. While representativeness would be assured by dividing the map into as many blocks as possible, yet a balance must be struck between desirable precision and the labor of acquiring it. For present purposes 10 square blocks, 2 inches to the side, are used. These are shown in the figure.

Since the sample is to cover 5 percent of the area, a practicable scheme, and one of minimum labor, follows from conceiving each block (these being 2 inches square) as made up of 40 contiguous strips, each 1/20-inch wide and extending the length of the block. It is handy, therefore, to consider the ultimate unit as a small square area, 1/20-inch on the side, and the random sampling unit as the sum of the 40 ultimate units of a single strip.

The pattern of cover types seems to extend vertically rather than horizontally. Accordingly, there should be less variability, hence small-

er sampling errors, between strips run horizontally. Then the remaining part of the observational program is the identification of the two strips, out of the 40 of each block, which are to supply the samples. This work is easily effected by means of any random sampling scheme, such as a page of random sampling numbers.

The strips which supply the samples are shown in Figure 16; and the observations, consisting of the number of 1/20-inch squares, according to type, out of the 40 of each strip, are given in Table 17. They contain their own checks. Since the length of each strip is 40 units, the sum of the type observations in each and every strip is 40. Likewise,

TABLE 17. Direct Observations on the Populations of Figure 16. Cover Type Areas According to Random Sampling Units (Strips) and Blocks. Units of 1/20-inch Squares

Block	Strip 1				Strip 2			
	Number of 1/20-inch squares according to cover type							
	A	*B*	*C*	*D*	*A*	*B*	*C*	*D*
1....	..	18	9	13	14	11	..	15
2....	5	..	..	35	..	..	8	32
3....	17	9	2	12	19	7	11	3
4....	12	6	9	13	5	8	11	16
5....	1	5	31	3	31	..	..	9
6....	11	22	..	7	8	32	..	..
7....	3	4	8	25	6	5	3	26
8....	11	3	3	23	14	2	7	17
9....	13	5	9	13	11	5	4	20
10....	33	6	..	1	32	1	4	3
Sum..	106	78	71	145	140	71	48	141

the sum of all the observed values over all 10 blocks is 800, since 800 units of strip were run.

Denoting the first and second random sampling units by subscripts 1 and 2, respectively, the sample sums of cover type areas in units of 1/20-inch squares are the following:

$$\text{Cover type } A\text{: } \overset{10}{S}(A_1+A_2)=106+140=246$$

$$\text{Cover type } B\text{: } \overset{10}{S}(B_1+B_2)=\ 78+\ 71=149$$

$$\text{Cover type } C\text{: } \overset{10}{S}(C_1+C_2)=\ 71+\ 48=119$$

$$\text{Cover type } D\text{: } \overset{10}{S}(D_1+D_2)=145+141=286$$

$$\text{All cover types: } 800$$

These estimates may be expressed, if so desired, either in proportions of total area, by dividing by 800; or in acres, by multiplying by the factor derived from the map scale; that is, since a 1/20-inch square corresponds to $(800/20)^2$ square feet, these estimates are to be multiplied by $(40)^2/43{,}560 = 0.03673$ to give acres of observed area.

For purposes of calculating their probable discrepancies from the true type areas, however, it is preferable for the present to maintain the units of 1/20-inch squares.

5.4 Systematic Reduction of Observations. The observed values of Table 17 are in handy form for drawing up a work sheet of calculations upon which to base the errors of estimate. There are two conditions to the experiment, however, and these are of the same kind as encountered previously; namely, (1) the samples are small, for each block supplies a single sample of just two random sampling units; and (2) the population is a finite one—an area 10x4 inches on the original map—of which 5 percent is contained in the observations, and variances are to be adjusted accordingly.

As before, the first of these conditions requires that strict account be kept of the *degrees of freedom* available for the estimate of the variances involved. There is a single degree of freedom between the two independent observations of any one or more types in each block. The estimate of the variances of the sum of two independent numbers, say (A_1+A_2), where these are the observed values of cover type A on the first and second strips, respectively, of any block, is the square of their difference (Sec. 4.3), that is,

$$V(A_1+A_2) = (A_1-A_2)^2$$

in which V denotes the variance of the enclosed terms following it. In like manner, the estimate of the variance of, say, the combined areas of cover types A and B in one block, may be written

$$V\Big[(A_1+B_1)+(A_2+B_2)\Big] = \Big[(A_1+B_1)-(A_2+B_2)\Big]^2$$

This may be expressed in a form to facilitate later numerical calculation as follows:

$$V\Big[(A_1+B_1)+(A_2+B_2)\Big] = \Big[(A_1-A_2)+(B_1-B_2)\Big]^2$$
$$= (A_1-A_2)^2+(B_1-B_2)^2+2(A_1-A_2)(B_1-B_2).$$

The terms in the expanded right-hand member are, in order, the variance of A, the variance of B, and twice the covariance of A and B, A and B denoting the observed block sum according to cover type. It is necessary to recognize and calculate all the variances and covariances

among the types of the present problem. For conciseness, notation may be simplified as follows: In a given block, let

$$A=(A_1+A_2);\text{ and } a=(A_1-A_2)$$
$$B=(B_1+B_2);\text{ and } b=(B_1-B_2)$$
etc.

Then in this block

$$V(A) = a^2$$
$$V(B) = b^2$$
$$V(A+B) = a^2+b^2+2ab$$
etc.

These equations hold for the pair of random sampling units of each block. In k blocks, the k pairs are independent of one another. Hence over all blocks, the variances of observed sums are

$$V\left[\overset{k}{S}(A)\right] = \overset{k}{S}(a^2)$$
$$V\left[\overset{k}{S}(B)\right] = \overset{k}{S}(b^2)$$
$$V\left[\overset{k}{S}(A+B)\right] = \overset{k}{S}(a^2)+\overset{k}{S}(b^2)+2\overset{k}{S}(ab)$$
etc.,

because the variance of the sum of independent quantities is the sum of their separate variances, the sampling of the individual blocks having been done independently and at random.

Table 18 is the work sheet upon which have been performed the necessary calculations leading to the estimates of the variances of the observed sums of each type area and of possible combinations of areas of

TABLE 18. Calculation of Variances and Covariances of the Data of Table 17

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Block	a	b	c	d	a^2	ab	ac	ad	b^2	bc	bd	c^2	cd	d^2
1...	−14	7	9	− 2	196	− 98	−126	28	49	63	−14	81	− 18	4
2...	5	0	− 8	3	25	0	− 40	15	0	0	0	64	− 24	9
3...	− 2	2	− 9	9	4	− 4	18	− 18	4	− 18	18	81	− 81	81
4...	7	− 2	− 2	− 3	49	− 14	− 14	− 21	4	4	6	4	6	9
5...	−30	5	31	− 6	900	−150	−930	180	25	155	−30	961	−186	36
6...	3	−10	0	7	9	− 30	0	21	100	0	−70	0	0	49
7...	− 3	− 1	5	− 1	9	3	− 15	3	1	− 5	1	25	− 5	1
8...	− 3	1	− 4	6	9	− 3	12	− 18	1	− 4	6	16	− 24	36
9...	2	0	5	− 7	4	0	10	− 14	0	0	0	25	− 35	49
10...	1	5	− 4	− 2	1	5	− 4	− 2	25	− 20	−10	16	8	4
Sum.	....	....	....	....	1,206	−291	−1,089	174	209	175	−93	1,273	−359	278

two or more types. In the columns 2 to 5 are listed the differences between the magnitudes of type areas on the two random sampling units of each block. A check is afforded in that

$$a+b+c+d=0$$

because of the constant strip length. As these numbers are multiplied through by the value of a in the same block, the resulting products are listed in columns 6 to 9. Again a check of the arithmetic is available; for

$$a(a+b+c+d)=0=a^2+ab+ac+ad.$$

In columns 10 to 12, operations of the form

$$b(a+b+c+d)=0=ab+b^2+bc+bd$$

are performed and checked, although the product ab is not repeated, since its numerical value has already been calculated and checked. In the remaining columns, operations of the same kind are performed by using as multipliers, c and d in turn. No product already performed and checked by means of the check sum zero, need be repeated.

The totals in the bottom line of the table are, accordingly, the estimates of variances and covariances of observed cover-type areas, although not yet corrected for the finite population sampled. The corrected values are 38/40 times these tabular totals, since in each block the sample size, $n=2$, and the population size, $N=40$, produce the factor

$$\frac{N-n}{N}=\frac{40-2}{40}=\frac{38}{40}.$$

Upon applying this factor, the corrected values are assembled in Table 19 in handy form for inspection and use.

By way of illustration, from Table 17 the observed area of type A is 246; its variance, from Table 19, is 1,146. The observed area of types $(B+C)$ is 268; its variance is

$$199+1,209+2(166)=1,740.$$

TABLE 19. Variances and Covariances of Area Sums.*
Corrected from the Limited Populations Sampled

	A	*B*	*C*	*D*
A	1,146	−276	−1,035	165
B	.....	199	166	− 88
C	.....	.....	1,209	−341
D	.....	.....	.....	264

*Numbers at intersections of columns and rows of like designation are variances; at intersections of unlike designations, they are covariances.

The observed area of types $(A+C)$ is 365; its variance is

$$1{,}146+1{,}209+2(-1{,}035)=285.$$

As a final illustration, the observed area in types $(A+B+C)$ is 514; its variance is

$$1{,}146+199+1{,}209+2(-276)+2(-1{,}035)+2(166)=264$$

It is to be noted that the variance of the sum for any three types is equal to the variance of the fourth. This follows from the fact that the true area of all four cover types is known exactly as 10x4 square inches, for which the sample supplied 800 square 1/20-inches. Inspection of Table 19 demonstrates that the sum of all variances plus twice the sum of all covariances tallied therein, is zero, that is, the sampling error of all types combined is zero.

The standard error of an area estimate is the square root of its sampling variance. A standard error is, accordingly, a measure of discrepancy of the occurrence of a cover type between the two strips within the blocks. Had each standard error been based upon many degrees of freedom—rather than the 10 actually available for its estimate—it would have implied that the real error would have exceeded the standard in 32 out of 100 sets of samples, based upon the same sampling design.

With but 10 degrees of freedom available for its estimate, as in this case, a standard error is, itself, subject to appreciable sampling error. Consequently, the discrepancy between the population parameter and the sampling estimate thereof as a ratio of its standard error—producing the statistic t—is distributed as t on 10 degrees of freedom. If the probability be fixed at 0.05 that the real discrepancies of these sums exceed their calculated values, the latter in each case is

$$ts\sqrt{kn}=2.228(s)\sqrt{(2)(10)}$$

for which 2.228 is the value of t at this probability on 10 degrees of freedom and $s\sqrt{kn}$ is the standard error of a particular estimated sum under discussion, s being the standard deviation of random sampling units within the blocks. Several of these are listed in Table 20 according to certain types and type classifications; in the second and third columns according to the units of observation, and in the fourth and fifth columns they are given as proportion of total area, by dividing the observed units by 800. These figures may be transcribed to acres by multiplying those of the second and third columns by 0.7346—that is, by 20(0.03673). In the fifth column the limits of probable discrepancies as percentages of observed areas are listed.

In conclusion, it is worth remarking that while this problem deals with areas as delineated on a map, no new principle would have been involved had the data been timber volume by types or age classes, reproduction counts by species classes, forage areas by density classes or types, or any other forestry data as observed by a representative sampling method in the field.

TABLE 20. Partial Summary of Results

Type	In 1/20-inch squares		Proportion of total area		Limit of discrepancy in percent of observed area*
	Observed area	Limit of discrepancy*	Observed proportion	Limit of discrepancy*	
A	246	±75.4	.308	±.094	±30.7
B	149	±31.4	.186	±.039	±21.2
C	119	±77.5	.149	±.097	±65.1
D	286	±36.2	.358	±.045	±12.7
A+*B*	395	±62.7	.494	±.078	±15.9
A+*C*	365	±37.6	.456	±.047	±10.3
B+*C*	268	±92.9	.335	±.116	±34.7

*The probability is 0.05 that the real discrepancies exceed those listed.

CHAPTER VI

THE METHOD OF SUB-SAMPLING

6.1 Distinctive Feature of the Method. This chapter treats of an extension of the representative sampling method, the distinctive feature of which is that the actual measurements, or observations, are taken from a portion only—and not the whole—of each random sampling unit. In other words, each random sampling unit drawn for observation from a block—and which is now termed *major random sampling unit*—is sampled in turn. This is effected by drawing therefrom, independently and at random, a portion of the *minor random sampling units* into which each major random sampling unit may be divided, and confining the direct observations, or measurements, to these. The minor random sampling unit may or may not be the ultimate unit.

In consequence of the sub-sampling procedure involved, the sampling error of the population estimate within the blocks is made up of contributions from two sources of variation, namely: (a) among major random sampling units of the same block, and (b) among minor random sampling units of the same major random sampling unit.

The sampling error will be discussed in some detail later. For the present, an illustration of the method as applied by Hasel[11] to the timber cruise may be helpful.

6.2 An Illustration of the Method. Nine square miles (sections) of the ponderosa pine type of California were divided by Hasel into 18 blocks, each a half-section of 320 acres (Fig. 17). The population is volume in M feet b.m. to the 2½ acre plot, and consists of 2,304 such volumes altogether, or 128 to each block. The major random sampling unit is the strip, 2½ by 80 chains in dimension, running the length of the block, there being 16 such strips to the block. Each strip is subdivided into eight plots of 2½ acres. The plot dimensions are 2½ by 10 chains and lie end to end on the strip.

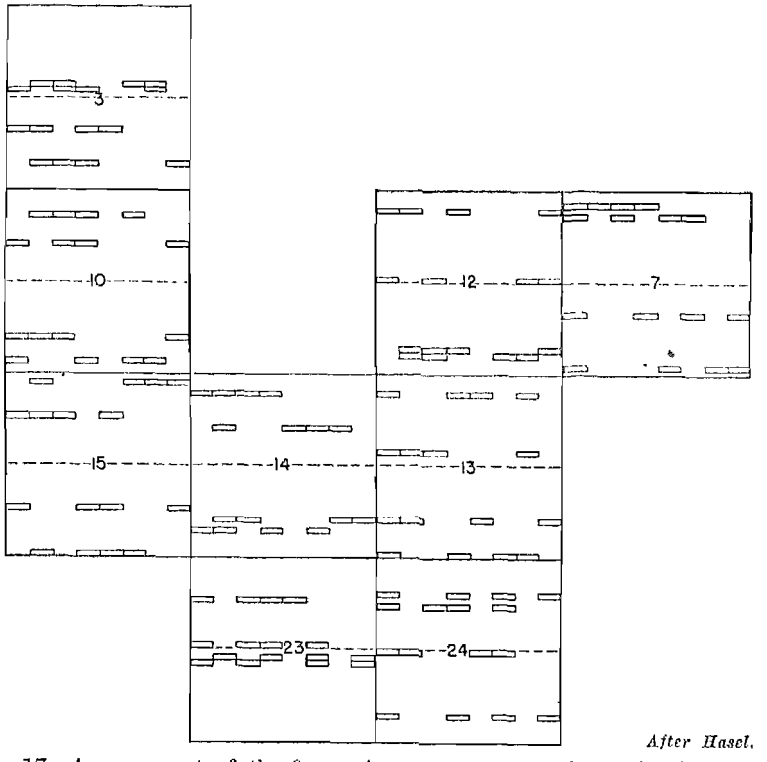
Hasel drew two strips, out of the 16 within each block, independently and at random. Each of these was then sampled in that only four of its eight plots were drawn, again independently and at random, their volumes in M feet b.m. supplying the observations.[12] The strip is thus the

[11] Hasel, A. A. Arrangement of cruise plots to permit a valid estimate of sampling error. California Forest and Range Experiment Station. *Multigraphed* report, 1937.

[12] As applied to timber estimation Hasel has called the method the "random line-plot cruise."

major random sampling unit, the plot is the minor random sampling unit and also the ultimate unit.

Figure 17 shows the ground plan of the 18 blocks as well as the location of the strips, and plots on each strip, which were drawn on one trial. As the sample cruise consists of volume on one-half the plots on one-eighth the strips, the cruise covers $\frac{1}{2}$ x $\frac{1}{8}$ or $6\frac{1}{4}$ percent of the population.



After Hasel.

FIG. 17. Arrangement of the four minor random sampling units (of plots) within each of the two major random sampling units (of strips) on 18 blocks in a timber cruise according to the method of sub-sampling.

In order to arrive at the sampling error appropriate to the method of sampling a limited population of plots within a limited population of strips of the same block, the components of such sampling error will first be discussed somewhat in detail.

6.3 Components of Sampling Error. Assume that a population is distributed over B blocks and that each block contains Q major divisions, say of contiguous strips; while each strip contains P

subdivisions, say of plots. Then the population of each block is distributed over QP plots.

Suppose, in the first case, that there is drawn from one of the blocks a random sample of p plots out of the P contained on a single strip. Then each *real* error, ϵ, between the plot observation, y, and the *true* mean, μ_q, of the strip is

$$\epsilon = y - \mu_q$$

and the variance among plots of the same strip may be denoted as σ_ϵ^2. It follows that the exact variance of the mean of p values of y is

$$\frac{\sigma_\epsilon^2}{p}\left(\frac{P-p}{P-1}\right)$$

including the adjustment to the finite population of P plots to the strip. This expression is taken from equation (7) of Sec. 3.4, the Appendix. The correction factor $\left(\frac{P-p}{P}\right)$ is not to be used here, since σ_ϵ^2 represents the population variance.

Suppose, as a second case, there is drawn a random sample of q *complete* strips out of the Q in the block, each strip value, denoted by μ_q, being expressed on the plot basis; that is, a random sample of q values of μ_q where each

$$\mu_q = \frac{1}{P}\overset{P}{S}(y).$$

Then if the *real* error, $\triangle$, between the strip mean, μ_q, and the *true* mean, μ_b, of the block be denoted

$$\triangle = (\mu_q - \mu_b),$$

the population variance among strips of the same block may be denoted as $\sigma_\triangle^2$, and the exact variance of the observed block mean, that is, the variance of

$$\frac{1}{q}\overset{q}{S}(\mu_q)$$

is

$$\frac{\sigma_\triangle^2}{q}\left(\frac{Q-q}{Q-1}\right)$$

including the adjustment to the finite population of Q strips to the block.

Suppose, finally, that for a given block, there is drawn a random sample of q values of y_q, each y_q being the mean of a random sample of p plots from the same strip. In this case, there are two sources of error contained in the variance among strips of the same block; for each

$$(y_q - \mu_b) = (\mu_q - \mu_b) + (y_q - \mu_q),$$

the first term on the right being the *real* error, Δ, among the true strip means, the population variance of which we have designated as σ^2_Δ; while the second term on the right represents the *real* error of the observed strip mean, the population variance of this mean being

$$\frac{\sigma^2_\epsilon}{p}\left(\frac{P-p}{P-1}\right)$$

as given above. As these two errors are independent, the variance among strips of the same block is the variance of

$$(\mu_q-\mu_b)+(y_q-\mu_q),$$

which may be expressed[13]

$$V(y_q)\rightarrow \sigma^2_\Delta+\frac{\sigma^2_\epsilon}{p}\left(\frac{P-p}{P-1}\right).$$

It follows that the sampling variance of the block mean, y_b, where

$$y_b=\frac{1}{q}\overset{q}{S}(y_q)=\frac{1}{qp}\overset{q}{S}\overset{p}{S}(y),$$

is the variance of y_q when divided by q, and then applying the factor $\left(\frac{Q-q}{Q-1}\right)$ to $\frac{\sigma^2_\Delta}{q}$. Hence

$$V(y_b)\rightarrow \frac{\sigma^2_\Delta}{q}\left(\frac{Q-q}{Q-1}\right)+\frac{\sigma^2_\epsilon}{qp}\left(\frac{P-p}{P-1}\right),$$

while the variance of the general mean of B block means, that is, the variance of

$$\bar{y}=\frac{1}{B}\overset{B}{S}(y_b)=\frac{1}{Bqp}\overset{B}{S}\overset{q}{S}\overset{p}{S}(y)$$

is the variance among block means when divided by the number of blocks, B. Hence

$$V(\bar{y})\rightarrow \frac{\sigma^2_\Delta}{Bq}\left(\frac{Q-q}{Q-1}\right)+\frac{\sigma^2_\epsilon}{Bqp}\left(\frac{P-p}{P-1}\right). \quad \ldots\ldots\ldots\ldots (1)$$

The estimates of σ^2_Δ and of σ^2_ϵ which this expression requires may be made in orderly fashion through the procedure known as the analysis of variance. This procedure will be treated next.

6.4 Analysis of Variation among Sampling Units. Given a single block made up of any number, Q, of major random sampling units, say of strips, each containing any number, P, of minor random sampling units, say of plots, let there be drawn q strips, independently and at random, and *on each of these* let there be drawn p plots, also independent-

[13] The symbol $\rightarrow$ is read "is an estimate of."

ly and at random. If, then, y represents the observation on a single plot, y_q the average of the p values of y on a single strip, then

$$y-y_b=(y_q-y_b)+(y-y_q) \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where y_b represents the block average, that is, the average y_q of the q strip averages. The quantity (y_q-y_b), being the deviation of the strip mean from the block mean, is the "strip effect"; and $(y-y_q)$ is the deviation of y from its own strip mean. Thus each residual $(y-y_b)$ about the block mean may be *analyzed* into two portions which are assignable, with more or less accuracy, to their causes; namely, to the average content of the strip from which y is drawn, and to the consistency with which individual values of y express the strip average.

The analysis of the data of the method of sub-sampling into these two classes of information—both of which are needed to evaluate sampling error—is readily performed through the arithmetical arrangement known as the analysis of variance.[14] Upon squaring the identity (2) for a single plot within a particular strip,

$$(y-y_b)^2=(y_q-y_b)^2+(y-y_q)^2+2(y_q-y_b)(y-y_q)$$

and after performing like operations upon each of the p plots of the same strip, and noting that (y_q-y_b) is identical over all the observations of this strip, their sum is the following:

$$\overset{p}{S}\left[(y-y_b)^2\right]=p(y_q-y_b)^2+\overset{p}{S}\left[(y-y_q)^2\right]+2(y_q-y_b)\overset{p}{S}(y-y_q).$$

The third term of the right-hand member is zero, since $\overset{p}{S}(y-y_q)$ is zero by definition.

Upon calculating sums of squares like the above according to each of the q strips of the block and adding together, we have for the block

$$\overset{q}{S}\overset{p}{S}\left[(y-y_b)^2\right]=p\overset{q}{S}\left[(y_q-y_b)^2\right]+\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right] \quad \ldots\ldots\ldots\ldots (3)$$

The total sum of squares of (3) is based upon qp observations and $(qp-1)$ degrees of freedom. These are divided into $(q-1)$ degrees of freedom among the q strips for the first term of the right-hand member,

[14] The analysis of variance, introduced by R. A. Fisher, is a general method of sorting out the various classes of information an experiment or investigation is designed to test. It provides estimates of experimental and sampling errors. Together with the known distribution of the statistic z—also due to Fisher—of which the statistic t is a special case, it provides tests of a great variety of statistical hypotheses. Professor Fisher (1936) (1938) treats elegantly of the analysis of variance.

Three well-known American authors who deal largely with the methods of Professor Fisher are the following: Snedecor (1937), Goulden (1938), and Rider (1939).

which is due to variation among the major random sampling units of strips; and into the q sets of $(p-1)$ degrees of freedom for the second term, which is due to variation among the minor random sampling units of plots of the same strip.

If, now, the population is made up of B blocks, each of which is sampled in the same way as this one, there will be B sums of squares, each of the form of equation (3). When these are summed over all B blocks, we have, finally,

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_b)^2\right]=p\overset{B}{S}\overset{q}{S}\left[(y_q-y_b)^2\right]+\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]$$

This identity may be tabulated in analysis of variance form, as in Table 21. The first two colums show the division of the total sum of squares of plots, within the blocks, into portions due, respectively, to variation between, and within, strips of the same block. The third column contains the degrees of freedom. These are each B times the number for a single block.

TABLE 21. Analysis of Variance Appropriate to the Method of Sub-Sampling

Source of variation	Sum of squares	Degrees of freedom	Mean square
Among strips, same block . . .	$p\overset{B}{S}\overset{q}{S}\left[(y_q-y_b)^2\right]$	$B(q-1)$	$C \rightarrow p\,\sigma^2_{\Delta}\left(\frac{Q}{Q-1}\right)+\sigma^2_{\epsilon}\left(\frac{P-p}{P-1}\right)$
Among plots, same strip	$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]$	$Bq(p-1)$	$D \rightarrow \sigma^2_{\epsilon}\left(\frac{P}{P-1}\right)$
Among plots, same block . . .	$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_b)^2\right]$	$B(qp-1)$	

The last column on the right contains the pertinent mean squares, symbolized as C and D. As these mean squares contain the estimates of σ^2_{Δ} and of σ^2_{ϵ} of equation (1) of the preceding section, they need to be analyzed into their components.

The mean square, D, among plots of the same strip, is the mean square of the residuals $(y-y_q)$, each of which is a part of the corresponding *real* (but unknown) error $(y-\mu_q)$, such that each

$$(y-\mu_q)=(y-y_q)+(y_q-\mu_q).$$

The second term on the right, in this expression, is the *true* (also unknown) error of the observed strip mean, y_q. Upon squaring and sum-

ming over the p plots of a single strip, and noting that the cross-product term of the expression is zero,

$$\overset{p}{S}\left[(y-\mu_q)^2\right]=\overset{p}{S}\left[(y-y_q)^2\right]+p(y_q-\mu_q)^2.$$

There is, of course, an expression of the same form for each of the q sampled strips in the particular block. Summing over all q expressions within a given block, and then over all B blocks,

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-\mu_q)^2\right]=\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]+p\overset{B}{S}\overset{q}{S}\left[(y_q-\mu_q)^2\right].$$

Finally, upon transposing so as to have the sum of squares of the observed residuals on the left

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]=\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-\mu_q)^2\right]-p\overset{B}{S}\overset{q}{S}\left[(y_q-\mu_q)^2\right]\ldots\ldots\ldots(4)$$

The left-hand member of equation (4) is the sum of squares among plots of the same strip as given in Table 21. The expressions on the right of equation (4) are sums of squares of *real* (but unknown) errors, the first containing the individual errors ϵ, and the second containing the mean of p such errors. Hence equation (4) may be expressed

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]\rightarrow Bqp\,\sigma_\epsilon^2-Bpq\frac{\sigma_\epsilon^2}{p}$$

provided the number of sampled plots, p, is a small proportion of P. If this proportion is not small, the adjustment $\left(\frac{P-p}{P-1}\right)$ is to be applied to the estimate of the variance of the strip means, y_q, in which case

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]\rightarrow Bqp\,\sigma_\epsilon^2-Bqp\,\frac{\sigma_\epsilon^2}{p}\left(\frac{P-p}{P-1}\right)$$

and this simplifies to the following:

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]\rightarrow Bq(p-1)\,\sigma_\epsilon^2\left(\frac{P}{P-1}\right).$$

Upon dividing by $Bq(p-1)$, the mean square, D, among plots of the same strip is

$$D=\frac{1}{Bq(p-1)}\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]\rightarrow\sigma_\epsilon^2\left(\frac{P}{P-1}\right)$$

as given in the mean square column of Table 21.

The mean square, C, among strips of the same block, in Table 21, is the mean square of the residuals $(y_q - y_b)$, each of which is a part of the corresponding *real* (but unknown) error $(y_q - \mu_b)$ among strips of the same block, such that each

$$(y_q - \mu_b) = (y_q - y_b) + (y_b - \mu_b).$$

The second term on the right is the *true* (also unknown) error of the observed block mean, y_b. Squaring and summing over the q strips of a single block

$$\overset{q}{S}\left[(y_q - \mu_b)^2\right] = \overset{q}{S}\left[(y_q - y_b)^2\right] + q(y_b - \mu_b)^2.$$

There is an expression of this same form for each block. Summing over all B expressions,

$$\overset{B}{S}\,\overset{q}{S}\left[(y_q - \mu_b)^2\right] = \overset{B}{S}\,\overset{q}{S}\left[(y_q - y_b)^2\right] + q\,\overset{B}{S}\left[(y_b - \mu_b)^2\right].$$

Upon transposing so as to have the sum of squares of the observed residuals on the left, and multiplying by p,

$$p\,\overset{B}{S}\,\overset{q}{S}\left[(y_q - y_b)^2\right] = p\,\overset{B}{S}\,\overset{q}{S}\left[(y_q - \mu_b)^2\right] - pq\,\overset{B}{S}\left[(y_b - \mu_b)^2\right] \quad \ldots\ldots (5)$$

The left-hand member of equation (5) is the sum of squares among strips of the same block, as given in Table 21. The expressions on the right of equation (5) are p times the sums of squares of the *real* (but unknown) errors, the first containing the individual strip errors of the form

$$(y_q - \mu_b) = (\mu_q - \mu_b) + (y_q - \mu_q)$$
$$= \triangle + \frac{1}{p}\overset{p}{S}(\epsilon)$$

so that

$$p\,\overset{B}{S}\,\overset{q}{S}\left[(y_q - \mu_b)^2\right] \rightarrow pBq\left[\sigma^2_{\triangle} + \frac{\sigma^2_{\epsilon}}{p}\right],$$

without adjustment to the finite populations of P plots to the strip; while the second term on the right in equation (5) is pq times the sum of squares of the *true* (also unknown) errors of the observed block means y_b. Now each

$$(y_b - \mu_b) = \frac{1}{q}\overset{q}{S}(\mu_q - \mu_b) + \frac{1}{q}\overset{q}{S}(y_q - \mu_q)$$

$$= \frac{1}{q} \overset{q}{S}(\triangle) + \frac{1}{qp} \overset{q}{S} \overset{p}{S}(\epsilon)$$

so that

$$pq \overset{B}{S}\left[(y_b - \mu_b)^2\right] \rightarrow pqB \frac{\sigma^2_{\triangle}}{q} + pqB \frac{\sigma^2_{\epsilon}}{qp}$$

although not yet adjusted to the finite populations, either of Q strips to the block, or of P plots to the strip.

Assembling these portions of the right-hand member of equation (5), we find that the sum of squares among strips of the same block, as listed in Table 21, contains an estimate of variances as follows:

$$p \overset{B}{S} \overset{q}{S}\left[(y_q - y_b)^2\right] \rightarrow pBq\left\{\left[\sigma^2_{\triangle} + \frac{\sigma^2_{\epsilon}}{p}\right] - \left[\frac{\sigma^2_{\triangle}}{q} + \frac{\sigma^2_{\epsilon}}{qp}\right]\right\}.$$

The adjustments to finite populations have not yet been applied. If the number of sampled strips, q, is not a negligibly small proportion of the entire number, Q, within the blocks, the factor $\left(\frac{Q-q}{Q-1}\right)$ is to be applied to the variance of block means, while if the corresponding proportion of plots to the strip is not small, the factor $\left(\frac{P-p}{P-1}\right)$ is to be applied to the variance of strip means. Applying these factors,

$$p \overset{B}{S} \overset{q}{S}\left[(y_q - y_b)^2\right] \rightarrow pBq\left\{\left[\sigma^2_{\triangle} + \frac{\sigma^2_{\epsilon}}{p}\left(\frac{P-p}{P-1}\right)\right] - \left[\frac{\sigma^2_{\triangle}}{q}\left(\frac{Q-q}{Q-1}\right) + \frac{\sigma^2_{\epsilon}}{qp}\left(\frac{P-p}{P-1}\right)\right]\right\}$$

which may be simplified to the following:

$$p \overset{B}{S} \overset{q}{S}\left[(y_q - y_b)^2\right] \rightarrow pB(q-1)\left[\sigma^2_{\triangle}\left(\frac{Q}{Q-1}\right) + \frac{\sigma^2_{\epsilon}}{p}\left(\frac{P-p}{P-1}\right)\right].$$

Finally, upon dividing by $B(q-1)$, the mean square, C, among strips of the same block is

$$C = \frac{1}{B(q-1)} p \overset{B}{S} \overset{q}{S}\left[(y_q - y_b)^2\right] \rightarrow p\sigma^2_{\triangle}\left(\frac{Q}{Q-1}\right) + \sigma^2_{\epsilon}\left(\frac{P-p}{P-1}\right)$$

as given in the mean square column of Table 21.

It should be noted at once, that one cannot separate, exactly, the two variances, $\sigma^2_{\triangle}$ and σ^2_{ϵ}, involved in the mean square, C, among the strips of the same block. The analysis of variance table, however, supplies an independent estimate of σ^2_{ϵ}, in the mean square among plots of the same strip, for which as shown above,

$$D=\frac{1}{Bq(p-1)}\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right]\rightarrow\sigma^2_{\epsilon}\left(\frac{P}{P-1}\right).$$

The *estimate* of σ^2_{Δ} may therefore be calculated from the mean square C, upon inserting $D\left(\frac{P-1}{P}\right)$ for σ^2_{ϵ} therein. With this substitution

$$C\rightarrow p\,\sigma^2_{\Delta}\left(\frac{Q}{Q-1}\right)+D\left(\frac{P-1}{P}\right)\left(\frac{P-p}{P-1}\right),$$

and accordingly,

$$\left[C-D\left(\frac{P-p}{P}\right)\right]\left[\frac{Q-1}{Q}\right]\rightarrow p\sigma^2_{\Delta}\quad\ldots\ldots\ldots\ldots(6)$$

We now have all the materials needed to estimate the variance of the general mean, $\bar{y}$. As given in equation (1) of the preceding section, it may be expressed as follows:

$$V(\bar{y})\rightarrow\frac{1}{Bqp}\left[p\,\sigma^2_{\Delta}\left(\frac{Q-q}{Q-1}\right)+\sigma^2_{\epsilon}\left(\frac{P-p}{P-1}\right)\right]$$

Upon substituting expression (6) for $p\sigma^2_{\Delta}$, and $D\left(\frac{P-1}{P}\right)$ for σ^2_{ϵ}, this becomes

$$V(\bar{y})=\frac{1}{Bqp}\left\{\left[C-D\left(\frac{P-p}{P}\right)\right]\left[\frac{Q-1}{Q}\right]\left[\frac{Q-q}{Q-1}\right]+D\left(\frac{P-1}{P}\right)\left(\frac{P-p}{P-1}\right)\right\}$$

and this may be simplified to the following:

$$V(\bar{y})=\frac{1}{Bqp}\left[C\left(\frac{Q-q}{Q}\right)+D\left(\frac{P-p}{P}\right)\left(\frac{q}{Q}\right)\right]\quad\ldots\ldots\ldots\ldots(7)$$

6.5 Application to an Insect Population. Table 22 shows the distribution of Colorado potato beetles (*Leptinotarsa decemlineata*) in a heavily infested field according to each 2-feet of row (the ultimate unit) of potatoes for entire rows in the field. Let it be required to estimate the population of beetles from an examination of 1/16 of the 2304 ultimate units according to a sampling design based upon the method of sub-sampling.

To this end the field is arbitrarily divided into 12 blocks of equal area, such that they constitute a system of four tiers—or rows of blocks in the table—of three blocks to the tier. Then each block consists of 12 rows of potatoes of 16 ultimate units to the row, or 192 to the block. Designating the potato row of a block as the major random sampling unit, we shall draw three of them, independently and at random, from

TABLE 22. Distribution of Colorado Potato Beetles *(Leptinotarsa decemlineata)* in a Potato Field in Ultimate Units According to Half Rows within Blocks*

Rows within blocks (major random sampling units)

	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	2	4	2	4	2	3	4	6	6	7	11	1	3	2	14	3	5	1	6	4	5	4	3	11
2	0	2	0	2	7	4	2	3	9	10	5	13	7	10	11	4	5	3	6	5	1	0	14	5
3	16	9	2	8	5	10	8	7	9	3	1	9	17	10	7	7	16	10	4	12	4	13	11	12
4	5	7	7	14	20	5	9	6	8	4	7	13	1	7	18	7	5	13	7	1	4	6	8	10
5	12	5	7	0	4	10	13	5	13	3	13	3	14	5	16	7	16	9	15	11	7	1	6	6
6	11	6	17	1	9	9	5	17	6	24	9	23	16	9	23	12	5	10	9	18	12	15	18	6
7	11	10	13	21	10	11	2	20	7	16	12	9	5	13	12	10	7	11	18	3	3	22	9	11
8	2	3	5	7	4	1	0	14	11	4	5	4	0	1	7	14	10	8	4	9	2	18	7	18
9	6	8	0	0	2	1	7	1	4	9	10	6	0	3	18	11	8	10	9	0	2	23	8	14
10	8	14	10	7	3	3	17	8	5	9	1	10	14	4	8	11	9	10	8	11	13	16	13	4
11	7	12	13	16	11	8	9	1	11	5	12	19	12	11	13	5	9	12	20	20	5	8	14	22
12	13	2	10	10	7	8	15	28	8	3	9	13	4	10	14	7	18	7	13	14	12	6	10	8
1	4	7	3	2	4	1	3	7	9	7	3	3	0	3	3	9	3	8	5	6	7	4	5	3
2	2	1	7	8	2	6	7	1	5	3	13	2	6	6	3	13	10	5	4	4	5	4	4	2
3	5	7	5	2	5	6	8	2	10	4	6	2	8	3	7	2	8	4	2	8	8	7	4	2
4	6	1	3	5	6	3	5	7	8	6	5	4	4	2	1	2	5	4	11	9	6	3	9	5
5	5	4	15	6	8	5	10	6	9	6	5	3	12	5	8	8	4	0	5	8	3	1	4	6
6	6	5	1	6	6	7	5	3	2	11	2	6	5	3	8	5	6	4	4	3	2	4	4	11
7	4	3	7	10	5	5	2	2	1	5	9	10	3	6	7	11	9	2	2	7	7	4	3	1
8	6	3	7	6	3	5	6	1	4	6	1	0	2	3	2	1	5	0	4	3	6	6	7	7
9	5	4	1	5	3	6	1	2	0	5	6	3	2	2	6	1	3	2	5	1	7	5	6	2
10	3	7	1	4	0	2	4	0	3	4	5	2	2	4	4	1	2	0	4	4	4	12	8	5
11	6	5	5	6	2	0	1	5	3	3	2	2	0	1	7	2	4	2	4	7	1	2	4	2
12	5	4	6	3	10	0	9	3	3	10	1	4	5	4	2	3	1	1	3	4	1	5	0	1
1	5	6	10	6	13	3	14	7	6	8	7	10	1	6	3	0	4	0	9	15	1	4	2	13
2	3	3	10	3	10	1	1	4	2	6	1	12	2	1	2	5	7	5	8	3	4	6	8	1
3	3	8	8	6	4	7	8	7	9	8	13	8	6	1	9	13	0	3	9	2	5	4	2	3
4	9	7	10	3	5	6	6	2	6	10	4	4	1	2	0	1	6	9	3	2	4	1	1	9
5	9	11	9	3	10	2	6	6	5	22	5	6	7	4	2	4	5	13	4	7	3	2	3	4
6	4	3	9	9	9	4	6	7	2	1	8	3	5	1	6	6	15	1	7	0	4	7	3	3
7	7	4	4	8	14	5	4	9	4	11	0	4	6	4	6	8	5	7	6	11	9	2	7	4
8	6	8	10	4	9	2	0	13	9	8	1	4	1	2	2	6	9	4	11	12	8	7	12	8
9	2	8	4	9	3	4	9	5	3	4	8	3	0	1	9	6	7	8	5	12	7	4	6	7
10	4	10	6	8	7	1	8	2	4	5	6	6	11	3	5	12	11	7	8	5	2	4	3	2
11	2	4	5	9	11	10	3	4	3	5	7	1	8	8	6	2	6	7	13	5	8	5	5	2
12	4	13	4	2	8	0	3	2	2	8	6	5	1	6	4	15	5	5	3	4	15	8	13	16
1	2	6	4	11	0	3	0	0	3	2	2	7	6	6	2	4	2	4	4	7	5	3	5	5
2	9	7	10	0	2	4	5	3	2	7	4	1	1	0	1	5	8	3	4	3	5	1	2	2
3	1	6	4	0	4	2	0	5	1	6	2	2	1	0	4	8	5	1	4	2	9	5	5	4
4	4	5	0	3	4	3	4	8	1	7	7	5	5	0	12	4	0	2	4	7	2	1	2	5
5	9	6	3	6	13	5	2	4	2	2	3	1	2	2	5	4	6	5	2	3	5	0	5	4
6	6	9	3	0	2	4	7	16	11	7	5	1	7	1	9	5	4	7	8	2	0	3	3	5
7	5	5	1	10	5	4	1	4	0	1	4	7	6	1	7	5	3	4	2	2	5	4	4	3
8	5	4	5	2	4	3	4	1	2	4	1	4	2	3	2	0	1	4	3	2	1	2	5	2
9	1	0	1	8	2	5	1	2	2	1	7	10	1	5	2	1	2	4	2	1	2	0	3	1
10	0	1	0	4	0	0	0	0	0	4	1	2	4	6	1	4	2	6	4	3	1	2	6	0
11	2	1	0	0	0	15	0	8	4	2	6	2	11	1	0	1	4	4	1	4	5	2	6	3
12	2	2	4	0	0	2	11	2	5	4	2	6	2	2	1	2	5	1	2	4	6	5	2	3

	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	4	6	2	12	1	1	1	2	4	4	0	4	6	5	1	3	1	0	4	3	2	6	0	3
2	0	2	0	0	2	3	2	4	4	1	1	1	0	2	8	1	0	0	1	6	1	2	0	0
3	0	2	4	1	4	1	2	2	1	2	1	1	3	7	3	2	1	1	3	4	0	0	1	0
4	3	2	5	0	3	3	2	1	6	6	4	1	3	11	7	0	3	5	4	7	0	2	0	3
5	5	7	3	4	7	2	3	5	3	1	4	6	9	0	5	3	1	1	2	4	2	0	1	0
6	14	11	6	6	3	2	3	7	13	1	4	9	4	5	0	3	5	4	0	1	6	1	7	2
7	10	7	4	7	2	6	11	1	5	6	0	3	5	7	4	8	4	1	3	2	6	5	1	0
8	6	4	2	4	7	6	4	3	1	0	0	0	4	2	0	1	2	7	1	3	0	1	2	1
9	5	9	5	4	5	14	1	7	11	4	3	5	2	1	0	4	0	3	9	0	0	0	0	6
10	6	7	4	10	7	2	5	9	7	9	2	8	0	4	3	5	3	3	2	4	3	5	3	2
11	4	9	5	9	5	10	8	4	3	4	11	8	1	9	4	5	8	2	4	9	8	0	3	2
12	3	9	10	10	6	1	3	1	2	0	2	6	10	5	11	3	4	5	2	0	5	2	4	1
1	25	9	8	1	17	9	11	23	9	2	12	13	4	9	5	12	5	5	3	15	8	7	5	4
2	13	9	3	17	12	15	11	18	4	3	3	4	7	6	8	8	2	9	5	2	5	7	3	1
3	6	8	7	4	10	13	4	8	6	7	14	5	10	16	0	5	3	7	2	5	7	2	1	4
4	11	6	6	9	3	6	0	8	11	2	8	10	7	6	8	3	5	6	9	3	7	7	2	6
5	7	2	5	5	11	3	10	9	2	10	9	7	7	7	4	4	9	2	1	12	3	5	0	1
6	1	19	9	4	10	3	6	5	4	8	0	8	6	7	3	9	7	9	2	6	2	4	6	0
7	1	2	5	11	5	4	4	3	0	1	1	1	3	5	4	4	4	8	5	4	4	7	4	3
8	7	3	6	4	2	14	1	3	6	3	7	7	8	0	7	2	12	6	7	1	13	2	1	2
9	4	5	3	7	2	6	7	4	12	2	6	8	3	4	7	4	6	6	5	1	3	4	5	4
10	5	9	10	8	3	7	2	13	4	5	6	18	6	11	14	6	6	8	8	3	1	8	5	0
11	3	6	7	2	4	4	6	8	2	6	4	1	5	7	2	2	11	5	3	6	5	10	3	5
12	6	1	10	3	5	4	6	6	15	6	9	8	9	3	3	8	7	6	1	6	3	7	4	1
1	6	6	2	0	3	5	1	4	3	2	4	1	0	2	3	0	8	0	1	5	2	0	0	2
2	0	1	2	4	2	1	2	4	5	1	3	4	3	2	1	1	0	2	2	2	4	1	0	1
3	4	6	5	7	1	6	5	5	2	4	2	5	3	5	3	2	0	3	2	3	3	0	13	3
4	6	0	2	1	3	0	4	2	5	4	3	4	3	4	4	2	6	2	0	2	1	1	1	1
5	8	4	5	3	0	6	3	1	5	6	5	1	6	4	3	2	3	3	3	5	1	4	2	3
6	1	2	2	5	5	3	4	7	4	4	0	5	2	3	2	5	2	3	5	2	2	3	3	4
7	8	6	4	6	9	3	6	3	0	7	2	4	3	4	0	6	8	7	6	2	2	3	0	4
8	6	4	5	0	7	3	4	4	2	9	4	4	2	1	2	3	3	6	2	1	4	8	0	3
9	5	5	4	2	4	2	3	2	1	2	5	5	5	3	6	7	1	3	1	3	5	6	4	9
10	1	7	6	1	7	1	4	5	6	5	3	4	3	2	0	4	4	2	2	4	2	9	0	4
11	1	8	5	2	6	10	6	10	4	6	9	1	4	1	4	3	3	2	2	2	1	4	5	0
12	4	5	1	4	3	4	1	4	3	4	8	12	1	4	1	4	1	0	5	1	2	7	0	1
1	4	4	0	1	3	3	3	4	0	3	1	3	4	8	0	0	8	7	1	1	10	7	6	6
2	6	1	1	1	2	1	1	3	2	3	0	2	13	6	6	2	5	5	7	9	4	8	7	2
3	3	1	1	0	3	6	3	3	1	4	1	0	2	3	8	3	7	8	2	12	2	3	9	3
4	1	1	2	3	3	3	0	0	0	2	0	1	4	2	6	10	0	2	3	3	4	4	5	7
5	3	1	3	4	1	2	1	1	1	0	2	2	2	2	4	2	2	3	6	5	2	5	4	4
6	3	3	1	1	0	2	0	3	3	3	0	0	5	1	4	1	2	6	7	2	5	0	7	1
7	0	1	1	0	0	0	2	0	4	2	1	2	0	2	2	3	1	4	4	7	3	4	5	0
8	2	1	0	1	1	1	1	1	0	2	0	0	1	5	1	2	3	6	2	4	2	2	1	3
9	3	1	0	1	2	1	3	5	0	1	2	1	7	5	2	0	2	4	1	6	5	1	1	0
10	1	1	1	0	4	1	1	0	0	1	0	0	7	4	2	3	1	2	2	5	4	3	4	0
11	2	0	0	1	0	0	0	0	3	2	2	2	3	2	3	1	3	2	3	3	2	4	1	1
12	2	1	1	0	0	0	1	1	3	1	2	2	0	3	5	2	2	4	2	4	1	0	2	0

← Direction of rows of potatoes →

*Beall, G. Methods of estimating the population of insects in a field. Biometrika. 30 : 422–439. 1939. Table VI.

TABLE 23. Observational Program for Sampling the Beetle Population of Table 22, According to the Method of Sub-Sampling

Random Sampling Unit Numbers

Major*	Minor†	Major	Minor	Major	Minor
2	4, 8	5	2, 5	3	4, 6
7	3, 4	7	4, 7	7	5, 8
9	3, 8	8	1, 6	11	2, 7
1	4, 6	5	1, 7	4	2, 3
3	4, 7	6	7, 8	5	1, 3
12	5, 6	9	2, 5	8	6, 8
1	3, 8	1	2, 5	4	4, 7
2	6, 8	6	1, 3	6	1, 8
8	1, 8	12	1, 4	9	3, 8
4	2, 8	4	4, 5	2	2, 6
7	2, 6	8	1, 4	5	5, 7
8	4, 8	9	1, 5	6	6, 8

*Row number within the block.

†Each minor random sampling unit number is the two ultimate units of the given designation in the row of a block.

TABLE 24. Sample Census of the Beetle Population of Table 22 According to the Program of Table 23

Observations	Sums	Observations	Sums	Observations	Sums
15, 7	22	16, 14	30	5, 7	12
25, 30	55	10, 20	30	11, 8	19
10, 12	22	16, 24	40	6, 7	13
	99		100		44
5, 4	9	11, 14	25	8, 17	25
4, 15	19	10, 16	26	11, 10	21
15, 4	19	7, 9	16	2, 4	6
	47		67		52
17, 7	24	6, 4	10	6, 5	11
2, 9	11	16, 9	25	6, 9	15
15, 19	34	9, 8	17	6, 16	22
	69		52		48
12, 12	24	10, 5	15	8, 14	22
6, 5	11	3, 3	6	4, 8	12
6, 1	7	5, 4	9	1, 2	3
	42		30		37

each block for sampling in turn. Let the minor random sampling unit be the two ultimate units of the same ordinal number in each half row of potatoes into which the row of 16 ultimate units has been divided. There are thus eight minor random sampling units, each of two ultimate units, to the potato row of a block. By drawing two of them, independently and at random, from each of the three potato rows which are to supply the samples of the blocks, the observations will comprise 3/12 x 2/8 of each block, or 1/16 of the entire population as required.

The observational program according to this scheme, worked out with the aid of a set of random sampling numbers, is given in Table 23, and the observations themselves are presented in Table 24. The grand sum is 687 or an average 9.542 among the 72 minor random sampling units observed, and an estimate of 16(687) or 10,992 beetles in the population. The sampling errors of these estimates are needed.

6.6 Analysis of Variance and the Sampling Error. It has been pointed out (Sec. 6.4) that the sampling variance appropriate to the method of sub-sampling may be derived from the arrangement of pertinent contributions thereto into an analysis of variance table. It has also been shown how the total sum of squares among minor random sampling units within the blocks is analyzed into its relevant portions, namely, the sum of squares among major random sampling units within the blocks, and the sum of squares among minor random sampling units within the major random sampling units. Symbolically

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_b)^2\right]=p\overset{B}{S}\overset{q}{S}\left[(y_q-y_b)^2\right]+\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_q)^2\right].$$

In the beetle sampling problem, the number of blocks, B, is 12; the number of major random sampling units of rows, q, is 3 in each block; and the number of minor random sampling units, p, is 2 in each sampled row.

The total sum of squares among minor random sampling units within blocks may be expressed as follows:

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_b)^2\right]=\overset{B}{S}\overset{q}{S}\overset{p}{S}(y^2)-qp\overset{B}{S}(y_b^2).$$

Squaring each of the 72 values of y, the grand sum is found to be 9,019, that is,

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}(y^2)=9{,}019.$$

The calculation of the correction to be applied to this, the numerical equivalent of

$$\overset{B}{qpS}(y_b^2)$$

may be simplified upon noting that

$$\overset{B}{qpS}(y_b^2) = \frac{1}{qp}\overset{B}{S}\left[(qpy_b)^2\right]$$

each value of qpy_b, being six times the block mean, is the block sum, each of which is given in Table 24. Upon performing the operation indicated, the numerical equivalents become

$$\frac{1}{2(3)}\left[(99)^2+(100)^2+\ldots.+(30)^2+(37)^2\right]=7{,}490.167$$

whence the total sum of squares among minor random sampling units within the blocks is

$$\overset{B}{S}\overset{q}{S}\overset{p}{S}\left[(y-y_b)^2\right]=9{,}019-7{,}490.167$$
$$=1{,}528.833$$

This is based upon a total of 60 degrees of freedom; that is, five degrees of freedom among the six minor random sampling unit observations in each of the 12 blocks. These values are entered in the bottom line of the analysis of variance Table 25.

Upon turning now to the sum of squares among the major random sampling units within the blocks, the expansion of its symbolic form shows that

$$p\overset{B}{S}\overset{q}{S}\left[(y_q-y_b)^2\right]=p\overset{B}{S}\overset{q}{S}(y_q^2)-\overset{B}{qpS}(y_b^2)$$

TABLE 25. Analysis of Variance of the Random Sampling Units of Table 24

Source of variation	Degrees of freedom	Sum of squares	Mean square
Among major rsu*, same block.....	24	985.333	$41.056=C$
Among minor rsu*, same major rsu.....	36	543.500	$15.097=D$
Total, among minor rsu*, same block...	60	1,528.833	

*Random sampling units.

giving the same correction factor as before, or 7,490.167; furthermore, the first term of the right-hand member may be simplified for numerical work, since

$$p \overset{B}{S} \overset{q}{S}(y_q^2) = \frac{1}{p} \overset{B}{S} \overset{q}{S}\left[(py_q)^2 \right]$$

the individual values py_q, being in the present case twice the row means, are the row sums as listed in Table 24. Numerically this is

$$\frac{1}{2}\left[(22)^2+(55)^2+(22)^2+\ldots.+(22)^2+(12)^2+(3)^2 \right] = 8{,}475.5.$$

Hence

$$p \overset{B}{S} \overset{q}{S}\left[(y_a - y_b)^2 \right] = 8{,}475.5 - 7{,}490.167$$
$$= 985.333.$$

This is based upon a total of 24 degrees of freedom; that is, two degrees of freedom among the three major random sampling unit values of rows in each of the 12 blocks. These are entered in the first line of Table 25.

The sum of squares among minor random sampling units of the same major random sampling unit—the middle line of Table 25—is calculated by subtraction.

The right-hand column contains the two pertinent mean squares. One may now calculate the sampling variance of the mean number of beetles to the minor random sampling unit—which in Sec. 6.5 was given as 9.542—directly from equation (7) of Sec. 6.4. The number of blocks, B, is 12; and in each block, $q=3$ and $Q=12$; while within each major random sampling unit, $p=2$ and $P=8$. Upon applying these numbers, as well as the pertinent mean squares to the equation,

$$V(\bar{y}) = \frac{1}{Bqp}\left[C\left(\frac{Q-q}{Q}\right) + D\left(\frac{P-p}{P}\right)\left(\frac{q}{Q}\right) \right],$$

the variance of the general mean is estimated to be the following:

$$V(9.542) = \frac{1}{72}\left[41.056\left(\frac{12-3}{12}\right) + 15.097\left(\frac{8-2}{8}\right)\left(\frac{3}{12}\right) \right]$$
$$= 0.467.$$

The square root of 0.467 is 0.683. Hence, on the minor random sampling unit basis, the mean number of beetles observed, together with its standard error, is

$$9.542 \pm 0.683$$

based upon 24 degrees of freedom.

6.7 Efficiency of the Method. In measuring the effectiveness of methods of estimation, an appropriate scale, proposed by R. A. Fisher (1937, Sec. 60) is provided by the reciprocal of the variance of the mean. Thus in an agricultural experiment it is convenient to consider a standard error of 10 percent of the mean as supplying one unit of information, and one giving 5 percent as supplying four units. Or, in general, if U is the number of units of information,

$$U = \frac{(\bar{y})^2}{100 \cdot V(\bar{y})}.$$

The number of units of information elicited in the beetle sampling is then

$$U = \frac{(9.542)^2}{100(0.467)} = 1.95.$$

An estimate of the efficiency of the method of sub-sampling, as used, might be based upon the comparison of these, the number of units actually elicited, to the number available had the beetles on the three major random sampling units of rows within each block been enumerated completely. Since the true variance among rows is σ^2_Δ of equation (6), Sec. 6.4, its value may be estimated upon inserting the pertinent numerical equivalents in

$$\frac{1}{p}\left[C - D\left(\frac{P-p}{P}\right)\right]\left[\frac{Q-1}{Q}\right] \rightarrow \sigma^2_\Delta.$$

This gives

$$13.628 \rightarrow \sigma^2_\Delta$$

whence the expected value of $V(\bar{y})$ as based upon the 36 rows is

$$\frac{13.628}{36} = 0.379$$

and the number units of information which should be expected to have been available under these conditions is

$$\frac{(9.542)^2}{100(0.379)} = 2.40.$$

Finally, then, the efficiency of the method used, in which only one-quarter of the row-lengths were observed with its resulting 1.95 units of information, is

$$\frac{1.95}{2.40} = 0.81$$

or 81 percent of what would have been expected had the sample rows been observed throughout their lengths. As this would have involved about four times as much field work, more information for the time expended might evidently be obtained by examining more rows rather than more complete examination of the rows which have supplied the samples.

CHAPTER VII

REPRESENTATIVE SAMPLING OF IRREGULAR BLOCKS

7.1 Proportional Sampling of Blocks of Known, but Diverse, Areas. Chapters IV, V and VI—with the exception of Sec. 4.4—treat of representative sampling within limited areas which have been divided, for purposes of assuring representativeness, into blocks of identical size and shape. The division into blocks with such ready nicety, however, is not practicable whenever the area of forest or range, which is to be sampled, is irregular in outline.

In the infrequent case when the areas of blocks, though diverse, are known in advance, sampling may be carried out without regard to equality, or proportionality, in number of random sampling units to the block. An illustration of disproportional sampling is that of Sec. 4.4.

It is usually preferable, however, that the number of random sampling units, drawn from each block, be proportional to block area. Representative sampling is then truly representative and as simple in conception as when blocks are identical in size and form. For if the total area of all k blocks, expressed in number of random sampling units, N, in the population, is expressed

$$N = N_1 + N_2 + \ldots . + N_k$$

where $N_1, N_2, \ldots, N_k$ represent the population number of random sampling units in the blocks individually; and if the representative set of samples is to make the proportion, p, of N, then a sample of n_1 random sampling units from block 1, n_2 from block 2, and so on, can be drawn such that

$$n_1 = pN_1; \quad n_2 = pN_2; \quad \ldots . n_k = pN_k.$$

Suppose the characteristic to be sampled is the timber volume on a forest property, the various compartments (blocks) of which are made up of known, but diverse, areas in number of random sampling units such as square chains. If y is then the volume on a random sampling unit, the total volume on the n_1 units of the sample from block 1 is the sum of the n_1 values of y, or

$$\overset{n_1}{S}(y).$$

The estimate of the variance of this sum, corrected at once for the finite population of the block is

$$V\left[\overset{n_1}{S}(y)\right] = \frac{n_1}{n_1-1}\overset{n_1}{S}\left[(y-y_b)^2\right]\left[\frac{N_1-n_1}{N_1}\right]$$

where y_b is the mean of the n_1 values of y.

The observed sum of y over all blocks is then p times the estimate of timber volume on the forest property. The estimate of the standard error of the observed sum is the square root of the sum of the variances of the block sums; there being k of these, the first one of which is written above. It is based upon $(n_1+n_2+\ldots.+n_k-k)$ degrees of freedom, and it is p times the estimate of the standard error of timber volume on the forest property.

There is a limitation to the feasibility of strictly proportional sampling of irregular blocks. For in order that each random sampling unit which is to enter the block sample be given exactly the same chance of draw, block maps, showing the location of all random sampling units on the ground, are prerequisite to the draw.

7.2 Proportional Sampling of Blocks of Diverse, but Unknown, Areas. As a consequence of the limitation just cited we shall consider a modification in which the *number* of random sampling units of the block samples are proportional to the existing *number* within the blocks, although in area the sampling may be more or less disproportional.

Usually it is quite feasible to divide the general area to be sampled, into blocks as diverse in area as might be, but with one side of constant length, as illustrated in Figure 18. The division is effected by a base line—real or imagined—across the general area, at equidistant points along which perpendicular lines, extending to the outside boundaries, divide the whole into two tiers of blocks of equal width.

It was required to design, from the original map of Figure 18, a sampling technique covering 5 percent—more or less—of the entire bounded universe. Its object was the estimation of the area of each of the four cover types, and of combinations of any two or three among them, so as to provide exact measures of the probable discrepancy between the true, but unknown, value and the sampling estimate thereof.

The scale of the original map of Figure 18, is 1 inch to 800 feet. In scale units, the base line through the length of the area is just 14 inches. Perpendicular lines at each 2-inch point and extending to the outside boundaries delimit 12 blocks as shown in the figure. It may be noted that a small portion of blocks 6 and 7 extend over the base line. Had these portions extended across the base line over the entire width of

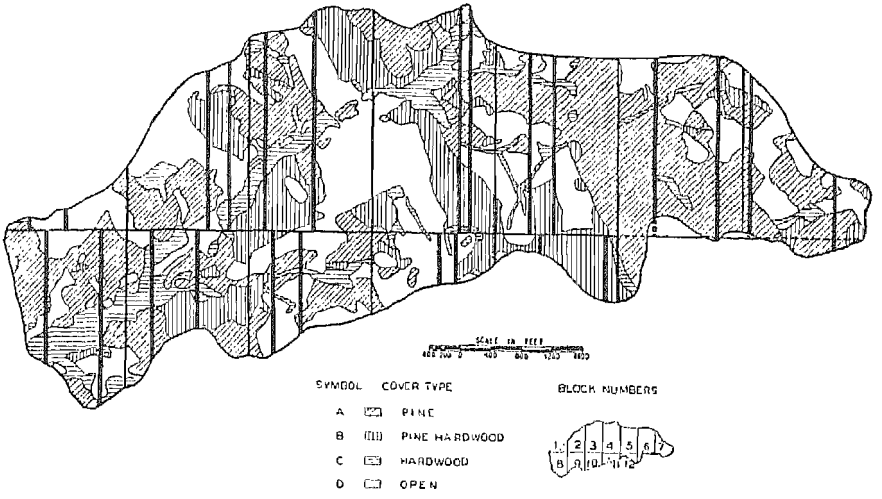


FIG. 18. A population of irregular boundaries subdivided into blocks of constant width, and showing the random sampling units of the block samples.

their respective blocks, they might have been considered as separate blocks, just as block 12 is separate from block 5.

Since each block is two inches wide, its area is conceived as the sum of areas of 40 contiguous strips, each $\frac{1}{20}$-inch wide, and extending the length of the block. If the random sampling unit is now defined as the strip, two of them supply an *estimate* of 5 percent of the block area from which they are drawn. It is not to be expected that they supply the exact proportion because variation in their lengths precludes the possibility that the preassigned proportion be free from sampling error. In other words, each block sample contains just 5 percent of the *number* of random sampling units of the block, but these are of different lengths.

This is an illustration of a class of sampling problems among the most common in forestry practice. Except in experimental work, it is not often that a forester or range examiner knows the precise area, one or more of whose characteristics of timber or range he is required to estimate. It may be the watershed of a small creek or large river; or it may be the area occupied by certain plant associations, such as timber type, the outlines of which have been but roughly sketched. A necessary condition however—and an obvious one—is that he recognizes the boundary of the area as he comes upon it.

Preliminary reconnaissance, even if cursory, should afford sufficient information as to the best position and direction of the base line from which the blocks emanate. The length of base line need not be an exact multiple of block width as in this illustration. If, for example, block width has been decided upon before the base line has been run, there will almost certainly be some remainder less than block width. The area traversed by this remaining part of the base line may then require a distinct sampling design and separate analysis, yet the final estimates of its characteristics and their variances can be combined with those of the main portion.

7.3 The Observations and the Estimate of the Population Mean. The requirements of randomization—that is, that the constituent parts of the observational program upon which estimates are to be based, be drawn independently and at random—are completely met by identifying the two strips, which are to supply the samples, out of the 40 of each block, by means of random sampling numbers.

The strips actually drawn are shown in Figure 18. The ultimate unit is taken as a square, $\frac{1}{20}$-inch to the side, and the observations are recorded in number of ultimate units according to cover type (A, B, C, and D) and all types (L) in each random sampling unit (strip) of the block sample, in Table 26.

TABLE 26. Direct Observations. Cover Type Areas According to Random Sampling Unit (Strip) and Block

Block	Strip 1					Strip 2				
	Number of 1/20-inch squares according to type									
	A_1	B_1	C_1	D_1	L_1	A_2	B_2	C_2	D_2	L_2
1.....	1			1	2				5	5
2.....	15	14	1	19	49	3	17	8	23	51
3.....	27	17	9	8	61	10	22	7	29	68
4.....	18	26	9	19	72	15	20	7	24	66
5.....	23	2	2	30	57	16	19	3	18	56
6.....	44		1	5	50	32	9		16	57
7.....	35	2		12	49	15			8	23
8.....	17	2	10	12	41	19	5	16	15	55
9.....	19		18	4	41	15	8	2	6	31
10.....	9	3	2	24	38	10		1	21	32
11.....				15	15		5		10	15
12.....		5			5		20			20
Sum.....	208	71	52	149	480	135	125	44	175	479

If the first and second random sampling units be denoted by subscripts 1 and 2, respectively, the numbers of ultimate units according to type, separately and combined, are the following:

$$\begin{aligned}
&\text{Cover type } A: && \overset{12}{S}\,(A_1+A_2)=208+135=343\\
&\text{Cover type } B: && \overset{12}{S}\,(B_1+B_2)=\ 71+125=196\\
&\text{Cover type } C: && \overset{12}{S}\,(C_1+C_2)=\ 52+\ 44=\ 96\\
&\text{Cover type } D: && \overset{12}{S}\,(D_1+D_2)=149+175=324\\
&\text{All cover types:} && \overset{12}{S}\,(L_1+L_2)=480+479=959
\end{aligned}$$

Upon dividing any one, or combination of two or more type sums, by 959, the estimate of the population mean on the ultimate unit basis is obtained. This is also the estimate of type area as a proportion of total area.

There is still required the estimate of the variances of such means.

7.4 The Weighted Mean of a Sample and the Estimate of Its Variance. In the present problem, each random sampling unit is based upon a different number of ultimate units. The random sampling units, accordingly, have different *weights*.

The meaning of weight is easily shown by a simple example. Given five values of equal reliability, say,

$$y_1,\ y_2,\ y_3,\ y_4,\ y_5;$$

their mean is

$$\bar{y}=\frac{1}{5}(y_1+y_2+y_3+y_4+y_5).$$

Suppose, now, that for some reason or other, these are recorded as only two separate observations, say, y_1 and y' where

$$y'=\frac{1}{4}(y_2+y_3+y_4+y_5).$$

Then y_1 and y' have different weights; y_1 having *unit weight*, and y' a weight of 4. The *weighted mean* of these

$$\bar{y}_w=\frac{y_1+4y'}{1+4}$$

is, obviously, the mean of the original five separate values.

If, then, y is the sum of the observations on the w ultimate units of a random sampling unit, the observed value on the *ultimate unit* basis is

$$y_w = \frac{y}{w}$$

and w is the weight of y_w. Let the block sample consist of n random sampling units of variable weight w. The weighted block mean is then

$$\bar{y}_w = \frac{\overset{n}{S}(wy_w)}{\overset{n}{S}(w)} = \frac{\overset{n}{S}(y)}{\overset{n}{S}(w)}.$$

This is the value of $\bar{y}_w$ which gives a minimum sum of *weighted squares of residuals;* that is, a minimum value to

$$\overset{n}{S}\left[w(y_w - \bar{y}_w)^2\right].$$

The first derivative of this expression with respect to $\bar{y}_w$ is

$$-2\overset{n}{S}\left[w(y_w - \bar{y}_w)\right].$$

Equating this to zero and dividing by 2, we have

$$\bar{y}_w \overset{n}{S}(w) = \overset{n}{S}(wy_w)$$

or

$$\bar{y}_w = \frac{\overset{n}{S}(wy_w)}{\overset{n}{S}(w)},$$

and this is the weighted mean as given above.

The sum of weighted squares of residuals, therefore, contains $V(y)$, the variance among observations of *unit weight*; and $V(\bar{y}_w)$, the variance of the weighted mean $\bar{y}_w$. The estimate of the first of these is the mean of the $(n-1)$ independent squares among the n, weighted squared residuals, or

$$V(y) = \frac{1}{n-1}\overset{n}{S}\left[w(y_w - \bar{y}_w)^2\right]$$

whence, the variance of the weighted mean is the variance of y of unit weight divided by the sum of the weights. Accordingly,

$$V(\bar{y}_w) = \frac{1}{\overset{n}{S}(w)}\left[V(y)\right]$$

is the estimate of the variance of the weighted mean. Or, if the variance of the sample sum is required, its estimate is

$$V\left[\overset{n}{S}(y)\right] = \overset{n}{S}(w)\left[V(y)\right]$$

that is, the product of the sum of the weights and the variance of unit weight.

For purposes of computation, the sum of weighted squares of residuals may be put in either of the two forms,

$$\overset{n}{S}\left[w(y_w-\bar{y}_w)^2\right] \quad \text{or} \quad \overset{n}{S}\left[\frac{1}{w}(y-w\bar{y}_w)^2\right]$$

which are identities.

7.5 Simplification of Computational Work with Samples of Two Random Sampling Units. In the special case when $n=2$, as in the observations of Table 26 for which the number of sample strips in each block is equal to 2, there is but one degree of freedom to the block; hence, the sum of weighted squares of residuals is also the variance of y of unit weight. Hence,

$$V(y)=\frac{1}{w_1}(y_1-w_1\bar{y}_w)^2+\frac{1}{w_2}(y_2-w_2\bar{y}_w)^2$$

in which the subscripts 1 and 2 refer to the first and second random sampling units, respectively. In this special case the numbers

$$(y_1-w_1\bar{y}_w) \text{ and } (y_2-w_2\bar{y}_w),$$

although differing in sign, are identical in *absolute* value; hence, their squares are identical, and therefore

$$V(y)=\frac{1}{w_1}(y_1-w_1\bar{y}_w)^2+\frac{1}{w_2}(y_2-w_2\bar{y}_w)^2=\left(\frac{1}{w_1}+\frac{1}{w_2}\right)(y_1-w_1\bar{y}_w)^2$$

as the estimate of the variance of y of unit weight. Then the estimate of the variance of the sample sum is the sum of the weights times the variance of unit weight, that is,

$$V(y_1+y_2)=(w_1+w_2)\left(\frac{1}{w_1}+\frac{1}{w_2}\right)(y_1-w_1\bar{y}_w)^2$$

$$=\left(2+\frac{w_2}{w_1}+\frac{w_1}{w_2}\right)(y_1-w_1\bar{y}_w)^2$$

By way of illustration consider the observations of block 7 of Table 26. The weights of the cover type observations in length of the two random sampling units differ considerably. For type A, in fact,

$$(A_1+A_2)=(35+15).$$

Here A_1 and A_2 correspond to y_1 and y_2 of the discussion above. In this same block, the total length of the random sampling units is

$$(L_1+L_2)=(49+23),$$

which corresponds to (w_1+w_2) of the preceding discussion. The weighted block mean, then

$$\bar{y}_w=\frac{A_1+A_2}{L_1+L_2}=\frac{35+15}{49+23}=0.694$$

to the ultimate unit. Hence the estimate of the variance of (35+15), that is,

$$V(A_1+A_2)=\left[2+\frac{L_2}{L_1}+\frac{L_1}{L_2}\right]\left[A_1-L_1\left(\frac{A_1+A_2}{L_1+L_2}\right)\right]^2$$

is numerically

$$\begin{aligned}V(50)&=\left(2+\frac{23}{49}+\frac{49}{23}\right)(35-34.0)^2\\&=4.60\end{aligned}$$

on one degree of freedom.

The operations to be performed upon the observations of Table 26, including the calculation of variances and covariances, may be stated most concisely by simplifying notation. For a given block, let

$$A=(A_1+A_2);\ B=(B_1+B_2);\ \text{etc.},$$

and let

$$a=A_1-L_1\left(\frac{A_1+A_2}{L_1+L_2}\right);\ b=B_1-L_1\left(\frac{B_1+B_2}{L_1+L_2}\right);\ \text{etc.},$$

where, of course, a, b, etc., may be either positive or negative. Finally, let

$$l=2+\frac{L_2}{L_1}+\frac{L_1}{L_2}$$

be the weighting factor. Then it follows that

$$V(A)=la^2;\ V(B)=lb^2;\ V(A+B)=la^2+lb^2+2lab;\ \text{etc.},$$

where lab is the covariance which may, of course, be positive or negative.

Table 27 is the work sheet upon which have been performed the calculations leading to the estimates of the variances of the observed sums of each cover type area and of combinations of areas of two or more

TABLE 27. Calculation of Variances and Covariances of Sample Sums*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Block	l	a	b	c	d	la	lb	lc	ld	la^2	lab	lac	lad	lb^2	lbc	lbd	lc^2	lcd	ld^2
1...	4.9	0.7	.....	.....	− 0.7	3.43	.....	.....	− 3.43	2.401	.....	.....	− 2.401	.....	.....	.....	.....	.....	2.401
2...	4.0	6.2	− 1.2	−3.4	− 1.6	24.80	− 4.80	−13.60	− 6.40	153.760	− 29.760	−84.320	− 39.680	5.760	16.320	7.680	46.240	21.760	10.240
3...	4.0	9.5	− 1.4	1.4	− 9.5	38.00	− 5.60	5.60	−38.00	361.000	− 53.200	53.200	−361.000	7.840	− 7.840	53.200	7.840	− 53.200	361.000
4...	4.0	0.7	2.0	0.7	− 3.4	2.80	8.00	2.80	−13.60	1.960	5.600	1.960	− 9.520	16.000	5.600	− 27.200	1.960	− 9.520	46.240
5...	4.0	3.3	− 8.6	−0.5	5.8	13.20	−34.40	− 2.00	23.20	43.560	−113.520	− 6.600	76.560	295.840	17.200	−199.520	1.000	− 11.600	134.560
6...	4.0	8.5	− 4.2	0.5	− 4.8	34.00	−16.80	2.00	−19.20	289.000	−142.800	17.000	−163.200	70.560	− 8.400	80.640	1.000	− 9.600	92.160
7...	4.6	1.0	0.6	....	− 1.6	4.60	2.76	.....	− 7.36	4.600	2.760	.....	− 7.360	1.656	.....	− 4.416	.....	.....	11.776
8...	4.1	1.6	− 1.0	−1.1	0.5	6.56	− 4.10	− 4.51	2.05	10.496	− 6.560	− 7.216	3.280	4.100	4.510	− 2.050	4.961	− 2.255	1.025
9...	4.1	−0.4	− 4.5	6.6	− 1.7	− 1.64	−18.45	27.06	− 6.97	0.656	7.380	−10.824	2.788	83.025	−121.770	31.365	178.596	− 46.002	11.849
10...	4.0	−1.4	1.4	0.4	− 0.4	− 5.60	5.60	1.60	− 1.60	7.840	− 7.840	− 2.240	2.240	7.840	2.240	− 2.240	0.640	− 0.640	0.640
11...	4.0	....	− 2.5	....	2.5	.....	−10.00	.....	10.00	.....	.....	.....	.....	25.000	.....	− 25.000	.....	.....	25.000
12...	6.3	....	.....	....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
Sum.		29.7	−19.4	4.6	−14.9	120.15	−77.79	18.95	−61.31	875.273	−337.940	−39.040	−498.293	517.621	− 92.140	− 87.541	242.237	−111.057	696.891
38/40 times the above										831.509	−321.043	−37.088	−473.378	491.740	− 87.533	− 83.164	230.125	−105.504	662.046
Multiplied by $\left(\frac{100}{959}\right)^2$										9.041	− 3.491	− 0.403	− 5.147	5.347	− 0.952	− 0.904	2.502	− 1.147	7.199

*As the numbers in columns 3 to 6 have been rounded off to a single decimal, it is no longer a necessary *numerical* condition that $a+b+c+d=0$. With these data, however, the condition has been preserved by altering the decimal of three numbers, namely, a in blocks 4 and 10, and b in block 9.

cover types. The numerical values of l, a, b, c, and d are listed in columns 2 to 6. It is worth noting that

$$a+b+c+d=0,$$

thus affording a check upon the calculation of these values. Columns 7 to 10 are merely computational steps, in which the values of a, b, c, and d are multiplied by the weighting factor, l, of the same line. A check on the arithmetic is again available, since

$$l(a+b+c+d)=0=la+lb+lc+ld.$$

Variances and covariances involving A, given in columns 11 to 14, are the products of a by la, lb, lc, and ld in turn. Again a check of the arithmetic is available; for

$$a(la+lb+lc+ld)=0=la^2+lab+lac+lad.$$

The operations indicated by the remaining columns are of the same kind, but using as multipliers, b, c, and d, in turn. No product already performed, and checked by means of the check sum, zero, need be repeated.

The totals in the third line from the bottom are the estimates of variances and covariances of the observed type sums, although not yet corrected for the finite population sampled. The corrected values are

$$\frac{N-n}{N}=\frac{40-2}{40}=\frac{38}{40}$$

times these totals. These are given in the second line from the bottom of the table.

It should be recalled, at this point, that although the recorded type areas have occurred on exactly two to each 40 of the random sampling units, or 5 percent, it is hardly to be expected that they represent exactly 5 percent of the total area; for the random sampling units used supply only an *estimate* of 5 percent of the total area. The sampling error of this estimate will be treated in Sec. 7.6.

On the other hand, cover type areas as percentages of total area are quite independent of the absolute magnitude of cover type areas. Since the percentage of any type area to the total area observed is 100 times the ratio of the former to the latter—that is, to 959—the variances and covariances of these percentages are the products of variances and covariances of the observed type areas to the square of $\frac{100}{959}$. They are recorded in the bottom line of Table 27, and they are assembled in handy form for inspection and use in Table 28.

By way of illustration, the observed proportion of cover type A is, from Table 26, $\frac{343}{959}$ or 35.8 percent of the total area; its variance, from Table 28, is 9.04. The observed proportion of cover types B and C combined is $\frac{292}{959}$ or 30.4 percent; its variance is

$$5.35+2.50+2(-0.95)=5.95.$$

The observed proportion of types $(A+B+C)$ is $\frac{635}{959}$ or 66.2 percent; the variance of this percentage is

$$9.04+5.35+2.50+2(-3.49)+2(-0.40)+2(-0.95)=7.21$$

which is also, except for the errors due to dropping decimals, the variance of the percentage in the remaining type D.

The square root of each of these variances is the standard error of the type percentage concerned.

TABLE 28. Variances and Covariances of Type-Area Percentages*

Type	A	B	C	D
A.....	9.04	−3.49	−0.40	−5.15
B.....	....	5.35	−0.95	−0.90
C.....	....	....	2.50	−1.15
D.....	....	....	....	7.20

*Numbers at intersections of columns and rows of like designation are variances; at intersections of unlike designation, they are covariances.

7.6 The Estimate of Total Area and Its Sampling Variance. The evaluation of sampling errors of cover type percentages is but part of the problem at hand. The estimate of type areas in absolute units is required, such as in ultimate units of $\frac{1}{20}$-inch squares. This is, evidently, the product of total area, in these same absolute units, and type proportion. Were the former known exactly, the standard error of area estimate of a given cover type would simply be the product of total area by the standard error of the type proportion. With the present data, however, the total area is itself subject to an error of estimate. As this enters into the calculation, we shall evaluate it at once.

Each block has supplied two random sampling units of block area, namely L_1 and L_2 of Table 26. Hence for a given block, the variance of the observed area is $V(L_1+L_2)=(L_1-L_2)^2$

on one degree of freedom. The variance of the observed area over all 12 blocks is then the sum of the variances of the individual blocks, and is based upon 12 degrees of freedom.

The calculations are performed in Table 29. Corrected for the finite population sampled, the estimate of the variance of the 959 ultimate units on all random sampling units is 1,312. The square root of this number is the estimate of the standard error of 959. Hence the estimate of the total area of the population is

$$20(959 \pm \sqrt{1{,}312}) = 19{,}180 \pm 724$$

in $\frac{1}{20}$-inch squares; and this may be converted to acres by multiplying by the conversion factor contained in the map scale.

TABLE 29. Calculation of the Variance of the Sum of 12 Sample Sums of Area

Block number	Observations L_1	Observations L_2	Variance of (L_1+L_2), i.e., $(L_1-L_2)^2$
1	2	5	9
2	49	51	4
3	61	68	49
4	72	66	36
5	57	56	1
6	50	57	49
7	49	23	676
8	41	55	196
9	41	31	100
10	38	32	36
11	15	15	0
12	5	20	225
Sum		959	1,381
38/40 of above			1,312

7.7 The Sampling Variance of Cover Type Areas. If M is the proportion that the area of a given cover type is of total area N, where N is in absolute units, such as the $\frac{1}{20}$-inch squares of our data, the area of this same type in these absolute units is MN. Furthermore, if M and N are independently subject to sampling errors, of variance $V(M)$ and $V(N)$ respectively, it is easily shown[15] that the variance of the product MN may be expressed

$$V(MN) = M^2\left[V(N)\right] + N^2\left[V(M)\right].$$

[15] The development is given in the Appendix, Sec. 7.7.

If this notation be referred to the present problem, the estimate of the variances of each cover type, or combination among them, is readily calculated in absolute units. In fact, the numerical equivalents of the second term of the above expression for the variance of the product, that is,

$$N^2\left[V(M)\right]$$

are given in the third line from the bottom of Table 27; while in the next line of the same table, they are adjusted for the finite population. These latter values would need no further adjustment if it were known that the random sampling units observed were not only 5 percent of the population number, but 5 percent of the population area as well. However, as the aggregate areas are but an estimate of 5 percent of the total area, the term

$$M^2\left[V(N)\right]$$

is to be added, after applying the factor $\frac{38}{40}$ on account of the limited population sampled. Thus for cover type A,

$$M=\frac{343}{959}=0.3577;\text{ and } V(N)=1{,}312,$$

whence

$$M^2\left[V(N)\right]=167.9.$$

This has been done according to individual cover types and certain combinations among them. The results are listed in Table 30. The numbers in the column headed $N^2\left[V(M)\right]$ would have been the estimates of the variances of the observed areas MN, had block areas been known exactly. Under the circumstances, however, the values $M^2\left[V(N)\right]$ are added thereto, the sum of the two terms being the estimate of the variance of MN.

The standard errors of the last column are based upon 12 degrees of freedom. By means of these, the observed cover type areas, MN, and the table of t, the usual probability statements may be made concerning

the range which encloses the population values. And these may be put on an acreage basis for the population by multiplying by a factor derived from the map scale. Thus each ultimate unit is the square of $\frac{1}{20}$-inch,

TABLE 30. Partial Summary of Observed Cover Type Areas, and Their Standard Errors. Units of 1/20-inch Squares

Type	Observed area MN	$N^2\left[V(M)\right]$	$M^2\left[V(N)\right]$	Variance of observed area $V(MN)$	Standard error of observed area $SE(MN)$
A	343	831.5	167.9	999.4	31.61
B	196	491.7	54.8	546.5	23.38
C	96	230.1	13.1	243.2	15.59
D	324	662.0	149.8	811.8	28.49
$A+B$	539	681.2	414.4	1,095.6	33.10
$A+C$	439	987.5	275.0	1,262.5	35.53
$B+C$	292	546.8	121.6	668.4	25.85

and as the map scale is 800 feet to the inch, an ultimate unit contains

$$\left(\frac{800}{20}\right)^2 \text{ square feet or } \frac{1}{43{,}560}\left(\frac{800}{20}\right)^2 \text{ acres.}$$

As 5 percent of the strips were taken, each ultimate unit of the samples represents

$$\frac{20}{43{,}560}\left(\frac{800}{20}\right)^2 = 0.7346 \text{ acres}$$

of the population.

When total area is not known precisely, some information is sacrificed. Calculation of the information lost may indicate at once whether it is trivial in quantity or whether steps should be taken to recover it, in whole or in part.

The amount of information provided by an estimate is proportional to the reciprocal of its variance; and the ratio of the amount extracted to the amount available under the condition to be tested is called the efficiency of the method of estimation under discussion. Consider cover type A as an illustration. Were the map area known exactly, the amount of information available for the estimate of A, under the sampling design used, would have been proportional to

$$\frac{1}{831.5}$$

where 831.5—taken from the third column of Table 30—is the estimate of the variance of A on the supposition that block areas are known. But on account of the sampling error to which the estimate of total area is subject, the amount of information actually obtained concerning the area of A is proportional to

$$\frac{1}{999.4}$$

where 999.4—taken from the fifth column of Table 30—includes the sampling variance of total area. The efficiency of this method of estimating A is thus

$$\frac{831.5}{999.4}$$

or 83 percent of what it would have been had the block areas been known precisely. The loss of information is 17 percent.

PART 3

INDIRECT ESTIMATES THROUGH REGRESSION

CHAPTER VIII

THE MEANING AND USE OF REGRESSION IN SAMPLING

8.1 The Problem of the Present Part. The sampling problems dealt with in Part 2 were based upon direct observations in particular populations; the estimate of one or more of whose parameters was needed. Thus in a timber cruise each of the random sampling units was regarded as supplying a measured quantity directly of timber volume; hence, timber volume was the only variable analyzed.

Direct measurement of timber volume, however, implies direct measurement not only of the diameters. but of the height of all trees which contribute volume to each random sampling unit observation as well. Now accurate measurement of tree height is time-consuming. Furthermore, if the measured volume of the trees is to be in board feet, considerable experience is required in order to recognize the limit of merchantability to which height is to be measured on the upper stem of individual trees. Consequently, the direct measurement of timber volume on random sampling units is a relatively expensive operation. In view of the variability of volume among random sampling units, the sample may appear too small to yield an estimate of the desired degree of accuracy.

It is known that the timber volume of a random sampling unit is proportional to the basal area—or sum of the squares of the diameters—of the trees thereon. Basal area is thus completely determined by the frequency distribution of tree diameter alone.

If, therefore, a portion of the random sampling units is made to supply both volume and concomitant basal area; and if this portion is used to determine the expression of volume in terms of basal area; and if, finally, a more accurate estimate of basal area is contained in the entire body of random sampling units; then the total information on volume is greater than the information on only that portion of the random sampling units upon which it is measured directly. The additional information on basal area is obtained with relatively little expense.

It is the purpose of the present part to show how such added information may be extracted from the samples.

8.2 The Regression Equation. Suppose that from each of n random sampling units of a block, direct measurements have been made on the variable y, say volume b.m., and on x, say basal area; and that it is required to express y in terms of x.

If variation in y is, in part at least, proportional to variation in x, this portion, denoted by Y, may be expressed in terms of x, as

$$Y=a+b(x-\bar{x}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where a and b are constants which may be determined from the sample of (y, x) by the method of least squares.

The coefficient, b, is the average rate of change of Y to unit change in x; and b is known, in the biological sciences, as the *regression coefficient.* The constant, a, is the value of Y when x is $\bar{x}$.

As y is expressed, partly at least, in term of x, y is called the *dependent variable,* for its calculated values depend upon given values of the *independent variable,* x.

The equation is called the *regression equation,* or the regression of y on x. The sample of (y, x) supplies the numerical equivalent of $\bar{x}$. The unknowns, a and b, may be calculated from the data by the method of least squares, if a and b are defined as numbers which will render a minimum sum of squares to the residuals, that is, a minimum value to

$$\overset{n}{S}\left[(y-Y)^2\right].$$

Upon substituting $a+b(x-\bar{x})$ for Y, this is equivalent to making

$$\overset{n}{S}\left\{\left[(y-a)-b(x-\bar{x})\right]^2\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

a minimum. Upon differentiating with respect to the unknown a and equating to zero, the expression becomes

$$2\overset{n}{S}\left\{\left[(y-a)-b(x-\bar{x})\right](-1)\right\}=0$$

whence

$$\overset{n}{S}(y-a)=b\overset{n}{S}(x-\bar{x}).$$

Now the sum of residuals of x around the mean of x is zero; hence, the right-hand member is zero. Accordingly,

$$\overset{n}{S}(y-a)=0$$

hence, as in Sec. 1.2,

$$a=\frac{1}{n}\overset{n}{S}(y)=\bar{y}.$$

Next, equation (2) is differentiated with respect to the unknown b, and equated to zero; that is,

$$2S^{n}\left\{\left[(y-a)-b(x-\bar{x})\right]\left[-(x-\bar{x})\right]\right\}=0$$

or

$$-S^{n}\left[(y-a)(x-\bar{x})\right]+bS^{n}\left[(x-\bar{x})^2\right]=0$$

and upon substituting $\bar{y}$ for a, and rearranging,

$$b=\frac{S^{n}\left[(y-\bar{y})(x-\bar{x})\right]}{S^{n}\left[(x-\bar{x})^2\right]}$$

If the denominator of this expression were divided by the $(n-1)$ degrees of freedom upon which it is based, it would be the estimate of the variance of x. In like manner, were the numerator divided by $(n-1)$, it would be the estimate of the covariance of x and y. The regression coefficient may thus be regarded as the ratio of the covariance of the two variables to the variance of the independent variable.

One may calculate at once the sum of squares of residuals upon expanding

$$S^{n}\left\{\left[(y-\bar{y})-b(x-\bar{x})\right]^2\right\},$$

the result of which is the following:

$$S^{n}\left[(y-\bar{y})^2\right]-2b\,S^{n}\left[(y-\bar{y})(x-\bar{x})\right]+b^2\,S^{n}\left[(x-\bar{x})^2\right].$$

This may be somewhat shortened since, from the definition of the regression coefficient b,

$$b^2\,S^{n}\left[(x-\bar{x})^2\right]=b\,S^{n}\left[(y-\bar{y})(x-\bar{x})\right],$$

so that the sum of squares of residuals may be expressed in any one of the following forms:

$$S^{n}\left[(y-\bar{y})^2\right]-b\,S^{n}\left[(y-\bar{y})(x-\bar{x})\right];\text{ or}$$

$$\overset{n}{S}\left[(y-\bar{y})^2\right]-b^2\overset{n}{S}\left[(x-\bar{x})^2\right]; \text{ or}$$

$$\overset{n}{S}\left[(y-\bar{y})^2\right]-\frac{\left\{\overset{n}{S}\left[(y-\bar{y})(x-\bar{x})\right]\right\}^2}{\overset{n}{S}\left[(x-\bar{x})^2\right]}$$

The left-hand term of each of these is, evidently, the total sum of squares of y around the mean of y. Each right-hand term is, consequently, that portion of the total sum of squares of y which is due to x.

These results may be put concisely in analysis of variance form as in Table 31. The degrees of freedom for the total sum of squares around the mean of y is, of course, $(n-1)$ among the n observations, as one is used in the estimate of $\bar{y}$. The residuals are then based upon $(n-2)$ degrees of freedom since the estimate of the regression coefficient b has also required a degree of freedom.

TABLE 31. Division of the Sum of Squares of y into Portions Due to, and Independent of, x, with Degrees of Freedom and Mean Squares

Due to	Sum of squares	Degrees of freedom	Mean square
Regression on x.	$b^2\overset{n}{S}\left[(x-\bar{x})^2\right]$	1	$\beta^2\overset{n}{S}\left[(x-\bar{x})^2\right]+\epsilon_b^2\overset{n}{S}\left[(x-\bar{x})^2\right]$
Residuals.....	$\overset{n}{S}\left\{\left[(y-\bar{y})-b(x-\bar{x})\right]^2\right\}$	$n-2$	$s^2_{y\cdot x}\begin{cases}\rightarrow n\sigma_d^2\\ \rightarrow \overset{n}{S}\left[(x-\bar{x})^2\right]\sigma_b^2\end{cases}$
Total.....	$\overset{n}{S}\left[(y-\bar{y})^2\right]$	$n-1$	s_y^2

The last column on the right contains the mean squares. As the test of significance of the regression coefficient, as well as the sampling variances of the statistics a and b, are based upon these mean squares, we need to consider them somewhat in detail. For this purpose the sample is considered as drawn at random from a population of *samples* of y which have the same values of the independent variable as represented by the sample of x at hand. The mean square due to regression may then be regarded as made up of two components, namely the *true* (but

unknown) portion of the sum of squares due to x, and the real (also unknown) error of its estimate. Numerically, these components cannot be separated with exactness, but symbolically the observed statistic, b, may be expressed

$$b=\beta+\epsilon_b$$

where β is the *true coefficient*, the unknown population parameter corresponding to the observed statistic, b; and ϵ_b is the *real* error of the estimate of β. This real error, ϵ_b, may, of course, be positive or negative. The sum of squares due to the regression of y on x may therefore be expressed

$$b^2 S\left[(x-\bar{x})^2\right]_n = (\beta+\epsilon_b)^2 S\left[(x-\bar{x})^2\right]_n$$

$$= \beta^2 S\left[(x-\bar{x})^2\right]_n + \epsilon_b^2 S\left[(x-\bar{x})^2\right]_n.$$

The cross-product term of the expanded form is not listed, since the population average of ϵ_b is zero. As the expression is based upon a single degree of freedom, it is recorded in the mean square column of Table 31.

The mean square independent of the regression, on the other hand, is the estimate of the variance of the residuals around the true equation; that is, it is an estimate of the variance of that part of the individual observations, y, which is independent of x. Symbolizing it as $s^2_{y\cdot x}$, it may be written

$$s^2_{y\cdot x}=\frac{1}{n-2} S\left[(y-Y)^2\right]_n.$$

The sampling variances of the statistics a and b of the regression equation

$$Y=a+b(x-\bar{x})$$

can be computed from $s^2_{y\cdot x}$. In the first place, the mean square of the residuals contains an estimate of the variance of a, that is

$$s^2_{y\cdot x} \rightarrow n\,\sigma_a^2$$

whence the sampling variance of a is represented by the expression

$$V(a)=\frac{s^2_{y\cdot x}}{n}.$$

It should be noted that although

$$a=\bar{y}$$

its sampling variance is not that of the mean of y when no information is at hand regarding x. For in the latter case, as in Sec. 1.6, the sampling variance of the mean of y may be expressed,

$$V(\bar{y}) = \frac{s_y^2}{n};$$

but the sampling variance of a is, rather, that of Y when x is $\bar{x}$.

In the second place, the mean square of the residuals is an estimate of

$$\epsilon_b^2 \, S\left[(x-\bar{x})^2\right]^n$$

which contains the real error, ϵ_b, of the calculated regression coefficient, b. Now the expected value of ϵ_b is zero; and the expected value of its square is the variance of b. Accordingly,

$$s_{y \cdot x}^2 \rightarrow \sigma_b^2 \, S\left[(x-\bar{x})^2\right]^n$$

where σ_b^2 denotes the population variance of b, having the same weight in sum of squares of x. The estimate of this variance—that is the sampling variance of b—is then the following:

$$V(b) = \frac{s_{y \cdot x}^2}{S\left[(x-\bar{x})^2\right]^n}$$

These developments will next be applied to a numerical example.

8.3 A Numerical Example. Volume in M feet b.m. and basal area in square feet of each of six half-acre random sampling units of a 40-acre tract of upland hardwood are listed in the first two columns of Table 32. The problem is the calculation of the regression of volume

TABLE 32. Calculation of Sums, and Sums of Squares and Products, of Volume and Basal Area among Six Random Sampling Units of Upland Hardwood

Basal area in square feet x	Volume in M feet b.m. y	x^2	xy	y^2
11	1.22	121	13.42	1.4884
14	1.43	196	20.02	2.0449
5	0.67	25	3.35	0.4489
11	1.28	121	14.08	1.6384
15	1.74	225	26.10	3.0276
18	1.62	324	29.16	2.6244
74	7.96	1012	106.13	11.2726

(y) on basal area (x) of these data, and the standard errors of volume involved in the regression equation.

For the sake of simplicity, adjustments of estimates to the finite population sampled will be neglected for the present. They will be taken up separately in a later section (Sec. 8.5).

The sum of products of deviations about the means of y and x, needed in the calculation of b, may be expressed, for purposes of numerical calculation, in one of the following ways:

$$\overset{n}{S}\left[(y-\bar{y})(x-\bar{x})\right] = \overset{n}{S}(yx)-\bar{y}\overset{n}{S}(x)$$

$$= \overset{n}{S}(yx)-\bar{x}\overset{n}{S}(y)$$

$$= \overset{n}{S}(yx)-\frac{1}{n}\left[\overset{n}{S}(y)\right]\left[\overset{n}{S}(x)\right]$$

Upon inserting the appropriate numbers from Table 32, this becomes

$$\overset{6}{S}\left[(y-\bar{y})(x-\bar{x})\right] = 106.13-\frac{1}{6}(74)(7.96)$$

$$= 7.9567$$

Similarly, since

$$\overset{n}{S}\left[(x-\bar{x})^2\right] = \overset{n}{S}(x^2)-\frac{1}{n}\left[\overset{n}{S}(x)\right]^2,$$

the sum of squares of basal area is

$$\overset{6}{S}\left[(x-\bar{x})^2\right] = 1{,}012-\frac{1}{6}(74)^2$$

$$= 99.333.$$

The regression coefficient, therefore, is

$$b=\frac{\overset{n}{S}\left[(y-\bar{y})(x-\bar{x})\right]}{\overset{n}{S}\left[(x-\bar{x})^2\right]}=\frac{7.9567}{99.333}=0.0801.$$

This means that there is an average increase of 0.0801 M feet b.m. to the square foot increase in basal area.

From Table 32, the numerical equivalents of $\bar{y}$ and $\bar{x}$ are as follows:

$$\bar{y}=\frac{1}{6}(7.96)=1.327$$

$$\bar{x}=\frac{1}{6}(74) \quad =12.333.$$

Upon inserting these means in the regression equation

$$Y=\bar{y}+b(x-\bar{x})$$

we have

$$Y=1.327+0.0801(x-12.333);$$

and this is presented graphically in Figure 19, together with the observed points upon which it is based.

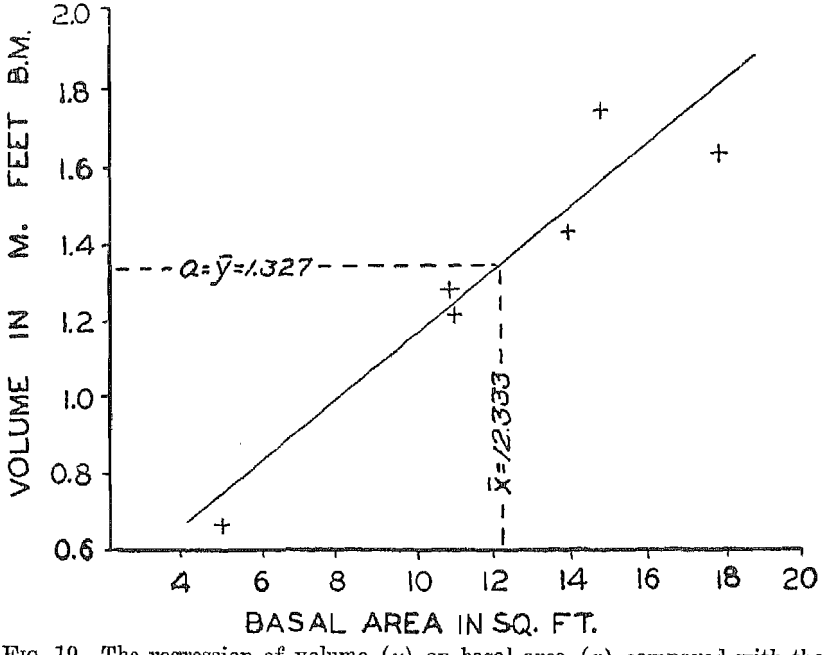


FIG. 19. The regression of volume (y) on basal area (x) compared with the direct observations of six half-acre random sampling units upon which it is based.

The sum of squares of the residuals,

$$S^{n}\left[(y-Y)^2\right]=S^{n}\left[(y-\bar{y})^2\right]-b\,S^{n}\left[(y-\bar{y})(x-\bar{x})\right],$$

is calculated by deducting from the total sum of squares, that is, from

$$\overset{n}{S}\left[(y-\bar{y})^2\right] = 11.2726 - \frac{1}{6}(7.96)^2$$
$$= 0.7123,$$

the portion due to the regression on x, which with the present data is

$$b\,\overset{n}{S}\left[(y-\bar{y})(x-\bar{x})\right] = 0.0801(7.9567)$$
$$= 0.6373.$$

These quantities are tabulated in the analysis of variance of Table 33, together with their mean squares and degrees of freedom upon which they rest. The mean square of the total is the estimate of the variance of volume without any regard whatever to basal area. The mean square of the residuals, on the other hand, is the estimate of the variance of that part of the sample plot volumes which is quite independent of basal area. It is thus the variance of volume to be expected of random sampling units which have identical basal area.

The variance of the regression coefficient, 0.0801, expressed as in the preceding section

$$V(b) = \frac{s^2_{y\cdot x}}{\overset{n}{S}\left[(x-\bar{x})^2\right]}$$

is, numerically,

$$V(b) = V(0.0801) = \frac{0.01875}{99.333} = 0.000189$$

on the four degrees of freedom used in the calculation of $s^2_{y\cdot x}$; the estimate of the standard error of b being the square root of 0.000189, or 0.01375 M feet b.m.

TABLE 33. Analysis of Variance of the Volume Data of Table 32

Due to	Degrees of freedom	Sum of squares	Mean square
Regression.....	1	0.6373	0.6373
Residuals.....	4	0.0750	0.01875
Total.....	5	0.7123	0.14246

8.4 Application of the Distribution of t to the Regression Coefficient. In preceding chapters use was made of the distribution of

the statistic t (Table 7) for the purpose of delimiting the range which, on a given probability, encloses the true mean of the sampled population. These methods apply equally to regression coefficients, in which case

$$t=(b-\beta)\Bigg/\sqrt{\frac{s^2_{y\cdot x}}{S\left[(x-\bar{x})^2\right]}}$$

where the denominator, as given in Sec. 8.2, denotes the estimate of the standard error of b. Although the parameter β is unknown, yet a range may be calculated, corresponding to any chosen probability, such that it contains—or does not—the true coefficient. If, for example, the probability be fixed at 0.05 that

$$\left|b-\beta\right|>t\sqrt{\frac{s^2_{y\cdot x}}{S\left[(x-\bar{x})^2\right]}}$$

one takes from the table of t the value corresponding to this probability and the number of degrees of freedom upon which the standard error of b is based. With the four degrees of freedom of the present data, $t=2.776$. Hence with probability of 0.95

$$\begin{aligned}\beta&=0.0801\pm(2.776)(0.01375)\\&=0.0801\pm0.03817.\end{aligned}$$

Previous experience with regressions of volume and basal area has established beyond question that the parameter β must be a positive number; but in many other problems involving regressions there is no advance knowledge of the magnitude of β. Under these circumstances it is customary to test the hypothesis that $\beta=0$ in the sampled population. By way of illustration the test will be performed on the volume-basal area data.

The analysis of variance of Table 33 contains the materials for the test. The ratio of the mean square due to regression, to the mean square of the residuals is, in fact,

$$t^2=\frac{b^2\,S\left[(x-\bar{x})^2\right]}{s^2_{y\cdot x}}$$

for, as given in the preceding section, the estimate of the sampling variance of b is

$$V(b)=\frac{s^2_{y\cdot x}}{S\left[(x-\bar{x})^2\right]}.$$

The observed value of t is thus

$$t = \sqrt{\frac{0.6373}{0.01875}} = 5.83.$$

Upon reference to the table of t, it is ascertained that with four degrees of freedom t is expected to exceed 4.604 only once in 100 such samples from the same population. Consequently, the hypothesis that $\beta = 0$ is untenable in the face of the observed t of 5.83.

8.5 The Variance of Y. Given the numerical equivalent of x, the corresponding value of Y of the regression equation

$$Y = a + b(x - \bar{x})$$

may be calculated readily. Often however, and particularly in sampling work, the estimate of the variance of the calculated Y is needed as well. Since Y is the sum of two independent quantities, a and $b(x-\bar{x})$, the variance of Y is the sum of the variances of these two quantities. Hence

$$V(Y) = V(a) + V\Big[b(x-\bar{x})\Big].$$

The second term of the right-hand member of the expression, $V(Y)$, is the variance of the product of the two factors, b and $(x-\bar{x})$. As both factors may be subject to sampling errors, the variance of their product is analogous to that of Sec. 7.7. Accordingly,

$$V\Big[b(x-\bar{x})\Big] = (x-\bar{x})^2\Big[V(b)\Big] + b^2\Big[V(x-\bar{x})\Big],$$

but as $\bar{x}$ is not variable in this equation, $V(x-\bar{x})$ is $V(x)$.

In general, then, if

$$Y = a + b(x - \bar{x})$$

the variance of Y may be expressed

$$V(Y) = V(a) + (x-\bar{x})^2\Big[V(b)\Big] + b^2\Big[V(x)\Big].$$

This expression brings out that the accuracy of an estimate, Y, depends, in the first place, upon the amount of information contained in the regression equation. If this is considerable, the variances of the equation constants, a and b, are relatively small and the estimates of Y are correspondingly precise. Now the amount of information on a statistic varies inversely with its variance; and since (Sec. 8.2)

$$V(a) = \frac{s^2_{y \cdot x}}{n}; \text{ and } V(b) = \frac{s^2_{y \cdot x}}{S\left[(x - \bar{x})^2\right]}$$

the amount of information on these constants is increased if the individual observations of y fit closely around the regression line, that is, if $s^2_{y \cdot x}$ is small; and also as n, the number of observations of (y, x) used in calculating the regression equation, is increased. Furthermore, the information on the regression coefficient, b, becomes greater as the *range* of the independent variable, x, is extended; for the magnitude of the sum of squares of n residuals, $(x - \bar{x})$ depends considerably upon the range encountered, as we shall see in Chapter IX.

In the second place, the accuracy of an estimate, Y, depends upon the particular value of x for which it is an estimate, and, in turn, upon the amount of information on this value of x. If one considers that x is given exactly, its variance—that is, $V(x - \bar{x})$—is zero, and the variance of Y, then, becomes least as x approaches the mean of x. But if x is itself an estimate, hence subject to sampling error, the estimate of Y is, of course, made at some additional sacrifice of precision.

The distinction between finite and hypothetically infinite populations is seldom recognized in estimating the variances of the regression constants a and b. In practice, the regression equation is rarely based upon an appreciable proportion of the population concerned. If needed, however, the usual factor

$$\frac{N - n}{N}$$

may be applied to the variance of both a and b.

8.6 The Variance of Y when x is Free of Error. Suppose the estimate is required of the average volume in M feet b.m. (Y) to the half-acre, of that part of the population of y for which basal area (x) is exactly 16 square feet. Upon putting this basal area for x into the regression equation of Sec. 8.3, which is,

$$Y = 1.327 + 0.0801(x - 12.333),$$

the equivalent volume then becomes

$$\begin{aligned} Y &= 1.327 + 0.0801(16 - 12.333) \\ &= 1.621 \text{ M feet b.m.} \end{aligned}$$

The variance of this estimate may be calculated by means of the general expression for $V(Y)$, for which

$$V(Y) = V(a) + (x - \bar{x})^2\left[V(b)\right] + b^2\left[V(x)\right].$$

However, $V(16-12.333)$ is zero, since x is 16 exactly; whence, from the preceding section

$$V(a)=\frac{s^2_{y\cdot x}}{n}=\frac{0.01875}{6}$$
$$=0.003125,$$

and

$$V(b)=\frac{s^2_{y\cdot x}}{n\left[S(x-\bar{x})^2\right]}=\frac{0.01875}{99.333}$$
$$=0.000189.$$

It follows, then, that

$$V(1.621)=0.003125+(16-12.333)^2(0.000189)$$
$$=0.005666$$

the square root of which, or 0.0753, is the estimate of the standard error. If preferred, one may compute a range such that

$$\left|Y-\mu\right|>ts_Y$$

with probability of, say, 0.05. In this, μ is, of course, the population mean of volume in M feet b.m. when basal area is exactly 16 square feet. With the four degrees of freedom upon which $s_Y=0.0753$ is based, $t=2.776$. Hence with probability of 0.95

$$\mu=1.621\pm(2.776)(0.0753)$$
$$=1.621\pm0.2090 \text{ M feet b.m.}$$

when basal area is exactly 16 square feet.

If corresponding limits, calculated in the same way, for other values of x are plotted on coordinate paper, as in Figure 20, the graph exhibits a band, covering the regression symmetrically, within which lies the true volume for given values of x according to the probability upon which it is constructed. This band, known as the *confidence band,* is relatively narrow for $x=\bar{x}$, and widens as the extremities of the range of x are approached, indicating that the accuracy of the estimate of Y is lessened as x diverges from its mean.

8.7 The Variance of Y when x is Subject to Sampling Error. Suppose, now, that by means of the regression equation it is required to estimate Y corresponding to the best estimate of x. This latter is, obviously,

$$\bar{x}=12.333$$

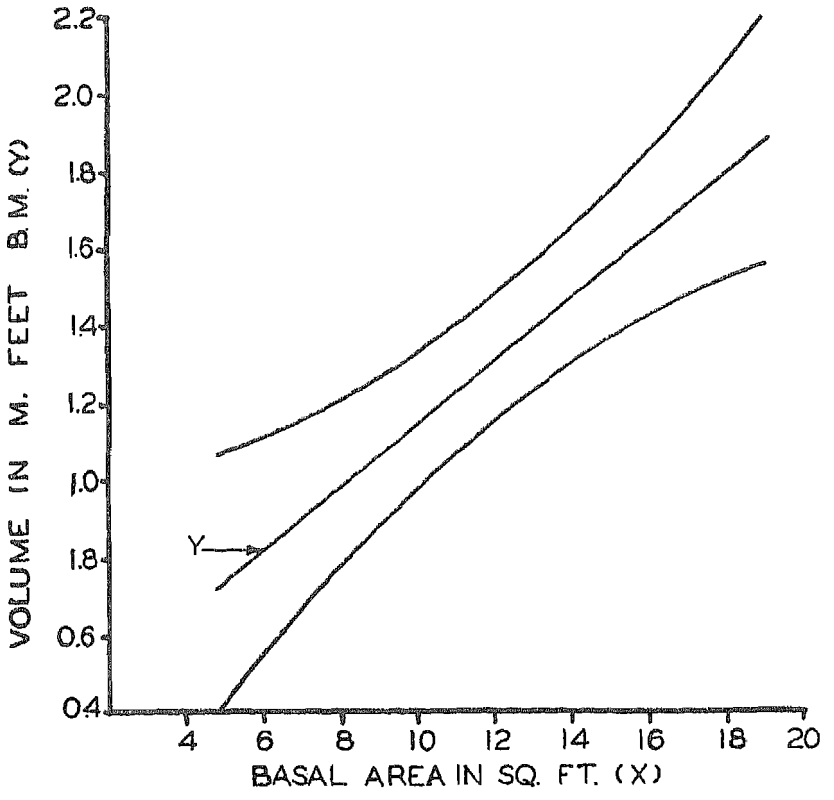


FIG. 20. The curve of volume, Y, on basal area, x, and the 95 percent confidence band.

and its sampling variance is contained in the sum of squares used in Sec. 8.3. Accordingly, upon applying the correction factor for the finite population of x,

$$V(\bar{x}) = \frac{6\left[S(x-\bar{x})^2\right]}{6(5)}\left(\frac{80-6}{80}\right) = \frac{99.333}{30}\left(\frac{74}{80}\right) = 3.063.$$

Upon putting $x = 12.333$ into the regression equation

$$Y = 1.327 + 0.0801(x - 12.333),$$

the calculated volume obviously becomes

$$Y = 1.327 \text{ M feet b.m.}$$

The variance of this estimate may be calculated from the general expression

$$V(Y) = V(a) + (x - \bar{x})^2 \left[V(b) \right] + b^2 \left[V(x) \right]$$

for which, however, the second term is zero, as $x = \bar{x}$. Then

$$\begin{aligned} V(1.327) &= \frac{0.01875}{6} + (0.0801)^2(3.063) \\ &= 0.003125 + 0.019652 \\ &= 0.022777. \end{aligned}$$

Now 1.327 M feet b.m. is, of course, the mean of y. It would seem that the common-sense approach to the estimate of its variance is to disregard x entirely. Under this condition the variance of $\bar{y}$ may be estimated directly from the formula

$$V(\bar{y}) = \frac{s_y^2}{n}\left(\frac{N-n}{N}\right) = \frac{S\left[(y-\bar{y})^2\right]}{n(n-1)}\left(\frac{N-n}{N}\right),$$

and upon taking the sum of squares from the bottom line of Table 33, the numerical equivalent reduces to

$$V(1.327) = \frac{0.7123}{6(5)}\left(\frac{74}{80}\right) = 0.021963.$$

This estimate is, in fact, somewhat better than that already derived strictly from the regression equation. The difference in its favor, that is,

$$0.022777 - 0.021963 = 0.000814,$$

is due to the fact that it does not at all involve the regression coefficient, b, which in itself is subject to sampling error.

The instructive feature of the application just illustrated is that the regression equation can add no new information concerning the population mean of y when no more is known about the independent variable, x, than is contained in the random sample of (y, x) upon which the regression is based. Under these circumstances there may be no point in calculating the regression equation at all.

Fullest use can be made of regression when the sampling work is so planned as to supply more information on the independent variable, x, than is contained in the sample upon which the regression is based.

It was remarked in Sec. 8.3 that the volume and basal area data of the regression were six half-acre random sampling units from a 40-acre tract of upland hardwood. As it happened, however, basal area was measured on 20 half-acres drawn by a random sampling device from the

80 half-acres of the tract; but the volume data were taken from just six out of the 20; these six values of volume and concomitant basal area supplying the information on the regression.

The problem now is the estimate of volume on the 40-acre tract from all the information at hand. According to the regression equation, mean volume in M feet b.m. to the half-acre is

$$Y = 1.327 + 0.0801(x - 12.333)$$

where x is basal area in square feet. Now the mean basal area among the 20 random sampling units is 13.853 square feet, and the variance of this mean, not corrected for the finite population sampled, is 0.550. Upon applying the correction,

$$V(13.853) = 0.550\left(\frac{80-20}{80}\right) = 0.4125.$$

Upon putting the best estimate of the population mean of basal area, 13.853, for x in the equation,

$$\begin{aligned} Y &= 1.327 + 0.0801(13.853 - 12.333) \\ &= 1.327 + 0.122 \\ &= 1.449 \text{ M feet b.m.} \end{aligned}$$

The variance of this estimate may be calculated directly from the variance of Y, that is,

$$V(Y) = V(a) + (x - \bar{x})^2\Big[\,V(b)\,\Big] + b^2\Big[\,V(x)\,\Big].$$

As given above, the necessary contributing variances, corrected to the finite populations, are the following:

$$\begin{aligned} V(x) &= V(13.853) = 0.4125 \\ V(a) &= V(1.327) \;\; = 0.003125 \\ V(b) &= V(0.0801) = 0.000189 \end{aligned}$$

and

$$(x - \bar{x}) = 1.520.$$

Finally, then,

$$\begin{aligned} V(1.449) &= 0.003125 + (1.520)^2(0.000189) + (0.0801)^2(0.4125) \\ &= 0.003125 + 0.000437 + 0.002647 \\ &= 0.006209 \end{aligned}$$

on four degrees of freedom. The square root of this is 0.0788 in M feet b.m. Hence mean volume with its standard error is

$$1.449 \pm 0.0788 \text{ M feet b.m.}$$

to the half-acre. By using the additional information on basal area the estimated mean volume has been changed from 1.327 to 1.449 M feet b.m.,

while its standard error has been reduced from 0.1482—that is, from $\sqrt{0.021963}$—to 0.0788 M feet b.m.

Since t on four degrees of freedom is 2.776 at the 5 percent level, then with probability of 0.95, the mean volume is

$$1.449 \pm 0.219 \text{ M feet b.m.}$$

to the half-acre. The estimate of the entire volume on the 40-acre tract is therefore 80 times this quantity, or

$$115.9 \pm 17.5 \text{ M feet b.m.}$$

8.8 The Utility of Regression in Sampling. It has been shown how the method of regression may be of service in estimating a population mean. It does not follow, however, that it always adds to the accuracy of estimate. There is no gain in the estimate of mean y through regression if it costs no more to measure y directly than it costs to measure the independent variable, x; or if the variance of mean y for constant x is not substantially less than the variance of mean y when x does not enter into its estimate.

The method of regression is of greatest utility, therefore, when the character y, the mean or aggregate of which is needed to be estimated, is difficult or expensive to measure directly, and where it is known that y is correlated with a second character, x, which, in turn, is relatively inexpensive to measure.

These conditions hold not infrequently in problems of sampling a forest or range. Foresters and range ecologists still need to rely largely upon eye-estimates of timber or forage crop, preliminary to administrative decisions pertaining thereto. In such cases the method of regression may be particularly valuable for the purpose of adjusting an eye-estimate. If on relatively few random sampling units the eye-estimate, x, is taken independently of the measured, y, the latter may be expressed in terms of the former by the regression equation

$$Y = a + b(x - \bar{x}).$$

Then if the mean of x is established by ocular estimate on the remaining random sampling units, the precision of the estimate, Y, may, indeed, be purchased cheaply.

Furthermore, the method of regression may often be usefully combined with the method of representative sampling, with an efficiency which exceeds the contribution of either method alone. This feature will be discussed in Chapter XI.

CHAPTER IX

PURPOSIVE SELECTION IN SAMPLING

9.1 Exemption of the Independent Variable from the Restriction of Randomization. It was shown in the preceding chapter that if one is required to estimate the population mean, or aggregate, of a variable, say y, which is difficult—and consequently expensive—to measure directly, it may be more expedient to confine the double sampling to a comparatively small sample of (y, x) and to gather the great bulk of observations on an easily measurable variable, say x; thence to use the mean of x indirectly to estimate the corresponding mean of y, provided that y can be expressed reliably in terms of x. The method implies the regression of y on x and the statistics pertaining thereto.

The illustrations used involved the regression of volume b.m. (y) on basal area (x) according to six half-acre sample plots, drawn independently and at random from the population of 80 half-acres.

Now it is a necessary condition that the estimate of the population mean of x—with which the regression equation is entered—be based upon one or more random samples of x. But it is not at all necessary that the basic data *upon which the regression is built* be drawn strictly at random from the population of sampling units. In fact, certain advantages—both theoretical and practical—may often be gained by purposive selection of these particular sampling units.

What is involved may, perhaps, be best brought out by considering afresh what is sought in sampling work, from the regression equation.

When y is to be expressed in terms of x, it is required (1) that the calculated value, Y, be the best estimate obtainable, from the sample of (y, x), of the *mean of y for a given value of x*; and (2) that the mean square of the residuals—which has been symbolized $s^2_{y \cdot x}$—be the best estimate obtainable from the sample of (y, x) of the variance of y when x is some given value.

It is to be noted that an estimate of the general mean of either of the associated populations is not required from the sample of (y, x). It follows, then, that one may choose at will the values of the independent variable x, to which the sample of (y, x) is to be confined, provided the observations taken on the associated y supply a random sample within each value of x selected. This provision is extremely important. The regression value, Y, corresponding to any x, cannot be expected to be an

unbiased estimate of the population mean of y for this class of x, unless the sample of y according to each class of x selected, has been drawn independently and at random.

The practical consequences of this limitation is the common preference for mechanical selection whereby the regression sample is drawn according to some prearranged, usually geometrical, pattern which has the virtue of assuring representativeness.

9.2 Effect on Pertinent Statistics. If each observation on the dependent variable, y, has the same precision; and if it is known that y varies directly with x, then the sampling variance of Y, that is, of the estimate of y corresponding to a particular x in the regression equation

$$Y = a + b(x - \bar{x})$$

is, as in Sec. 8.5,

$$V(Y) = V(a) + (x - \bar{x})^2 \left[V(b) \right] + b^2 \left[V(x) \right].$$

The sampling variance of the equation constant, a, and of the regression coefficient, b, in this expression are contained in the sample of (y, x) upon which the regression equation is based. As given in Sec. 8.2,

$$V(a) = \frac{s^2_{y \cdot x}}{n}; \qquad V(b) = \frac{s^2_{y \cdot x}}{n \left[S(x - \bar{x})^2 \right]}.$$

It was brought out in the preceding section that the purposive choice of x does not affect the mean square of the residuals, $s^2_{y \cdot x}$. Consequently, it does not affect the sampling variance of the constant, a; for this varies inversely only with sample size, n.

On the other hand, the sampling variance of the regression coefficient, b, decreases with increase in the range of x as well as with increase in the size of sample; for upon writing

$$V(b) = \frac{s^2_{y \cdot x}}{n \left[S(x - \bar{x})^2 \right]} = \frac{s^2_{y \cdot x}}{(n-1)s^2_x},$$

it is apparent, from the expression on the right, that the sampling variance of b varies inversely with the mean square, s^2_x, of the sample of x. The latter, in turn, increases approximately as the square of the range of x.

9.3 Experimental Verification. A two-variate population is chosen from which observations on y and x may be drawn by means of random sampling numbers. Let

$$x = 10(\text{sum of five random digits});$$
$$y = x + (\text{sum of five random digits}).$$

As the digits are

$$0, 1, 2, 3, 4, 5, 6, 7, 8, \text{ and } 9,$$

one may observe sums of five random ones at will, for example, in Tippett's Random Sampling Numbers. The population mean of the sum of five random digits (Sec. 1.3) is 22.5, and the population variance is 41.25; while the range is from 0 to 45.

These definitions involve the following conditions:

(1) The independent variable, x, which is limited between 0 and 450, occurs only according to the discrete values

$$0, 10, 20, \ldots .440, \text{ and } 450.$$

The population mean of x is 225, and the population variance of x is 4125.

(2) The population mean of y in any given class of x—which is symbolized, $\mu_{y \cdot x}$—may be expressed exactly in terms of x; that is,

$$\mu_{y \cdot x} = 22.5 + x.$$

Thus the population value of the regression coefficient is unity.

(3) The true variance of y in each class of x is 41.25, and this is, of course, the population value of the mean square of the residuals about the regression.

The object, then, is to compare the observed regression coefficients with their true value, unity, and to note how the dispersion of individual coefficients is affected by the purposive choice of x, within which random observations on y are confined. To do this, two random observations are drawn from each of two arrays of x; thus each regression coefficient is based upon four observations.

Choosing for the first comparison

$$x = 130, \quad \text{and} \quad x = 340$$

four sums of five digits are read from Tippett's Random Sampling Numbers as follows:

$$33, 25, 18, 29.$$

Thus the two observed values of y when $x = 130$, are

$$130 + 33 = 163, \quad \text{and} \quad 130 + 25 = 155,$$

the average of which is 159.0.

The corresponding observations when $x = 340$, are

$$340 + 18 = 358. \quad \text{and} \quad 340 + 29 = 369,$$

with an average of 363.5. As the regression line must pass through the mean y of these two x-arrays there are two observation equations,

$$a + b(130) = 159.0$$
$$a + b(340) = 363.5$$

whence

$$b = \frac{363.5 - 159.0}{340 - 130} = 0.974.$$

Altogether, ten regression coefficients were calculated in the same way for these two values of x, and they are shown graphically in Figure 21A.

In the second comparison, two observations were taken on y when $x = 170$; and two when $x = 300$. Repeated in ten independent samplings, the regression coefficients are presented in Figure 21B.

In the third comparison, two observations were taken on y when $x = 210$, and when $x = 260$. Ten regression coefficients resulting from as many samplings are presented in Figure 21C.

Finally, ten samples of four were drawn, each sample consisting of one observation of y corresponding to each of four random values of x, The ten regression coefficients are plotted in Figure 21D.

All of the regression coefficients of Figure 21 show satisfactory clustering around the true value of unity. The dispersion, however, of individual coefficients in the several groups, is evidently markedly influenced by the range in the selected x, the gain in precision becoming particularly effective as the range of sampled x is lengthened.

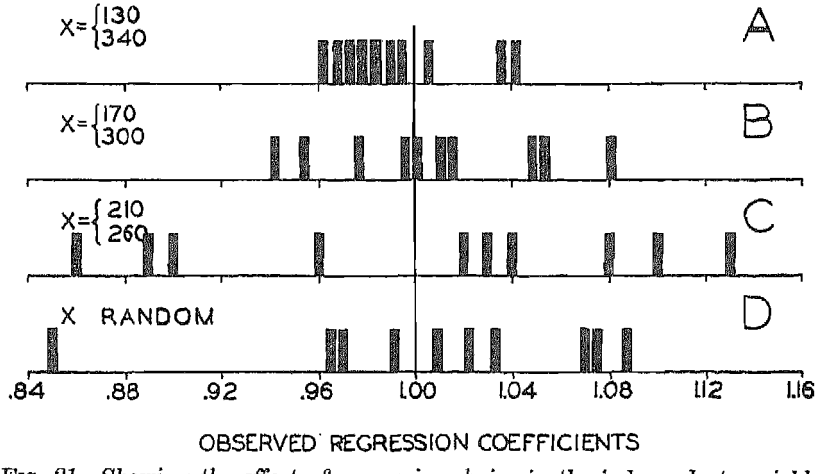


FIG. 21. Showing the effect of purposive choice in the independent variable on the precision of regression coefficients. The range of x is 210 in A, 130 in B, 50 in C; while in D the observed values of x are random.

9.4 Limitation to Purposive Selection. These experiments show that under the conditions to which they apply, increase in precision of the estimate of the dependent variable, y, for given values of the in-

dependent variable, x, may be gained by confining the sample of (y, x) to outlying values of x. The gain, however, is at the sacrifice of other information on both variables.

A completely random sample of the population (y, x) contains information on (1) the general mean of x; (2) the general mean of y; (3) the variance of x; (4) the variance of y; (5) the regression of x on y; (6) the variance of the x-residuals, that is, $s^2_{x \cdot y}$; (7) the regression of y on x; and (8) the variance of the y-residuals, that is, $s^2_{y \cdot x}$.

By the purposive choice of x information is sacrificed on the first six of these items in order to gain precision on the seventh—the regression of y on x. The eighth item, $s^2_{y \cdot x}$, is not affected by the method of selecting x.

As the object of sampling is to arrive at the general mean of the dependent variable, y, the general mean of x is estimated from a new set of one or more random samples of the population of x; and it is inserted in the regression equation of y on x. The operation supplies the estimate of the general mean of y.

There is no value to the regression of x on y, as it is not pertinent to the estimate of y.

As the purposively selected sample does not contain estimates of either the population means or the population variances of the variables concerned, purposive selection should be resorted to only when the estimate of the regression equation of y on x and of the variance of the residuals $(s^2_{y \cdot x})$ are all that are required of the sample of (y, x).

When practiced, it is not to be recommended that only two values of x be selected as was done in the preceding section; for it is not always known with sufficient assurance that y varies directly with x. It is usually preferable, therefore, to sample y according to each of several classes of x, and to plot the mean of y in each class of x on coordinate paper before calculating the regression equation. The true form of the regression equation—straight line or curve—is best indicated when each plotted mean of y is based upon the same number of observations.

It should be kept in mind, of course, that with certain kinds of data, the variation of y around the regression curve is not constant over all values of x. In these special cases the variation in y, as well as the number of observations upon which the means are based, enter into the weights of the plotted points.

The use of weights in regression will be treated in the following chapter.

CHAPTER X

CONDITIONED REGRESSION AND THE USE OF WEIGHTS

10.1 The Sample Census of a Forest Nursery. This chapter deals with aspects of the sample census of a forest-tree nursery as an illustration of the use of regression in sampling when special conditions are imposed by the nature of the data or by the choice of sampling design.

In certain districts the season for planting forest-tree stock is relatively short. If quantities of seedlings, of the order of hundreds of thousands, are to be planted, the administrative planning of the planting program becomes an exceedingly important part of the project. Precise information on the seedling production of a forest nursery is required, by species and grade of stock, prior to the time of commercial lifting of the stock from the seedbeds.

A standard nursery seed bed is 4 feet wide and 50 or more feet long. The common random sampling unit—which is also the ultimate unit—is the strip, one foot wide by four long, extending across the width of the bed.

The sample census of "plantable" seedlings consists of two operations, as follows:

(1) The establishment of the proportion plantable, based upon the count of the number plantable, and of total number of seedlings, on relatively few random sampling units. The sampling unit is lifted, and a skilled inspector identifies the plantable seedlings according to specifications regarding root as well as shoot. This operation is destructive in part, and takes time; hence, it is comparatively expensive.

(2) An independent sample census of the entire number of seedlings on a larger body of random sampling units. Care, but no degree of skill is required, as the sampling units are not lifted. This operation is comparatively cheap.

The first of these operations has for purpose the establishment of the regression of number of plantable seedlings, y, on entire number of seedlings, x, of the sampling units. As it gives rise to a regression of condition, and to observations of variable weights, it will be treated at once.

10.2 Conditioned Regression and the Weights Involved. The number of plantable seedlings, y, is plotted in Figure 22, on total number of seedlings, x, according to each of 54 sampling units, one from each seed bed of 1-year-old longleaf pine in a given nursery. If the

plantable number varies directly with total number, the regression equation of the form used heretofore, that is,

$$Y = a + b(x - \bar{x})$$

is subject to a special condition; for it is certain that each y is limited in the values it can take, between zero and corresponding x. Consequently, when

$$x = 0, \quad Y = 0.$$

Upon putting this condition into the regression equation,

$$0 = a + b(0 - \bar{x}),$$

whence,

$$a = b\bar{x}$$

and the equation takes the simpler form

$$Y = bx.$$

The next question has to do with the weights to be assigned to the observed coordinates of the sample (y, x) in Figure 22. Since an estimate of the absolute numbers of plantable seedlings is required, the regression equation should be made to satisfy the condition that the sum of the estimated values of y be equal to the sum of the actual values; that is, that

$$b \overset{n}{S}(x) = \overset{n}{S}(y),$$

for which, accordingly,

$$b = \frac{\overset{n}{S}(y)}{\overset{n}{S}(x)}.$$

But if each observed y be given unit weight, then the ensuing regression coefficient, b', is a proportionality factor such that

$$\overset{n}{S}\left[(y - Y)^2\right] = \overset{n}{S}\left[(y - b'x)^2\right] \text{ is minimum.}$$

Upon differentiating this sum of squares of residuals with respect to b', and equating to zero,

$$2\overset{n}{S}\left[(y - b'x)(-x)\right] = 0 = b'\overset{n}{S}(x^2) - \overset{n}{S}(xy)$$

whence

$$b' = \frac{\overset{n}{S}(xy)}{\overset{n}{S}(x^2)}.$$

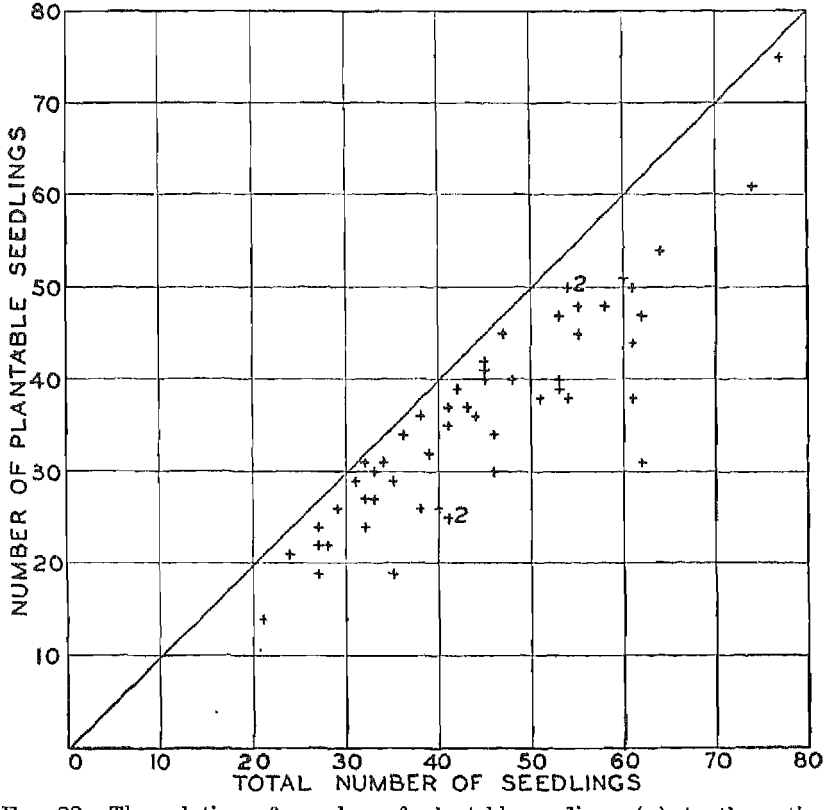


FIG. 22. The relation of number of plantable seedings (y) to the entire number of seedings (x) on 54 selected sampling units of nursery seed bed. The 45-degree line expresses the upper limit of plantable seedlings.

This does not satisfy the condition imposed, for [16]

$$b' \overset{n}{S}(x) \neq \overset{n}{S}(y)$$

but its result is, rather, that

$$b' \overset{n}{S}(x^2) = \overset{n}{S}(xy).$$

On the other hand, if each observed y be given a weight of $\frac{1}{x}$ —provided $x \neq 0$—the regression coefficient, b, is a proportionality factor such that

$$\overset{n}{S}\left[\frac{1}{x}(y-Y)^2\right] = \overset{n}{S}\left[\frac{1}{x}(y-bx)^2\right] \text{ is minimum.}$$

[16] The symbol $\neq$ is read "is not equal to."

Upon differentiating this sum of *weighted* squares of residuals with respect to b, and equating to zero,

$$2\overset{n}{S}\left[\frac{1}{x}(y-bx)(-x)\right]=0=b\overset{n}{S}\left(\frac{x^2}{x}\right)-\overset{n}{S}\left(\frac{xy}{x}\right)$$

and

$$b=\frac{\overset{n}{S}(y)}{\overset{n}{S}(x)},$$

which is the needed regression coefficient, for it satisfies the condition $b\overset{n}{S}(x)=\overset{n}{S}(y)$.

If the outcome of the above discussion be compared with that of Sec. 7.4, it will be recognized at once that the regression coefficient, b, is nothing more than the weighted mean number of plantable trees, per tree of total production. Indeed, the problem of sampling irregular blocks might have been treated from the viewpoint of regression.

In the present case, however, the notion of regression is fundamental to the extension of the problem of sampling plantable trees when the proportion plantable is not constant (Sec. 10.4).

The sampling variance of b is derived from the sum of the weighted squares of residuals,

$$\overset{n}{S}\left[\frac{1}{x}(y-bx)^2\right]=\overset{n}{S}\left(\frac{1}{x}y^2\right)+b^2\overset{n}{S}(x)-2b\overset{n}{S}(y).$$

This may be somewhat shortened for purposes of numerical calculation Since, according to the definition of the regression coefficient, b,

$$b\overset{n}{S}(x)=\overset{n}{S}(y),$$

the above is conveniently expressed as follows:

$$\overset{n}{S}\left[\frac{1}{x}(y-bx)^2\right]=\overset{n}{S}\left(\frac{1}{x}y^2\right)-b^2\overset{n}{S}(x).$$

This identity may be put up in analysis of variance form as in Table 34. The degrees of freedom for the total of the weighted squares of y around zero, are n. As the regression coefficient, b, has used one degree of freedom, the weighted squares of residuals rest upon the remaining $(n-1)$ degrees of freedom.

TABLE 34. Division of the Sum of Weighted Squares of y into Portions Due to, and Independent of, x, with Degrees of Freedom and Mean Squares

Due to	Sum of squares	Degrees of freedom	Mean square
Regression on x..	$b^2 \overset{n}{S}(x)$	1	$\beta^2 \overset{n}{S}(x) + \epsilon_b^2 \overset{n}{S}(x)$
Residuals.....	$\overset{n}{S}\left[\frac{1}{x}(y-bx)^2\right]$	$n-1$	$s^2_{y\cdot x}$ (of unit weight) $\rightarrow \sigma_b^2 \overset{n}{S}(x)$
Total.....	$\overset{n}{S}\left(\frac{1}{x}y^2\right)$	n	

The last column on the right shows pertinent mean squares quite analogous to the unweighted mean squares discussed in connection with Table 31. Accordingly, the mean square of the residuals of unit weight, as given in Table 34, contains the sampling variance of the regression coefficient; hence

$$V(b) = \frac{s^2_{y\cdot x}}{\overset{n}{S}(x)} \text{ (of unit weight).}$$

These results will next be applied to the nursery census.

10.3 Application to the Forest Nursery Sample Census. The data of Figure 22 are given in Table 35. With regard to the independent variable, x, they were not taken at random. A sampling unit, 1 x 4 feet, extending across the bed, was lifted from each bed at a place, designated by the inspector, who immediately counted the entire number of seedlings, and—upon examination of root and shoot—the number of these he judged plantable. The arbitrary choice of situation of each sampling unit was adopted in an effort to have approximately equal representation of sampling units according to classes of total production, x. The resulting sample of 54 pairs is called the "Sample of Plantables."

Concurrently, *two random sampling units*, of the same size, drawn by means of a random sampling device, were examined on each of the 54 beds. As the beds were 100 feet long, these samples make up 2 percent of the entire population of random sampling units. In this part of the job, the population of total production alone was sampled, and the random sampling units were not lifted. This set of 54 samples of two random sampling units each, is called the "Samples of Density."

The analysis of variance of the samples of density is given in Table 36.

All the materials are now available for the estimate of the production of the 54 seed beds in total population of seedlings, as well as in the population of plantable seedlings.

TABLE 35. Total Number of Seedlings (x), and the Number of These Considered Plantable (y), on Sampling Units of Four Square Feet, Taken One from Each of 54 Seed Beds of Longleaf Pine

x	y	x	y	x	y	x	y	x	y	x	y
21	14	32	27	38	36	45	40	53	39	60	51
24	21	32	31	39	32	45	41	53	40	61	38
27	19	33	27	40	26	45	42	53	47	61	44
27	22	33	30	41	25	46	30	54	38	61	50
27	24	34	31	41	35	46	30	54	50	62	31
28	22	35	19	41	37	46	34	54	50	62	47
29	26	35	29	42	39	47	45	55	45	64	54
31	29	36	34	43	37	48	40	55	48	74	61
32	24	38	26	44	36	51	38	58	48	77	75

$\overset{54}{S}(x) = 2{,}413$; $\overset{54}{S}(y) = 1{,}954$; $\overset{54}{S}\left(\frac{1}{x}y^2\right) = 1{,}613.24$

TABLE 36. Analysis of Variance of the Samples of Density*

Source of variation	Degrees of freedom	Sum of squares	Mean square
Among beds	53	11,534.75	
Between random sampling units (same bed)	54	6,260.50	115.94
Total, among random sampling units	107	17,795.25	

*Total number of seedlings over all 108 random sampling units, 4,527.

Turning first to the samples of density, one finds the observed number of all seedlings on the 108 random sampling units is 4,527, as given at the bottom of Table 36. The sampling variance of this number is 108 times the sampling variance of the random sampling units within the beds, that is,

$$V(4{,}527) = 108(115.94) = 12{,}522,$$

the standard error being the square root of this, or 112. As the observed random sampling units are but 2 percent of the entire area of the 54 seed beds, the sum, and its standard error, should be multiplied by 50, that is, by $\left(\frac{100}{2}\right)$. The correction for the finite population within the beds—

that is, $\sqrt{\frac{100-2}{100}}$—is negligible, and is not applied. The best estimate, then, of the population of all seedlings is

$$50(4{,}527 \pm 112) = 226.4 \pm 5.60 \text{ M seedlings.}$$

The estimate of the number of plantable seedlings is to be derived from the regression of plantable number, y, on entire number, x. From the totals of Table 35, the regression coefficient is

$$b = \frac{\overset{n}{S}(y)}{\underset{n}{S}(x)} = \frac{1{,}954}{2{,}413} = 0.8098$$

whence

$$Y = 0.8098x.$$

The division of the sum of weighted squares of y into portions due to, and independent of, x, is given in Table 37. The mean square of the weighted residuals is the variance of unit weight, and therefore

TABLE 37. Analysis of Regression of Plantable Seedlings on Total Seedlings

Due to	Sum of squares	Degrees of freedom	Mean square
Regression.....	$b^2 \overset{54}{S}(x)$ = 1,582.31	1	
Residuals.....	$\overset{54}{S}\left[\frac{1}{x}(y-bx)^2\right]$ = 30.93	53	0.5836
Total.....	$\overset{54}{S}\left(\frac{1}{x}y^2\right)$ = 1,613.24	54	

$$s^2_{y \cdot x} \text{ (of unit weight)} = 0.5836$$

on 53 degrees of freedom. Then the sampling variance of the regression coefficient, b, which was given in the preceding section as

$$V(b) = \frac{s^2_{y \cdot x} \text{(of unit weight)}}{\underset{n}{S}(x)}$$

is

$$V(0.8098) = \frac{0.5836}{2{,}413} = 0.000242.$$

The number of plantable seedlings corresponding to the observed total number of the density samples is

$$Y = 0.8098(4527)$$
$$= 3{,}666 \text{ plantable seedlings.}$$

Since this estimate is the product of two numbers, each of which is subject independently to sampling error, its sampling variance is the sampling variance of Y. Hence

$$V\left[{}^{108}_{}S\ (Y)\right] = V\left[b\ {}^{108}_{}S\ (x)\right] = \left[{}^{108}_{}S\ (x)\right]^2\left[V(b)\right] + b^2\left[V\left\{{}^{108}_{}S\ (x)\right\}\right]$$

and

$$V(3{,}666) = (4{,}527)^2(0.000242) + (0.8098)^2(12{,}522)$$
$$= 13{,}171$$

its standard error being the square root of this, or 115. As before, this estimate is based upon 2 percent of the entire area of the 54 seed beds. Accordingly, the estimate of the population of plantable seedlings is

$$50(3{,}666 \pm 115) = 183.3 \pm 5.75 \text{ M seedlings.}$$

Had the correction factor on account of the finite population of random sampling units within the beds been applied, the standard error of 5.75 M seedlings would have been multiplied by the factor

$$\sqrt{\frac{100-2}{100}}$$

an adjustment of less than 1 percent. This appears entirely negligible.

10.4 The Introduction of a Second Independent Variable. It was pointed out in Sec. 10.2 that the regression coefficient which expresses the number of plantable seedlings in terms of the entire number of seedlings is merely the weighted mean number of plantable seedlings per tree of total production; and that the notion of regression is not essential when the proportion plantable is constant, that is, when it is independent of the entire number on the sampling units.

It is common nursery experience, however, that the proportion of plantable seedlings falls off as seedling density becomes excessive. In such cases, the regression of the number plantable, y, on the entire number, x, is not a straight line.

Figure 23 shows the coordinates of 23 sampling units of slash pine. They do not represent a random sample of x. Effort was made to collect an approximately equal number of sampling units according to class of x, so that any nonlinearity which might characterize the true relationship of y to x, would be emphasized. The broken line is the representation of the regression line,

$$Y = 0.7694x$$

for which

$$0.7694 = b = \frac{S(y)}{S(x)}.$$

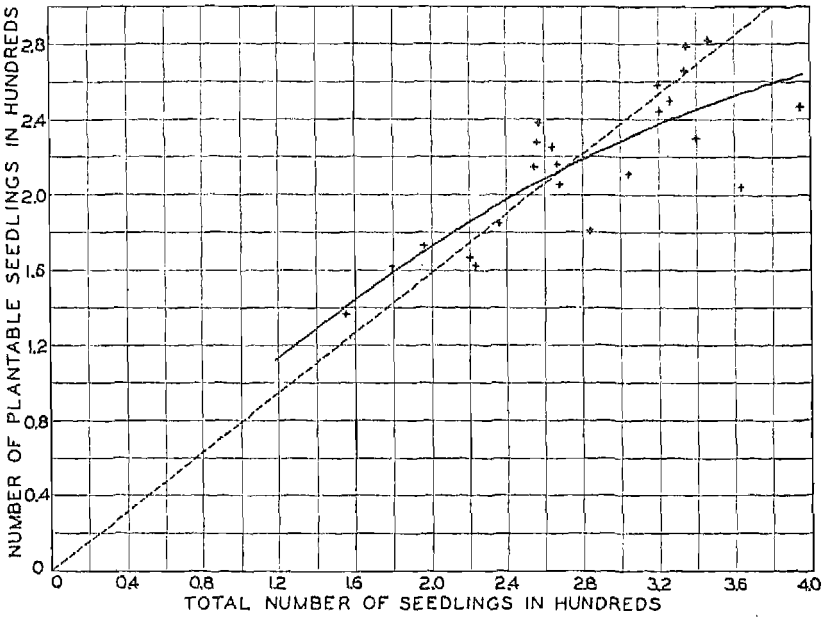


FIG. 23. The relation of number of plantable seedings (y) to the entire number of seedlings (x) on 23 selected sampling units of nursery seed bed. The broken line represents the best fit on the supposition that the proportion plantable is independent of the entire number.

By comparison with the plotted points upon which it is based, it seems too low for low density and too high for high density. It is to be expected, therefore, that a better estimate should be obtained were a *regression curve* of the form

$$Y = b_1 x + b_2 x^2$$

fitted to the data, subject to the conditions used previously; namely, that the sum of the estimated plantables be identical with the sum of the observed plantables upon which the regression is founded; hence that

$$b_1 \overset{n}{S}(x) + b_2 \overset{n}{S}(x^2) = \overset{n}{S}(Y) = \overset{n}{S}(y).$$

This condition will be fulfilled, if b_1 and b_2 are so chosen that the sum of the *weighted* squares of residuals

$$\overset{n}{S}\left[\frac{1}{x}(y-Y)^2\right] \text{ is minimum.}$$

This is equivalent to making

$$\overset{n}{S}\left[\frac{1}{x}(y-b_1x-b_2x^2)^2\right] \text{ a minimum.}$$

Upon differentiating this expression with respect to each of the unknowns, b_1 and b_2, in turn and equating to zero—following the process as explained in Sec. 10.4(A) of the Appendix—there are the two normal equations,

$$b_1 \overset{n}{S}(x)+b_2 \overset{n}{S}(x^2)= \overset{n}{S}(y)$$
$$b_1 \overset{n}{S}(x^2)+b_2 \overset{n}{S}(x^3)= \overset{n}{S}(xy).$$

The simultaneous solution of these equations affords the regression coefficients.

The division of the sum of weighted squares of y into portions due to and independent of the regression on x, is effected by expanding the sum of weighted squares of residuals. For, as developed in Sec. 10.4(B) of the Appendix,

$$\overset{n}{S}\left[\frac{1}{x}(y-b_1x-b_2x^2)^2\right]= \overset{n}{S}\left(\frac{1}{x}y^2\right)-b_1 \overset{n}{S}(y)-b_2 \overset{n}{S}(xy).$$

It is convenient, for purposes of computation, to put this identity in analysis of variance form, as in Table 38. It is to be noted that the sum of weighted squares of residuals is based upon $(n-2)$ degrees of freedom among the n observations of (y, x) as one degree of freedom is used in the estimate of each regression coefficient.

TABLE 38. Division of Sum of Weighted Squares of y into Portions Due to, and Independent of, the Regression on x, when the Regression Takes the Form $Y = b_1x + b_2x^2$

Source of variation	Sum of squares	Degrees of freedom
Regression on x and x^2	$b_1\overset{n}{S}(y)+b_2\overset{n}{S}(xy)$	2
Residuals independent of x and x^2	$\overset{n}{S}\left[\frac{1}{x}(y-b_1x-b_2x^2)^2\right]$	$n-2$
Total	$\overset{n}{S}\left(\frac{1}{x}y^2\right)$	n

TABLE 39. Total Number of Seedlings in Hundreds (x), and the Number of These (also in Hundreds) Considered Plantable (y), on Sampling Units of Four Square Feet, Taken One from Each of 23 Seed Beds of Slash Pine

x	y	x	y	x	y	x	y
1.55	1.37	2.55	2.28	2.84	1.81	3.36	2.79
1.80	1.62	2.55	2.15	3.05	2.11	3.40	2.30
1.96	1.73	2.58	2.38	3.20	2.58	3.47	2.82
2.21	1.67	2.64	2.25	3.21	2.44	3.64	2.04
2.24	1.62	2.67	2.16	3.27	2.50	3.95	2.46
2.36	1.85	2.68	2.05	3.34	2.66		

The result of the above discussion will next be applied to the observations presented in Figure 23. These are listed in Table 39. Upon performing the operations on these observations which lead to quantities to be substituted in the normal equations,

$$\overset{23}{S}(x) = 64.52; \qquad \overset{23}{S}(y) = 49.64;$$

$$\overset{23}{S}(x^2) = 189.4010; \qquad \overset{23}{S}(xy) = 143.6000;$$

$$\overset{23}{S}(x^3) = 577.5079.$$

The normal equations, then, are the following:

$$64.5200b_1 + 189.4010b_2 = 49.6400$$
$$189.4010b_1 + 577.5079b_2 = 143.6000$$

whence

$$b_1 = 1.0587$$
$$b_2 = -0.09855.$$

The regression equation of the number of plantable seedlings on the entire number on the sampling units is

$$Y = 1.0587x - 0.09855x^2,$$

and this is the curve of Figure 23.

The results of calculations leading to the mean square of the residuals —that is, the variance of y of unit weight—are given in Table 40. The bottom three lines are the numerical equivalents corresponding to Table 38. The sum of weighted squares of y, without any regard to the independent variables, is

$$\overset{n}{S}\left(\frac{1}{x}y^2\right) = 38.7771;$$

while the portion of this which is due to x and x^2, is

$$b_1 \overset{n}{S}(y) + b_2 \overset{n}{S}(xy) = 1.0587(49.64) - 0.09855(143.60)$$
$$= 38.4021.$$

The sum of the weighted squares of residuals, therefore, is

$$\overset{n}{S}\left[\frac{1}{x}(y - b_1x - b_2x^2)^2\right] = 38.7771 - 38.4021$$
$$= 0.3750$$

on 21 degrees of freedom. The variance of the residuals (of unit weight) is 0.01786. These values are given in Table 40.

TABLE 40. Analysis of Regression of Plantable Seedlings on Total Seedlings; and Test of Curvilinearity of Regression

Due to	Degrees of freedom	Sum of squares	Mean square
Regression on x alone.....	1	38.1917	
Additonal effect of x^2.....	1	0.2104	0.2104*
Regression on x and x^2.....	2	38.4021	
Residuals.....	21	0.3750	$0.01786^* = s^2_{6.12}$
Total.....	23	38.7771	

*Observed $t = \sqrt{\frac{0.2104}{0.01786}} = 3.43$; expected t at 1 percent level, 2.83.

The question might logically be raised: Is the difference between the curve and the straight line as fitted to these data—both of which are presented in Figure 23—a significant difference? In other words, is the contribution of the second independent variable, x^2, real, and not merely an accident of sampling? The test will be performed at once.

If the relationship were adequately described by the straight line, the regression equation—the broken line of Figure 23—would be of the form

$$Y = bx$$

where

$$b = \frac{S(y)}{S(x)} = \frac{49.64}{64.52} = 0.7694$$

and the sum of squares due to x alone would be

$$b^2 \overset{23}{S}(x) = \frac{\left[\overset{23}{S}(y)\right]^2}{\overset{23}{S}(x)} = \frac{(49.64)^2}{64.52} = 38.1917$$

on one degree of freedom. These values are listed in the top line of Table 40. From the same table, the sum of squares

Due to x and $x^2 = 38.4021$ on 2 degrees of freedom
Due to x alone $= 38.1917$ on 1 degree of freedom.

Consequently, the contribution of x^2, over and above that ascribable to x alone, is the difference, that is,

Due to $x^2 = 0.2104$ on 1 degree of freedom,

and this is listed in the second line of the table. Were this merely a chance contribution, its value, in the average, should be the same as the mean square of the residuals, 0.01786. In fact, however, the discrepancy is such that

$$t = \sqrt{\frac{0.2104}{0.01786}} = 3.43$$

on 21 degrees of freedom. Referring to the table of t, one notes that, due to sampling only, t should exceed 2.831 on 21 degrees of freedom, only once in 100 trials. The observed t of 3.43 is therefore highly significant, and the contribution of x^2 is unquestionably real. The best expression, then, for the number of plantable seedlings in terms of the entire number is

$$Y = 1.0587x - 0.09855x^2,$$

and this will be applied to the independent estimate of x and x^2 of the 23 seed beds.

10.5 The Variance of the Conditioned Regression Curve and Its Application. The application of the regression involves the variance of the calculated Y. Developments leading to $V(Y)$ are somewhat lengthy to give here; hence, they are presented in Sec. 10.5 of the Appendix. There it is demonstrated—Sec. 10.5(D)—that the variance of Y is made up of contributions from two sources, one of which, symbolized as s^2R, is ascribable to the sampling errors of the regression equation itself, while the second, symbolized as S^2, is ascribable to the sampling errors of the values of the independent variables inserted into the equation. Thus in the regression equation of the preceding section, of the form

$$Y = b_1x + b_2x^2$$

we have

$$V(Y) = s^2R + S^2$$

in which s^2 is the mean square of the residuals (of unit weight) independent of the regression; and in which

$$R = c_{11}x^2 + c_{22}(x^2)^2 + 2c_{12}(xx^2)$$

also, in which

$$S^2 = b_1^2\left[\ V(x)\ \right] + b_2^2\left[\ V(x^2)\ \right] + 2b_1b_2\left[\ Cov(xx^2)\ \right].$$

If the variance of the residuals of unit weight be symbolized as s^2, then it is shown in Sec. 10.5(C) of the Appendix that the estimates of the variances and covariance of the regression coefficients b_1 and b_2 may be stated

$$V(b_1) = s^2c_{11}; \quad V(b_2) = s^2c_{22}; \quad Cov(b_1b_2) = s^2c_{12}.$$

The derivation of the c-multipliers, c_{11}, c_{22}, and c_{12}, is given in Sec. 10.5(B) of the Appendix. Their numerical equivalents are computed from the sums of squares and products among the independent variables only. In the case of the data of the preceding section, they are to be evaluated from the following two sets of expressions:

$$c_{11}\overset{n}{S}(x) + c_{12}\overset{n}{S}(x^2) = 1$$
$$c_{11}\overset{n}{S}(x^2) + c_{12}\overset{n}{S}(x^3) = 0$$

and

$$c_{12}\overset{n}{S}(x) + c_{22}\overset{n}{S}(x^2) = 0$$
$$c_{12}\overset{n}{S}(x^2) + c_{22}\overset{n}{S}(x^3) = 1.$$

Turning, now, to the numerical work, the c-multipliers c_{11} and c_{12} are given by

$$64.5200c_{11} + 189.4010c_{12} = 1$$
$$189.4010c_{11} + 577.5079c_{12} = 0$$

whence

$$c_{11} = 0.416051; \quad c_{12} = -0.136449;$$

and upon calculating c_{12} (as a check) and c_{22} from the expressions

$$64.5200c_{12} + 189.4010c_{22} = 0$$
$$189.4010c_{12} + 577.5079c_{22} = 1$$

the results are,

$$c_{12} = -0.136449; \quad c_{22} = 0.046482.$$

If the values of x (and x^2) which are inserted into the regression equation

$$Y = b_1x + b_2x^2$$

are free of error, then $S^2 = 0$, and

$$V(Y) = s^2R = s^2\left[\ c_{11}x^2 + c_{22}(x^2)^2 + 2c_{12}(xx^2)\ \right]$$

the square root of which is the standard error of the regression function. This has been done for a number of values of x in the regression equation

$$Y = 1.0587x - 0.09855x^2.$$

The 95 percent confidence band, which has been derived therefrom, is presented in Figure 24.

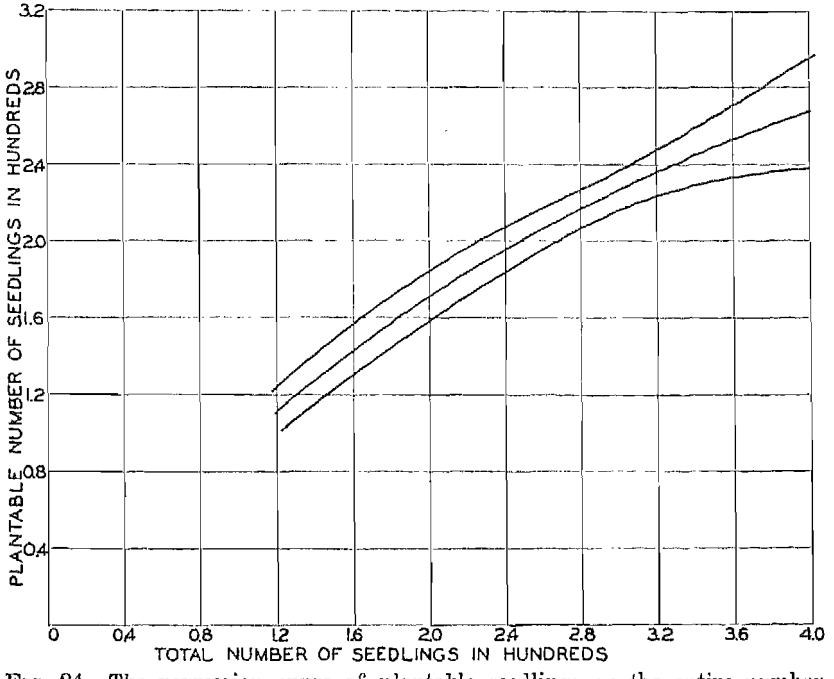


FIG. 24. The regression curve of plantable seedlings on the entire number, and the 95 percent confidence band.

The regression equation is to be applied to the independent samples of density taken on the 23 seed beds. Two random sampling unit observations of density were taken on each of the beds, of number of seedlings (x), and of the square of this number (x^2). These are listed in columns 2-5 of Table 41. Subscripts 1 and 2 refer to the first and second random sampling unit, respectively, of each bed.

Calculations leading to the estimates of the variances and covariance of the sample sums are also shown in the table. If one denotes the individual sums of x as X_1, and of x^2 as X_2, the sample sums over all beds, which are 2 percent of the estimates of the population aggregates, are the following:

$$\sum_{23}(X_1) = 58.20 + 57.82 = 116.02$$

$$\overset{23}{S}(X_2) = 159.7296 + 155.1162 = 314.8458.$$

Upon substituting these for x and x^2 of the regression equation, one obtains

$$Y = 1.0587(116.02) - 0.09855(314.8458)$$
$$= 91.802$$

in hundreds of plantable seedlings. The variance of this estimate may now be calculated readily. From the values of the c-multipliers, given above and from $s^2 = 0.01786$ from Table 40.

$$s^2R = 0.01786\Big[0.416051(116.02)^2 + 0.046482(314.8458)^2 + 2(-0.136449)(116.02)(314.8458)\Big]$$
$$= 4.276.$$

TABLE 41. Samples of Density; Number of Seedlings (x), and their Squares (x^2), on Each of Two Random Sampling Units to the Seed Bed. And Calculation of Variances and Covariance Within Beds

Bed number	Random sampling units of x		Random sampling units of x^2				$V(X_1)$	$Cov.(X_1X_2)$	$V(X_2)$
	x_1	x_2	x_1^2	x_2^2	(x_1-x_2)	$(x_1^2-x_2^2)$	$(x_1-x_2)^2$	$(x_1-x_2).(x_1^2-x_2^2)$	$(x_1^2-x_2^2)^2$
1.....	2.42	3.03	5.8564	9.1809	-0.61	-3.3245	0.3721	2.0279	11.0523
2.....	1.33	2.57	1.7689	6.6049	-1.24	-4.8360	1.5376	5.9966	23.3869
3.....	2.12	1.43	4.4944	2.0449	0.69	2.4495	0.4761	1.6902	6.0001
4.....	2.04	1.62	4.1616	2.6244	0.42	1.5372	0.1764	0.6456	2.3630
5.....	3.06	1.97	9.3636	3.8809	1.09	5.4827	1.1881	5.9761	30.0600
6.....	1.23	1.72	1.5129	2.9584	-0.49	-1.4455	0.2401	0.7083	2.0895
7.....	2.32	1.47	5.3824	2.1609	0.85	3.2215	0.7225	2.7383	10.3781
8.....	2.88	2.33	8.2944	5.4289	0.55	2.8655	0.3025	1.5760	8.2111
9.....	3.62	3.91	13.1044	15.2881	-0.29	-2.1837	0.0841	0.6333	4.7685
10.....	1.76	3.25	3.0976	10.5625	-1.49	-7.4649	2.2201	11.1227	55.7247
11.....	4.07	3.43	16.5649	11.7649	0.64	4.8000	0.4096	3.0720	23.0400
12.....	1.63	2.94	2.6569	8.6436	-1.31	-5.9867	1.7161	7.8426	35.8406
13.....	2.87	2.48	8.2369	6.1504	0.39	2.0865	0.1521	0.8137	4.3535
14.....	3.56	2.31	12.6736	5.3361	1.25	7.3375	1.5625	9.1719	53.8389
15.....	2.49	2.62	6.2001	6.8644	-0.13	-0.6643	0.0169	0.0864	0.4413
16.....	2.22	1.85	4.9284	3.4225	0.37	1.5059	0.1369	0.5572	2.2677
17.....	2.48	2.62	6.1504	6.8644	-0.14	-0.7140	0.0196	0.1000	0.5098
18.....	2.69	3.24	7.2361	10.4976	-0.55	-3.2615	0.3025	1.7938	10.6374
19.....	3.42	2.33	11.6964	5.4289	1.09	6.2675	1.1881	6.8316	39.2816
20.....	3.21	2.67	10.3041	7.1289	0.54	3.1752	0.2916	1.7146	10.0819
21.....	1.68	3.33	2.8224	11.0889	-1.65	-8.2665	2.7225	13.6397	68.3350
22.....	2.88	2.62	8.2944	6.8644	0.26	1.4300	0.0676	0.3718	2.0449
23.....	2.22	2.08	4.9284	4.3264	0.14	0.6020	0.0196	0.0843	0.3624
Totals .	58.20	57.82	159.7296	155.1162	0.38	4.6134	15.9252	79.1946	405.0692

From Table 41,

$$V\left[\overset{23}{S}(X_1)\right]=15.9252; \quad V\left[\overset{23}{S}(X_2)\right]=405.0692;$$

$$Cov\left[\overset{23}{S}(X_1X_2)\right]=79.1946;$$

neglecting the trivial adjustment to the finite seed bed populations, since the samples contain only 2 percent thereof. The numerical equivalent of S^2 is, therefore,

$$\begin{aligned} S^2 &= (1.0587)^2(15.9252)+(-0.09855)^2(405.0692) \\ &\quad +2(1.0587)(-0.09855)(79.1946) \\ &= 5.258. \end{aligned}$$

Upon combining the two contributions one has

$$\begin{aligned} V(91.802) &= 4.276+5.258 \\ &= 9.534 \end{aligned}$$

on the 21 degrees of freedom upon which the estimate $s^2=0.01786$ has been based. The standard error is the square root of this variance, or 3.088. Hence, the population estimate of the number of plantable seedlings in the 23 seed beds is

$$50(91.802\pm 3.088)=4{,}590\pm 154$$

plantable seedlings in hundreds. With the 21 degrees of freedom available, $t=2.080$ at the 5 percent level; hence, the probability is 0.95 that the population aggregate consists of

$$4{,}590\pm 320$$

in hundreds; or

$$459\pm 32$$

in thousands of plantable seedlings.

10.6 Certain Remarks Concerning Regression in Sampling. It should be pointed out that the data from which the regression equations of this chapter were derived, were taken in such a way that the effect of variation among the blocks (beds) could not be eliminated from the regressions; for only a single sampling unit, out of the numbers upon which the regressions were based, was taken from each bed. Hence the sums of weighted squares of plantable seedlings contained, in each case, a portion due to variation among the beds.

This portion, the numerical equivalent of which is unknown, was, nevertheless, believed to be negligibly small in each of the nursery problems cited. It would have been eliminated from the sampling error had

it been feasible to dig up and inspect a minimum of two sampling units to the bed rather than the single one as actually inspected. Such an expanded program, however, would have at least doubled the inspection labor. Under the circumstances, any additonal gain in precision expected thereby was considered more costly than warranted.

The next chapter treats of the problem of eliminating variation among blocks from the regression equation.

CHAPTER XI

REGRESSION IN REPRESENTATIVE SAMPLING

11.1 The Problem. The present chapter deals with a timber cruise for which the sampling was so designed that the effect of variation among the blocks is removable from the regression equation; and the latter then is used to adjust the ocular estimate of timber volume.

A quarter-section of pine-hardwood timber is divided into eight blocks of 20 acres each, the block dimensions being 10 x 20 chains. In each block, two cruise strips, each 1 x 20 chains over the length of its block, have been selected, independently and at random, from among the 10 in the block area, and the ocular estimate of hardwood volume confined to these. Separate record, however, was kept according to each quarter-strip, or subplot of 1 x 5 chains, within the sample strips. The results of this cruise are listed in the left half of Table 42, each entry being the whole-strip sum. Figure 25 shows the distribution of the strips over the blocks.

Upon the completion of the ocular estimates,one of the quarter-strips (1 x 5 chains) of each sample strip was re-run and the hardwood volume thereon measured carefully. Thus the volume of two quarter-strips of each block is represented by a direct measurement (y) and a concomitant ocular estimate (x). Their locations are also shown in Figure 25, while the right-half of Table 42 contains the observations.

The problem is to calculate from the data submitted, the best estimate of the hardwood volume in M feet b.m. together with an evaluation of its probable accuracy.

From a practical standpoint, a simpler sampling design might, indeed, have proven just as efficient. Had the volumes over the entire lengths of the two strips of each block been carefully measured, there need have been no ocular estimate; hence no adjustments by regression. And while the entire field time might have doubled, the job is a relatively small one and the difference in field time perhaps of little consequence.

The principle of the example, however, is useful in practice. If, for instance, a large tract of timber has been cruised by a group of inexperienced men (such as student assistants), the chief of party, or perhaps an independent check cruiser, may need to re-run a sample of the cruised strips in order to determine the accuracy of the work. If necessary, a correction factor may be evaluated so as to eliminate such part of the

TABLE 42. Ocular Estimate of Hardwood Volume in M feet b.m. on Two Random Strips in Each of Eight Blocks; and Concomitant Observations of Hardwood Volume in M feet b.m. According to Direct Measurement (y), and Ocular Estimate (x), on Two Selected Quarter-Strips of the Block

Block	On whole strips	On quarter-strips	
	Ocular estimate	Ocular estimate x	Direct Measurement y
1	19.10	5.02	4.63
	16.50	4.03	3.82
2	8.74	1.45	1.63
	5.12	0.83	0.60
3	21.22	5.37	4.36
	12.56	2.67	3.38
4	6.88	1.99	1.67
	7.51	1.63	1.58
5	16.56	3.60	3.49
	19.93	4.52	4.07
6	6.22	1.01	0.97
	0.96	0.00	0.00
7	9.60	2.41	2.00
	13.42	3.67	3.55
8	8.06	1.39	2.38
	9.40	2.52	2.33
Sum	181.78	42.11	40.46
Mean	11.361	2.632	2.529

variation as might be ascribed to the idiosyncrasies of the different cruisers.

Or, again, suppose a large number of woodlots have been hastily examined by a fairly reliable cruiser. His estimate of the aggregate volumes may be adjusted to measured volumes, from which the variation between neighboring groups of woodlots have been eliminated; provided only that a small percentage of the area of the woodlots within each group is revisited, and the timber thereon accurately measured.

Whenever a population to be sampled is subdivided into blocks, variation between blocks can be eliminated from the regression analyses by a procedure known as the analysis of covariance. This procedure is discussed in the next section.

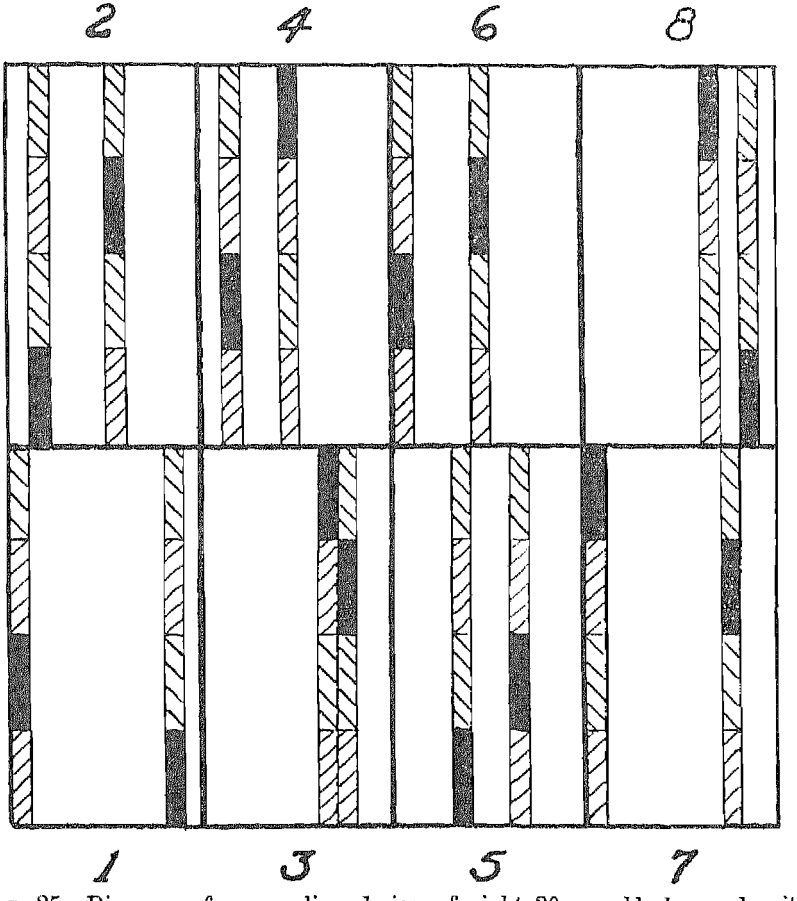


FIG. 25. Diagram of a sampling design of eight 20-acre blocks, each with two random sampling units of whole-strips upon which ocular estimates of timber volume have been made. On the shaded quarter of each strip, volume has also been measured.

11.2 The Analysis of Covariance. The covariance of two variables, say x and y, has been used in earlier chapters (e.g., Secs. 5.2 and 8.2). It is derived from the sum of products of paired residuals. Thus the sum of products

$$S_{n}\left[(x-\bar{x})(y-\bar{y})\right]$$

when divided by the number of degrees of freedom involved, is an estimate of the covariance of x and y.

If a sample of n values of y and of concomitant x is drawn from each of k strata, or blocks, of a general 2-variate population, the deviations of each from its general mean, $\bar{y}$ (or $\bar{x}$), may be analyzed into a portion due to the stratification, and a remaining portion independent of the stratification. Thus

$$\left.\begin{aligned}(y-\bar{y}) &= (y-y_b)+(y_b-\bar{y})\\ (x-\bar{x}) &= (x-x_b)+(x_b-\bar{x})\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1),$$

in which y_b and x_b represent the block means. The sum of squares of each of these over all kn observations are the identities used in the analysis of variance of each variable. As developed in Sec. 6.4, it follows that

$$\left.\begin{aligned}\overset{k}{S}\overset{n}{S}\left[(y-\bar{y})^2\right] &= \overset{k}{S}\overset{n}{S}\left[(y-y_b)^2\right]+n\overset{k}{S}\left[(y_b-\bar{y})^2\right]\\ \overset{k}{S}\overset{n}{S}\left[(x-\bar{x})^2\right] &= \overset{k}{S}\overset{n}{S}\left[(x-x_b)^2\right]+n\overset{k}{S}\left[(x_b-\bar{x})^2\right]\end{aligned}\right\} \quad \ldots\ldots\ldots (2).$$

If, however, equations (1) are multiplied together and summed over all kn values,

$$\begin{aligned}\overset{k}{S}\overset{n}{S}\left[(y-\bar{y})(x-\bar{x})\right] &= \overset{k}{S}\overset{n}{S}\left[(y-y_b)(x-x_b)\right]+\overset{k}{S}\left[(y_b-\bar{y})\overset{n}{S}(x-x_b)\right]\\ &\quad +\overset{k}{S}\left[(x_b-\bar{x})\overset{n}{S}(y-y_b)\right]+n\overset{k}{S}\left[(y_b-\bar{y})(x_b-\bar{x})\right].\end{aligned}$$

But since

$$\overset{n}{S}(x-x_b)=\overset{n}{S}(y-y_b)=0,$$

the second and third terms of the right-hand member are zero; hence

$$\overset{k}{S}\overset{n}{S}\left[(y-\bar{y})(x-\bar{x})\right]=\overset{k}{S}\overset{n}{S}\left[(y-y_b)(x-x_b)\right]+n\overset{k}{S}\left[(y_b-\bar{y})(x_b-\bar{x})\right] \ldots (3).$$

The sum of products of x and y, then, like the sum of squares of each, may be divided into two portions, as follows:

(1) A portion independent of the stratification; that is, the first term of the right-hand member of equations (2) and (3).

(2) A remaining portion due to the stratification.

Equations (2) and (3) may be conveniently assembled as in Table 43. Inspection of the table makes evident that there must be a minimum of

two sampling units to the block in order to permit the division illustrated. For if $n = 1$, the first and third lines of each column would be numerically identical, and the middle line would become zero.

TABLE 43. Division of Sum of Squares of Each of Two Correlated Variables, and of Their Sum of Products, into Portions Due to, and Independent of, Stratification

Source of variation	Degrees of freedom	Sum of squares		Sum of products
		x^2	y^2	xy
Among blocks.	$k-1$	$n \overset{k}{S}\left[(x_b-\bar{x})^2\right]$	$n \overset{k}{S}\left[(y_b-\bar{y})^2\right]$	$n \overset{k}{S}\left[(x_b-\bar{x})(y_b-\bar{y})\right]$
Within blocks.	$k(n-1)$	$\overset{k}{S}\overset{n}{S}\left[(x-x_b)^2\right]$	$\overset{k}{S}\overset{n}{S}\left[(y-y_b)^2\right]$	$\overset{k}{S}\overset{n}{S}\left[(x-x_b)(y-y_b)\right]$
Total.....	$kn-1$	$\overset{k}{S}\overset{n}{S}\left[(x-\bar{x})^2\right]$	$\overset{k}{S}\overset{n}{S}\left[(y-\bar{y})^2\right]$	$\overset{k}{S}\overset{n}{S}\left[(x-\bar{x})(y-\bar{y})\right]$

Such was the case in the regressions of the nursery problems of the preceding chapter.

But when variation can be eliminated from the sum of squares of both variables, and from the sum of their cross-products, it can also be eliminated from the regression equation of y on x, by the simple device of calculating the regression from those portions which have been freed from block effects.

11.3 **The Adjusted Estimate and Its Variance.** It is to be expected, of course, that the measured quarter-strip volume of Table 42 varies directly with corresponding ocular estimate; hence, the regression of y on x is of the form

$$Y = a + b(x - \bar{x})$$

where a and $\bar{x}$ are, respectively, the general means of y and x. Their numerical equivalents, from the data of Table 42, are the following:

$$a = \bar{y} = \frac{40.46}{16} = 2.529 \text{ M feet b.m.}$$

$$\bar{x} = \frac{42.11}{16} = 2.632 \text{ M feet b.m.}$$

If these be substituted into the general regression equation, we have

$$Y = 2.529 + b(x - 2.632)$$

and the regression coefficient alone is needed in order to complete it.

The sums of squares and products of the quarter-strip data are listed in Table 44. Confining the calculations to those portions which have been freed from block effects, one obtains the weighted average regression coefficient of y on x, that is

$$b=\frac{3.7644}{6.7575}=0.5571,$$

and the sum of squares due to the regression is

TABLE 44. Division of Sums of Squares, and Sum of Products, of the Quarter-Strip Data of Table 42, into Portions Due to, and Independent of, the Blocks

Source of variation	Degrees of freedom	Sum of squares		Sum of products
		y^2	x^2	xy
Among blocks...	7	26.6501	31.1997	28.5535
Within blocks...	8	3.1839	6.7575	3.7644
Total.....	15	29.8340	37.9572	32.3179

$$\frac{(3.7644)^2}{6.7575}=2.0970,$$

the analysis of the regression being completed in Table 45. The equation, then, is the following:

$$Y=2.529+0.5571(x-2.632)$$

in which Y is volume in M feet b.m. to the quarter-strip.

The ocular estimate may now be adjusted by inserting the general mean of the whole-strip volumes for x, though we need to express the latter in the quarter-strip unit of area so as to be consistent with the area unit of the regression equation. From Table 42 this value is

$$\frac{181.78}{4(16)}=2.840 \text{ M feet b.m.}$$

to the quarter-strip. Hence the best estimate of the hardwood volume of the tract from the data submitted is

$$\begin{aligned} Y&=2.529+0.5571(2.840-2.632)\\ &=2.645 \text{ M feet b.m.} \end{aligned}$$

to the quarter-strip, or half-acre of area.

The sampling variance of this estimate follows at once. It was shown in Sec. 8.7 that if

$$Y=a+b(x-\bar{x})$$

the sampling variance of Y may be expressed

$$V(Y) = V(a) + (x - \bar{x})^2 \left[V(b) \right] + b^2 \left[V(x) \right].$$

TABLE 45. Division of Sum of Squares of Measured Volume within the Blocks into Portions Due to, and Independent of, Ocular Estimate of Volume

Source of variation	Degrees of freedom	Sum of squares
Regression	1	2.0970
Residuals	7	1.0869
Within blocks	8	3.1839

The variance of a may be estimated from the data of Table 45; that is,

$$V(a) = V(\bar{y}) = \frac{1.0869}{7(16)} = 0.009704$$

while for the variance of b, which in Sec. 8.2 was expressed

$$V(b) = \frac{s^2_{y \cdot x}}{S\left[(x - \bar{x})^2 \right]},$$

the sum of squares of x is taken from Table 44, and the mean square of the residuals independent of x, from Table 45. Accordingly,

$$V(b) = \frac{1.0869}{7(6.7575)} = 0.022978.$$

It should be noted that $V(a)$ and $V(b)$ might be corrected to the finite block populations sampled. But as only two quarter-strips, out of a total of 40 within each block, constitute the regression data, the correction factor, $\left(\frac{38}{40}\right)$, has not been applied.

The variance of the general mean of the whole-strip volumes is taken from the analysis of variance of these ocular estimates, presented in Table 46. But as it is to be expressed on the quarter-strip basis, the variance of the general mean of the whole-strips is to be divided by the square of 4. Therefore

$$V(2.840) = \frac{9.4168}{16(4^2)} \left(\frac{10 - 2}{10} \right) = 0.029428$$

to the quarter-strip, and corrected for the finite population of 10 strips to the block.

TABLE 46. Analysis of Variance of Ocular Estimate of the Whole-Strip Volume of Table 42

Source of variation	Degrees of freedom	Sum of squares	Mean square
Among blocks.....	7	447.4843	63.9263
Between strips, same block.	8	75.3347	9.4168
Total, among strips.....	15	522.8190	

Finally, then, the estimate of the variance of the adjusted volume, $Y = 2.645$ M feet b.m., which is expressed

$$V(Y) = V(a) + (x - \bar{x})^2 \left[V(b) \right] + b^2 \left[V(x) \right]$$

is, numerically,

$$\begin{aligned} V(2.645) &= 0.009704 + (2.840 - 2.632)^2(0.022978) \\ &\quad + (0.5571)^2(0.029428) \\ &= 0.009704 + 0.000994 + 0.009133 \\ &= 0.019831 \end{aligned}$$

on seven degrees of freedom. To the quarter-strip of half-acre in area, therefore, the best estimate from the data submitted is

$$2.645 \pm 0.1408 \text{ M feet b.m.}$$

on seven degrees of freedom, for which $t = 2.365$ corresponding to a probability of 0.05. Hence with probability of 0.95, the tract of 160 acres contains

$$320 \left[2.645 \pm 0.1408(2.365) \right] = 846 \pm 107 \text{ M feet b.m.}$$

of hardwood volume.

11.4 The Adjustment of Ocular Estimates of Correlated Populations. The sampling unit observations of the timber cruise are usually tallied according to certain pertinent species groups within mixed types. Accordingly, the adjustment of ocular estimate by the method of regression, as illustrated in the preceding section, when applied to species groups in combination as well as singly, involves not only the variances of the adjusted volumes of individual groups but also the covariance among them.

The data of the preceding section, taken according to the observational plan of Figure 25, were confined to the hardwoods of the pine-hardwoods timber-type. The complete volume record, however, is that

of pine as well as hardwoods, within the same sampling units, for which the quarter-strip data are presented in Table 47; the ocularly estimated volume (x) and the measured volume (y), are listed according to hardwoods (subscript H), and pine (subscript P).

The problem now is the adjustment of the ocular estimate of volume according to the species groups, singly and combined, for the entire tract of 160 acres.

TABLE 47. Concomitant Observations of Volume in M feet b.m. to the Quarter-Strip, in Pine and Hardwoods, According to Ocular Estimate (x) and Direct Measurement (y)

Block	Ocular estimate		Direct measurement	
	Hardwoods x_H	Pine x_P	Hardwoods y_H	Pine y_P
1	5.02	4.49	4.63	4.27
	4.03	2.87	3.82	2.97
2	1.45	1.13	1.63	1.22
	0.83	1.94	0.60	1.95
3	5.37	2.84	4.36	2.73
	2.67	3.49	3.38	3.30
4	1.99	1.16	1.67	1.43
	1.63	0.70	1.58	0.90
5	3.60	6.87	3.49	4.74
	4.52	3.60	4.07	3.73
6	1.01	3.17	0.97	2.34
	0.00	1.80	0.00	1.59
7	2.41	4.52	2.00	3.92
	3.67	4.12	3.55	3.49
8	1.39	1.71	2.38	1.36
	2.52	3.30	2.33	3.48
Sum	42.11	47.71	40.46	43.42
General mean	2.632	2.982	2.529	2.714

Having already the solution for hardwoods alone, it may seem that the problem merely implies the adjustment of the ocular estimates of pine by the same method. Such would, indeed, be the simplest solution if the main interest was centered in either group alone, with no more regard for the possible influence of other groups of timber upon that group than the cruiser normally has for the influence of grass, brush, or timber reproduction upon the merchantable volume of a given group.

As adjustment is to be applied to the ocular estimate of both groups together, as well as of each group separately, the three regression equations required should be made to express the adjustments in terms of the same independent variables; namely, the ocular estimate of hardwood, and of pine.

The method to be used contains the supposition that measured pine may be associated with the eye-estimate of hardwoods; and that measured hardwoods may be associated with the eye-estimate of pine. While such suppositions may seem to border on the ridiculous, since anyone can distinguish a pine from a hardwood at a glance, yet it is easy to imagine conditions for which the associations might be expected. If, instead of pine and hardwood groups, we were concerned with two groups of pine, say loblolly and shortleaf pines, and if the ocular estimator tended to confuse large shortleaf pine with loblolly pine, there would then be such an association.

In the present case the regression for adjusting the ocular estimate of hardwood is of the form

$$Y_H = \bar{y}_H + h_H(x_H - \bar{x}_H) + h_P(x_P - \bar{x}_P)$$

in which the regression coefficients are symbolized by h, since hardwood volume is the dependent variable; the subscripts thereto (H or P) referring to the associated independent variable (hardwoods or pine, as the case may be). The corresponding regression for adjusting the ocular estimate of pine, in terms of the same independent variables, is

$$Y_P = \bar{y}_P + p_H(x_H - \bar{x}_H) + p_P(x_P - \bar{x}_P)$$

in which the regression coefficients are symbolized by p, since pine is the dependent variable. The regression for adjusting the ocular estimate of hardwoods and pine together, is then the sum of these two, or

$$(Y_H + Y_P) = (\bar{y}_H + \bar{y}_P) + (h_H + p_H)(x_H - \bar{x}_H) + (h_P + p_P)(x_P - \bar{x}_P)$$

and this is in terms of the same independent variables. Consequently, only two of these three equations need be calculated.

Applications of the equations involve, at one stage or another, all the sum of squares and products, within blocks, among the four variables of Table 47. These are presented in Table 48. Each entry is, of course, based upon eight degrees of freedom.

In order to simplify the algebra, let x_H, x_P, y_H, and y_P denote deviations from block means. Then the coefficients in the first regression equation above, for which

$$(Y_H - \bar{y}_H) = h_H x_H + h_P x_P$$

may be derived from the two normal equations

$$\begin{aligned} h_H \overset{n}{S}(x_H^2) \quad &+ h_P \overset{n}{S}(x_H x_P) = \overset{n}{S}(x_H y_H) \\ h_H \overset{n}{S}(x_H x_P) &+ h_P \overset{n}{S}(x_P^2) \quad = \overset{n}{S}(x_P y_H) \end{aligned}$$

TABLE 48. Sums of Squares and Products within the Blocks, among the Quarter-Strip Observations of Table 47*

	x_H	x_P	y_H	y_P
x_H	6.7575	−0.4099	3.7644	0.5842
x_P		9.5863	−0.6924	5.5923
y_H			3.1839	−0.4203
y_P				4.5453

*Sums of squares are at intersections of rows and columns of like designation; sums of products among variables are at intersections of unlike designations.

Whenever the same set of values of the independent variables applies to more than one dependent variable, it is usually preferable to solve for the c-multipliers, which involve the independent variables only, and to use them to obtain the regression coefficients and the necessary variances.

This scheme has been used in Sec. 10.5. The theory behind it, for the case when the regression equation constant, a, is zero, is discussed in Sec. 10.5(B) of the Appendix. Further development which provides for the present case—in which a is not zero—is treated in Sec. 11.4 of the Appendix.

Solving first for c_{HH} and c_{HP}, we have

$$\left.\begin{aligned} c_{HH} \overset{n}{S}(x_H^2) \quad &+ c_{HP} \overset{n}{S}(x_H x_P) = 1 \\ c_{HH} \overset{n}{S}(x_H x_P) &+ c_{HP} \overset{n}{S}(x_P^2) \quad = 0 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots (4a)$$

and for the solution of c_{HP} (as a check) and c_{PP}

$$\left.\begin{aligned} c_{HP} \overset{n}{S}(x_H^2) \quad &+ c_{PP} \overset{n}{S}(x_H x_P) = 0 \\ c_{HP} \overset{n}{S}(x_H x_P) &+ c_{PP} \overset{n}{S}(x_P^2) \quad = 1 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots (4b)$$

whence the regression coefficients for the hardwood equation are obtained as follows:

$$\left.\begin{aligned} h_H &= c_{HH} \overset{n}{S}(x_H y_H) + c_{HP} \overset{n}{S}(x_P y_H) \\ h_P &= c_{HP} \overset{n}{S}(x_H y_H) + c_{PP} \overset{n}{S}(x_P y_H) \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots (5)$$

In the numerical work, one takes the sums of squares and products among x_H and x_P from Table 48. Then c_{HH} and c_{HP} are calculated from the equations (4a)

$$6.7575c_{HH} - 0.4099c_{HP} = 1$$
$$-0.4099c_{HH} + 9.5863c_{HP} = 0$$

whence

$$c_{HH} = 0.148369; \qquad c_{HP} = 0.006344.$$

The multipliers c_{HP} and c_{PP} are next calculated from the equations (4b)

$$6.7575c_{HP} - 0.4099c_{PP} = 0$$
$$-0.4099c_{HP} + 9.5863c_{PP} = 1$$

whence

$$c_{HP} = 0.006344; \qquad c_{PP} = 0.104587.$$

Now if one takes the sums of products, $\overset{n}{S}(x_H y_H)$ and $\overset{n}{S}(x_P y_H)$ from Table 48, the regression coefficients for the hardwood equation are, from equation (5), the following:

$$h_H = (0.148369)(3.7644) + (0.006344)(-0.6924) = \quad 0.55413$$
$$h_P = (0.006344)(3.7644) + (0.104587)(-0.6924) = -0.04853.$$

Thus the regression equation for the hardwood adjustment is

$$Y_H = 2.529 + 0.55413(x_H - 2.632) - 0.04853(x_P - 2.982)$$

on the quarter-strip basis; the means $\bar{y}_H$, $\bar{x}_H$, and $\bar{x}_P$ having been taken from Table 47.

The regression coefficients for the pine equation are calculated from the c-multipliers, and the sums of products $\overset{n}{S}(x_H y_P)$ and $\overset{n}{S}(x_P y_P)$ of Table 48. We have

$$p_H = (0.148369)(0.5842) + (0.006344)(5.5923) = 0.12215$$
$$p_P = (0.006344)(0.5842) + (0.104587)(5.5923) = 0.58859.$$

Then the regression equation for the pine adjustment is

$$Y_P = 2.714 + 0.12215(x_H - 2.632) + 0.58859(x_P - 2.982)$$

also on the quarter-strip basis. Finally, the regression equation for the adjustment of the combined groups is the sum of the two separate equations, so that

For the hardwoods:

$$Y_H = 2.529 + 0.55413(x_H - 2.632) - 0.04853(x_P - 2.982)$$

For the pine:

$$Y_P = 2.714 + 0.12215(x_H - 2.632) + 0.58859(x_P - 2.982)$$

For both groups:

$$Y_H + Y_P = 5.243 + 0.67628(x_H - 2.632) + 0.54006(x_P - 2.982).$$

The ocular estimates of hardwoods and pine, as taken on the eight random samples of whole-strips, are listed in Table 49. As these represent a separate and more accurate estimate of each independent variable, they are symbolized by X_H and X_P for hardwoods and pine, respectively. Upon inserting the quarter-strip means as given in the bottom line of the table—2.840 for hardwoods, and 2.446 for pine—for x_H and x_P, respectively, in each of the three regression equations, the adjusted volumes are the following:

For the hardwoods:	$Y_H = 2.670$ M feet b.m.
For the pine:	$Y_P = 2.424$ M feet b.m.
For both:	$Y_H + Y_P = 5.094$ M feet b.m.

Each of these is according to the quarter-strip of half-acre in area.

TABLE 49. Ocular Estimate of Volume in M Feet b.m., According to Hardwood and Pine Groups, on two Random Whole Strips in Each of Eight Blocks

Block	Hardwoods X_H	Pine X_P
1	19.10 16.50	15.37 10.32
2	8.74 5.12	7.61 7.70
3	21.22 12.56	10.54 13.17
4	6.88 7.51	7.70 2.72
5	16.56 19.93	11.54 8.27
6	6.22 0.96	6.46 5.23
7	9.60 13.42	19.03 15.84
8	8.06 9.40	3.41 11.62
Whole-strip means	11.361	9.783
Quarter-strip means	2.840	2.446

The next section treats of the estimates of the variances of these adjusted volumes.

11.5 Variances of the Adjusted Estimates. The three regression equations of the preceding section are of a common form,

$$Y = a + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) \ldots\ldots\ldots\ldots (6).$$

The variance of the calculated value—that is, $V(Y)$—in an equation of this form may be expressed in the same general form used in Sec. 10.5, that is,

$$V(Y) = s^2R + S^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

where s^2R is the contribution to $V(Y)$ ascribable to the sampling error of the regression equation itself; and S^2 is the contribution to $V(Y)$ ascribable to the sampling errors of the values x_1 and x_2 inserted into the equation.

From the development given in Sec. 11.5(A) of the Appendix, one may write

$$s^2\,R = s^2\left[\frac{1}{n} + c_{11}(x_1 - \bar{x}_1)^2 + c_{22}(x_2 - \bar{x}_2)^2 + 2c_{12}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)\right] \ldots (7a)$$

as the contribution ascribable to the sampling error of the regression function itself. Furthermore,

$$S^2 = b_1^2\left[V(x_1)\right] + b_2^2\left[V(x_2)\right] + 2b_1b_2\left[Cov(x_1x_2)\right] \ldots\ldots\ldots\ldots (7b)$$

is the additional contribution ascribable to the sampling errors of the values x_1 and x_2 inserted into the equation.

As each of the three regression equations of the preceding section are not only based upon the same independent variates, but as there were also inserted into them the same values—the quarter-strip means of the independently sampled ocular estimates—the numerical equivalents of $(x_1 - \bar{x}_1)$ and $(x_2 - \bar{x}_2)$ of equation (6) are, in each case,

$$(x_H - 2.632) = (2.840 - 2.632) = \quad 0.208$$

and

$$(x_P - 2.982) = (2.446 - 2.982) = -0.536,$$

respectively. Furthermore, c_{11}, c_{12}, and c_{22} of equation (7a) are c_{HH}, c_{HP}, and c_{PP} of the preceding section; and, finally, since each of the three regression equations were calculated from the data of the same 16 sampling units of quarter-strips, $n = 16$. Consequently, the numerical value

of R of equation (7a) applies equally to the three regression equations; hence

$$R=\frac{1}{16}+(0.148369)(0.208)^2+(0.104587)(-0.536)^2 \\ +2(0.006344)(0.208)(-0.536) \\ =0.097551.$$

This leaves the coefficient of R—that is, s^2—alone to be determined in order to arrive at the contribution to $V(Y)$ of the sampling errors of the regression equations. In equations (7) and (7a), s^2 denotes the mean square of the residuals of unit weight about the regression equation under consideration. With the three regression equations there are, then, three separate equivalents of s^2, one with respect to each equation. These are derived from the sums of squares and products of residuals about the hardwood and pine regression equations, as listed in the middle line of Table 50. Here the sum of squares of the hardwood residuals is the numerical value of

$$\overset{n}{S}\left[(y_H-h_H x_H-h_P x_P)^2\right]$$

in which the variables are taken as deviations from block means. Upon expanding and simplifying according to Sec. 10.4(B) of the Appendix, this may be written

$$\overset{n}{S}(y_H^2)-h_H\overset{n}{S}(x_H y_H)-h_P\overset{n}{S}(x_P y_H)$$

for which the numerical equivalents of the summations are taken from Table 48.

In like manner, the sum of squares of pine residuals independent of the pine regression is

$$\overset{n}{S}\left[(y_P-p_H x_H-p_P x_P)^2\right]=\overset{n}{S}(y_P^2)-p_H\overset{n}{S}(x_H y_P)-p_P\overset{n}{S}(x_P y_P).$$

The sum of products of corresponding residuals about the regressions

$$\overset{n}{S}\left[(y_H-h_H x_H-h_P x_P)(y_P-p_H x_H-p_P x_P)\right]$$

contains the covariance of the hardwood-pine residuals. Upon expansion and simplifying according to Sec. 11.5(B) of the Appendix, it may be expressed in either of the following forms:

$$\overset{n}{S}(y_H y_P)-h_H\overset{n}{S}(x_H y_P)-h_P\overset{n}{S}(x_P y_P)$$

$$\overset{n}{S}(y_H y_P)-p_H\overset{n}{S}(x_H y_H)-p_P\overset{n}{S}(x_P y_H).$$

TABLE 50. Division of Sums of Squares and Products of Measured Volume within Blocks, of Hardwood (y_H) and Pine (y_P), into Portions Due to, and Independent of, Regression on Ocular Estimate of Volume of the Same Quarter-Strips

Due to	Degrees of freedom	Sum of squares		Sum of products
		y_H^2	y_P^2	$y_H y_P$
Regression on x_H and x_P..	2	2.1196	3.3629	0.0523
Residuals.....	6	1.0643	1.1824	−0.4726
Total, within blocks.....	8	3.1839	4.5453	−0.4203

With six degrees of freedom, then, the mean squares, s^2, may be calculated at once; accordingly,

For the hardwood: $s^2 = \frac{1}{6}(1.0643) = 0.1774$

For the pine: $s^2 = \frac{1}{6}(1.1824) = 0.1971$

For both: $s^2 = \frac{1}{6}\left[1.0643 + 1.1824 + 2(-0.4726)\right] = 0.2169.$

The numerical equivalents of S^2, appropriate to the three regression equations, are next required in order to complete the estimates of the variances of the adjusted volumes. Each S^2 contains the regression coefficients, which are already available, together with the quarter-strip variances and covariance among the whole-strip means. These latter are calculated from the whole-strip data of Table 49, the sums of squares and products within blocks, on the whole-strip basis, being the following:

$$\overset{n}{S}(X_H^2) = 75.3347; \qquad \overset{n}{S}(X_P^2) = 73.5069;$$

$$\overset{n}{S}(X_H X_P) = -9.4218.$$

As these are each based upon eight degrees of freedom among the 16 whole-strip observations, the estimates of the variances and covariance of the means *to the quarter-strip*, including the correction for samples of two random sampling units from blocks of 10 are as follows:

$$V(X_H) = \frac{75.3347}{8(16)(4^2)}\left(\frac{10-2}{10}\right) = \quad 0.029428;$$

$$V(X_P) = \frac{73.5069}{8(16)(4^2)}\left(\frac{10-2}{10}\right) = \quad 0.028714;$$

$$Cov(X_H X_P) = \frac{-9.4218}{8(16)(4^2)}\left(\frac{10-2}{10}\right) = -0.003680;$$

whence, for the numerical equivalents of S^2 in equation (7b), For the hardwoods:

$$\begin{aligned} S^2 &= (0.55413)^2(0.029428) + (-0.04853)^2(0.028714) \\ &\quad + 2(0.55413)(-0.04853)(-0.003680) \\ &= 0.009302; \end{aligned}$$

For the pine:

$$\begin{aligned} S^2 &= (0.12215)^2(0.029428) + (0.58859)^2(0.028714) \\ &\quad + 2(0.12215)(0.58859)(-0.003680) \\ &= 0.009858; \end{aligned}$$

For both:

$$\begin{aligned} S^2 &= (0.67628)^2(0.029428) + (0.54006)^2(0.028714) \\ &\quad + 2(0.67628)(0.54006)(-0.003680) \\ &= 0.019146. \end{aligned}$$

Upon assembling the values for the solution of the estimates

$$V(Y) = s^2 R + S^2$$

appropriate to each of the three regression equations,

$$\begin{aligned} V(Y_H) &= V(2.670) = 0.1774(0.097551) + 0.009302 = 0.026608 \\ V(Y_P) &= V(2.424) = 0.1971(0.097551) + 0.009858 = 0.029085 \\ V(Y_H + Y_P) &= V(5.094) = 0.2169(0.097551) + 0.019146 = 0.040305 \end{aligned}$$

The square roots of these variances are the estimates of the standard errors of volume in M feet b.m. to the quarter-strip of half-acre. Each is based upon the six degrees of freedom appropriate to the estimate of s^2 of equation (7a). To the half-acre, then, the estimates of the population means are

For hardwoods:	2.670 ± 0.163 M feet b.m.
For pine:	2.424 ± 0.171 M feet b.m.
For both:	5.094 ± 0.201 M feet b.m.

while the volumes on the entire tract of 160 acres are 320 times these.

11.6 Reconciliation of the Conflicting Requirements of Mapping and Sampling in Forest Surveys. The two major objectives of forest surveys of many properties are the estimation of the timber volume and the construction of a contour map. Field work is commonly carried out according to the plan known as the *two-run map-cruise*. Each 40-acre square is traversed by two lines, at 10-chain interval, with the aid of staff-compass, Abney hand-level, and 2-chain trailer tape. These lines serve the twofold purpose: (1) to establish locations

and elevations for fitting contours on a sketch-map of the area traversed, to 5 chains on either side of the line, and (2) to determine for the timber cruiser—a member of the party—the location of sampling units of timber volume.

In certain surveys the sampling unit is the continuous strip, usually one chain wide, centered along the survey line; in others it is the sample plot, perhaps ¼-acre in area, and a series of such plots are located at uniform distances along the survey lines. Consequently, the sampling units describe a *systematic* pattern of strips, or line-plots, on the map of the property in question.

Under such conditions the probable discrepancy between the volume as estimated from the direct measurements and the corresponding true but unknown volume cannot be assessed unequivocably; for the mathematical requirements for the solution of the problem of probable dis-

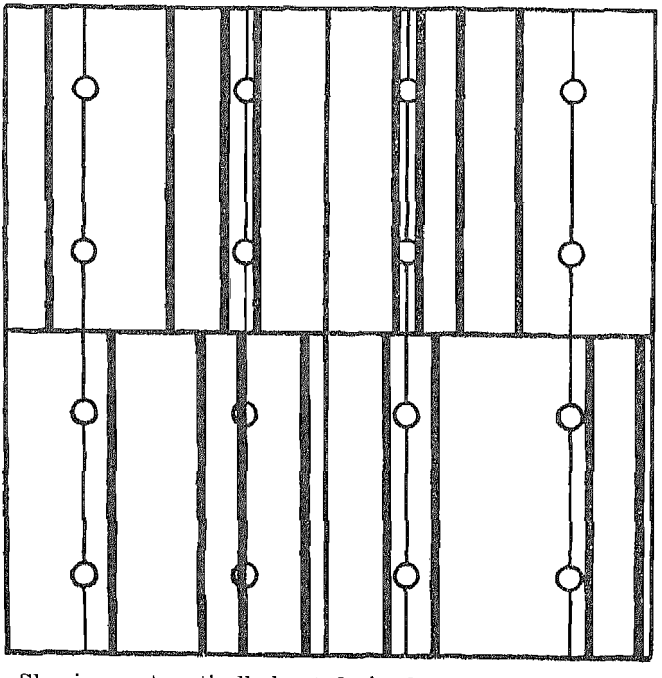


FIG. 26. Showing systematically-located circular sample plots along survey lines, 10 chains apart; and the location of four random strips, ½-chain wide, in each block. The data of the circular plots are used only for the regression of volume on basal area.

crepancy imply that the constituent parts upon which sampling error is to be based be located independently and at random.

Now it has frequently been urged, and it is rather generally accepted, that when the timber cruise and map construction are joint projects, the exigencies of the latter leave no practical alternative to the systematic pattern of sampling units.

Regression, however, offers the opportunity to reconcile the opposing requirements. One scheme is illustrated in Figure 26. In this case there are four 40-acre squares, with two lines at 10-chain interval across each. If the circles represent systematically located sample plots upon which volume has been measured, basal area will have been measured as well. These plots contain the materials for the regression of volume on basal area. Furthermore, if as many *random* sampling units as feasible —pictured in Figure 26 as strips, ½-chain wide—are run in each 40, these supply the independent set of samples of basal area, with which the regression equation is entered.

Sampling designs which are efficient and at the same time adapted to the practical requirements of a forest survey, have rarely, if ever, been tried. Yet there is every reason to suppose a great variety of them waiting to be explored.

CHAPTER XII

ON CERTAIN PRACTICAL ASPECTS OF SAMPLING

12.1 Definition of Sampling Objectives. Before starting the field work for a sample inventory of any considerable population it is essential that objectives be clearly defined, and that definitions of terms appropriate thereto be specific and free from ambiguity of interpretation.

If, for instance, information is needed on the present stocking of two square miles of area that were planted to slash pine several years previously, the general objective "information on stocking" is, in itself, not very clear. It is not sufficiently free from misinterpretation on the part of the field men. Should the latter confine their observations to planted slash pine, the samples would contain nothing concerning the number of natural seedlings.

More specifically, the objectives of the sample inventory might be as follows:

"To obtain an estimate of the number of seedlings to the acre of (1) planted slash pine, (2) natural slash pine, and (3) natural longleaf pine. Assurance is required that the estimated number of planted slash pine be within 10 percent of the true number, with probability of 0.95."

When several correlated populations are to be sampled simultaneously, as in this case, the standard of precision is usually referred explicitly to one of them, or a combination of two or more. Others then fall in line, their precisions depending upon their frequencies of occurrence, and their variances and covariances of unit weight. In this case, planted slash pine has been selected to bear the test of precision.

The next step is to learn what one can, perhaps by preliminary reconnaissance, concerning the population to be sampled. Such information is often helpful in sampling design.

12.2 Bias. A constant error that affects all observations alike is called bias. Its magnitude is not lessened with increase in sample size, for it may be encountered in the complete enumeration as well as in the sample survey.

Bias may be introduced into measurements through instrumentation, the personal equation, or instability of the population being sampled.

Bias of instrumentation is the effect either of improper use of an instrument, or use of an instrument not in adjustment. Thus if the diameters of trees are measured with a diameter tape and reasonable

care is not exercised in holding the instrument horizontally, diameters will be overestimated. Bias of this kind, improper use of an instrument, can readily be corrected by training in attention to detail. If, however, diameters are measured with a caliper that is out of adjustment, bias is again introduced into all diameter measurements. This same type of error may be encountered in the use of a hypsometer, or, indeed, with one or more volume tables which do not apply to the timber at hand; for the volume table itself may be regarded as an intrument. Such errors can be eliminated by frequent instrumental checks.

The personal equation as a source of bias appears frequently in the estimate of timber volume. Should the volume on sampling units be recorded according to eye-estimate, it is very likely that the cruiser will consistently over- or under-estimate the real volume, unless he is familiar with the tree-form, cull, and utilization practices of the region. It was shown in preceding chapters that systematic error of ocular estimation may be eliminated through the method known as regression. The practice of check cruising, however, goes a long way toward the elimination of another common systematic error of ocular estimate—that of judging the boundaries of sampling units, and recording, in consequence, estimates of volume, the errors of which vary directly with the corresponding errors of sampling unit areas.

The third type of bias, instability of the population, requires careful consideration. For instance, the viability of seed sown in most nurseries is estimated from samples of the seed prior to sowing time. But viability tests are carried out under conditions more nearly ideal than are encountered in the field. They are usually performed in the greenhouse. In seedbed sowing, then, due allowance should be made for the effects of differences in growing conditions as well as of storage on the viability of the seed.

The forest-tree nursery problem serves to illustrate the instability of certain populations. The sample census of a nursery, or parts thereof, supplies an estimate of the number of seedlings available for distribution. This estimate, however, is made a month or so before the stock is lifted. In the interim, some of the seedlings, culled because of size, may have become plantable. Insofar, then, as the sample census is taken as a forecast, it is subject to the error of forecast as well as to sampling error.

Tree-growth data are often collected as part of the sample inventory of forest properties, primarily by means of increment cores. But the use of the increment borer is time-consuming; hence it is not feasible to bore all trees on the sampling units. Sub-samples of trees are therefore

used in determining growth. How, then, should each be taken? Should the borings be taken on all trees of a sub-plot, or will it do as well to take a core from one tree to the plot, let us say that one which is nearest the plot center? If the latter alternative is chosen, a higher proportion of measurements from relatively open-grown trees will fall in the samples than occurs in the population, and an over-estimate of growth would be the result.

One is occasionally tempted to discard increment core measurements of certain hardwood species with indistinct annual rings. But slow growth is often associated with difficulty in measuring it; hence the resulting bias of the practice is apparent.

When increment core measurements have been taken properly, the growth within recent decades on trees of various sizes may be estimated unambiguously. However, the forecast of future growth from that of the past must rest upon certain assumptions, and it will be markedly influenced by the particular assumptions chosen.

12.3 Size, Shape, and Structure of Sampling Units. Two classes of sampling units have been defined (Sec. 3.2) as follows:

Ultimate unit. The smallest plot, or area, that is not subdivided.

Random sampling unit. A constituent part of the sample which is drawn independently and at random. It consists of one or more ultimate units.

If the population to be sampled is homogeneous, there might be little reason for choosing one particular size, or shape, of either the ultimate unit or the random sampling unit, in preference to any other. As a general rule, however, the populations of forest and field are heterogeneous to such an extent that the shape of the ultimate unit and the structure and size of the random sampling unit may easily affect the precision of the work. In choosing these two units one should make use of all information available concerning the pattern and causes of variation within the population.

Imagine, for example, a nursery seedbed, 4 by 100 feet in dimension, for which the number of plantable seedlings is known for each square-foot ultimate unit of the 400. Variation in the yields of the ultimate units may be due to

(1) Variation across the bed;

(2) Variation along the bed;

(3) Residual variation.

Now it is known from the nurseryman's experience that there are more seedlings to the square foot in the interior of the bed than there are near

the bed boundaries, though they may not be the sturdiest stock. Hence variation in number of plantable seedlings across the bed is commonly of considerably greater magnitude than along its length; and the logical random sampling unit is the narrow plot of four ultimate units extending across the bed.

It thus appears that the greater the variation among ultimate units within the random sampling units, the less, in general, becomes variation among random sampling units themselves.

In the forest plantation, as another example, variation in survival of individual trees may be assigned, with more or less accuracy, to

(1) The larger subdivisions, or blocks;
(2) Topographic position within blocks:
(3) Care of planting;
(4) Residual variation.

Planting of nursery-grown stock is commonly done by crews of 10 to 15 men, each crew member being assigned a row, in a set of 10 to 15 parallel rows across the principal drainage.

Suppose a 40-acre tract had been planted in parallel rows six feet apart, and also with 6-foot spacing between trees of the same row. Assuming a square pattern, the plantation would contain 220 rows of 220 trees in each row. If planted by an 11-man crew, the crew as a whole would have planted 20 sets of rows, of 11 rows to the set. Thus the variation in survival between sets is due primarily to heterogeneity of soil fertility and moisture, and the occurrence of competing vegetation among the 2-acre subdivisions upon which the sets had been planted. The 20 sets may therefore be taken as 20 separate blocks, and if a sample is drawn from each and every set, variation among sets is completely eliminated from the estimate of the number, or proportion, of survivors as well as from its sampling error.

Variation among the 11 rows within each set, however, is due primarily to care of planting on the part of the individual planters, as each has contributed one row to the set. Variation among trees of the same row, on the other hand, is assignable to topographic position within the set, as each row extends alike across the drainage.

The problem is to define the ultimate unit, and the random sampling unit, so as to eliminate, in so far as practicable, these two sources of variation from the estimate of survival, and from its sampling error. One might, for example, define the ultimate unit as an area 6 by 66 feet extending across the rows of a set, so as to include one tree planted by each crew member. Effects of variation in care in planting would thus be

eliminated. The ultimate unit observation would then be the number of survivors out of 11 trees planted.

If marked changes in topography occur from one end of a set to another, much of the variation among ultimate units of the same set is assignable to topographic position. And this should also be eliminated, in so far as practicable, by sampling design.

Suppose it were feasible to make direct observations on 10 ultimate units from among the 220 in each set. As one alternative, then, we might conceive the ultimate unit and the random sampling unit as identical. Under this condition, a single random sample of 10 units from each set would contain all the variation due to topography. Obviously it would be preferable to eliminate most of this by dividing each set into five blocks of 44 units each, and then draw an independent random sample of two units from each block. Sampling the entire plantation in this manner would involve 100 samples of two sampling units each, and the sampling error would be founded upon one degree of freedom from each block, or 100 in all.

Should such a procedure seem somewhat too detailed, one might define the random sampling unit as the sum of the five ultimate units having the same ordinal number, over the five blocks of a set. Thus if 6 and 32 represent two random numbers out of 44, the first random sampling unit is the sum of the ultimate units number 6 over the five blocks of a set. By this device the variation assignable to topography is mostly within the random sampling units, and not between them. The sampling job of the plantation would then involve a sample of two random sampling units from each of the 20 sets, and the sampling error of the number, or proportion, of survivors would rest upon 20 degrees of freedom.

Sometimes it may happen that the sampled areas are irregular in shape, such as those shown in Figure 18. For that illustration the ultimate unit was a square of 1/20-inch on the side, and the random sampling unit was a line of these ultimate units. The resulting inequality in weight was due to variation in the number of ultimate units to the random sampling unit.

A certain amount of generalization may be drawn from the above discussion. Sampling units should be formed in such a manner as will eliminate as much heterogeneity as practicable—to be accomplished by long, narrow plots that extend across variability trends, or by using more or less complex random sampling units the ultimate units of which lie at various intervals along this trend. If the population were strictly homogeneous, the shape of the ultimate unit would not affect the accuracy of the results.

Irregular areas can be sampled in the same manner as rectangular areas. The disproportionality in weighting of random sampling units arises, not on account of variation in *size* of ultimate unit—for ultimate units are always constant in size—but from variation in the *number* of ultimate units to the random sampling unit.

12.4 The Sample. A sample may consist of a set of random sampling units drawn from the supply of such units in the population as a whole; or, in representative sampling (Sec. 4.1) from each stratum, block or sub-population into which the general population has been separated.

In all the illustrations thus far used, the sample was drawn in such a way that each and every part of the population, or sub-population, had an equal chance of being included in the sample. Insofar as this condition of randomization is fulfilled, the sample statistic—such as the mean—supplies not only an unbiased estimate of the corresponding characteristic of the population, but it also supplies a valid estimate of the probable discrepancy between the true, but unknown, population characteristic and the sampling estimate thereof.

It may be argued that the purposive choice of such sampling units which, by eye-estimate, seem to contain better approximations to the population characteristic should also supply a better estimate of it than is contained in any random sample. If the sampler has had considerable experience with particular kinds of populations, and is not subject to personal bias, he may, indeed, be very successful in sampling them by purposive choice. But should the sample of purposive choice be considered a random sample, the hypothetical sampling error calculated therefrom would foster overconfidence by its abnormally low value. For the very high—and the very low—sampling unit observations would have been denied the chance of inclusion in the sample.

On the other hand, certain practical sub-sampling designs may impose restrictions whereby a large proportion of the population be denied the chance to appear in the samples. The significance of such restrictions is, perhaps, best shown by a concrete illustration. A county of North Carolina contains 280 square miles of farmland and woodland. This area was sampled so as to provide an estimate of its forest area and timber volume according to major forest types. To start with, an excellent map of the county was available. Based upon an aerial survey, it portrayed the location of farm buildings, schools, churches and highways, as well as the network of secondary and woods roads. One can go readily to any designated point marked on the map.

Each of the 280 square miles was taken as a block, and each block

was conceived as subdivided into 64 square plots—as on a chessboard—of 10 acres. Two of the 64 plots of each block were then drawn, independently and at random, and examined in the field to ascertain timber volume and subdivision of land area thereon, according to major forest type. Each random sampling unit of 10 acres had, of course, precisely the same chance of making the samples.

Each of the selected random sampling units, however, was not examined throughout its extent, but was *systematically sub-sampled* by confining the direct measurements to two parallel strips, of ½ by 10 chains in dimension, and 5 chains apart. Thus each sub-sample was made up of but 1 acre out of the 10 of the plot, its location within the plot fixed by the sampling design; and the remaining 9 acres were simply denied any chance whatever of being included in the observations representing the random sampling unit concerned.

It has been assumed, in this case, that failure to randomize the sub-samples can introduce only a negligible bias to the estimates and to their sampling variances. Should the assumption have been considered unwarranted, the possibility of bias might have been completely eliminated, upon selecting a random sub-sample of 1 acre, from among the 10 acres of each plot, independently. In either case the variance of the general mean is contained in the mean square between plots within the blocks.

The use of circular plots in sampling forest and field populations is another example of systematic sub-sampling of random sampling units. Thus while the latter unit might be an area of 2 chains square, there are practical advantages in confining observations to the inscribed circle of 3.1416 square chains, or 78.4 percent of the entire random sampling unit. Any loss in information which might adhere to this scheme should, indeed, be more than recovered by the additional number of circular plots it may be made to provide.

The importance of random selection lies in the fact that the sample supplies all the information necessary to evaluate its own accuracy. One or more systems of sampling, more accurate, perhaps, than random sampling, might be designed; but unless the sampling distributions of their statistics are known, the only way to test accuracy is by comparison with the complete canvass, that is, with corresponding population parameters. This can be done for a few populations only. Thus any use of statistics obtained from a system of sampling, other than random sampling, must be predicated on the similarity between the population sampled, and others for which an accurate check is at hand. Samples collected by such procedures do not, therefore, supply all of the informa-

tion necessary for evaluating sampling reliance, and inferences drawn from them are weakened.

12.5 The Determination of Sampling Intensity. The amount of time and funds to be applied to sampling any considerable population is usually preassigned. It is then the sampler's job to provide a sampling design which will provide the maximum amount of information as well as an estimate of the precision of the sample statistics.

Often, however, it is required to estimate in advance the amount of time and funds which will purchase statistics of given precision. Thus if the variance of unit weight in a population is known, the expression of the variance of the mean of weight n is

$$\frac{\sigma_y^2}{n}$$

and this may be used to determine n. Should one strive for a standard error of the mean of y equal to 10 percent of the mean itself, then

$$V(\bar{y}) = (0.1\bar{y})^2.$$

Upon substituting this in the expression above, and solving for n,

$$n = \frac{\sigma_y^2}{(0.1\bar{y})^2}.$$

Obviously this expression calls for an estimate both of σ_y^2 and $\bar{y}$.

A rough estimate of the standard deviation may be had by making use of an ocular estimate of the range, obtained, perhaps, in a preliminary reconnaissance of the population. In the normal curve of error the range which encloses 99 percent of the distribution is 2.576σ on either side of the mean, where 2.576, taken from Table 7, is t at the 1 percent level when based upon any considerable number of degrees of freedom.

One may, accordingly, take the range, R, to be $2(2.576\sigma)$. If, then, a rough estimate of R is obtained by direct observation,

$$\frac{R}{5.152} \text{ is a rough estimate of } \sigma.$$

By way of illustration, the random sampling unit of the plantation population of Sec. 12.3, contains 55 trees, divided among the survivors and the dead. In counting the survivors, according to the random sampling unit, the range in possibilities is from 0 to 55. Hence a rough estimate of σ may be taken to be

$$\frac{55}{5.152} = 10.7.$$

Such rough schemes should be used only in the absence of better information.

In order to complete the estimate of what sampling intensity to apply, there is needed an estimate of the mean number of survivors in 55 trees planted. This estimate may be had in the same preliminary reconnaissance which includes the location of the plantation boundaries. Suppose one guesses the survival to be between 35 and 55 percent, that is, between 19 and 30 survivors to the random sampling unit of 55 trees. If the standard error of the mean is to be of the order of 10 percent of the mean, then, roughly,

$$SE(\bar{y}) \text{ is to be between 2 and 3 trees,}$$

whence the number of random sampling units to be included in the 20 sets of samples—one sample from each block—lies between

$$\left(\frac{10.7}{2}\right)^2 \text{ and } \left(\frac{10.7}{3}\right)^2$$

or between 13 and 29. One judges, therefore, that if each of the 20 samples is made up of two random sampling units, the resulting precision will be at least as good as expected.

12.6 Allocation of Costs in Double Sampling. The use of regression is to be recommended in sampling whenever (1) the direct measurement of the variate, y, whose mean is to be estimated, is relatively costly; and (2) the variate y is associated with, or dependent upon, another variate, x, which costs relatively little to observe, or measure directly.

Under these conditions comparatively few observations are taken on the 2-variate population of (y, x), and the regression of y on x is calculated therefrom. Then a random sample, or a set of random samples, is drawn from the population of x, and the general mean of x is inserted into the regression equation, the solution affording the best estimate of the mean of y under the conditions.

The decision to employ the method of double sampling implies that the cost of the field work is to be allocated between the two parts of the job. Suppose, by way of illustration, that a sample census of the plantable seedlings of a given species in a nursery is required. The regression of the number of plantables, y, according to the sampling unit, on the entire number of seedlings, x, of the form used in Sec. 10.3, was

$$Y = bx$$

for which

$$\sigma_b^2 = \frac{\sigma_{y \cdot x}^2}{n\bar{x}}$$

based, in that case, on 54—or, in general, on n—sampling units from the 2-variate population of (y, x). There is substituted for x, in the re-

gression equation, the general mean of a set of random samples from the population of x, based upon m random sampling unit observations over the entire set. Thus if x' denotes this general mean,

$$V(x') \rightarrow \frac{\sigma_x^2}{m}.$$

The best estimate, Y, of the mean of y is then

$$Y = bx',$$

and the variance of this estimate may be expressed (Sec. 10.3) as

$$\sigma_Y^2 = \frac{b^2\,\sigma_x^2}{m} + \frac{(x')^2\,\sigma_{y\cdot x}^2}{n\bar{x}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1).$$

Now it should be kept in mind that, at best, the allocation of sampling costs rests upon one's judgment and experience with the particular kind of population concerned. In the present case, it seems reasonable that any difference between the mean of x in the regression sample—that is, $\bar{x}$—and the general mean of the independent random samples of x—that is, x'—should be negligible. If, then, $\bar{x}$ be substituted for x', one may, for present purposes, simplify equation (1) to the following:

$$\sigma_Y^2 = \frac{b^2\,\sigma_x^2}{m} + \frac{\bar{x}\,\sigma_{y\cdot x}^2}{n} \text{ approximately} \ldots\ldots\ldots\ldots (1a).$$

Consider next the second phase of the problem. The field cost of sampling, say T, is to be distributed between (1) the n observations on the relatively costly sampling for the regression, the charge for which, is, say, c_n to the sampling unit observed; and (2) the m observations on the relatively inexpensive work of collecting the independent set of random samples of the independent variate, x, at a cost of c_m to the random sampling unit observed. It follows, then, that

$$T = nc_n + mc_m \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2).$$

The values m and n are to be chosen, so that equation (1a) is minimum, subject to the condition of equation (2). From the latter equation

$$m = \frac{T - nc_n}{c_m},$$

and substituting this for m in equation (1a), the latter may be expressed

$$\sigma_Y^2 = \frac{b^2\,\sigma_x^2\,c_m}{T - nc_n} + \frac{\bar{x}\,\sigma_{y\cdot x}^2}{n} \text{ approximately.}$$

The value of n which makes this a minimum is required. Upon differentiating with respect to n, and equating to zero, it follows that

$$\frac{c_m c_n b^2 \sigma_x^2}{(T - nc_n)^2} - \frac{\bar{x}\sigma_{y\cdot x}^2}{n^2} = 0$$

and after substituting mc_m for $(T-nc_n)$, this may be expressed as

$$\frac{m}{n}=\sqrt{\frac{c_n}{c_m}}\left(\frac{b\sigma_x}{\sigma_{y\cdot x}\sqrt{\bar{x}}}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots (3).$$

It is now apparent that the most efficient allocation of costs rests to a large extent upon experience with the particular kind of population concerned, since the numerical equivalents of the factors within the parenthesis are to be obtained only after the completion of the job. But budgeting of funds requires some advance guess of their expected values.

The forester in charge of the sample census in Sec. 10.3 estimated the parenthesized factors of equation (3) to be $\frac{2}{3}$. From past experience, he knew that

$$c_n=15 \text{ and } c_m=2$$

to be a satisfactory approximation, so that by applying equation (3) he arrived at

$$\frac{m}{n}=\frac{2}{3}\sqrt{\frac{15}{2}}=2 \text{ approximately.}$$

This turns out to be not the best ratio, however, for the observed equivalents of the parenthesized factors of equation (3) obtained *after* the completion of the field work, taken from Sec. 10.3, are the following:

$$116 \rightarrow \sigma_x^2; \qquad 0.81 \rightarrow b;$$
$$0.58 \rightarrow \sigma_{y\cdot x}^2; \qquad \frac{2{,}413}{54}=45=\bar{x};$$

whence

$$\frac{b\sigma_x}{\sigma_{y\cdot x}\sqrt{\bar{x}}}=\sqrt{\frac{(0.81)^2(116)}{(0.58)(45)}}=1.7$$

from which it appears that

$$\frac{m}{n}=1.7\sqrt{\frac{15}{2}}=5 \text{ approximately.}$$

Thus a set of random samples containing, in all, five times as many random sampling units as the number of sampling units in the regression sample—instead of twice as many, as actually obtained—might have been a happier choice.

Another type of adjusting equation, applicable to timber cruising, is the regression of measured volume on either basal area, or the eye-estimate of the same volume, for which

$$Y=a+b(x'-\bar{x}).$$

The variance of a and of the inserted mean (x') of the independent set of samples of x, are, in this case,

$$\sigma_a^2 = \frac{\sigma_{y\cdot x}^2}{n}\ ; \qquad V(x') = \frac{\sigma_x^2}{m}\ ;$$

for which, as before, n is the number of sampling units of the regression, and m is the total number of random sampling units upon which x' is based. The variance of Y as estimated from the equation is, from Sec. 8.5, the following:

$$\sigma_Y^2 = \frac{\sigma_{y\cdot x}^2}{n} + b^2\,\frac{\sigma_x^2}{m} + (x' - \bar{x})^2\,\sigma_b^2.$$

In practice, the third term of the right-hand member is discarded as it should be close enough to zero to be negligible. Then

$$\sigma_Y^2 = \frac{\sigma_{y\cdot x}^2}{n} + \frac{b^2\,\sigma_x^2}{m} \text{ approximately} \ldots\ldots\ldots\ldots (4).$$

If, as before,

$$T = nc_n + mc_m$$

one may replace m of equation (4) by

$$\frac{T - nc_n}{c_m}$$

and (4) may be expressed as follows:

$$\sigma_Y^2 = \frac{\sigma_{y\cdot x}^2}{n} + \frac{b^2\,\sigma_x^2\,c_m}{T - nc_n} \text{ approximately.}$$

Upon equating the first derivative of this (with respect to n) to zero and then substituting mc_m for $(T - nc_n)$, as before, one may express the result as follows:

$$\frac{m}{n} = \sqrt{\frac{c_n}{c_m}}\left(\frac{b\sigma_x}{\sigma_{y\cdot x}}\right).$$

Experience in timber cruising indicates that the numerical equivalent of the parenthesized factor is 2, approximately, provided the regression is one of measured volume on basal area of the same sampling units. Should the regression equation express measured volume in terms of eye-estimate of the same volume, the parenthesized factor varies from 1 to 4, depending upon the accuracy and consistency of the eye-estimate of volume.

APPENDIX

Technical Notes

The section numbers correspond to the sections of the text wherein reference is first made to this Appendix.

3.4 The Sampling Variance of the Mean of a Random Sample of n Values from a Finite Population of N. Let each of the n observations be expressed as errors, that is, let each

$$(y - \mu) = \epsilon$$

be a deviation from the population mean. Then the mean error of a sample of size n is

$$\bar{\epsilon} = \frac{1}{n}(\epsilon_1 + \epsilon_2 + \ldots + \epsilon_i + \ldots + \epsilon_j + \ldots + \epsilon_n),$$

and its square is

$$(\bar{\epsilon})^2 = \frac{1}{n^2}(\epsilon_1 + \epsilon_2 + \ldots + \epsilon_i + \ldots + \epsilon_j + \ldots + \epsilon_n)^2$$

$$= \frac{1}{n^2}\left[\overset{n}{S}(\epsilon_i^2) + 2\,\overset{{}_nC_2}{S}(\epsilon_i\epsilon_j)\right] \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

in which

$${}_nC_2 = \frac{n!}{(n-2)!\,2!}$$

Now in a finite population of N values of ϵ there are ${}_NC_n$ possible values of $\bar{\epsilon}$, each based upon n observations. Upon summing (1) over all ${}_NC_n$ possibilities, we have

$$\overset{{}_NC_n}{S}(\bar{\epsilon})^2 = \frac{1}{n^2}\left\{\overset{{}_NC_n}{S}\left[\overset{n}{S}(\epsilon_i^2)\right] + 2\,\overset{{}_NC_n}{S}\left[\overset{{}_nC_2}{S}(\epsilon_i\epsilon_j)\right]\right\} \ldots\ldots\ldots (2).$$

The first term within the braces of equation (2) is the sum of $({}_NC_n)\,(n)$ values of ϵ_i^2. Since there are only N distinct values of ϵ, some of these have been used more than once. They have, in fact, been used $\frac{({}_NC_n)\,(n)}{N}$ times. Thus one may write

$$\overset{{}_NC_n}{S}\left[\overset{n}{S}(\epsilon_i^2)\right] = \frac{n}{N}({}_NC_n)\overset{N}{S}(\epsilon_i^2) \ldots\ldots\ldots\ldots\ldots (3).$$

Furthermore, the second term within the braces of equation (2) may be expressed

$$\overset{{}_NC_n}{S}\left[\overset{{}_nC_2}{S}(\epsilon_i\epsilon_j)\right] = \frac{({}_nC_2)\,({}_NC_n)}{{}_NC_2}\,\overset{{}_NC_2}{S}(\epsilon_i\epsilon_j) \ldots\ldots\ldots (4)$$

because it is the sum of $({}_nC_2)({}_NC_n)$ values, there being ${}_NC_2$ possible combinations of $(\epsilon_i \epsilon_j)$, and each product is used $\frac{1}{{}_NC_2}({}_NC_n)({}_nC_2)$ different times.

Upon putting the forms (3) and (4) into equation (2),

$$\overset{{}_NC_n}{S}\left[(\bar{\epsilon})^2\right] = \frac{1}{n^2}\left\{\frac{n}{N}({}_NC_n)\overset{N}{S}(\epsilon_i^2) + 2\frac{({}_nC_2)({}_NC_n)}{{}_NC_2}\overset{{}_NC_2}{S}(\epsilon_i\epsilon_j)\right\}$$

the average of which is the exact value of the variance of means of weight n. This average is

$$\frac{1}{{}_NC_n}\overset{{}_NC_n}{S}\left[(\bar{\epsilon})^2\right] = \frac{1}{n^2}\left\{\frac{n}{N}\overset{N}{S}(\epsilon_i^2) + 2\frac{{}_nC_2}{{}_NC_2}\overset{{}_NC_2}{S}(\epsilon_i\epsilon_j)\right\} \quad \ldots\ldots (5).$$

The part within braces may be simplified; for

$$\frac{{}_nC_2}{{}_NC_2} = \frac{n!}{(n-2)!2!}\frac{(N-2)!2!}{N!} = \frac{n(n-1)}{N(N-1)}.$$

Furthermore,

$$\overset{N}{S}(\epsilon_i^2) + 2\overset{{}_NC_2}{S}(\epsilon_i\epsilon_j) = \left[\overset{N}{S}(\epsilon_i)\right]^2 = 0$$

since $\overset{N}{S}(\epsilon_i)$ is the sum of residuals by definition. Hence

$$2\overset{{}_NC_2}{S}(\epsilon_i\epsilon_j) = -\overset{N}{S}(\epsilon_i^2).$$

Equation (5) may therefore be written as follows:

$$\frac{1}{{}_NC_n}\overset{{}_NC_n}{S}\left[(\bar{\epsilon})^2\right] = \frac{1}{n^2}\left\{\frac{n}{N}\overset{N}{S}(\epsilon_i^2) - \frac{n(n-1)}{N(N-1)}\overset{N}{S}(\epsilon_i^2)\right\} \quad \ldots\ldots (5a).$$

Now

$$\frac{1}{N}\overset{N}{S}(\epsilon_i^2) = \sigma_y^2 \quad \ldots\ldots (6).$$

Hence, the exact value of the variance of means of n observations of y, from the finite population of N, is

$$\frac{1}{n^2}\left\{n\sigma_y^2 - \frac{n(n-1)}{(N-1)}\sigma_y^2\right\} = \frac{\sigma_y^2}{n}\left(\frac{N-n}{N-1}\right) \quad \ldots\ldots (7).$$

In practice, one cannot obtain, from a sample of n values, the exact σ_y^2 of equation (6) which this expression requires. Neither can one substitute, for it, the estimate

$$s_y^2 = \frac{1}{n-1}\overset{n}{S}\left[(y-\bar{y})^2\right]$$

which applies to hypothetically infinite populations. Consequently, it is necessary to consider the estimate s_y^2 of σ_y^2 afresh.

Regardless of whether the sampled population is finite or hypothetically infinite, each real error $(y-\mu)$ may be expressed as the sum of two contributions, as follows:

$$(y-\mu)=(y-\bar{y})+(\bar{y}-\mu).$$

Upon squaring, then adding over all n values of the sample, and remembering that the error of the sample mean $(\bar{y}-\mu)$ is constant for the sample,

$$\overset{n}{S}\left[(y-\mu)^2\right]=\overset{n}{S}\left[(y-\bar{y})^2\right]+n(\bar{y}-\mu)^2+2(\bar{y}-\mu)\overset{n}{S}(y-\bar{y}).$$

The third term on the right is zero, as one of its factors, being the sum of residuals, is zero. Upon dropping this term and then transposing so as to express the errors in terms of the residuals,

$$\overset{n}{S}\left[(y-\bar{y})^2\right]=\overset{n}{S}\left[(y-\mu)^2\right]-n(\bar{y}-\mu)^2.$$

It is known, from Sec. 1.7, that the first term on the right is an unbiased estimate of $n\sigma_y^2$; and that $(\bar{y}-\mu)^2$ is an unbiased estimate of $\frac{1}{n}\sigma_y^2$. Thus, the above equality may be written

$$\overset{n}{S}\left[(y-\bar{y})^2\right]\rightarrow n\sigma_y^2-n\frac{\sigma_y^2}{n}.$$

But should the sample be drawn from a *finite* population of just N values of y, the variance of the sample mean requires the adjustment factor

$$\frac{N-n}{N-1}$$

of equation (7) as developed above. Therefore the sum of squares obtained from a random sample of size n, when drawn from a finite population of size N, should be expressed

$$\overset{n}{S}\left[(y-\bar{y})^2\right]\rightarrow n\sigma_y^2-n\frac{\sigma_y^2}{n}\left(\frac{N-n}{N-1}\right)\ldots\ldots\ldots\ldots\ldots(8)$$

and this may be written

$$\overset{n}{S}\left[(y-\bar{y})^2\right]\rightarrow(n-1)\,\sigma_y^2\left(\frac{N}{N-1}\right);$$

whence, upon dividing by $(n-1)\left(\frac{N}{N-1}\right)$, the estimate of the variance of the finite population may be expressed,

$$\left(\frac{N-1}{N}\right)\frac{S\left[(y-\bar{y})^2\right]}{n-1}=\left(\frac{N-1}{N}\right)s_y^2 \rightarrow \sigma_y^2.$$

Upon inserting this estimate of σ_y^2 into equation (7) above,

$$V(\bar{y})=\frac{s_y^2}{n}\left(\frac{N-n}{N}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9).$$

This is the estimate of the variance of the mean of a random sample of size n, drawn from a finite population of N values of y.

7.7 The Variance of the Product MN, when M and N Are Independently Subject to Sampling Error. Suppose the observed M and N contain errors ϵ_M and ϵ_N, respectively, such that

$$M=\mu_M+\epsilon_M$$
$$N=\mu_N+\epsilon_N$$

in which μ_M and μ_N are the true characteristics of the populations of M and N. The real errors ϵ_M and ϵ_N are independent of one another, and each is, of course, as likely to be positive as negative. Then the product is

$$\begin{aligned} MN &= (\mu_M+\epsilon_M)(\mu_N+\epsilon_N) \\ &= \mu_M\mu_N+\mu_M\epsilon_N+\mu_N\epsilon_M+\epsilon_M\epsilon_N \end{aligned}$$

and the error ϵ_{MN} of the product is

$$\epsilon_{MN}=(MN-\mu_M\mu_N)=\mu_M\epsilon_N+\mu_N\epsilon_M+\epsilon_M\epsilon_N.$$

The average value of the square of this expression is the variance of the product. But the average value of the square of $(\epsilon_M\epsilon_N)$ is negligible by comparison with the average value of the square of $\mu_M\epsilon_N$ and of $\mu_N\epsilon_M$; furthermore, the errors are independent; hence in the average

$$\sigma_{MN}^2=\mu_M^2\sigma_N^2+\mu_N^2\sigma_M^2.$$

In practice, M and N are estimates of μ_M and μ_N, respectively. If M and N are independently distributed, and $V(M)$ and $V(N)$ are estimates of their sampling variances, then

$$V(MN)=M^2\left[V(N)\right]+N^2\left[V(M)\right]$$

in which $V(MN)$ is the estimate of the sampling variance of MN.

10.4(A) Derivation of Normal Equations. If an estimate, Y, corresponding to an observed dependent variate, y, is to be expressed

in terms of, say, the three independent variates x_1, x_2 and x_3, such that

$$Y = b_1x_1 + b_2x_2 + b_3x_3$$

the numerical equivalents of the regression coefficients, b_1, b_2 and b_3, may be calculated according to the method of least squares. The application of the method of least squares consists in the evaluation of the unknowns, b_1, b_2 and b_3, such that the sum of squares of residuals

$$\overset{n}{S}\left[(y-Y)^2\right]$$

is minimum. Given n independent sets of observations on the population (y, x_1, x_2, x_3), this is equivalent to setting

$$\overset{n}{S}\left[(y-b_1x_1-b_2x_2-b_3x_3)^2\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

to a minimum; or if the observations do not all have the same weight, but each is assigned a weight w, then the sum of weighted squares of residuals, that is,

$$\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)^2\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

is to be minimum. With the three unknowns of equations (1) and (2), their sums of squares are based upon $(n-3)$ degrees of freedom. Consequently, n must exceed three observation equations, in these cases, in order to provide for a mean square of the residuals. As equation (1) is a special case of equation (2) with $w=1$ throughout, the latter will be used for purposes of illustration.

From the calculus it is known that the minimum value of (2) will be obtained if its first derivatives with respect to b_1, b_2, *and* b_3, are zero. Upon differentiating equation (2) with respect to each unknown in turn, and equating each first derivative to zero, we have

$$2\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)(-x_1)\right]=0$$

$$2\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)(-x_2)\right]=0$$

$$2\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)(-x_3)\right]=0$$

and upon dividing by 2, and carrying through the products and summations indicated, the three normal equations are the following:

$$\left.\begin{aligned} b_1\left[\overset{n}{S}(wx_1^2)\right]+b_2\left[\overset{n}{S}(wx_1x_2)\right]+b_3\left[\overset{n}{S}(wx_1x_3)\right]&=\left[\overset{n}{S}(wx_1y)\right] \\ b_1\left[\overset{n}{S}(wx_1x_2)\right]+b_2\left[\overset{n}{S}(wx_2^2)\right]+b_3\left[\overset{n}{S}(wx_2x_3)\right]&=\left[\overset{n}{S}(wx_2y)\right] \\ b_1\left[\overset{n}{S}(wx_1x_3)\right]+b_2\left[\overset{n}{S}(wx_2x_3)\right]+b_3\left[\overset{n}{S}(wx_3^2)\right]&=\left[\overset{n}{S}(wx_3y)\right] \end{aligned}\right\} \quad \ldots(3).$$

Quantities derived from the observations are enclosed within brackets. The simultaneous solution of these three equations renders the unknowns b_1, b_2 and b_3.

Should all the observation equations be assigned the same weight, then w may be taken as unity, and deleted from each term of the normal equations (3). The remaining part would, indeed, be the normal equations as derived directly from equation (1).

Should there be but two unknowns such that the regression equation takes the form

$$Y=b_1x_1+b_2x_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4),$$

one would merely delete from the normal equations (3), the third column and the third line. In this case one should have the two normal equations,

$$\left.\begin{aligned} b_1\left[\overset{n}{S}(wx_1^2)\right]+b_2\left[\overset{n}{S}(wx_1x_2)\right]&=\left[\overset{n}{S}(wx_1y)\right] \\ b_1\left[\overset{n}{S}(wx_1x_2)\right]+b_2\left[\overset{n}{S}(wx_2^2)\right]&=\left[\overset{n}{S}(wx_2y)\right] \end{aligned}\right\} \quad \ldots\ldots\ldots(5).$$

Specifically, the regression equation of Sec. 10.4 in Chapter X is that of equation (4) above, with the difference that x_1 and x_2 of this equation are x and x^2 of the nursery seed bed data; and the weights, w, of equations (5) are the values $\frac{1}{x}$ of the seed bed data. Inserting these for w, x_1 and x_2 of the normal equations (5), the latter take on the form used in Sec. 10.4 of Chapter X.

10.4(B) The Sum of Squares Independent of the Regression. This is the sum of squares of equation (1) or (2) above. Upon expanding equation (2),

$$\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)^2\right]=\overset{n}{S}(wy^2)+b_1^2\overset{n}{S}(wx_1^2)+b_2^2\overset{n}{S}(wx_2^2)$$
$$+b_3^2\overset{n}{S}(wx_3^2)-2b_1\overset{n}{S}(wx_1y)-2b_2\overset{n}{S}(wx_2y)-2b_3\overset{n}{S}(wx_3y)$$
$$+2b_1b_2\overset{n}{S}(wx_1x_2)+2b_1b_3\overset{n}{S}(wx_1x_3)+2b_2b_3\overset{n}{S}(wx_2x_3).$$

This is rather unwieldy, but, fortunately, it submits to considerable simplification. The above expression may be written

$$\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)^2\right]=\overset{n}{S}(wy^2)+b_1\left[b_1\overset{n}{S}(wx_1^2)\right.$$
$$\left.+b_2\overset{n}{S}(wx_1x_2)+b_3\overset{n}{S}(wx_1x_3)\right]$$
$$+b_2\left[b_1\overset{n}{S}(wx_1x_2)+b_2\overset{n}{S}(wx_2^2)+b_3\overset{n}{S}(wx_2x_3)\right]$$
$$+b_3\left[b_1\overset{n}{S}(wx_1x_3)+b_2\overset{n}{S}(wx_2x_3)+b_3\overset{n}{S}(wx_3^2)\right]$$
$$-2b_1\overset{n}{S}(wx_1y)-2b_2\overset{n}{S}(wx_2y)-2b_3\overset{n}{S}(wx_3y).$$

Now the expressions of the right-hand member which are enclosed within brackets are, respectively, the left-hand members of the normal equations (3) in Sec. 10.4(A) above. Thus

$$\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)^2\right]=\overset{n}{S}(wy^2)+b_1\overset{n}{S}(wx_1y)+b_2\overset{n}{S}(wx_2y)+b_3\overset{n}{S}(wx_3y)$$
$$-2b_1\overset{n}{S}(wx_1y)-2b_2\overset{n}{S}(wx_2y)-2b_3\overset{n}{S}(wx_3y),$$

and the sum of weighted squares of residuals may be stated as follows:

$$\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)^2\right]=\overset{n}{S}(wy^2)-b_1\overset{n}{S}(wx_1y)-b_2\overset{n}{S}(wx_2y)-b_3\overset{n}{S}(wx_3y).$$

In this form it becomes evident that the sum of weighted squares of y

$$\overset{n}{S}(wy^2)$$

on n degrees of freedom is divisible into a portion due to regression, that is,

$$b_1\overset{n}{S}(wx_1y)+b_2\overset{n}{S}(wx_2y)+b_3\overset{n}{S}(wx_3y)$$

on three degrees of freedom, and the residue portion, independent of the regression,

$$\overset{n}{S}\left[w(y-b_1x_1-b_2x_2-b_3x_3)^2\right]$$

on $(n-3)$ degrees of freedom.

Should there be but two unknowns, as in Sec. 10.4 of Chapter X for which x_1 and x_2 of the above discussion are x and x^2, respectively; and for which the weights, w, are taken as $\frac{1}{x}$, it follows that the division of total sum of weighted squares of y, into portions due to, and independent of, the regression, takes the form presented in Table 38.

10.5 The Variance of the Regression Function $Y=b_1x_1+b_2x_2+b_3x_3$, and Developments Leading Thereto. Many applications of multiple regression involve certain coefficients, known as c-multipliers, which depend upon the independent variables only (Fisher, 1936. Secs. 29 and 29.1). But in order to show the meaning and use of the c-multipliers, it is desirable to develop (1) the solution of normal equations by determinants; (2) a convenient method of calculating the c-multipliers; (3) through them, the estimate of variances and covariances among regression coefficients; and (4) the variance of the regression function. These will be taken up in order.

10.5(A) Solution of the Normal Equations by Determinants. Given the regression equation

$$Y=b_1x_1+b_2x_2+b_3x_3,$$

notation may be changed for purposes of condensation, to the following:

Let 0, 1, 2, and 3, denote y, x_1, x_2, and x_3, respectively, such that

$$\begin{array}{llll} \overset{n}{S}(y^2)=(00); & \overset{n}{S}(yx_1)=(01); & \overset{n}{S}(yx_2)=(02); & \overset{n}{S}(yx_3)=(03); \\ & \overset{n}{S}(x_1^2)=(11); & \overset{n}{S}(x_1x_2)=(12); & \overset{n}{S}(x_1x_3)=(13); \\ & & \overset{n}{S}(x_2^2)=(22); & \overset{n}{S}(x_2x_3)=(23); \\ & & & \overset{n}{S}(x_3^2)=(33); \end{array}$$

the parentheses of the right-hand members signifying summation over all the n independent pairs. In this notation the normal equations may be expressed as follows:

$$\left.\begin{array}{rl} \text{I.} & b_1(11)+b_2(12)+b_3(13)=(01) \\ \text{II.} & b_1(12)+b_2(22)+b_3(23)=(02) \\ \text{III.} & b_1(13)+b_2(23)+b_3(33)=(03) \end{array}\right\} \quad \ldots\ldots\ldots\ldots (1).$$

Now the *determinant of the system*, involving only the independent variables, is

$$D=\begin{vmatrix}(11) & (12) & (13)\\ (12) & (22) & (23)\\ (13) & (23) & (33)\end{vmatrix}=(11)A_{11}+(12)A_{12}+(13)A_{13}\ldots\ldots\ldots(2)$$

in which A_{11}, A_{12}, and A_{13} are cofactors, respectively, of the elements (11), (12), and (13).

Upon multiplying I, II, and III of equation (1), by A_{11}, A_{12}, and A_{13}, respectively,

$$\left.\begin{aligned} b_1(11)A_{11}+b_2(12)A_{11}+b_3(13)A_{11}&=(01)A_{11}\\ b_1(12)A_{12}+b_2(22)A_{12}+b_3(23)A_{12}&=(02)A_{12}\\ b_1(13)A_{13}+b_2(23)A_{13}+b_3(33)A_{13}&=(03)A_{13}\end{aligned}\right\}\ldots\ldots(3).$$

If these three equations are added together,

$$b_1\Big[(11)A_{11}+(12)A_{12}+(13)A_{13}\Big]+b_2\Big[(12)A_{11}+(22)A_{12}+(23)A_{13}\Big]+b_3\Big[(13)A_{11}+(23)A_{12}+(33)A_{13}\Big]=(01)A_{11}+(02)A_{12}+(03)A_{13}.$$

Now the second and third of the bracketed terms are zero because they represent a determinant in which two rows are identical. Hence equation (3) may be expressed

$$b_1\Big[(11)A_{11}+(12)A_{12}+(13)A_{13}\Big]=(01)A_{11}+(02)A_{12}+(03)A_{13}$$

and

$$b_1=\frac{(01)A_{11}+(02)A_{12}+(03)A_{13}}{(11)A_{11}+(12)A_{12}+(13)A_{13}}$$

or, in determinant form,

$$b_1=\frac{\begin{vmatrix}(01) & (12) & (13)\\ (02) & (22) & (23)\\ (03) & (23) & (33)\end{vmatrix}}{\begin{vmatrix}(11) & (12) & (13)\\ (12) & (22) & (23)\\ (13) & (23) & (33)\end{vmatrix}}=\frac{\begin{vmatrix}(01) & (12) & (13)\\ (02) & (22) & (23)\\ (03) & (23) & (33)\end{vmatrix}}{D}.$$

The regression coefficients b_2 and b_3 may be derived along the same lines; for the determinant of the system of equation (2) may also be expressed in either of the following forms:

$$D=(12)A_{12}+(22)A_{22}+(23)A_{23}$$
$$D=(13)A_{13}+(23)A_{23}+(33)A_{33}.$$

These equations lead to the determinant expressions for b_2 and b_3 in which

$$b_2=\frac{1}{D}\begin{vmatrix}(11) & (01) & (13)\\ (12) & (02) & (23)\\ (13) & (03) & (33)\end{vmatrix};$$

and

$$b_3=\frac{1}{D}\begin{vmatrix}(11) & (12) & (01)\\ (12) & (22) & (02)\\ (13) & (23) & (03)\end{vmatrix}.$$

These expressions may be put in slightly different form, for purposes of immediate use, so that

$$b_1=\frac{1}{D}\left[(01)A_{11}+(02)A_{12}+(03)A_{13}\right] \quad \ldots\ldots\ldots (4a)$$

$$b_2=\frac{1}{D}\left[(01)A_{12}+(02)A_{22}+(03)A_{23}\right] \quad \ldots\ldots\ldots (4b)$$

$$b_3=\frac{1}{D}\left[(01)A_{13}+(02)A_{23}+(03)A_{33}\right] \quad \ldots\ldots\ldots (4c)$$

10.5(B) Calculation of the c-Multipliers. The c-multipliers may be defined in terms of the symbols used above. For, in fact,

$$c_{11}=\frac{A_{11}}{D};\quad c_{12}=\frac{A_{12}}{D};\quad c_{13}=\frac{A_{13}}{D};$$
$$c_{22}=\frac{A_{22}}{D};\quad c_{23}=\frac{A_{23}}{D};$$
$$c_{33}=\frac{A_{33}}{D}.$$

Hence, upon substituting the c-multipliers for their equivalents as given in equations (4a), (4b) and (4c) above, these equations may be expressed in the following form:

$$\left.\begin{aligned} b_1&=c_{11}(01)+c_{12}(02)+c_{13}(03)\\ b_2&=c_{12}(01)+c_{22}(02)+c_{23}(03)\\ b_3&=c_{13}(01)+c_{23}(02)+c_{33}(03)\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots (5).$$

The b-regression coefficients may thus be readily calculated provided the numerical equivalents of the c-multipliers are known. These may be deduced from the discussion above. From equations (2) and (3),

$$(11)A_{11}+(12)A_{12}+(13)A_{13}=D$$
$$(12)A_{11}+(22)A_{12}+(23)A_{13}=0$$
$$(13)A_{11}+(23)A_{12}+(33)A_{13}=0.$$

Upon dividing each of these by D and remembering that

$$\frac{A_{11}}{D}=c_{11}; \quad \frac{A_{12}}{D}=c_{12}; \quad \frac{A_{13}}{D}=c_{13},$$

one may write

$$\left.\begin{aligned} c_{11}(11)+c_{12}(12)+c_{13}(13)&=1\\ c_{11}(12)+c_{12}(22)+c_{13}(23)&=0\\ c_{11}(13)+c_{12}(23)+c_{13}(33)&=0 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6a).$$

These equations, involving only the independent variates, may thus be made to supply the numerical equivalents of c_{11}, c_{12} and c_{13}. In like manner it can be shown that

$$\left.\begin{aligned} c_{12}(11)+c_{22}(12)+c_{23}(13)&=0\\ c_{12}(12)+c_{22}(22)+c_{23}(23)&=1\\ c_{12}(13)+c_{22}(23)+c_{23}(33)&=0 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6b)$$

and that

$$\left.\begin{aligned} c_{13}(11)+c_{23}(12)+c_{33}(13)&=0\\ c_{13}(12)+c_{23}(22)+c_{33}(23)&=0\\ c_{13}(13)+c_{23}(23)+c_{33}(33)&=1 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6c).$$

10.5(C) Variances and Covariances of Regression Coefficients. As an example the estimate of the variance of b_1—that is, $V(b_1)$—will be derived in detail.

In the regression equation

$$Y=b_1x_1+b_2x_2+b_3x_3$$

each observed value of the dependent variable, y, may be conceived as the sum of two components, namely, (1) the true value of the dependent variable—which we may designate $\mu_{0.123}$—corresponding to a given combination of the independent variables, x_1, x_2 and x_3; and (2) the real error, $\epsilon_{0.123}$, of the observed y. Thus each

$$y=\mu_{0.123}+\epsilon_{0.123}.$$

Let b_1 be defined according to equation (5) of Sec. 10.5(B) above, that is,

$$b_1 = c_{11}(01) + c_{12}(02) + c_{13}(03).$$

The sums of products in this equation may, for the present purposes, be expressed as follows:

$$(01) = \overset{n}{S}\left[x_1(\mu_{0.123} + \epsilon_{0.123}) \right]; \qquad (02) = \overset{n}{S}\left[x_2(\mu_{0.123} + \epsilon_{0.123}) \right];$$

$$(03) = \overset{n}{S}\left[x_3(\mu_{0.123} + \epsilon_{0.123}) \right].$$

The statement for b_1, above, may now be given as

$$b_1 = \left[c_{11}\overset{n}{S}(x_1\mu_{0.123}) + c_{12}\overset{n}{S}(x_2\mu_{0.123}) + c_{13}\overset{n}{S}(x_3\mu_{0.123}) \right] + \left[c_{11}\overset{n}{S}(x_1\epsilon_{0.123}) + c_{12}\overset{n}{S}(x_2\epsilon_{0.123}) + c_{13}\overset{n}{S}(x_3\epsilon_{0.123}) \right]$$

for which the expression within the first brackets is the true (population) regression coefficient, β_1; while the expression within the second brackets represents the real error of b_1, being the exact value of $(b_1 - \beta_1)$. Thus

$$(b_1 - \beta_1) = c_{11}\overset{n}{S}(x_1\epsilon_{0.123}) + c_{12}\overset{n}{S}(x_2\epsilon_{0.123}) + c_{13}\overset{n}{S}(x_3\epsilon_{0.123})$$

and this may be written,

$$(b_1 - \beta_1) = \overset{n}{S}\left\{ \epsilon_{0.123}\left[c_{11}x_1 + c_{12}x_2 + c_{13}x_3 \right] \right\}$$

Now the variance of b_1 is the average of the square of $(b_1 - \beta_1)$ over all sets of samples, each of size n and with the same distribution of independent variables as this one. Upon squaring the above equation and remembering that the square of a sum of independent values is the sum of their squares,

$$(b_1 - \beta_1)^2 = \overset{n}{S}\left\{ \epsilon^2_{0.123}\left[c_{11}x_1 + c_{12}x_2 + c_{13}x_3 \right]^2 \right\}$$

and this, upon expansion, may be written as follows:

$$(b_1 - \beta_1)^2 = \overset{n}{S}\Big[\epsilon^2_{0.123}\Big\{ c_{11}\left[c_{11}x_1^2 + c_{12}x_1x_2 + c_{13}x_1x_3 \right] + c_{12}\left[c_{11}x_1x_2 + c_{12}x_2^2 + c_{13}x_2x_3 \right] + c_{13}\left[c_{11}x_1x_3 + c_{12}x_2x_3 + c_{13}x_3^2 \right] \Big\} \Big].$$

Since the individual values of $\epsilon_{0.123}$ are independent of the independent variables, the above may be expressed as the *average* $\epsilon^2_{0.123}$, or $\sigma^2_{0.123}$, multiplied by the sums of the remaining terms. Expressing these latter in the notation of Sec. 10.5(B) above, that is

$$\overset{n}{S}(x_1^2) = (11); \qquad \overset{n}{S}(x_1x_2) = (12); \qquad \text{etc.},$$

the above may be written

$$\begin{aligned}(b_1-\beta_1)^2 = \sigma^2_{0.123}\Big\{ & c_{11}\Big[c_{11}(11)+c_{12}(12)+c_{13}(13)\Big] \\ & +c_{12}\Big[c_{11}(12)+c_{12}(22)+c_{13}(23)\Big] \\ & +c_{13}\Big[c_{11}(13)+c_{12}(23)+c_{13}(33)\Big]\Big\} \\ = c_{11}\,\sigma^2_{0.123}\end{aligned}$$

since, from equation (6a) of Sec. 10.5(B), the quantity within the first square bracket is unity, and the others are zero.

In practice the exact $\sigma^2_{0.123}$ is not known; but the mean square of the residuals around the regression equation—that is, $s^2_{0.123}$—is an unbiased estimate of it. Hence the estimate of the sampling variance of the regression coefficient, b_1, is

$$V(b_1) = c_{11}s^2_{0.123}.$$

In like manner it can be shown that

$$V(b_2) = c_{22}s^2_{0.123}; \qquad V(b_3) = c_{33}s^2_{0.123};$$
$$Cov(b_1b_2) = c_{12}s^2_{0.123}; \qquad Cov(b_1b_3) = c_{13}s^2_{0.123}; \qquad Cov(b_2b_3) = c_{23}s^2_{0.123}.$$

10.5(D) The Variance of the Regression Function. In the regression equation

$$Y = b_1x_1+b_2x_2+b_3x_3$$

the variance of the calculated Y is the variance of the function, that is,

$$V(Y) = V(b_1x_1+b_2x_2+b_3x_3).$$

Since this is the variance of the sum of three terms which are not necessarily independent of one another, it may be written

$$\begin{aligned}V(Y) = V(b_1x_1)+V(b_2x_2)+V(b_3x_3)+2\ Cov\Big[(b_1x_1)(b_2x_2)\Big] \\ +2\ Cov\Big[(b_1x_1)(b_3x_3)\Big]+2\ Cov\Big[(b_2x_2)(b_3x_3)\Big] \quad \ldots\ldots (7).\end{aligned}$$

Suppose, on the one hand, that the values of x_1, x_2 and x_3 with which the equation is entered—and which supplies the particular value of Y whose variance is sought—are themselves free of sampling error. It would then follow, for example,

$$V(b_1x_1) = x_1^2\Big[\ V(b_1)\ \Big] = x_1^2 c_{11} s_{0.123}^2$$

or, as another example,

$$Cov\Big[\ (b_1x_1)(b_2x_2)\ \Big] = x_1x_2\Big[\ Cov(b_1b_2)\ \Big] = x_1x_2c_{12}s_{0.123}^2$$

and the remaining terms would submit to analogous expressions.

It thus follows that when each independent variable is free of sampling error, the variance of Y in the regression equation

$$Y = b_1x_1 + b_2x_2 + b_3x_3$$

is

$$V(Y) = s_{0.123}^2\Big[\ c_{11}x_1^2 + c_{22}x_2^2 + c_{33}x_3^2 + 2c_{12}x_1x_2 + 2c_{13}x_1x_3 + 2c_{23}x_2x_3\ \Big] \quad \ldots\ldots\ldots (8)$$
$$= s^2R$$

where, in general, s^2 denotes the mean square of the residuals, of unit weight, which are independent of the regression, and R denotes the quantity within brackets. Thus s^2R represents the contribution to $V(Y)$ of the sampling errors ascribable to the regression equation itself.

If the independent variables with which the equation is entered are subject to sampling errors whose variances and covariances are

$$\begin{array}{lll} V(x_1); & V(x_2); & Cov(x_1x_2); \\ & V(x_3); & Cov(x_1x_3); \\ & & Cov(x_2x_3); \end{array}$$

then the terms of equation (7) represent variances or covariances of products, both factors of which are subject to sampling error. One should then have, for example,

$$V(b_1x_1) = x_1^2\Big[\ V(b_1)\ \Big] + b_1^2\Big[\ V(x_1)\ \Big]$$

and, as another example,

$$Cov\Big[\ (b_1b_2)(x_1x_2)\ \Big] = x_1x_2\Big[\ Cov(b_1b_2)\ \Big] + b_1b_2\Big[\ Cov(x_1x_2)\ \Big]$$

the remaining terms supplying analogous expressions, the first term of each right-hand member being already contained in equation (8). The sum of the second terms of the right-hand members, symbolized as S^2, is the following:

$$S^2 = b_1^2\left[V(x_1)\right] + b_2^2\left[V(x_2)\right] + b_3^2\left[V(x_3)\right] + 2b_1b_2\left[Cov(x_1x_2)\right] + 2b_1b_3\left[Cov(x_1x_3)\right] + 2b_2b_3\left[Cov(x_2x_3)\right] \quad \ldots\ldots\ldots (9).$$

It is apparent that S^2 is due to the sampling errors of the independent variates.

In general, then, the variance of Y, where

$$Y = b_1x_1 + b_2x_2 + b_3x_3$$

is given by the addition of equation (9) to equation (8), that is,

$$V(Y) = s^2R + S^2.$$

11.4 The *c*-Multipliers Appropriate to the Regression Function, $Y = a + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + b_3(x_3 - \bar{x}_3)$. The regression equation

$$Y = a + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + b_3(x_3 - \bar{x}_3)$$

differs essentially from the regression equation treated above only in the appendage of the constant a. As in Chapter XI, however, when written in this form,

$$a = \bar{y}.$$

Consequently, the regression equation may be written in the alternative form,

$$(Y - \bar{y}) = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + b_3(x_3 - \bar{x}_3)$$

and upon changing notation, such that each $(y - \bar{y})$ be denoted by 0, each $(x_1 - \bar{x}_1)$ be denoted by 1, etc., the sums of squares and products among the variables may be expressed as follows:

$$\overset{n}{S}\left[(y - \bar{y})^2\right] = (00); \qquad \overset{n}{S}\left[(y - \bar{y})(x_1 - \bar{x}_1)\right] = (01); \text{ etc.}$$

$$\overset{n}{S}\left[(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)\right] = (12); \text{ etc.}$$

The parentheses again signify summation over all the n independent pairs. The normal equations, then, appropriate to the solution of the

regression coefficients are of the same form as those of equation (1) of Sec. 10.5(A) above. In consequence,

$$\begin{aligned} &\text{I.} && b_1(11)+b_2(12)+b_3(13)=(01)\\ &\text{II.} && b_1(12)+b_2(22)+b_3(23)=(02)\\ &\text{III.} && b_1(13)+b_2(23)+b_3(33)=(03). \end{aligned}$$

The c-multipliers may therefore be calculated quite in accordance with equations (6a), (6b), and (6c) above.

11.5(A) The Variance of the Regression Function, $Y=a+b_1(x_1-\bar{x}_1)+b_2(x_2-\bar{x}_2)+b_3(x_3-\bar{x}_3)$. Given the regression equation

$$Y=a+b_1(x_1-\bar{x}_1)+b_2(x_2-\bar{x}_2)+b_3(x_3-\bar{x}_3),$$

it follows that the variance of Y may be expressed:

$$V(Y)=V(a)+V\left[b_1(x_1-\bar{x}_1)+b_2(x_2-\bar{x}_2)+b_3(x_3-\bar{x}_3)\right]$$

that is, it is completely given as the variance of a plus the variance of the remaining portion of the regression equation, since the constant a is independent of the remaining portion.

Now the variance of a is the variance of Y when x_1, x_2, and x_3 are $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$, respectively. Hence

$$V(a)=\frac{s^2_{0.123}}{n}$$

where $s^2_{0.123}$ is the mean square of the residuals of unit weight. The variance of the remaining portion of the regression equation, that is,

$$V\left[b_1(x_1-\bar{x}_1)+b_2(x_2-\bar{x}_2)+b_3(x_3-\bar{x}_3)\right]$$

being the variance of the sum of terms which are not necessarily independent of one another, may be expanded to the following:

$$\begin{aligned} &V\left[b_1(x_1-\bar{x}_1)\right]+V\left[b_2(x_2-\bar{x}_2)\right]+V\left[b_3(x_3-\bar{x}_3)\right]\\ &\qquad +2\,Cov\left[b_1b_2(x_1-\bar{x}_1)(x_2-\bar{x}_2)\right]\\ &+2\,Cov\left[b_1b_3(x_1-\bar{x}_1)(x_3-\bar{x}_3)\right]+2\,Cov\left[b_2b_3(x_2-\bar{x}_2)(x_3-\bar{x}_3)\right]. \end{aligned}$$

If the values of x_1, x_2, and x_3 with which the regression equation is entered, are subject to sampling error, each of the above terms is the

variance (or covariance) of a product, both factors of which have sampling error. Consequently,

$$\left.\begin{aligned}
V\left[b_1(x_1-\bar{x}_1)\right] &= (x_1-\bar{x}_1)^2\left[V(b_1)\right]+b_1^2\left[V(x_1)\right] \\
V\left[b_2(x_2-\bar{x}_2)\right] &= (x_2-\bar{x}_2)^2\left[V(b_2)\right]+b_2^2\left[V(x_2)\right] \\
V\left[b_3(x_3-\bar{x}_3)\right] &= (x_3-\bar{x}_3)^2\left[V(b_3)\right]+b_3^2\left[V(x_3)\right] \\
Cov\left[b_1b_2(x_1-\bar{x}_1)(x_2-\bar{x}_2)\right] &= (x_1-\bar{x}_1)(x_2-\bar{x}_2)\left[Cov(b_1b_2)\right]+b_1b_2\left[Cov(x_1x_2)\right] \\
Cov\left[b_1b_3(x_1-\bar{x}_1)(x_3-\bar{x}_3)\right] &= (x_1-\bar{x}_1)(x_3-\bar{x}_3)\left[Cov(b_1b_3)\right]+b_1b_3\left[Cov(x_1x_3)\right] \\
Cov\left[b_2b_3(x_2-\bar{x}_2)(x_3-\bar{x}_3)\right] &= (x_2-\bar{x}_2)(x_3-\bar{x}_3)\left[Cov(b_2b_3)\right]+b_2b_3\left[Cov(x_2x_3)\right]
\end{aligned}\right\} \quad (1).$$

Upon adding the column of first terms of the right-hand members, after multiplying the covariances by 2, and remembering that

$$V(b_1)=c_{11}s^2_{0.123}; \qquad Cov(b_1b_2)=c_{12}s^2_{0.123}; \qquad \text{etc.},$$

the sum may be expressed:

$$s^2_{0.123}\Big[c_{11}(x_1-\bar{x}_1)^2+c_{22}(x_2-\bar{x}_2)^2+c_{33}(x_3-\bar{x}_3)^2+2c_{12}(x_1-\bar{x}_1)(x_2-\bar{x}_2) +2c_{13}(x_1-\bar{x}_1)(x_3-\bar{x}_3)+2c_{23}(x_2-\bar{x}_2)(x_3-\bar{x}_3)\Big].$$

This expression, plus the variance of a, for which

$$V(a)=\frac{s^2_{0.123}}{n},$$

is the variance of the regression function, provided the independent variables with which the equation is entered are free of sampling error. Combining them,

$$s^2R = s^2_{0.123}\left[\frac{1}{n} + c_{11}(x_1-\bar{x}_1)^2 + c_{22}(x_2-\bar{x}_2)^2 + c_{33}(x_3-\bar{x}_3)^2\right.$$

$$\left. + 2c_{12}(x_1-\bar{x}_1)(x_2-\bar{x}_2) + 2c_{13}(x_1-\bar{x}_1)(x_3-\bar{x}_3) + 2c_{23}(x_2-\bar{x}_2)(x_3-\bar{x}_3)\right]$$

as the contribution to $V(Y)$ ascribable to the sampling errors of the regression equation.

The column sum of second terms of the right-hand members of equations (1), after multiplying covariances by 2, is due to the sampling errors of the independent variables; whence, upon denoting it by S^2,

$$S^2 = b_1^2\left[V(x_1)\right] + b_2^2\left[V(x_2)\right] + b_3^2\left[V(x_3)\right] + 2b_1b_2\left[Cov(x_1x_2)\right]$$
$$+ 2b_1b_3\left[Cov(x_1x_3)\right] + 2b_2b_3\left[Cov(x_2x_3)\right].$$

11.5(B) The Covariance of Paired Residuals Which Are Independent of Regressions on Identical Independent Variates. Given the regression of y_H on x_H and x_P, such that the sum of squares of residuals is

$$\overset{n}{S}\left[(y_H - h_Hx_H - h_Px_P)^2\right]$$

it was shown in Sec. 10.4(B) above that the expansion leads to the expression

$$\overset{n}{S}(y_H^2) - h_H\overset{n}{S}(x_Hy_H) - h_P\overset{n}{S}(x_Py_H)$$

that is, the sum of squares of the residuals is the residue after deducting the portion

$$h_H\overset{n}{S}(x_Hy_H) + h_P\overset{n}{S}(x_Py_H)$$

which is due to the regression on x_H and x_P, from the total sum of squares of y_H.

The sum of products of corresponding residuals as applied in Chapter XI rests upon the expansion of

$$\overset{n}{S}\left[(y_H - h_Hx_H - h_Px_P)(y_P - p_Hx_H - p_Px_P)\right];$$

and this may be expressed

$$\overset{n}{S}(y_H y_P) - h_H \overset{n}{S}(x_H y_P) - h_P \overset{n}{S}(x_P y_P) + h_H p_H \overset{n}{S}(x_H^2) + h_P p_H \overset{n}{S}(x_H x_P)$$
$$- p_H \overset{n}{S}(x_H y_H) - p_P \overset{n}{S}(x_P y_H) + h_H p_P \overset{n}{S}(x_P x_H) + h_P p_P \overset{n}{S}(x_P^2)$$

or in alternative form:

$$\overset{n}{S}(y_H y_P) + \left[h_H p_H \overset{n}{S}(x_H^2) + h_P p_H \overset{n}{S}(x_H x_P) - p_H \overset{n}{S}(x_H y_H) \right]$$
$$+ \left[h_H p_P \overset{n}{S}(x_P x_H) + h_P p_P \overset{n}{S}(x_P^2) - p_P \overset{n}{S}(x_P y_H) \right]$$
$$- h_H \overset{n}{S}(x_H y_P) - h_P \overset{n}{S}(x_P y_P) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1).$$

Now the two bracketed expressions of the latter form are zero, as becomes evident upon factoring p_H out of the first, and p_P out of the second, and comparing with the normal equations. Therefore,

$$\overset{n}{S}\left[(y_H - h_H x_H - h_P x_P)(y_P - p_H x_H - p_P x_P) \right]$$
$$= \overset{n}{S}(y_H y_P) - h_H \overset{n}{S}(x_H y_P) - h_P \overset{n}{S}(x_P y_P),$$

or, by another rearrangement of the terms in expression (1), one may also write

$$\overset{n}{S}\left[(y_H - h_H x_H - h_P x_P)(y_P - p_H x_H - p_P x_P) \right]$$
$$= \overset{n}{S}(y_H y_P) - p_H \overset{n}{S}(x_H y_H) - p_P \overset{n}{S}(x_P y_H).$$

TABLE 7. Table of t. Values of t, Outside the Range of Which in Both Tails Lie Selected Proportions of the Total Area*

Degrees of freedom	Relative Area in Both Tails											
	.9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01
	Values of t											
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.756
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.160
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	.12566	.25335	.38532	.52440	.67449	.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582

*This table is taken by consent from Statistical Methods for Research Workers by Professor R. A. Fisher, published at 15/- by Oliver and Boyd, Edinburgh. Attention is drawn to the larger collection in Statistical Tables by Professor R. A. Fisher and F. Yates, published by Oliver and Boyd, Edinburgh.

Selected List of References

I. BOOKS CONTAINING DISCUSSION OF SAMPLING THEORY OR APPLICATION

Brunt, David. 1931. The combination of observations. 2d ed. 239 pp. Cambridge University Press. London.

Deming, W. E. 1938. Some notes on least squares. 181 pp. U. S. Dept. Agri. Graduate School. Washington, D. C.

Fisher, R. A. 1936. Statistical methods for research workers. 6th ed. 339 pp. Oliver and Boyd. Edinburgh.

——— 1937. The design of experiments. 2d ed. 260 pp. Oliver and Boyd. Edinburgh.

———, **and F. Yates.** 1938. Statistical tables for biological, agricultural and medical research. 90 pp. Oliver and Boyd. Edinburgh.

Goulden, C. H. 1939. Methods of statistical analysis. 277 pp. John Wiley and Sons. New York.

Neyman, J. 1938. Lectures and conferences on mathematical statistics. 160 pp. U. S. Dept. Agri. Graduate School. Washington, D. C.

Richardson, C. H. 1934. An introduction to statistical analysis. Enlarged ed. 312 pp. Harcourt Brace and Co. New York.

Rider, P. R. 1939. An introduction to modern statistical methods. 220 pp. John Wiley and Sons. New York.

Shewhart, W. A. 1931. Economic control of quality of manufactured product. 501 pp. Van Nostrand. New York.

Snedecor, G. W. 1937. Statistical methods applied to experiments in agriculture and biology. 341 pp. Collegiate Press. Ames, Iowa.

Tippett, L. H. C. 1931. The methods of statistics. 222 pp. Williams and Norgate. London.

II. BULLETINS, AND PAPERS IN SCIENTIFIC JOURNALS

Beall, G. 1939. Methods of estimating the population of insects in a field. *Biometrika* **30**:422-439.

Bowley, A. L. 1936. The application of sampling to economic and sociological problems. *Jour. Amer. Stat. Assoc.* **31**:474-480.

Carver, H. C. 1930. Fundamentals of the theory of sampling. *Annals of Math. Stat.* 1:101-121.

——— 1930. Fundamentals of the theory of sampling. *Annals of Math. Stat.* 1:260-274.

Clapham, A. R. 1929. The estimation of yield in cereal crops by sampling methods. *Jour. Agri. Sci.* **19**:214-235.

——— 1931. Studies in sampling technique: Cereal experiments. I. Field technique. *Jour. Agri. Sci.* 21:366-371.

——— 1931. Studies in sampling technique: Cereal experiments. III. Results and discussion. *Jour. Agri. Sci.* 21:376-390.

Cochran, W. G. 1936. Statistical analysis of field counts of diseased plants. *Supplement Jour. Royal Stat. Soc.* 3:49-67.

——— 1939. The use of the analysis of variance in enumeration by sampling. *Jour. Amer. Stat. Assoc.* 34:492-510.

———, **and D. J. Watson.** 1936. An experiment on observer's bias in the selection of shoot-heights. *Empire Jour. Expt. Agri.* 4:69-76.

Craig, A. T. 1939. On the mathematics of representative method of sampling. *Annals of Math. Stat.* 10:26-34.

Gallup, George. 1938. Government and the sampling referendum. *Jour. Amer. Stat. Assoc.* 33:131-142.

Grumell, E. S. 1935. Statistical methods in industry, with special reference to the sampling of coal and other materials (with discussion). *Supplement Jour. Royal Stat. Soc.* 2:1-26.

Hasel, A. A. 1941. Estimation of vegetation-type areas by linear measurement. *Jour. Forestry.* 39:34-40.

Heath, O. V. S. 1934. Sampling and growth observations in plant development studies in cotton. Report and Summary of Proceedings, Empire Cotton Growing Corporation. Second Conference on Cotton Growing Problems, pp. 96-110.

Holmes, Irvin. 1939. Results of four methods of sampling individual farms. *Jour. Farm Econ.* 21:365-374.

Immer, F. R. 1932. A study of sampling technique with sugar beets. *Jour. Agri. Research.* 44:633-647.

Irwin, J. O., W. G. Cochran, and J. Wishart. 1938. Crop estimation and its relation to agricultural meteorology (with discussion). *Supplement Jour. Royal Stat. Soc.* 5:1-45.

Jennett, W. J., and B. P. Dudding. 1936. Application of statistical principles to an industrial problem (with discussion). *Supplement Jour. Royal Stat. Soc.* 3:1-28.

Jensen, Adolph. 1926. Report on the representative method in statistics. *Bulletin de l'Institut International de Statistique.* 22:359-377.

——— 1926. The representative method in practice. *Bulletin de l'Institut International de Statistique.* 22:381-439.

Jessen, R. J. 1939. An experiment in the design of agricultural surveys. *Jour. Farm Econ.* 21:856-863.

Justesen, S. H. 1932. Influence of size and shape of plots on the precision of field experiments with potatoes. *Jour. Agri. Sci.* 22:366-372.

Kalamkar, R. J. 1932. Experimental error and the field-plot technique with potatoes. *Jour. Agri. Sci.* 22:373-385.

——— 1932. A study in sampling technique with wheat. *Jour. Agri. Sci.* 22:783-796.

King, A. J., and E. H. Jebe. 1940. An experiment in pre-harvest sampling of wheat fields. *Iowa Agri. Expt. Sta. Research Bull.* 273:624-649.

McKay, A. T. 1934. Sampling from batches. *Supplement Jour. Royal Stat. Soc.* 1:207-216.

Meyers, M. T., and L. H. Patch. 1937. A statistical study of sampling in field surveys of the fall population of the European corn borer. *Jour. Agri. Research.* 55:849-871.

Miner, J. R. 1931. The standard error of a multiple regression equation. *Annals Math. Stat.* 2:320-323.

Neyman, J. 1938. Contribution to the theory of sampling human populations. *Jour. Amer. Stat. Assoc.* 33:101-116.

Pechanec, J. F. 1941. Sampling error in range surveys of sagebrush-grass vegetation. *Jour. Forestry.* 39:52-54.

Pearson, E. S. 1934. Sampling problems in industry (with discussion). *Supplement Jour. Royal Stat. Soc.* 1:107-151.

Roper, Elmo. 1940. Sampling public opinion. *Jour. Amer. Stat. Assoc.* 35:325-334.

Sarle, C. F. 1939. Future improvement in agricultural statistics. *Jour. Farm Econ.* 21:838-845.

——— 1940. The possibilities and limitations of objective sampling in strengthening agricultural statistics. *Econometrika.* 8:45-61.

Schoenberg, E. H., and Mildred Parten. 1937. Methods and problems of sampling presented by the urban study of consumer purchases. *Jour. Amer. Stat. Assoc.* 32:311-322.

Schultz, Henry. 1930. The standard error of a forecast from a curve. *Jour. Amer. Stat. Assoc.* 25:139-185.

Snedecor, G. W. 1939. Design of sampling experiments in the social sciences. *Jour. Farm Econ.* 21:846-855.

Stephan, F. F. 1939. Representative sampling in large-scale surveys. *Jour. Amer. Stat. Assoc.* 34:343-352.

———, **W. E. Deming, and M. H. Hansen.** 1940. The sampling procedure of the 1940 population census. *Jour. Amer. Stat. Assoc.* 35:615-630.

Student. 1908. The standard error of the mean. *Biometrika.* 6:1-25.

Sukhatme, P. V. 1935. Contribution to the theory of the representative method. *Supplement Jour. Royal Stat. Soc.* 2:253-268.

Wishart, J., and A. R. Clapham. 1929. A study in sampling technique: The effect of artificial fertilizers on the yield of potatoes. *Jour. Agri. Sci.* 19:600-618.

Yates, F. 1935. Some examples of biased sampling. *Annals of Eugenics.* 6:202-213.

———, **and D. J. Watson.** 1935. Observer's bias in sampling observation on wheat. *Empire Jour Exp. Agri.* 3:174-177.

———, **and I. Zacopanay.** 1935. The estimation of the efficiency of sampling, with special reference to sampling for yield in cereal experiments. *Jour. Agri. Sci.* 25:545-577.

Youden, W. J., and A. Mehlich. 1937. Selection of efficient methods for soil sampling. *Contrib. Boyce Thompson Inst.* 9:59-70.