ESTIMATION OF VOLUME IN TIMBER STANDS BY STRIP SAMPLING

By A. A. Hasel

California Forest and Range Experiment Station[1]

1. Introduction. The present paper is the second of a proposed series, in which it is intended to present a systematic study of the properties of several methods of sampling timber stands and statistical treatments of the samples.

The effects of size, shape, and arrangement of sampling units on the accuracy of sample estimates of timber stand volume were reported in the earlier paper [1] for 5,760 acres of the Blacks Mountain Experimental Forest. With complete inventory data, the nature of stand variation was shown to be such that 2.5-acre plots, the smallest size tested, were more efficient sampling units than larger plots, i.e., for a given intensity of sampling the sampling error was smaller. Long, narrow plots were more efficient than square plots of the same size. Line-plot sampling units consisting of two or more equally spaced plots along lines of fixed length were as efficient as single-plot sampling units and more efficient than strips consisting of plots contiguous end to end. Improvement in the accuracy of estimates was obtained by subdividing the area into rectangular blocks of equal size, and sampling each block to the same intensity. By systematic sampling, whereby the center lines of parallel line-plot or strip sampling units were spaced equidistant, the sample estimates of stand volume were improved over estimates from comparable random samples. Treatment of the volumes on individual plots of systematic samples as random sampling observations, however, as is sometimes done in practice, was shown to give seriously biased estimates of sampling error.

In the present paper we shall be concerned with sample estimates from strip samples taken within blocks of irregular shape, and consequently with sampling units which vary in length within samples. The methods will be equally applicable to line plot samples.

Following the general ideas expressed by Neyman [2] it is felt that: (1) If the formulae of the theory of probability have to be applied at all to the treatment of samples, the theoretical model of sampling must involve some element of randomness. (2) This element of randomness may conveniently be introduced by a random selection of the sample, but may also be assumed present in the distribution of deviations of timber stand volumes in the area sampled from a postulated pattern. (3) Many attempts to treat systematic arrangements statistically are faulty because the treatment consists in applying to systematic arrangements formulae that are deduced under the assumption of randomness. If the arrangement of sampling is a systematic one, and random errors are

[1] Maintained by the U. S. Department of Agriculture at Berkeley, in cooperation with the University of California.

ascribed to Nature, then the treatment of the data should be based on formulae deduced under explicit assumption of the systematic arrangement of sampling and of some random element in the material. An example of this kind of treatment is provided by Neyman's method of parabolic curves [2] devised for the treatment of systematically arranged agricultural experiments. (4) Lastly, a mathematical treatment of any practical problem is useful only if the predictions of the theory are in satisfactory agreement with the empirical facts. Whether the method of sampling is random or systematic, the mathematical theory of sampling always involves certain elements that are postulated, either in respect to the method of sampling itself or in respect to the material sampled. To have a reasonable certainty that a particular mathematical treatment is useful in practice it is necessary to make empirical studies to find out whether the deviations from postulates of the theory that may occur in actual situations do or do not seriously affect the validity of the predictions.

2. Notation and definitions. Before proceeding to the main subject of this paper it may be useful to explain the meaning of certain statistical terms and symbols. Following Neyman, a sharp distinction is made between three different conceptions that are frequently confused by the practical statistician.

Definition 1: If $u_1, u_2, \cdots, u_N$ are any fixed numbers, whether provided by some already completed experiment involving randomness, or just arbitrarily selected, Karl Pearson's term "standard deviation" of these numbers and the letter S will be used to denote the expression $S = \sqrt{\Sigma(u_i - \bar{u})^2/N}$ in which $\bar{u} = \Sigma u_i/N$ is the mean of the u's.

Now let X denote a random variable, that is a variable the value of which is going to be determined by a chance experiment. Thus X may be the timber volume on a strip that is going to be selected at random from an area. Denote by $E(X)$ the mathematical expectation of variable X capable of possessing values $u_1, u_2, \cdots, u_n$. Then

$$E(X) = u_1p_1 + u_2p_2 + \cdots + u_np_n,$$

in which the p's are the respective probabilities of all possible different values of X.

Definition 2: The words "standard error of X" and the letter σ_x will be used to denote the expression

$$\sigma_x = \sqrt{E[X - E(X)]^2}.$$

It will be noticed that the standard error of a random variable X may have its value equal to the standard deviation of some numbers u but that this does not mean that the two conceptions are identical or even similar. The $E(X)$, and consequently σ_x, can be calculated only when the probability law of X is known, and are constant for the population from which samples are drawn. On the other hand, S can be calculated for any sample of the population and changes in value from one sample of u's to another.

Before proceeding to the third conception, that of an estimate of the standard error, which is occasionally confused with the standard deviation or the standard error, the unbiased estimate of a parameter must be defined [3].

Consider a set of n random variables $X_1, X_2, \cdots, X_n$. These may be, for example, the volumes of timber to be observed on n strips that are going to be selected from some area by one random method or another. Denote by θ a parameter involved in the probability law of the X's. For example, θ may be the total volume of timber in the area.

Let F be any function of the X's.

DEFINITION 3: If it happens that the mathematical expectation of F is identically equal to θ, then it will be said that F is an unbiased estimate of θ.

Usually there will be an infinity of unbiased estimates of a parameter θ. They may be classified by the nature of the function F. Thus linear estimates may be considered such that

$$F = \lambda_0 + \lambda_1 X_1 + \cdots + \lambda_n X_n$$

in which the λ's stand for some fixed numbers.

DEFINITION 4: It will be said that a linear unbiased estimate of θ is the best linear unbiased estimate (B. L. U. E.) if its standard error is smaller than or, at most, equal to that of any other linear unbiased estimate.

It happens frequently that, while it is possible to determine the best linear unbiased estimate F of a parameter, it is not possible to calculate the value of its standard error, σ_F. For this purpose it would be necessary to know the whole population sampled. In such cases an unbiased estimate of the square, σ_F^2, is calculated. An unbiased estimate of the square of the standard error of F will be denoted by μ_F^2. This is the third of the conceptions mentioned above.

The reason for the extensive use of the linear unbiased estimates and of their standard errors considered as measures of accuracy is the so-called Theorem of Liapounoff. Its content can roughly be explained as follows: If the variables $X_1, X_2, \cdots, X_n$ are independent and the number n not too small, then the probability that $F - \theta$ will exceed a fixed multiple of σ_F is approximately equal to the probability as determined by the normal law. The above conclusion remains true whatever the probability distribution of the X's that is likely to be met in practice and also in certain cases where the X's are mutually dependent, for example, when they are determined by sampling a finite population without replacement [4].

The above conclusions do not apply to estimates that are biased in the sense of the above definition. . Also the standard error of such an estimate would not be a satisfactory measure of its accuracy.

3. Description of data. Complete inventory data from the Blacks Mountain Experimental Forest, located in the Lassen National Forest, provide suitable material for testing the applicability of sampling theory to timber cruising.

The timber is a virgin, all-aged stand, classed as pure pine type, with more than 90 per cent of the volume in ponderosa pine and Jeffrey pine. Most of the volume is in over-mature trees, i.e., trees over 300 years in age. The stand is considered to be fairly representative of the medium and the poor site qualities of the northeastern California plateau.

With the exception of a few localities, all of the area was mapped as of uniform timber type according to the standards commonly used. Being fairly uniform also with respect to site quality, it may therefore be considered as a single *stratum*. Variability of stand volume from place to place within a stratum may be generally expected to be less, on the average than variability between places in different strata. Likewise, within a stratum, variability within compact subdivisions may be expected to be less than average variability within the whole. Heterogeneity can therefore be controlled somewhat by subdividing the stratum into blocks and treating each block as a separate population.

More frequently than not, in practice, volume estimates are needed both for the total timbered area and for separate working units or compartments within the area. In general, working unit boundaries are defined by roads, ridge tops, drainage channels, and regular land subdivision lines. These working units can be taken conveniently as blocks, or if large enough, may be subdivided into two or more blocks. Such is the basis used for subdividing the area in the present study.

The complete inventory data for these blocks are given in Table I. All the strips are $2\frac{1}{2}$ chains in width and extend in an east-west direction. The length, X, is given in 10-chain units, and the volume, Y, is given in units of 1,000 feet board measure.

4. Method of estimation based on correlation between volume and strip length. The usual practice in sampling timber stand volume is to take measurements on plots or strips that are either regularly spaced or selected at random from all possible plots or strips within blocks. Oftener than not blocks are irregular in shape, and the number of plots along lines or the lengths of strips will vary. This variation introduces the matter of proper "weighting" in calculating sample statistics. Such is the case in 15 of the 20 Blacks Mountain blocks.

If we let Y_i represent the volume on the ith strip of length X_i, with length expressed say in 10-chain units, and assume that the entire block contains a population of N strips, then the average volume to the unit of strip is $\beta = \sum_{i=1}^{N} Y_i \Big/ \sum_{i=1}^{N} X_i$. It is obvious that, if $\sum X_i$ is known, and this is assumed to be true, the problem of estimating β is equivalent with that of estimating the total volume. The usual procedure of estimating is this:

Out of the N strips within the block a sample of n is taken, giving n pairs of numbers selected out of the X's and Y's. Let us denote them by

$$x_1, y_1; x_2, y_2; \cdots; x_n, y_n.$$

The ratio $b = \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i$, is then considered an estimate of β, so that the estimate of the total volume in the block is $b \sum_{i=1}^{N} X_i$.

Our purpose now will be to study the above estimate b from the point of view of unbiasedness. In this paper it is assumed that the sampling of strips is purely random. To find out whether b is an unbiased estimate or not, its expectation must be calculated. This will be done in two steps. To begin with assume that the *values* $x_1, x_2, \cdots, x_n$ are chosen in one way or another and fixed. The value of b will then depend on the y's only. It is possible that to a given value of x, say x_1, there will correspond just one value of y_1 in the block, but generally there will be several strips of the same length x_1 with varying volumes of timber. The selection of any strip of this group to be included in the sample will keep the denominator of b constant, but will cause some variation in the numerator. The expectation of b calculated under the assumption that the x's are fixed is

$$(1) \qquad E(b \mid x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} E(y_i \mid x_i) \Big/ \sum_{i=1}^{n} x_i,$$

in which $E(y_i \mid x_i)$ denotes the expectation of y_i calculated under the assumption that x_i has a fixed value. Obviously $E(y_i \mid x_i)$ will be what is called the regression function of y on x, or of volume on the length of strip.

It is safe to say that the graph of $E(y \mid x)$ would almost always be rather irregular. On the other hand, it is known that the substitution of smooth curves representing the regressions for the true irregular polygons frequently gives results that are surprisingly accurate. Therefore it would not be unreasonable to use the assumption that $E(y \mid x)$ can be represented by a polynomial of some moderate degree,

$$E(y \mid x) = A_0 + A_1 x + A_2 x^2 + \cdots + A_s x^s.$$

Substitution of this expression in (1) gives

$$E(b \mid x_1, x_2, \cdots, x_n) = \frac{nA_0}{\sum_{i=1}^{n} x_i} + A_1 + A_2 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} + \cdots + A_s \frac{\sum_{i=1}^{n} x_i^s}{\sum_{i=1}^{n} x_i}.$$

But this is the conditional expectation of b, calculated under the assumption of fixed x's, is only an intermediate stage in the calculations. We need an absolute expectation, calculated under the assumption that the x's are selected at random. This gives

$$(2) \quad E(b) = A_0 E\left(\frac{n}{\sum_{i=1}^{n} x_i}\right) + A_1 + A_2 E\left(\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i}\right) + \cdots + A_s E\left(\frac{\sum_{i=1}^{n} x_i^s}{\sum_{i=1}^{n} x_i}\right)$$

TABLE I

Complete inventory data for 15 blocks of the Blacks Mountain Experimental Forest

Strip[2] number	Block number															
	1		2		3		4		5		6		8		9	
	x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y
1	12	762.4	9	470.6	4	63.6	7	174.3	5	121.4	9	331.1	7	247.5	3	143.6
2	12	651.0	9	426.1	4	73.7	7	159.2	5	169.5	9	377.7	7	286.4	3	159.2
3	12	461.4	9	448.5	4	91.7	7	156.6	6	315.8	9	295.1	7	339.5	4	209.2
4	12	521.1	9	401.5	4	61.4	7	139.6	7	307.1	9	237.8	7	360.2	4	209.5
5	12	652.7	9	372.1	4	35.5	7	198.0	7	318.9	9	305.0	8	336.2	4	227.2
6	12	543.7	9	372.2	4	82.8	7	168.7	7	366.0	9	284.4	8	332.7	5	247.0
7	12	541.5	9	410.6	4	109.2	7	127.1	9	445.0	9	362.9	8	330.1	5	277.0
8	12	589.6	9	322.8	4	109.6	7	181.3	9	406.4	9	352.4	8	403.6	5	371.9
9	11	532.6	9	380.6	5	114.9	7	155.8	9	427.1	9	345.1	8	378.9	6	303.5
10	11	516.9	9	429.9	5	101.6	7	207.3	9	448.7	9	354.0	6	279.5	6	207.7
11	11	519.5	9	434.1	5	94.4	7	181.4	9	381.6	9	401.2	4	206.5	7	305.0
12	11	538.8	9	394.5	5	104.3	7	121.3	9	243.3	9	381.0	2	130.1	7	353.4
13	10	508.6	9	542.9	6	161.2	7	124.6	9	360.0	8	408.4	1	45.3	8	368.7
14	10	448.6	9	606.6	6	191.6	7	86.0	9	343.9	8	374.5			8	361.1
15	10	492.5	8	416.5	6	223.5	3	115.3			8	374.3			8	332.8
16	10	498.3	8	325.9	6	164.8	3	128.8			8	337.3			8	289.6
17																
18																
19																
20																
21																
22																
23																
24																
25																
26																
27																
28																
29																
30																
31																
32																
Total......	180	8,779.2	142	6,755.4	76	1,783.8	104	2,425.3	109	4,654.7	140	5,522.2	81	3,676.5	91	4,366.4

TABLE I (*Continued*)

Strip number	Block number													
	12		13		14		15		18		19		20	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
1	1	26.1	1	21.1	6	184.1	10	300.7	8	286.0	2	34.4	2	104.8
2	1	48.5	1	23.3	6	198.1	9	170.1	8	306.1	2	43.7	2	104.3
3	4	137.8	2	40.4	7	240.9	9	270.7	8	319.0	2	33.3	1	20.7
4	8	329.9	3	43.1	7	348.4	9	386.7	8	285.5	2	57.0	1	36.7
5	10	354.8	4	50.5	7	315.3	7	246.7	8	282.4	3	89.1	1	56.1
6	12	401.0	4	95.9	7	262.4	7	215.8	8	281.8	3	72.1	1	42.4
7	14	499.7	5	145.4	7	243.6	7	248.4	7	218.8	3	43.0	1	22.9
8	15	554.5	5	155.6	7	312.7	7	265.9	7	244.2	3	81.4	1	75.5
9	15	534.0	5	126.9	7	287.2	7	292.4	6	171.7	3	91.0	1	62.5
10	15	519.0	5	161.0	7	254.4	7	221.6	5	127.5	3	94.3	1	39.5
11	15	570.7	5	213.9	7	303.8	7	268.4	2	58.4	3	141.4	1	40.8
12	9	402.8	5	143.0	7	338.9	7	291.3	2	79.9	3	107.6	1	72.8
13	9	354.5	5	201.0	7	303.3	7	279.0	1	28.0	3	108.8	1	34.5
14	9	363.1	5	198.3	7	233.6	7	312.6	1	13.8	3	117.3	1	35.7
15	9	561.9	5	173.1	7	284.6	7	303.2			3	122.3	1	22.3
16	6	223.8	5	219.9	7	306.4	7	278.5			3	97.5	1	52.1
17	6	245.5	5	296.8	7	298.4	7	301.1			3	126.8	1	37.0
18	5	255.7	5	150.5	7	218.7	7	260.3			3	163.0	1	36.5
19	5	244.8	5	157.5	6	230.6	7	258.5			3	111.9	2	61.0
20	5	245.1	5	162.6	6	187.8	7	261.6			3	140.8	2	72.2
21	4	248.0	5	140.7	6	189.4	7	256.0			2	87.9	2	95.1
22	4	250.3	5	212.1	6	179.8	6	178.2			2	89.0	2	99.3
23	4	200.4	5	142.8	6	124.6	6	242.9					2	79.8
24	2	98.1	5	224.5	5	151.5	4	193.1					2	82.1
25	2	151.8	6	174.0	3	109.2	3	149.8					2	56.7
26	1	40.8	6	187.2	2	72.8	2	58.5					2	77.3
27	1	45.8			1	31.5	1	36.9					2	68.5
28													2	73.4
29													2	31.5
30													2	33.6
31													2	44.5
32													2	105.5
Total.........	191	7,908.4	117	3,861.1	165	6,212.0	178	6,548.9	79	2,703.1	60	2,053.6	48	1,877.6

[2] Numbered in order from north to south within blocks.

The value of β has the form of (2), except that instead of $E(\Sigma x_i^m/\Sigma x_i)$ it contains $\Sigma X_i^m/\Sigma X_i$. Since in general the former does not necessarily equal the latter, for the unbiasedness of b it is necessary and sufficient that $A_0 = A_2 = \cdots = A_s = 0$. This condition implies that the regression line of y on x is a straight line and passes through the origin of coordinates,

$$(3) \qquad E(y \mid x) = A_1 x.$$

Whether (3) is satisfied is a question of fact and can be authoritatively answered only by direct studies of regressions on some extensive inventory data. It may be noted also that in order to presume that (3) is *usually* satisfied, it should be established for a large number of areas. On the contrary, if a study of only a few areas shows that (3) is not true, then it would not be wise to take it for granted when attempting to make a sampling inventory of an unfamiliar area.

To investigate this point, linear regression equations of volume on the length of strip were calculated for 15 blocks of the Blacks Mountain Experimental Forest and it was found that the constant terms were both positive and negative with their absolute values varying from 12 to 677. The conclusion drawn is that the usual estimate b of the average volume per unit of strip is likely to be biased and that there is justification in looking for an alternative method leading to unbiased estimates.

5. Best linear unbiased estimate of volume, based on the linear regression of volume on length of strip. In this section will be suggested a method of estimating the total volume, say θ, of a timber stand, which could be considered as an improvement on the one considered above. The new method consists of using a linear unbiased estimate of θ. In order to deduce the form of this estimate, certain assumptions have to be taken for granted concerning the timber stands to be sampled, and if it happens that these assumptions are unsatisfied in a particular case, the new estimate will not necessarily possess the desired property of unbiasedness.

In deducing .the estimate F it will be assumed that the timber stand to be sampled satisfies the following conditions: (1) That the regression of timber volume on length of strip, X, be (approximately) linear and (2) that the variability of the Y's for a fixed X is precisely known. It will not be assumed, however, that the linear regression line passes through the origin of coordinates, and this will allow F to be unbiased in such cases, as exhibited above, where b is biased. Following the Markoff method [5], [3] it can easily be shown that there is an infinity of linear estimates of θ which are unbiased under condition (1). It follows that a choice can be made among them so as to diminish the standard error. This, however, is possible only when something is known about the variability of the Y's when the value of X is fixed. For the present we shall assume condition (2) concerning this particular point, but in practice this will generally be quite impossible. This point will be considered further in Section 6.

Consider then a sure or non-random variable[3] X able to assume the particular values $X_1, X_2, \cdots, X_s$. Assume that there is a finite population π_i of N_i numbers $u_{i1}, u_{i2}, \cdots, u_{iN_i}$ corresponding to each value X_i, $i = 1, 2, \cdots, s$. Assume that the mean $u_{i\cdot}$ of the population π_i is a linear function of X_i, i.e., for any i,

$$u_{i\cdot} = A + BX_i,$$

with some unknown values of A and B.

Assume that out of each population π_i there is selected without replacement a random sample of n_i individuals, with $0 \leq n_i \leq N_i$, and denote by $y_{i1}, y_{i2}, \cdots, y_{in_i}$ the values of the u's to be drawn.

If the regression of the amount of timber on the length of the strip is linear, then the problem of estimating the total stand is equivalent to that of estimating

$$\theta = \sum_{i=1}^{s} N_i(A + BX_i) = A \sum_{i=1}^{s} N_i + B \sum_{i=1}^{s} N_i X_i.$$

Since the length of the strip, X, could be measured from any arbitrarily chosen origin, no generality will be lost by assuming $\sum_{i=1}^{s} N_i X_i = 0$, so that $\theta = A \sum_{i=1}^{s} N_i = AN$ (say). Weighting the y_{ij} equally for each fixed i the B. L. U. E. of θ may be denoted by

$$F = \sum_{i=1}^{s} n_i \lambda_i y_{i\cdot}, \tag{4}$$

in which $y_{i\cdot} = \Sigma y_{ij}/n_i$. Here the λ's must satisfy the conditions of unbiasedness,

$$E(F) \equiv \theta, \tag{5}$$

and of optimum,

$$\sigma_F^2 = \text{minimum}. \tag{6}$$

It may easily be shown that condition (5) will be fulfilled by (4) if the λ_i's are so selected that

$$\sum_{i=1}^{s} n_i \lambda_i = N; \qquad \sum_{i=1}^{s} n_i \lambda_i X_i = 0. \tag{7}$$

[3] This is an English translation from an excellent French term "nombre certain" and "fonction certaine" to denote a non-random number and non-random function, invented by Fréchet.

Condition (6) may now be considered. From the general formula for the variance of a linear function of several random variables and the fact that y_{ij} is independent of y_{kl},

$$\begin{aligned} \sigma_F^2 &= \sum_{i=1}^{s} \{n_i\lambda_i^2 S_i^2 + n_i(n_i - 1)\lambda_i^2 E[(y_{ik} - u_{i\cdot})(y_{il} - u_{i\cdot})]\} \\ &= \sum_{i=1}^{s} \left[n_i\lambda_i^2 S_i^2 - \frac{S_i^2}{N_i - 1}(n_i^2\lambda_i^2 - n_i\lambda_i^2) \right] \qquad (8) \\ &= \sum_{i=1}^{s} S_i^2 \frac{n_i(N_i - n_i)}{N_i - 1}\lambda_i^2 = \sum_{i=1}^{s} A_i^2\lambda_i^2 \text{ (say)}, \end{aligned}$$

in which S_i^2 stands for the (S.D.)^2 of the population π_i, i.e.,

$$S_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i} (u_{ij} - u_{i\cdot})^2.$$

In addition to satisfying equations (7), the λ_i's must be selected so as to minimize (8).

Using the method of Lagrange, we find

$$(9) \qquad \lambda_i = \frac{n_i}{A_i^2}(\alpha + \beta X_i),$$

for the case where $0 < n_i < N_i$ and $A_i \neq 0$. If $n_i = N_i$, then $A_i = 0$ and $\alpha + \beta X_i = 0$.

Assume first that all $n_i < N_i$, $i = 1, 2, \cdots, s$. Then α and β are obtained from equations (7) after substituting in them (9), namely

$$(10) \qquad \begin{cases} \alpha \sum_i w_i + \beta \sum_i w_i X_i = N \\ \alpha \sum_i w_i X_i + \beta \sum_i w_i X_i^2 = 0, \end{cases}$$

where, for simplicity

$$(11) \qquad w_i = \frac{n_i^2}{A_i^2} = \frac{(N_i - 1)n_i}{(N_i - n_i)S_i^2}; \qquad \sum_{i=1}^{s} w_i = W.$$

If w_i is considered as the weight of the observations at $X = X_i$, it will be convenient to introduce a weighted mean and weighted S.D. of X's as follows:

$$(12) \qquad \bar{x} = \frac{\sum_{i=1}^{s} w_i X_i}{W}; \qquad S_x^2 = \frac{\sum_{i=1}^{s} w_i X_i^2}{W} - \bar{x}^2.$$

With this notation equations (10) can be rewritten and easily give

$$(13) \qquad \begin{cases} \beta = -\dfrac{N\bar{x}}{WS_x^2}, \\ \alpha = \dfrac{N(S_x^2 + \bar{x}^2)}{WS_x^2}. \end{cases}$$

Substituting these values into (4), simple transformations give

$$F = N(\bar{y} - \bar{x}b_0), \tag{14}$$

in which $\bar{y} = \sum_i w_i y_{i\cdot}/W$, and b_0 represents the unbiased estimate of B and is given by

$$b_0 = [(1/W) \sum_i w_i X_i y_{i\cdot} - \bar{x}\bar{y}]/S_x^2.$$

The next step is to calculate σ_F^2. Substituting (9) in (8) and using (11) and (12) gives

$$\sigma_F^2 = W(\alpha + \beta\bar{x})^2 + W\beta^2 S_x^2.$$

Using (13) gives finally

$$\sigma_F^2 = \frac{N^2}{W}\left(1 + \frac{\bar{x}^2}{S_x^2}\right). \tag{15}$$

If X is the length of a given strip in any chosen units and $\bar{X}$ the average of such X's for a given block, then (14) and (15) may be written

$$\begin{cases} F = N[\bar{y} + b_0(\bar{X} - \bar{x})] \\ \sigma_F^2 = \dfrac{N^2}{W}\left[1 + \dfrac{(\bar{X} - \bar{x})^2}{S_x^2}\right]. \end{cases} \tag{16}$$

Similarly for the case where one of the n_i's equals N_i, for example, $n_1 = N_1$, we find

$$F = N[y_{1\cdot} + \bar{b}(\bar{X} - X_1)], \tag{17}$$

in which

$$\bar{b} = \frac{\sum_{i=2}^{s} w_i(X_i - X_1)(y_{i\cdot} - y_{1\cdot})}{\sum_{i=2}^{s} w_i(X_i - X_1)^2}.$$

Also

$$\sigma_F^2 = \frac{N^2(\bar{X} - X_1)^2}{\sum_{i=2}^{s} w_i(X_i - X_1)^2}. \tag{18}$$

It should be emphasized that X_1 in the above formulae does not necessarily represent the smallest of the X's but the one of them for which $n_i = N_i$.

The case where two or more of the n_i's are respectively equal to the corresponding N_i's need not be considered in detail. Together with the assumption of a strict linearity of regression such an assumption, for example, that $n_1 = N_1$, and $n_2 = N_2$, would lead to the conclusion that the regression of volume on the length of strip is accurately known and that the estimation of θ could be made without error. Owing to the fact that the hypothesis about the linearity of regression is, at best, only approximately correct, the errors of estimation will always be present and it is imperative either to arrange the sampling so as to have at most one of the n_i's equal to the corresponding N_i, or to base the statistical treatment of the sample on a theory different from the one considered here.

6. Additional hypotheses concerning S_i^2. The formulae (16) with the w_i's determined by (11) are impossible to apply in practice because we do not know the values of the S_i^2. The best we can do is to make plausible guesses as to what may be the values of the S_i^2. These guesses are bound to be at most approximately correct and therefore the estimates of θ that one can apply in practice will be only "approximately best." It is easy to see, however, that we may keep them unbiased.

Suppose that we denote by r_i^2 the presumed value of S_i^2. Substituting this value in place of S_i^2 in (8) we should repeat all the calculations, leading us to such λ_i' that will assure the unbiasedness of, say

$$F_g = \sum_{i=1}^{s} n_i \lambda_i' y_{i\cdot},$$

and also a minimum value of, say

$$\sigma_g^2 = \sum_{i=1}^{s} r_i^2 \frac{n_i(N_i - n_i)}{N_i - 1} \lambda_i'^2.$$

The values of the λ_i' will be obtained from the same formulae as those of λ_i, except that instead of S_i^2 they will depend on r_i^2. Consequently F_g will have the same form as F,

$$F_g = N[\bar{y}' + b_0'(\bar{X} - \bar{x}')], \tag{19}$$

with the difference that $\bar{x}'$, $\bar{y}'$, S_x', and b_0' will now have to be calculated with different weights, say

$$v_i = \frac{(N_i - 1)n_i}{(N_i - n_i)r_i^2}; \qquad V = \sum_{i=1}^{s} v_i.$$

If the form of the unbiased estimate F_g is as that of F, the square of its standard error σ_g^2 is more complicated. In order to calculate it we have to go back to (8) and substitute into it the new values of λ_i' obtained from the guessed weights v_i,

$$\lambda_i' = \frac{N_i - 1}{(N_i - n_i)r_i^2}(\alpha' + \beta' X_i),$$

with

$$\left\{\begin{aligned} \alpha' &= \frac{N}{VS_x'^2}(S_x'^2 + \bar{x}'^2) \\ \beta' &= -\frac{N\bar{x}'}{VS'^2}, \end{aligned}\right. \tag{20}$$

we have

$$\sigma_g^2 = \sum_{i=1}^{s} S_i^2 \frac{n_i(N_i - n_i)}{N_i - 1} \lambda_i'^2 \tag{21}$$
$$= \sum_{i=1}^{s} v_i \rho_i (\alpha' + \beta' X_i)^2,$$

where $\rho_i = S_i^2/r_i^2$. It will now be helpful to introduce notation for another kind of weighted mean and weighted (S.D.) of the X's, with weights equal to $v_i\rho_i$. So let us write

$$\bar{x}'' = \frac{\sum_i v_i \rho_i X_i}{\sum_i v_i \rho_i}; \qquad S_x''^2 = \frac{\sum_i v_i \rho_i X_i^2}{\sum_i v_i \rho_i} - \bar{x}''^2. \tag{22}$$

Expanding (21) and using (20), we have

$$\sigma_g^2 = \frac{N^2 \sum_i v_i \rho_i}{V^2} \left\{ \left[1 + \frac{\bar{x}'(\bar{x}' - \bar{x}'')}{S_x'^2} \right]^2 + \frac{\bar{x}'^2 S_x''^2}{S_x'^4} \right\}. \tag{23}$$

Formula (23) refers to the case where the X's are measured from their population, mean, $\bar{X}$. In order to reduce it to the case where the X's are given in their original values we have to substitute $(\bar{x}' - \bar{X})$ for $\bar{x}'$ and $(\bar{x}'' - \bar{X})$ for $\bar{x}''$. Thus

$$\sigma_g^2 = \frac{N^2 \sum_i v_i \rho_i}{V^2} \left\{ \left[1 + \frac{(\bar{x}' - \bar{X})(\bar{x}' - \bar{x}'')}{S_x'^2} \right]^2 + \frac{(\bar{x}' - \bar{X})^2}{S_x'^2} \frac{S_x''^2}{S_x'^2} \right\}. \tag{24}$$

Applying a similar procedure to the case where $n_1 = N_1 = 1$, but $n_i < N_i$ for $i = 2, 3, \cdots, s$, we easily find

$$F_g = N \left[y_{1\cdot} - (X_1 - \bar{X}) \frac{\sum_{i=2}^{s} v_i (X_i - X_1)(y_{i\cdot} - y_{1\cdot})}{\sum_{i=2}^{s} v_i (X_i - X_1)^2} \right], \tag{25}$$

and

$$\sigma_g^2 = N^2 (X_1 - \bar{X})^2 \frac{\sum_{i=2}^{s} v_i \rho_i (X_i - X_1)^2}{\left[\sum_{i=2}^{s} v_i (X_i - X_1)^2 \right]^2}. \tag{26}$$

This formula will help us to test the appropriateness of guesses about the values of S_i^2. It will be noticed that the λ's contain S_i^2 or r_i^2 in the same powers in the numerator and in the denominator. It follows that all we need to guess

TABLE II

Values of S_i^2

X_i	1		2		3		4		5		6		7		8		9		10		11		12		15	
Block	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2	N_i	S_i^2
1																			4	525	4	82	8	7,894		
2															2	2,052	14	4,931								
3							8	561	4	54	4	625														
4					2	46							14	1,027												
5									2	578			3	647			8	4,028								
6															4	633	12	2,031								
8													4	1,961	5	879										
9					2	61	3	71	3	2,834	2	2,294	2	586	4	961										
12	4	75	2	721			4	2,093	3	26	2	118					4	6,989							4	387
13	2	1					2	515	18	1,731	2	44														
14											7	851	16	1,380												
15											2	1,047	17	691			3	7,832								
18	2	50	2	116									2	161	6	198										
19			6	538	16	814																				
20	16	261	16	554																						
Total	24		26		20		17		30		19		58		21		41		4		4		8		4	
Weighted average		191		529		662		830		1,370		814		1,026		765		4,319		525		82		7,894		387

is a system of numbers proportional to S_i^2 and not the S_i^2 themselves. Our problem will be to test a few such guesses on the data of the Blacks Mountain Experimental Forest and see which of them gives generally a smaller value of σ_g^2.

Table II gives values of the S_i^2, calculated for 15 blocks, together with the corresponding X_i. In a few cases $N_i = 1$ and consequently $S_i = 0$. These cases are not included in the table. Using the values of S_i^2 from Table II and assuming systems of the n_i's, the values of σ_F^2 were calculated for these blocks. These would be the true (S.E.)^2 of the best linear estimates of the total timber volume in each block, but it would never be possible to calculate them from sample data.

The σ_g^2's were calculated using the following guesses concerning the S_i^2: (1) That they do not depend on X_i, (2) that they are proportional to X_i, and (3) that they are proportional to $\sqrt{X_i}$. The ratios σ_F^2/σ_g^2 for all blocks taken together were found to be .770 for guess (1), .769 for guess (2), and .777 for guess (3). It is seen that, on the average, the guess that the S_i^2 are proportional to X_i gives the smallest average value of σ_g^2. It is interesting, however, to note that the differences between the three guesses are, for all practical purposes, negligible.

Ratios like σ_F^2/σ_g^2 are sometimes described as the "amount of information" in F_g as compared with that in the best linear unbiased estimate F. This expression was introduced by R. A. Fisher. In certain cases it has the following property which justifies the term used: Let n be the size of the sample which serves for calculating F_g, then, if it were possible to calculate the best linear unbiased estimate F, the same accuracy of estimation would be obtained by using a smaller sample size $n\sigma_F^2/\sigma_g^2$. In the case considered in the present paper the above circumstance does not occur. Still, the ratio σ_F^2/σ_g^2 seems to be convenient to describe the situation.

7. Another scheme for estimating θ.

It will be noticed that the ignorance of what are the S_i^2 is not the only circumstance which makes it difficult to apply the above formulae. There is also another one connected with the values of N_i. We have $N_i = 1$ in several blocks and for several different strip lengths. True this might have been avoided by defining block boundaries in such a way that $N_i \geq 2$, but it was considered best to conform strictly to the practical situation where the N_i's may be smaller. In such cases we may include in our sample all the strips of a given length, say X_1. If we apply to such samples the above formulae, deduced under the explicit assumption that the regression of Y on X is strictly linear, we shall force the fitted regression line through the point (X_1, $Y_1.$). As the assumption of strict linearity is obviously not exact and the exhaustion of strips of length X_1 is possible only when there are very few such strips, the whole procedure may lead to serious inaccuracies in the final estimate. One safeguard against this is never to exhaust strips of any given length when dealing with formulae deduced from finite populations.

The fact that the true regression point (X_1, $Y_1.$) does not actually lie on a straight line makes it uncertain whether taking into account the finiteness of populations of strips of the same length is beneficial to the accuracy of the finite estimate. In the preceding sections we worked on the assumption that there is but a finite number of strips of the same length and on an inaccurate assumption that the regression is strictly linear. In the present section the first assumption will be dropped, having in mind that the effect of the inaccuracy of the second assumption may thereby be reduced.

The assumption that each of the N_i is infinite will be made only in deducing the λ_i and will be reflected in weights. Formula (11) will now reduce to $w_i = n_i/S_i^2$. If we assume further that $S_i^2 = X_i^\gamma/k$, where γ and k are some constants, then

$$\bar{w}_i = \frac{kn_i}{X_i^\gamma}; \qquad \bar{W} = \sum_i \bar{w}_i = k \sum_i \frac{n_i}{X_i^\gamma},$$

and the final estimate is

$$F = N[\bar{y} + \bar{b}_0(\bar{X} - \bar{x})]. \tag{27}$$

The square of the standard error of F has again the same form as in (16),

$$\sigma_F^2 = \frac{N^2}{\bar{W}}\left[1 + \frac{(\bar{X} - \bar{x})^2}{\bar{S}_x^2}\right], \tag{28}$$

the only differences being in $\bar{W}$, $\bar{x}$, and $\bar{S}_x^2$. If $\gamma = 0$, so that the S_i^2 are assumed to be constant, then

$$\bar{w}_i = kn_i; \qquad \bar{W} = k \sum_i n_i,$$

and all the symbols $\bar{x}$, $\bar{y}$, and $\bar{S}_x^2$ assume their customary meaning of ordinary means and of ordinary (S.D.)^2.

It would be easy to deduce explicit formulae for $\gamma = 1/2$, etc., but they are not elegant and, if the necessity arises, the calculations could be carried through by starting with $\bar{w}_i = 1/X_i^\gamma$. The omission of k does not influence the form of F.

The question whether the combination of one true hypothesis about the N_i being finite, with another incorrect one that the regression is strictly linear, is better than that of two incorrect hypotheses, will be studied by means of a sampling experiment in Section 9.

8. Unbiased estimates of σ_F^2. While it may not be unreasonable to hope that a guess of a system of numbers proportional to the S_i^2 may be successful, it is entirely hopeless to try to guess the actual values of the S_i^2. It follows that, if it is desired to obtain from the sample some sort of measure of the accuracy of F, we have to calculate an estimate of σ_F^2.

We shall treat the problem by assuming that the regression of Y on X is strictly linear and that the S_i^2 are proportional to X_i^γ and that the N_i are all finite. It will be noticed that they will enter the formulae by means of the ratios $(N_i - 1)/(N_i - n_i)$. If it is desired to obtain formulae referring to the assumption of infinite N_i's, it will be sufficient to replace these ratios by unity. Of course, the symbol N will always represent the total number of strips in the actual block on which it is desired to estimate the volume of timber and will not be affected by the assumption of the N_i's being infinite.

On these assumptions

$$E(y_{i.}) = A + BX_i,$$

$$\sigma^2_{y_{ij}} = E(y_{ij} - A - BX_i)^2 = S_i^2 = kX_i^\gamma,$$

with some value of γ supposed to be accurately guessed, which however need not be specified, and with an unknown factor of proportionality, k. The square of the standard error of $y_{i.}$ is then known to be

$$\sigma^2_{y_{i.}} = \frac{S_i^2}{n_i}\,\frac{N_i - n_i}{N_i - 1} = k\,\frac{X_i^\gamma(N_i - n_i)}{n_i(N_i - 1)}\,. \tag{29}$$

The right-hand member of (29) is equal to the reciprocal of what we have formerly denoted by w_i and described as the weight of the observations at $X = X_i$. We have mentioned above that the formula giving F does not depend on the values of the w_i, but on proportions between the w_i. In other words, if we drop the unknown factor k and denote by w_i the ratio

$$\frac{n_i(N_i - 1)}{X_i^\gamma(N_i - n_i)} = w_i, \tag{30}$$

which involves only known quantities, these new weights will lead to exactly the same value of F as the original weights. It will now be convenient to alter the definition of weight and use formula (30). With this new meaning of w_i, (29) could be rewritten $\sigma^2_{y_{i.}} = k/w_i$.

Let us further use the letter m to denote the number of those X_i's for which we have at least one observation. In other words m will be the number of *different lengths* of strips in the sample and also the number of *different* $y_{i.}$'s that we are going to calculate from it.

Now let us go back to formula (16) giving the square of the standard error of F. We notice that, while $\bar{x}$ and S_x^2 do not depend on the unknown factor of proportionality, k, the sum W of the original weights does depend on it and with our new meaning of w_i,

$$W = \frac{1}{k}\sum_i w_i\,.$$

It follows that σ^2_F should now be written in the form

$$\sigma^2_F = \frac{kN^2}{\sum_i w_i}\left[1 + \frac{(\bar{X} - \bar{x})^2}{S_x^2}\right], \tag{31}$$

and that, in order to estimate σ_F^2 it is sufficient to get an estimate of k. We easily get an unbiased estimate of k by merely applying the second part of the Markoff Theorem [3]. According to it an unbiased estimate of k, based on $m - 2$ degrees of freedom is given by the ratio

$$\sum_{i=1}^{m} \frac{[y_{i\cdot} - \bar{y} - b_0(X_i - \bar{x})]^2}{m - 2} w_i, \tag{32}$$

in which $\bar{x}$, $\bar{y}$, and b_0 are calculated according to the assumptions made regarding N_i and γ. It may be expected, however, that the estimate (32) will not be a very accurate one because the number of degrees of freedom on which it is based may be very small.

In an attempt to find a better estimate of k we shall proceed by analogy and calculate the expectation of a sum similar to the one in the numerator of (32) but depending explicitly on the particular y_{ij}'s, namely of

$$S_0^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} [y_{ij} - \bar{y} - b_0(X_i - \bar{x})]^2 \frac{w_i}{n_i}.$$

It will be noticed that if the N_i are finite, y_{ij} and y_{il} are dependent and that the Theorem of Markoff does not apply to S_0^2. Introduce the notation

$$\eta_{ij} = y_{ij} - A - BX_i,$$

$$\eta_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} \eta_{ij} = y_{i\cdot} - A - BX_i.$$

Easy, but somewhat long calculations show that S_0^2 can be rewritten in the form

$$S_0^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{w_i}{n_i} \eta_{ij}^2 - \frac{\left(\sum_{i=1}^{m} w_i \eta_{i\cdot}\right)^2 + \frac{1}{S_x^2}\left[\sum_{i=1}^{m} w_i (X_i - \bar{x}) \eta_{i\cdot}\right]^2}{\sum_{i=1}^{m} w_i},$$

which is most convenient for calculating the expectation sought. We notice first that

$$E(\eta_{ij}^2) = kX_i^\gamma,$$

$$E(\eta_{i\cdot}^2) = \sigma_{y_{i\cdot}}^2 = \frac{k}{w_i}.$$

Further, as $y_{i\cdot}$ and $y_{j\cdot}$ are mutually independent if $i \neq j$, the same is true for $\eta_{i\cdot}$ and $\eta_{j\cdot}$. It follows that

$$E(y_{i\cdot}\eta_{j\cdot}) = 0, \qquad i \neq j.$$

Consequently

$$E\left(\sum_{i=1}^{m} w_i \eta_{i\cdot}\right)^2 = E\left(\sum_{i=1}^{m} w_i^2 \eta_{i\cdot}^2\right) = \sum_{i=1}^{m} w_i^2 E(\eta_{i\cdot}^2) = k \sum_{i=1}^{m} w_i.$$

Similarly and for the same reason

$$E\left[\sum_i w_i(X_i - \bar{x})\eta_{i\cdot}\right]^2 = kS_x^2 \sum_i w_i .$$

It follows that

$$E(S_0^2) = k\left[\sum_{i=1}^{m} \frac{n_i(N_i - 1)}{N_i - n_i} - 2\right],$$

and that the ratio

$$\frac{S_0^2}{\sum_{i=1}^{m} \frac{n_i(N_i - 1)}{N_i - n_i} - 2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} [y_{ij} - \bar{y} - b_0(X_i - \bar{x})]^2 \frac{w_i}{n_i}}{\sum_{i=1}^{m} \frac{n_i(N_i - 1)}{N_i - n_i} - 2}, \tag{33}$$

is an unbiased estimate of k. In cases where either all $n_i = 1$ or all N_i are infinite the denominator of (33) reduces to the number of degrees of freedom in S_0^2, equal to $\Sigma n_i - 2$. In other cases the denominator of (33) is greater than the number of degrees of freedom. Whether the numerical difference is large or small depends on the fractions $(N_i - 1)/(N_i - n_i)$. We may expect that in many practical cases it will be small.

We shall write

$$S_y^2 = \sum_{i=1}^{m} \frac{w_i}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \Big/ \sum_{i=1}^{m} w_i ,$$

$$r = \frac{\sum_{i=1}^{m} w_i(X_i - \bar{x})y_{i\cdot} \Big/ \sum_{i=1}^{m} w_i}{S_x S_y} .$$

It follows that

$$S_0^2 = \sum_{i=1}^{m} w_i S_y^2(1 - r^2).$$

Substituting this formula into (33) and then the result of this substitution in place of k in (31), we finally get

$$\mu_F^2 = N^2 \frac{S_y^2(1 - r^2)}{\sum_{i=1}^{m} \frac{n_i(N_i - 1)}{N_i - n_i} - 2}\left[1 + \frac{(\bar{X} - \bar{x})^2}{S_x^2}\right]. \tag{34}$$

The case where one of the n_i is equal to the corresponding N_i, e.g., where $n_1 = N_1 = 1$ is treated in a similar manner. Using formula (18) and the notation adopted above, we can write

$$\sigma_F^2 = k \frac{N^2(X_1 - \bar{X})^2}{\sum_{i=2}^{m} w_i(X_i - X_1)^2} .$$

The unbiased estimate μ_F^2 of σ_F^2 will differ from this expression in that instead of the unknown factor k it will contain its unbiased estimate. To find this estimate we proceed exactly as above and calculate the expectation of

$$S_0^2 = \sum_{i=2}^{m} \sum_{j=1}^{n_i} [y_{ij} - y_{1\cdot} - b_0(X_i - X_1)]^2 \frac{w_i}{n_i},$$

with

$$b_0 = \frac{\sum_{i=2}^{m} w_i(X_i - X_1)(y_{i\cdot} - y_{1\cdot})}{\sum_{i=2}^{m} w_i(X_i - X_1)^2}.$$

The unbiased estimate of σ_F^2 is

$$\mu_F^2 = \frac{S_0^2}{\sum_{i=2}^{m} \frac{n_i(N_i - 1)}{N_i - n_i} - 1} \frac{N^2(X_1 - \bar{X})^2}{\sum_{i=2}^{m} w_i(X_i - X_1)^2}. \tag{35}$$

The number of degrees of freedom, f, on which μ_F^2 is based is

$$f = \sum_{i=2}^{m} n_i - 1.$$

9. Empirical tests of the preceding theory. Applications of any mathematical theory involve certain assumptions about the phenomena studied that are not exactly true. In order to have a reasonable hope that the predictions of the theory will be comparable to the actual facts, we must perform empirical tests and see whether such deviations from the assumptions underlying the theory as are *usually* met in practice influence materially or not the working of a given theory. Our object in the present section will be to test whether and to what extent such deviations influence the applicability of the theory. For that purpose it will be useful to enumerate the more important uses of the theory that are likely to be made.

The first point refers to the choice of the standard error σ_F of the best linear estimate F as the measure of accuracy with which F estimates the unknown volume of timber, θ. If all the assumptions were true, the Theorem of Liapounoff would guarantee that, when the size of the sample, Σn_i, is only moderately large, the frequency distribution of the ratio

$$(F - \theta)/\sigma_F, \tag{36}$$

would be very approximately normal about zero with unit S.E. If this were actually true then the value of σ_F would be a justifiable basis for the choice between various alternative estimates of θ. However, the discrepancies between the hypothesis underlying the theory and the actual facts may easily produce a bias in F, or may deprive σ_F of the above important property.

Therefore, the first thing that we have to test is whether in such conditions as are actually met in practice the ratio (36) is in fact distributed in repeated sampling in a way that is comparable with the normal law. The data of the 100 per cent survey of the Blacks Mountain Experimental Forest will serve us for the test.

The second important application of the theory is connected with the use of μ_F. The purpose of calculating μ_F is to characterize the accuracy of the value of F obtained from the sample. The most appropriate way of doing so is to calculate the confidence interval for θ. This has the form [5]

$$F - t_\alpha \mu_F \leq \theta \leq F + t_\alpha \mu_F \tag{37}$$

in which t_α denotes the "Student"-Fisher t taken in accordance with the number of degrees of freedom in μ_F and the chosen value of P. The confidence interval has the property that, if calculated for a great number of samples, the frequency with which the true value of θ will lie between the limits $F \pm t_\alpha \mu_F$ will approach the value $\alpha = 1 - P$ defined as the confidence coefficient.

The above statement concerning the confidence coefficient is strictly true if, apart from the various hypotheses that were enumerated, the distribution of the y's is normal. As a result of a theorem by Kozakiewicz [6] the same statement will be approximately true also for non-normally distributed y_{ij}'s, on condition that the sample sizes are considerable. In the situation where the above theory is to be applied all these assumptions are not satisfied. Still the formula for the confidence interval may well work, but before accepting this we have to have some experimental evidence. The crucial point that it must cover is whether the ratio, say

$$t = (F - \theta)/\mu_F, \tag{38}$$

does or does not follow in repeated sampling a distribution which is sufficiently close to the theoretical one, known as "Student's" distribution. If the empirical distribution of t does approach "Student's" law, then the frequency of correct statements concerning θ in the form (37) will be approximately equal to the chosen α, and conversely.

The n_i's for this experiment were fixed according to the systems shown in Table III, with all X's having a chance of appearing in the samples, and the n_i's quite closely proportional to the N_i's and approximately 25 per cent of the latter. Random sampling numbers [7] were used in making the selections of n_i strips out of any group of strips. A total of 150 block samples were drawn, equally distributed among the 15 blocks.

There are 95 samples for the case where all $n_i < N_i$. For these, formula (19) was used to calculate F and formula (24) for σ_F^2, using the guess that the S_i^2 are constant over all strip lengths. On the hypothesis that the ratio (36) is normally distributed about zero with unit S.E., we divide the range of variation of possible values of (36) into 20 intervals such that, if the hypothesis is true, then the probability of an observed value falling in any particular interval is

TABLE III

Systems of n_i's for sampling experiment

Block	X_i	N_i	n_i
1	10	4	1
	11	4	1
	12	8	2
Total		16	4
2	8	2	1
	9	14	3
Total		16	4
3	4	8	2
	5	4	1
	6	4	1
Total		16	4
4	3	2	1
	7	14	3
Total		16	4
5	5	2	1
	6	1	
	7	3	1
	9	8	2
Total		14	4
6	8	4	1
	9	12	3
Total		16	4
8	1	1	1
	2	1	
	4	1	
	6	1	
	7	4	1
	8	5	1
Total		13	3

Block	X_i	N_i	n_i
9	3	2	1
	4	3	
	5	3	1
	6	2	1
	7	2	
	8	4	1
Total		16	4
12	1	4	1
	2	2	2
	4	4	
	5	3	1
	6	2	
	8	1	1
	10	1	
	12	1	
	14	1	
	9	4	1
	15	4	1
Total		27	7
13	1	2	1
	2	1	
	3	1	
	4	2	1
	6	2	
	5	18	4
Total		26	6

Block	X_i	N_i	n_i
14	1	1	1
	2	1	
	3	1	
	5	1	
	6	7	2
	7	16	4
Total		27	7
15	1	1	2
	2	1	
	3	1	
	4	1	
	6	2	
	7	17	4
	9	3	1
	10	1	
Total		27	7
18	1	2	1
	2	2	
	5	1	1
	6	1	
	7	2	
	8	6	2
Total		14	4
19	2	6	1
	3	16	4
Total		22	5
20	1	16	4
	2	16	4
Total		32	8

equal to .05. For 95 samples then, the expected frequency in each interval is 4.75. The observed frequencies are shown in Table IV.

The agreement between the observed and the hypothetical distribution is tested by means of the fourth order smooth test for goodness of fit [8]. The test is designed so as to be particularly sensitive to such deviations from the hypothetical distribution that could be described as "smooth." It is used here because it is expected that, if the actual distribution of the ratios considered does differ markedly from the normal or from "Student's" one, then still the curve representing this actual distribution would be a "smooth" one, presumably with a single mode, and would cross the hypothetical curves at only a few points. There is empirical evidence to show [9] that in such cases the smooth test of fourth order is more powerful than the usual χ^2 test.

TABLE IV

Frequency distribution[4] *of* $(F - \theta)/\sigma_F$ *and* $(F - \theta)/\mu_F$ *calculated under various assumptions*

Assumption of finite population of strips				Assumption of infinite population of strips		
$(F - \theta)/\sigma_F$		$(F - \theta)/\mu_F$		$(F - \theta)/\mu_F$		
All $n_i < N_i$	One $n_i = N_i$	All $n_i < N_i$	One $n_i = N_i$	All $n_i < N_i$	One or more $n_i = N_i$	Total
n_k	n_k	n_k	n_k	n_k	n_k	n_k
5	11	4	9	3	4	7
3	1	5	1	4	2	6
5	2	6	2	7	2	9
8	1	4	0	2	1	3
3	2	4	2	5	4	9
4	2	5	3	4	3	7
8	1	6	2	6	0	6
3	2	4	2	5	4	9
3	2	5	2	8	2	10
7	0	6	1	5	3	8
1	0	3	0	4	0	4
4	1	3	0	4	4	8
6	1	5	2	7	5	12
5	1	6	3	3	7	10
5	1	6	1	7	1	8
2	1	6	3	9	5	14
5	2	8	4	6	1	7
4	6	5	3	4	0	4
10	1	3	2	2	6	8
4	6	1	2	0	1	1
Total 95	44	95	44	95	55	150
ψ_4^2 1.326	33.812	5.463	13.091	9.055	2.572	8.764
P .85	<.01	.25	.01	.06	.63	.07
$P(\chi^2)$ .57	<.01	.63	.09	.18	.12	.21

[4] By 20 intervals of equal probability.

The criterion used in the smooth test of the fourth order is denoted by ψ_4^2. If the hypothesis tested is true, then ψ_4^2 is distributed, approximately, as χ^2 with 4 degrees of freedom. To calculate ψ_4^2 we proceed as follows:

Let x be a random variable and H denote the hypothesis that the distribution of x is given by a perfectly specified function $f(x)$. The range of variation of x is divided into $2s = 20$ intervals,

$$(-\infty, a_1), (a_1, a_2), \cdots, (a_{19}, +\infty),$$

so that, if H is true then the probability of x falling within any such interval is exactly equal to .05. Such a subdivision can frequently be made easily from appropriate tables for $f(x)$. We associate with these intervals a variate z whose value corresponding to the kth interval will be

$$z_k = \frac{2k-1}{4s} - \frac{1}{2} = \frac{2(k-s)-1}{4s}, \qquad k = 1, 2, \cdots, 2s.$$

It will be seen that if we start at the point a_s and follow up the intervals to the right and to the left, then the corresponding values of z will be

$$z = \pm\frac{1}{4s}, \pm\frac{3}{4s}, \cdots, \pm\frac{2s-1}{4s}.$$

Consideration of the variable z is then substituted for that of the observed values $x_1, x_2, \cdots, x_n$ of x. If any value x_m falls in the kth interval $a_{k-1} < x_m \leq a_k$, then this is interpreted as an observation of x which yielded the value z_k. Let n_k denote the number of observed x's which fall in the interval (a_{k-1}, a_k) and let the Gaussian symbol $[z^i]$ stand for the sum $[z^i] = \sum_{k=1}^{2s} n_k z_k^i$. To apply the fourth order smooth test such sums have to be calculated for $i = 1, 2, 3, 4$. Then they are substituted into the equations below, deduced under the assumption that the number of intervals of subdivision of the range of x is equal to $2s = 20$.

$$u_1^2 = n^{-1}(3.468{,}440[z])^2,$$

$$u_2^2 = n^{-1}(13.500{,}884[z^2] - 1.122{,}261n)^2,$$

$$u_3^2 = n^{-1}(53.857{,}548[z^3] - 8.031{,}507[z])^2,$$

$$u_4^2 = n^{-1}(218.148{,}007[z^4] - 46.239{,}587[z^2] + 1.139{,}500n)^2.$$

Finally $\psi_4^2 = u_1^2 + u_2^2 + u_3^2 + u_4^2$. If the calculated value of ψ_4^2 exceeds the tabled value of χ^2 with four degrees of freedom, corresponding to the chosen level of significance, then the hypothesis tested, H, should be rejected.[5]

[5] The above expressions for the u's differ a little from those published in the original paper on the smooth test because in the latter the test was designed to apply only to ungrouped observations. The present formulae obtained in the Statistical Laboratory of the University of California appear in print for the first time. Obviously if the number of intervals $2s$ is increased, the formulae for grouped data will approach those for ungrouped ones.

The agreement between the observed distribution and the expected distribution is shown to be excellent in Table IV, the probability of a greater difference occurring through errors of random sampling alone being .85. The corresponding P for the χ^2 test, where consecutive pairs of intervals are combined to make 10 intervals in all, is .57.

For the case where one $n_i = N_i = 1$ there are 44 samples. For these samples the value of F was calculated from formula (25), and the value of σ_F^2 from formula (26), again taking the values of the S_i^2 as being constant over strip length within blocks. In this case the deviation from expectation shown in Table IV is greater than can be attributed to chance alone. These results are also obtained by the χ^2 test, which gives $\chi^2 = 25.091$ and $P < .01$ on 9 degrees of freedom.

The conclusions we draw from these results where one of the assumptions made is that the population of strips is finite, are that the block boundaries should be so defined that all $N_i > 1$, or if this is not done, that the systems of n_i's be such that no sampling is done from strips where $N_i = 1$. The fact that some $n_i = 0$ when the corresponding $N_i \geq 1$ has no appreciable effect on the working of the theory. In the previously described test for samples in which $n_i < N_i$ the N used in formulae (19) and (24) always referred to all strips in the block, regardless of the fact that strips of some specified lengths X_i did not appear in particular samples.

Using the same samples, the distribution of $(F - \theta)/\mu_F$ is compared in a manner parallel to that described above, to the distribution of the "Student"-Fisher t, taking into account the number of degrees of freedom.

The formulae used for estimating σ_F^2, namely for calculating μ_F^2, are (34) where all $n_i < N_i$, and (35) where one $n_i = N_i$. The estimates of θ, namely F, remain unchanged from those previously calculated.

The results from this second application of the theory as judged by the smooth test in Table IV lead to the same conclusions as were made from the first application of the theory, namely, that under the assumption that the population of strips is finite no N_i should be exhausted in the sampling.

It is interesting to note that the application of the χ^2 test to the observed distribution of $(F - \theta)/\mu_F$ corresponding to samples with one $n_i = N_i = 1$, did not reject the hypothesis that it follows "Student's" law. In this case the range of t was divided into 10 intervals of equal probability and the value of χ^2 obtained was 15.091. With 9 degrees of freedom this gives P of the order of .09.

The ratio (36) cannot be determined under the assumption that the population of strips is infinite where one $n_i = N_i$ because the values of S_i^2/r_i^2 cannot be obtained for such strips. Under this assumption it is impossible to calculate σ_F by the formulae deduced in the present study and the first use of the theory must be omitted. However, the estimate of σ_F^2 from samples can be calculated and the ratio (38) determined.

The estimates F were calculated using formula (27), taking $n_i = w_i$. This same formula applies whether or not one or more of the N_i are exhausted. Each sample from Block 15 and one sample from Block 12 exhausted two or more strip lengths and their estimates could not be calculated under the assumptions made heretofore, but these can now be obtained under the present assumptions. The estimates μ_F^2 were obtained from (34) for all samples, taking the S_i^2 as constant over all strip lengths and $n_i = w_i$. The fact that one or more N_i are exhausted does not change the procedure for such samples in any way.

For the case where all $n_i < N_i$ in Table IV, the value of $P = .06$ obtained by the ψ_4^2 test indicates that the agreement of the observed distribution with expectation, although not close, is acceptable. When the data are regrouped into 10 classes and the χ^2 test is applied, we get $P = .18$ on 9 degrees of freedom.

The ψ_4^2 test applied to the distribution of $(F - \theta)/\mu_F$ for samples where one or more $n_i = N_i$ indicates that the correspondence with expectation is good. This result is in marked contrast to the corresponding results in previous tables and bears out the belief previously expressed in Section 7, based on intuitive considerations, that by dropping the assumption of finiteness of number of strips of a given length, the error of the assumption of strict linearity of regression would be compensated for to some extent. On the basis of these findings we can add the conclusion that if, in sampling, the number of strips of a given length are exhausted, the assumption of finiteness should be dropped and the sample estimates calculated from formulae deduced under the assumption that all N_i are infinite.

There remains some question as to statistical treatment of samples in which all $n_i < N_i$, that is, whether to use formulae deduced for finite or infinite populations. The final choice can best be based on the relative size of the confidence interval (37). Where all $n_i = 1$ the estimates are the same under both assumptions. For estimates of all blocks taken together the finite population estimates tended to be within 5.5 percent of θ in 95 out of 100 trials, while the corresponding percentage for infinite population estimates was 6.0. We therefore conclude that it is better to use the assumption of finiteness of N_i where all $n_i < N_i$.

The method of sampling considered here is what could be called restricted random. The restriction consists in that we group together the sampling units of the same size, select nonrandomly several such groups, and only then proceed to draw at random n_i units of a group of N_i. Frequently the strips of the same size will be situated within the block close to one another. In those cases the restricted sampling considered will assure that the sample will contain elements more or less uniformly distributed over the area of the block.

10. Summary.

Several methods of sampling timber stands and statistical treatment of the samples were considered. Data from a complete inventory of the Blacks Mountain Experimental Forest served for testing the methods in practice.

It was found that the usual method of estimating from strip samples taken within nonrectangular blocks of timber gave biased estimates, unless the linear

regression of volume on strip length passed through the origin of coordinates. It was shown that this condition was not a safe one to assume. Consequently methods of estimation were sought which were freed from this restriction.

The appropriate formulae for the best linear unbiased estimates were deduced under various combinations of the following assumptions:

(1) That the regression of timber volume on strip length is strictly linear, but may or may not pass through the origin of coordinates.

(2) That the values of the $(S.D.)^2$ of timber volumes on strips of equal lengths are (a) constant for different strip lengths, (b) proportional to strip length, and (c) proportional to the square root of strip length.

(3) That the number of strips of a given length in each block is (a) finite, and (b) infinite. Assumption (b) was based on intuitive considerations which indicated that this assumption, though known to be false, might compensate for another false assumption, namely, that of strict linearity of regression.

It was empirically found that assumption (b) of (2) gave better results than either (a) or (c). However, the advantage was small and, in the author's opinion, did not justify the extra labor in calculations which are simpler when assumption (a) is made. Therefore all other calculations were made on that assumption.

An extensive sampling experiment was made to test whether the smallness of the samples combined with the conflicts between assumptions of the theory and the actual facts, influenced the validity of the normal theory.

Whenever the sample did not exhaust strips of a given length, it was found that the formulae based on the assumptions that the populations of such strips are finite and that they are infinite both work satisfactorily, generating distributions similar to those determined by the normal theory. However, the confidence intervals based on the true assumption that the populations of strips of equal length are finite, proved to be narrower. Consequently, whenever the sample does not exhaust all strips of any given length in the block, the true hypothesis concerning the number of such strips should be used. Formulae (19) and (34) are therefore the appropriate ones, using weights based on finite populations.

In cases where the sample did exhaust the strips of a given length, the treatment of the number of such strips as finite, combined with the inaccuracy of the assumption that the regression of timber volume on length of strip is linear, resulted in marked disagreement between the actual distributions of statistics and those based on normal theory. This disagreement was not found to exist in statistics calculated with formulae (27) and (34) used on the assumption of an infinity of strips of a given length. This suggests the conclusion that the exhaustion of strips of a given length by the sample should be avoided and, when this is impossible, then the formulae based on the assumption of an infinity of strips of a given length should be used.

The formulae deduced can be applied equally well to line plots as to strips. With the formulae deduced the most efficient sampling will be obtained when the average sample strip length is close to the average for the population, but where the variation among sample strip lengths is the maximum.

The author is deeply indebted to Prof. J. Neyman for guidance and advice, and to Miss Evelyn Fix and Miss Phyllis Burleson for help in computations.

REFERENCES

[1] A. A. HASEL, *Jour. Agric. Res.*, Vol. 57 (1938), p. 713.

[2] J. NEYMAN, *Lectures and Conferences on Mathematical Statistics*, Washington, D. C. (1937).

[3] F. N. DAVID and J. NEYMAN, *Stat. Res. Mem.*, Vol. 2 (1938), p. 105.

[4] F. N. DAVID, *Stat. Res. Mem.*, Vol. 2 (1938), p. 69.

[5] J. NEYMAN, *Roy. Stat. Soc. Jour.*, Vol. 97 (1934), p. 558.

[6] W. KOZAKIEWICZ, *Ann. Soc. Polon. Math.*, Vol. 13 (1934), p. 24.

[7] L. H. C. TIPPETT, *Random Sampling Numbers*, Tracts for computers, No. 15, Cambridge University Press (1927).

[8] J. NEYMAN, *Skand. Aktuar. Tidskr.*, (1937), p. 149.

[9] J. NEYMAN, *Annals of Math. Stat.*, Vol. 11 (1940), p. 478.