Ewing

For Limited Distribution

SAMPLE SURVEY TECHNIQUES

By

W. G. Cochran

Prepared by

Institute of Statistics
North Carolina State College

and

Bureau of Agricultural Economics
United States Department of Agriculture
Cooperating

Raleigh, North Carolina
June, 1948

TABLE OF CONTENTS

ERRATA

Page 20. Line 5 from foot. In the expression for p, insert "k" before the square root.

Page 27. Line 9 from foot. In the expression for C, for "$\sqrt{n}$" read "$\sqrt{n_j}$".

Page 28. Line 14. For "(1939)" read "(1934)".

Page 40. Line 11. In the expression for $V(\bar{y}_n)$, for $\dfrac{\Sigma N_j^2 \sigma_j^2}{n_j}$ read $\Sigma \dfrac{N_j^2 \sigma_j^2}{n_j}$

Page 47. Line 4. For "written" read "within".

Page 61. Line 7 from foot. Delete "$N^2 =$" in the expression for the variance of the estimated mean.

Page 114. Line 4 from foot. For "twon" read "town".

Page 124. Formula (157). In the first expression for $V(Y_{Rs})$, for "$\dfrac{N_j}{n_j}$" read "$\dfrac{N_j^2}{n_j}$".

Page 145. Line 3 from foot. In formula (187), insert ")" after $s^2_{y.x}$.

Page 149. Line 11 from foot. For "dute" read "due".

PREFACE

These notes form the basis of a one-quarter course of lectures on sampling techniques delivered at North Carolina State College to graduate students who are specializing in statistics. The main object of the lectures is to present the principal techniques in current use, with the theory from which they are derived. For reading the notes, facility in elementary algebra and a good knowledge of elementary statistical theory are required: calculus is used only to a slight extent. Occasionally, proofs are given in a condensed form, since it is desired to concentrate attention on results rather than on details of proof.

In the preparation of the notes, generous assistance has been given by E. H. Jebe and A. L. Finkner, Resident Collaborators, Agricultural Estimates, Bureau of Agricultural Economics. Mr. Jebe prepared the first draft of most of Chapters 1 to 5, while Mr. Finkner prepared that of Chapter 9: both have taken major responsibility in supervising the later stages of mimeographing. My best thanks are due to Mrs. Jessie M. Gray for the typing, and to Miss Mary Ruth Reavis who did the mimeographing.

INTRODUCTION

1.1 Within recent years sampling has been increasingly used for obtaining information. The principal advantages claimed for the sampling method are:

(1) Reduced cost. If data are secured from only a small fraction of the population, expenditures will be smaller than if a complete count were attempted.

(2) Greater speed. For the same reason, the data can be collected and summarized more quickly with a sample than with a complete count. This may be a vital consideration when the information is urgently needed.

(3) Greater accuracy. A sample may actually give more accurate results than the kind of complete count that it is feasible to take. Since a much smaller field force is needed for a sample, it may be possible to engage personnel of higher quality and to give them more thorough training.

1.2 General procedure in sampling. In order to indicate the scope of this course, it is convenient to indicate briefly the steps that are usually involved in the planning and execution of a sample survey. These steps will be grouped rather arbitrarily under eight headings.

(1) Definition of the population to be sampled. This may present no problem, as for instance when sampling a given batch of 1,000 electric light bulbs in order to estimate the average length of life of a bulb. On the other hand, in sampling a population of farms, rules must be set up to define what constitutes a farm, and borderline cases will arise. It is important that these rules be usable in practice: that is, the enumerator should be able to decide without much hesitation whether a doubtful case belongs to the population or not. Further, the population sampled should coincide with the population about which information is

wanted. Sometimes this will not be feasible. For example, in taking a sample of voter's opinions in order to predict the result of an election, the population that it is desired to sample is the population of voter's opinions when they go to the polls. Since the sample must be taken several days before election day, all that can be sampled is the population of opinions of intending voters some days before election. Both their opinions and their intention to vote may change.

(2) Determination of the data to be collected. The data needed depend on the purpose of the inquiry. It is well to have this purpose clearly defined, and to verify that all the data are relevant to the purpose and that no essential data are omitted. There is frequently a tendency to attempt to collect too much data, some of which is never subsequently examined. Sometimes data that would be desirable are impossible to collect, at least in an accurate form. For instance, people may be unable to recall accurately their opinions or the details of their business transactions at some previous time.

The construction of the schedule or questionnaire on which the data are to be recorded often presents difficult problems, which have been the subject of specialized study in recent years. A few of the devices that have been found useful are given below.

(i) The questionnaire should be reviewed by disinterested persons.

(ii) The questionnaire should be tested in the field before the survey itself begins. This pre-test should reveal questions that are ambiguous or not clearly worded, questions that the respondent finds difficult to answer, and the types of query that the respondent may make about the meaning of certain questions.

(iii) In questions of opinion, every attempt should be made to ensure that the wording is'neutral':i.e., that it does not influence the respondent to give one kind of answer rather than another. If it is not clear which of two wordings is preferable, each may be tried in half the schedules.

(iv) Sometimes the questions asked are of little or no interest to the respondent. In such cases it may help to insert additional questions that will evoke the respondent's interest, even though they are rather irrelevant to the main prupose of the sample.

(3) Choice of sampling-unit. The sampling units are the elements into which the population is divided. Sometimes the appropriate unit is obvious, as in the case of the sample of light bulbs, where the unit would be a single bulb. In sampling a town population, however, the unit might be an individual, a household or a city block. In sampling a field of corn, the unit might be a single plant, a single hill, a group of four hills, or perhaps some larger group of hills. The best size of unit is that which will give the desired degree of accuracy in the estimates at the smallest cost. If a fixed percentage of the population is to be sampled, it usually is found that sampling costs are lower when the unit is large. On the other hand, the accuracy obtained through the use of larger units tends to be lower.

(4) Method of selecting the sample. There is now quite a variety of procedures by which the sample may be selected. In the choice of a method, the general principle is the same as that used in the choice of size of unit: the method selected should provide the desired degree of accuracy at minimum cost. The question of the size of sample also arises here. As will be seen later, the size needed can be estimated, at least roughly, when the method of sampling has been selected and its sampling properties have been studied.

(5) Method of collecting the data: After the members of the sample have been chosen there arises the question of how to obtain the information from them. This may be done by mail, by telephone, telegraph or by direct enumeration, i.e., an interviewer seeking out the sample members and eliciting the information. A combination of indirect, say mail, and direct enumeration may be employed. Efficient combination then must be considered.

(6) Organization of the field work. Here many problems of business administration are involved which lie outside the field of statistics. It cannot be too strongly emphasized, however, that the success of any survey depends on competent field work. The personnel must be qualified to cope with the task of enumeration, and must receive training in the purpose of the survey and in the methods to be employed. Supervision of the field work and checks on its quality are essential.

(7) Summary and analysis of the data. The first step is to 'edit' the schedules, in the hope of amending recording errors, or at least of deleting data that are obviously erroneous. Difficult questions of judgement may be met. Thereafter the tabulations leading to the estimates are performed. Different methods of estimation may be available on the same data, and a superior method sometimes results in a substantial increase in accuracy.

(8) Information gained for future surveys. The best method of sampling depends on the type of variation that exists among the units in the population. In general the only sources of information about this variation are the results either of samples or of complete censuses. Consequently any sample is potentially a valuable guide to the conduct of future sampling investigations. Given the results of a sample, it is often possible to investigate the accuracy that would have been obtained from alternative methods of sampling that were

considered but not used. The cost of such alternatives may be estimated from cost data. Thus each sample of a given type of population should lead to more efficient sampling in the future.

1.3 Scope of the course. The theory of sample surveys has been mainly concerned with items (3) choice of sampling unit, (4) method of selecting the sample, (7) summary and analysis of the data, and (8) information gained for future surveys. This course will likewise deal mainly with these topics. It should be realized, however, that the other items——definition of the population, determination of the necessary data and method of collecting it, and organization of the field work——are equally important: poor field work, for instance, may ruin an otherwise admirable survey.

The various topics will be discussed in the order that seems easiest for expository purposes, rather than in the order in which they are encountered in practice when a sample survey is undertaken.

1.4 General principle. In deciding whether to choose one sampling procedure rather than another, the following principle, which has already been mentioned, is being increasingly used. The principle is to select the method that gives the desired accuracy at the lowest cost; or alternatively the maximum accuracy at a given cost. In the practical use of this principle, we must be able to predict both the accuracy and the cost of each procedure before we can decide which to select. With samples of the sizes that are common in practice, there is usually good reason to believe that the sample estimates will be approximately normally distributed. Consequently, the sampling variance of the estimate is used to provide the measure of its accuracy. A considerable part of the work in this course will consist of the calculation of formulas for the sampling variances of estimates obtained by various procedures. These formulas usually contain one or more unknown parameters that describe properties of the population. In order to make a prediction of the sampling variance, values must be inserted for the unknown parameters. It is at this point that knowledge obtained from previous sampling of the same or similar populations is very helpful.

The prediction of probable costs may also require data obtained from previous surveys. Some rather simple types of cost function which have been used will be discussed later, though knowledge of cost functions is still rather scanty.

1.5 Errors of sample surveys. In connection with this general principle, various writers (Mahalanobis, Hotelling, Deming and Stephan) have discussed sources of error that will affect the accuracy of a sample. Among these sources, three may be indicated here:

(i) Sampling variations, that is, errors arising from the fact that only a portion of the population has been examined.

(ii) Recording mistakes. These comprize errors made in recording the data on the schedule. They might arise from either the enumerator or the respondent, and might be the result of mistakes, biases or dishonesty.

(iii) Physical fluctuations. There may be an inherent indefiniteness about the quantity that is being measured, e.g., the total production of a crop will vary according to the moisture content, which will depend on the weather. Similarly, many quantities change with time, such as voter's intentions or the population of a city, and when a survey extends over several weeks it is not clear exactly what has been measured.

This classification leads to some interesting conclusions. First, while a complete count avoids error (i), it is just as subject to errors

(ii) and (iii) as a sample. In fact, it may be more subject to (ii) than a sample if a lower quality of enumerator must be used. Secondly, the size of the physical fluctuations imposes a limit to the accuracy which it is worth-while trying to achieve by reducing sampling fluctuations and recording mistakes. Thirdly, if recording errors are large they may contribute much more than the sampling variations to the total error. If this is the situation, a marked increase in accuracy can be secured only by reducing the recording errors, and not by taking a larger sample in order to diminish still further the sampling variations.

REFERENCES

W. Edwards Deming

(1) "On Errors in Surveys" American Sociological Review. Vol. IX, No. 4, pp. 359-369, Aug. 1944.

(2) "On Training in Sampling" Journal of American Statistical Association, Vol. 40, pp. 307-316, 1945.

(3) (with F. F. Stephan) "On the Interpretation of Census as Samples" Journal of the American Statistical Association, Vol. 36, pp. 45-49, 1941.

(4) "Some Criteria for Judging the Quality of Surveys" Reprinted from The Journal of Marketing, Vol. XII, pp 145-157, October, 1947.

Mahalanobis, P. C.

(5) "On Large-Scale Sample Surveys" Phil. Trans. Royal Society, London, B, 231, 1944.

BASIC THEORY

2.1 Sample surveys deal with samples drawn from populations that contain a finite number N of units. The values of the item that is being measured are denoted by $y_1, y_2, \ldots y_N$. In general, no particular form of frequency distribution is assumed for these values. In practical applications it is, however, frequently taken for granted that the means of samples of size n are approximately normally distributed. This assumption implies that the original values are not too far removed from a normal distribution.

2.2 For the population these relations are defined:

$$\text{The Mean:} \quad \bar{y}_p = \frac{y_1 + y_2 + \ldots . y_N}{N} \qquad (1)$$

$$\text{The Variance:} \quad \sigma^2 = \frac{\Sigma(y_i - \bar{y}_p)^2}{N-1} \qquad (2)$$

Note: Some writers use N as a divisor when defining the variance as is usually done in the mathematical theory of finite populations. The definition given above makes it easier to use the concepts of the analysis of variance.

2.3 Simple Random Sampling: First it is to be noted that a sample of n distinct elements can be chosen in ${}_N C_n$ ways from the population. In factorial notation this is expressed as N! / (N-n)! n! ways.

Simple random sampling is defined as: A method of selecting n items out of N so that it gives every one of the ${}_N C_n$ groups an equal chance of being chosen. As an illustration consider an example: N = 5, a population of 5 elements and n = 3, samples of 3 items to be drawn from the population. There are 10 possible samples of 3 items. They are:

ABC	ABD	ABE	ACD	ADE
ACE	BCD	BCE	BDE	CDE

Note: If the elements are drawn, one by one, without replacement, and if at any stage, or any draw, all undrawn elements have an equal chance of selection, this process gives a simple random sample. Applied to our example the process gives an equal chance for obtaining any one of the 10 possible samples listed.

2.4 Let $\bar{y}_n$ denote the mean of a simple random sample of size n. Consider E ( ) as the average over all the ${}_NC_n$ possible samples. Observe that the operator E is used here as in the discrete case in formal probability theory, e.g. to the expectation of the throw of a single die.

<u>Theorem 1a</u>: $E(\bar{y}_n) = \bar{y}_p$ (3)

$$\text{For } E(\bar{y}_n) = \frac{1}{n} E\,(y_1 + y_2 \ldots + y_n)$$

Since every unit appears in an equal number of samples, $E\,(y_1 + \ldots + y_n)$ must be some multiple of $(y_1 + \ldots y_N)$. Further, the multiplier must be n/N, since the first expression contains n terms and the second N terms.

Hence:

$$E\,(\bar{y}_n) = \frac{1}{n} \frac{n}{N} (y_1 \ldots + y_N) = \bar{y}_p.$$

2.5 <u>Theorem 1b</u>:

$$E\,(\bar{y}_n^2) = \frac{1}{Nn} \frac{N-n}{N-1} \sum_1^N y_i^2 + \frac{N(n-1)}{n(N-1)} \bar{y}_p^2 \qquad (4)$$

This theorem is proved in order that it may be applied in the proof of later theorems.

Proof:

$$\bar{y}_n^2 = \frac{(y_1 + y_2 + \ldots y_n)^2}{n^2}$$

$$= \frac{y_1^2 + y_2^2 + \ldots y_n^2}{n^2} + \frac{2}{n^2} (y_1 y_2 + \ldots + y_{n-1} y_n)$$

By symmetry,

$$E\,(y_1^2 \ldots , + y_n^2) = \frac{n}{N} (y_1^2 \ldots , + y_N^2)$$

and $E\,(y_1\, y_2 \ldots + y_{n-1}\, y_n) = \frac{n(n-1)}{N(N-1)}\,(y_1\, y_2 + \ldots y_{N-1}\, y_N)$

Hence,

$$E\,(\bar{y}_n^2) = \frac{1}{nN} \sum_1^N y_i^2 + \frac{2(n-1)}{nN(N-1)}\,(y_1\, y_2 + \ldots y_{N-1}\, y_N)\ .$$

But

$$2(y_1\, y_2 + \ldots y_{N-1}\, y_N) = y_1\,(y_2 \ldots + y_N) + y_2\,(y_1 + y_3 \ldots + y_N) + y_N(y_1 + y_2 \ldots + y_{N-1})$$

$$= y_1\,(N\bar{y}_p - y_1) + y_2\,(N\bar{y}_p - y_2) + \ldots + y_N\,(N\bar{y}_p - \bar{y}_N)$$

$$= N\bar{y}_p(y_1 + y_2 + \ldots y_N) - y_1^2 - y_2^2 \ldots - y_N^2$$

$$= N^2\, y_p^2 - \sum_1^N y_i^2\ .$$

Introducing this last reduction in $E\,(\bar{y}_n^2)$, we obtain

$$E\,(\bar{y}_n^2) = \frac{1}{nN}\left\{1 - \frac{n-1}{N-1}\right\} \Sigma\, y_i^2 + \frac{N(n-1)}{n(N-1)}\ \bar{y}_p^2$$

$$= \frac{1}{nN}\left(\frac{N-n}{N-1}\right) \Sigma\, y_i^2 + \frac{N(n-1)}{n(N-1)}\ \bar{y}_p^2$$

2.6 Theorem 2. Variance of the mean of a random sample.

$$E\,(\bar{y}_n - \bar{y}_p)^2 = \frac{N-n}{N}\ \frac{\sigma^2}{n} \tag{5}$$

Proof: Expand the above, obtaining

$$E\,(\bar{y}_n^2) - 2E\ \bar{y}_n\ \bar{y}_p + \bar{y}_p^2$$

$= E\,(\bar{y}_n^2) - \bar{y}_p^2$, by Theorem 1a.

Substitution from Theorem 1b gives

$$\frac{1}{nN}\left(\frac{N-n}{N-1}\right) \sum_1^N y_i^2 + \left\{\frac{N(n-1)}{n(N-1)} - 1\right\}\ \bar{y}_p^2$$

$$= \frac{1}{nN}\ \frac{N-n}{N-1} \sum_1^N y_i^2 - \frac{N-n}{n(N-1)}\ \bar{y}_p^2$$

$$= \frac{1}{n} \frac{N-n}{N} \left\{ \frac{1}{N-1} (\Sigma\, y_i^2 - N\, \bar{y}_p^2) \right\}$$

$$= \frac{N-n}{N} \frac{\sigma^2}{n}$$

The quantity $\frac{N-n}{N}$ is usually called the finite population correction.

Note: If $\frac{n}{N} < .05$ (i.e. less than 5% sampled), $\sigma^2_{\bar{y}_n}$ depends primarily on n, and not on $\frac{n}{N}$. For instance, if σ^2 is the same in the two cases, a sample of 500 out of a population of size 200,000 will have a mean almost as accurate as that of a sample of 500 out of a population of size 10,000.

2.7 Theorem 3. Estimation of σ^2 from the sample data.

$$s^2 = \frac{\Sigma\, (y_i - \bar{y}_n)^2}{n - 1} \quad \text{is an unbiased estimate of } \sigma^2. \qquad (6)$$

Proof:

$$E\,(s^2) = \frac{1}{n-1}\, E\, (\sum_1^n y_i^2 - n\, \bar{y}_n^2)$$

$$= \frac{1}{n-1} \left[\frac{n}{N} \sum_1^N y_i^2 - \frac{N-n}{N(N-1)} \sum_1^N y_i^2 - \frac{N(n-1)}{N-1} \bar{y}_p^2 \right], \text{ from Theorem 1b.}$$

Combining the first two terms in brackets, this reduces to

$$= \frac{1}{n-1} \left[\frac{N(n-1)}{N(N-1)} \sum_1^N y_i^2 - \frac{N(n-1)}{N-1} \bar{y}_p^2 \right]$$

$$= \sigma^2$$

Hence, the estimated standard error of $\bar{y}_n$ is

$$s_{\bar{y}_n} = \frac{\sqrt{(N-n)}}{\sqrt{N}} \frac{s}{\sqrt{n}} \qquad (7)$$

CONFIDENCE LIMITS AND ESTIMATION OF SAMPLE SIZE

(SIMPLE RANDOM SAMPLING)

3.1 Confidence limits: If n is reasonably large and $\frac{n}{N}$ is not too large, $\bar{y}_n$ will be assumed approximately normally distributed about $\bar{y}_p$. Thus, approximate confidence limits may be constructed in the ordinary way by writing

$$\bar{y}_p = \bar{y}_n \pm t(\alpha,n-1) \sqrt{\frac{N-n}{N}} \; \frac{s}{\sqrt{n}} \qquad (8)$$

where $t(\alpha,n-1)$ is the value of t corresponding to a significance level α, for $(n-1)$ degrees of freedom.

3.2 Size of sample needed. Before the sample is taken, it is useful to be able to obtain some idea of the size of sample that will be needed in order to attain a desired standard of accuracy. The accuracy required is usually defined by specifying a probability level α (e.g., .05, .10, .20) and a margin of error d allowable in the sample mean. That is, we want

$$P \left\{ \left| \bar{y}_n - \bar{y}_p \right| \geqslant d \right\} = \alpha$$

If this equation holds, the probability that the sample mean lies within a distance d of the population mean is $(1-\alpha)$, and can be made as close to certainty as we like by making α sufficiently small. The equation simply states that the confidence interval is of width 2d. Two cases must be considered.

3.3 Case 1. The value of n cannot be predicted without some knowledge of the standard error σ in the population. In Case 1, σ is estimated from previous sampling of a similar population, or simply by intelligent guesswork. Since the estimated σ is likely to be itself in error, we cannot expect more than a rough estimate of n. If σ were to be correct, the value of d would be given by

$$d = \left| \bar{y}_n - \bar{y}_p \right| = t(\alpha,\infty) \sqrt{\frac{N-n}{N}} \cdot \frac{\sigma}{\sqrt{n}} \tag{9}$$

where $t(\alpha,\infty)$ is the normal deviate corresponding to the significance level α. Solving for n, we have

$$Nn = (N-n)\ \sigma^2\ \ t^2(\alpha,\infty)/d^2$$

$$\text{or}\quad n = \frac{N\sigma^2\ t^2(\alpha,\infty)/d^2}{N + \dfrac{\sigma^2\ t^2(\alpha,\infty)}{d^2}} = \frac{\sigma^2\ t^2(\alpha,\infty)/d^2}{1 + \dfrac{1}{N}\ \dfrac{\sigma^2 t^2(\alpha,\infty)}{d^2}} \tag{10}$$

If N is very large, the second term in the denominator can be neglected, and we obtain

$$n_0 = \sigma^2\ t^2\ /d^2. \tag{11}$$

The procedure is as follows: First calculate n_0. If n_0/N is an appreciable fraction (say greater than .05), take

$$n = \frac{n_0}{1 + \dfrac{n_0}{N}} \tag{12}$$

The value of n will then be the correct solution of equation (10).

When the sample is actually taken, the confidence interval will be calculated by means of the t distribution rather than of the normal distribution: that is, by equation (8) rather than by (9). A further refinement that is sometimes introduced is to adjust n so as to take account of the fact that the t value for (n-1) degrees of freedom, which appears in (8), is larger than the corresponding normal deviate which appears in (9). For instance, if n turned out to be 16, it may be verified that n would have to be increased to 18 for this reason. The refinement, however, is hardly worth-while unless the initial estimate of σ is good and n is less than 20.

3.4 Example: An example illustrating application of the formula for determining sample size: The data were obtained from a planting of silver maple seedlings in a bed 430' long. The sampling unit was a one foot strip across the bed. By complete enumeration of the bed, the following population values were obtained for the number of seedlings per unit.

$$\bar{y}_p = 19 \quad \text{and} \quad \sigma^2 = 85.6$$

Assuming simple random sampling, how many sampling units must be enumerated to estimate $\bar{y}_p$ within 10% with a confidence probability of .95? Applying equation (9), we obtain

$$n_0 = \frac{\sigma^2\, t^2}{d^2} = \frac{(85.6)\ (4)}{(1.9)^2} = 95$$

since d = (19) x (0.1).

Then,

$$n = \frac{95}{1 + 95/430} = 78.$$

The result shows that about 20% of a whole bed has to be counted to obtain the accuracy desired.

3.5 Case II. The methods given for Case I do not guarantee that the confidence interval will be of the required width, for the initial estimate of σ may turn out to be wrong, and even if σ is correct, the s that is found when the sample is taken will differ from σ. All that the procedure attempts to do is to ensure that the interval will be about the desired length. If an exact interval is wanted, the information about σ must be obtained from the population that is being sampled. A method that guarantees a more exact confidence interval is due to Charles Stein ("A Two Sample Test. . . ." Annals of Math. Stat., Vol. 16, pp. 243-258, 1945). Stein's approach considers taking the sample in two parts.

The first part of the sample, of size n_1, say, supplies an estimate s_1 of σ, calculated in the usual way, and also a preliminary estimate of the mean. When the first part has been taken, Stein shows how to calculate the number of additional observations needed in order to have a specified confidence interval. Note that both parts must be samples from the population about which information is desired. Thus, if the population changes with time, the time interval between the first and second parts must be sufficiently small that no appreciable change will have occurred.

Since Stein's method was developed for infinite populations, the case where n/N is negligible will be considered first. When the first sample has been obtained, a confidence interval for $\bar{y}_p$ can be calculated. The half-width of this interval is (by equation (8), with n/N negligible)

$$t(\alpha, n_1 - 1)s/\sqrt{n_1} \ .$$

If this quantity is less than or equal to d, the desired half-width, the sample is already sufficiently large. If the quantity exceeds d, take additional observations so that the total size of sample n is at least as great as

$$s^2 t^2(\alpha, n_1 - 1)/d^2 \qquad (13)$$

Then, if $\bar{y}_n$ is the mean of the whole sample

$$P\left\{ \left| \bar{y}_n - \bar{y}_p \right| \geq d \right\} \leq \alpha \ . \qquad (14)$$

Sketch of proof. The proof assumes that the observations, $y_1, y_2, \ldots y_n$, are normally distributed about $\bar{y}_p$. Throughout the proof, d,α and n_1 are assumed to be fixed quantities. The total sample size n is not fixed, but is a random variate, since its value depends on the value of s that turns up in the first sample. Nevertheless, for fixed s, n is fixed, and the quantity

$$\sqrt{n}\,(\bar{y}_n - \bar{y}_p)$$

is normally distributed with mean zero and variance σ^2. Hence, this quantity follows the normal distribution whether s is fixed or not. Moreover, the distribution is independent of that of s. Consequently,

$$\sqrt{n}\,(\bar{y}_n - \bar{y}_p)/s$$

follows the t distribution with (n_1-1) d.f.. By definition of $t(\alpha,n_1-1)$, it follows that

$$P\left\{\left|\sqrt{n}\,(\bar{y}_n - \bar{y}_p)/\ s\right| \geqslant t(\alpha,n_1-1)\right\} = \alpha \qquad (15)$$

This is the key result in the proof. Further, by the way in which the value of n was calculated, we always have

$$\sqrt{n} \geqslant st(\alpha,n_1-1)/d, \text{ or } \sqrt{n}/s \geqslant t(\alpha,n_1-1)/d, \qquad (16)$$

so that

$$\left|\sqrt{n}\,(\bar{y}_n - \bar{y}_p)\ \ /s\right| \ \geqslant \left|\ t(\bar{y}_n - \bar{y}_p)/d\right|$$

Hence, from (15)

$$P\left\{\left|t(\bar{y}_n - \bar{y}_p)/d\right| > t\right\} \leqslant \alpha$$

$$\text{i.e. } P\left\{\left|\ \bar{y}_n - \bar{y}_p\right| \ \geqslant d\right\} \leqslant \alpha\ .$$

The average value of n that is required in a given situation depends on the choice of n_1. Exact information about the optimum value of n_1 is not yet available, the optimum being that value which leads to the smallest average n. It appears, however, that the optimum n_1 is such that a second part will usually be necessary. In other words, if it is convenient to take the sample in two parts, n_1 should be chosen as somewhat less than the size that seems to be needed. On the other hand, if it is troublesome to take the sample in two parts, n_1 may be chosen at about the expected size, or perhaps a little larger if a few unnecessary observations do not matter.

Example. Suppose that $d = 10$, α .05. From previous information, σ is guessed as about 50 (though this guess may be seriously in error). With this value of σ, it appears from (13) that a sample of about

$$(2{,}500)\ (1.96)^2/100, \text{ or } 96,$$

will be needed. Assuming no difficulty in taking the sample in two parts, n_1 might be chosen as 50.

In this case $t(.05,49) = 2.01$. s^2 is found to be 1,938. We find that

$$ts/\sqrt{n_1} = (2.01)\ (44.02)/\ 7.0711 = 12.51,$$

so that a sample of 50 gives a confidence interval of half-width 12.51, which is larger than desired. Finally, n is chosen so that

$$n \geqslant t^2 s^2/d^2 = (4.040)\ (1{,}938)/100 = 78.3$$

That is, 29 additional observations are taken to make the total n = 79. If the finite population correction must be applied, the only change is to choose n so that it is at least as large as

$$\frac{\frac{t^2\ s^2}{d^2}}{1 + \frac{1}{N} \cdot \frac{t^2 s^2}{d^2}}$$

SAMPLING FROM "BINOMIAL TYPE" POPULATIONS

4.1 Suppose that the data to be taken divide the sample into two classes or groups, say A and A' (those not in A). The result of the sampling may be expressed as a percentage. Examples are a pre-election poll to determine the proportion of voters favoring a certain candidate, or a survey to measure the proportion of housewives listening to a radio program. This type of sampling resembles ordinary binomial sampling except that the individuals measured come from a finite population.

The results already obtained can be applied if the data are coded in the following manner: For the members of the sample $y_1, y_2 \ldots y_n$, or population, $y_1, y_2 \ldots y_N$, mark 1 for each y in A and 0 for each y not in A. Then the sample population proportion,

$$\bar{y}_n = \frac{\text{Number in Sample in A}}{n} = p_n \text{, and the}$$

population proportion,

$$\bar{y}_p = \frac{\text{Number in Population in A}}{N} = p$$

4.2 Theorem 4. The definition of the "Binomial Type" population variance:

$$\sigma^2 = \frac{N}{N-1} \; p \; q \quad \text{where } q = 1-p. \qquad (17)$$

Proof: By definition,

$$\sigma^2 = \frac{1}{N-1} (\Sigma \; y_i^2 - N \; \bar{y}_p^2) \qquad (i = 1,2 \ldots N).$$

$$= \frac{1}{N-1} (N \; p - N \; p^2) = \frac{N}{N-1} \; pq \quad \text{from the coding}$$

and definition of p given in (4.1).

4.3 Theorem 5. Variance of the sample proportion from a simple random sample is $\frac{N-n}{N-1} \; \frac{p \; q}{n}$ (18)

Proof:

This follows at once from the previous results, (Sec. 2.6). Theorem 2 gave

$$V(\bar{y}_n) = \frac{N-n}{N} \; \frac{\sigma^2}{n}$$

By substitution, using Theorem 4,

$$V(\bar{y}_n) = \frac{N-n}{N} \; \frac{1}{n} \; \frac{N}{N-1} \quad p\, q$$

$$= \frac{N-n}{N-1} \; \frac{p\, q}{n} \; .$$

4.4 Estimation of the variance of the sample proportion: By substituting the sample values, we obtain

$$V(p_n) = \frac{N-n}{(N-1)n} \; p_n \, q_n \quad \text{where } q_n = 1-p_n \tag{19}$$

It is to be noted, however, that $E(p_n\, q_n) = \dfrac{N(n-1)}{n(N-1)} \; pq$

Therefore, an unbiased estimate of

$$V(p_n) = \frac{N-n}{(N-1)n} \; \frac{n(N-1)}{N(n-1)} \; p_n \, q_n = \frac{N-n}{N(n-1)} \; p_n \, q_n \tag{20}$$

For any reasonable size n the correction for bias is negligible and either (19) or (20) may be used.

4.5 Confidence limits for the sample proportion: If normal theory can be applied the confidence limits are

$$p = p_n \pm t(\alpha) \; \frac{\sqrt{N-n}}{\sqrt{N-1}} \; \frac{\sqrt{p\;\; q}}{\sqrt{n}} \; . \tag{21}$$

This relation is still not in usable notation since p and q are unknown. Substitution of estimated values from the sample gives

$$p_n \pm t(\alpha) \frac{\sqrt{N-n}}{\sqrt{N-1}} \; \frac{\sqrt{p_n \;\; q_n}}{\sqrt{n}}$$ where $t(\alpha)$ is taken with ∞ degrees of freedom. (22)

When p is near .5 the normal approximation gives satisfactory results. With increasing sample size the normal theory may be applied even though the sample proportion deviates considerably from .5. The relation is indicated in the following abbreviated table:

Observed Proportion p_n	Sample size for normal theory to apply
.4 or .6	50
.3 or .7	100
.2 or .8	400
.1 or .9	1,000

4.6 Confidence limits when Normal Theory does not apply: Several procedures are available in this situation. One procedure is to construct charts for determining the confidence limits. These charts are based on a summation of the terms in the binomial expansion with varying p and n by use of the Incomplete Beta function. A good set of charts is given in Simon's "An Engineer's Manual of Statistical Methods". Other sources of charts are Clopper and Pearson and the Statistical Research Group (see references). Tables may also be prepared in place of charts. A useful table is given in Snedecor, pp. 4-5 (adapted from Clopper and Pearson).

A direct approach, which appears to be a more useful procedure, has been suggested by M. S. Bartlett. Bartlett considers the normal theory confidence limit equation (21) of (Sec. 4.5) and proceeds to solve it for p. Ignoring the finite population correction the quadratic solution for p can be expressed as

$$p = \frac{p_n + k \pm \sqrt{1 + 2p_n\, q_n/k}}{1 + 2k} \qquad \text{where } k = t^2(\alpha)/2n \text{ and } q_n = 1 - p_n . \qquad (23)$$

As an illustration of the results obtained by the various methods, let us consider the following sample results: Four hundred individuals were asked a given question to which "yes" or "no" answers were recorded. Seventy persons answered "yes", so with n = 400, p_n = 70/400 = .175.

Method	99% Confidence Limits	
	Lower	Upper
Standard "Normal"	.126	.224
Bartlett "Normal"	.131	.229
Simon Chart	.130	.228
Snedecor Table (by interpolation)	.121	.237

In this table the Simon Chart result probably is the "best" answer. The standard "normal" procedure is to be criticized for placing the limits symmetrically about the observed proportion p_n. The advantage of the Bartlett "Normal" method is that it gives an improved answer without the use of charts or tables. On the other hand, the Snedecor Table provides a fair approximation without much calculation.

4.7 Estimation of sample size required: Considering that d is one-half the width of the confidence interval, as in (Sec. 3.2), and that normal theory can be applied, a solution may be obtained for n, the required sample size, for a specified accuracy when sampling the "Binomial Type" population. The solution may be expressed as

$$n = \frac{t^2 \; p_n \; q_n/d^2}{1 - \frac{1}{N}\left\{1 - \frac{t^2 \; p_n \quad q_n}{d^2}\right\}} \qquad (24)$$

When the finite population correction can be ignored the solution becomes simply $n = t^2 \; p_n \; q_n/d^2$. When p is near 0 or 1, the use of the normal approximation will require a big sample. A study of the charts or tables (refer Sec. 4.6) will give a good approximation to the sample size required when normal theory does not apply.

4.8 Extension to more than 2 classes in the Population: There are a number of sampling situations in which the population divides itself into more than 2 classes. We are then confronted with a "Multinomial Type" estimation problem. As an example, suppose a survey has yielded these results in answer to a given question:

	Class			
Reply:	Yes	No	Don't know	No answer
Number giving the reply:	c_1	c_2	c_3	c_4

Then $n = c_1 + c_2 + c_3 + c_4$, the sums of the numbers in the classes. Other definite groupings may be envisaged. Ratios or percentages are then computed from such data. At this stage, 2 cases may be distinguished.

4.9 Case I. We calculate

$$p_n = \frac{\text{Number in any one class}}{n} \quad \text{or}$$

$$p_n = \frac{\text{Number in a combination of classes}}{n} .$$

From the above illustration, we might take the number "Yes" or combine the "Yes" and "No". Then, $p_n = c_1/n$ or $p_n = (c_1 + c_2)/n$. The theory as already presented applies to this case. That is,

$$V(p_n) = \frac{N-n}{(N-1)n} \quad p\ q$$

4.10 Case II. Suppose we take

$$p_{n'} = \frac{\text{Number in one or more classes in the sample}}{n'}$$

$$= \frac{\text{Number in one or more classes in the sample}}{n - (\text{number in certain classes omitted})}.$$

Now the denominator does not include all the classes, e.g, we might omit "no answer" and "don't know" in (Sec. 4.8) and calculate $p_n = c_1/(c_1 + c_2)$, the ratio of "yes" to "yes" plus "no".

Since the denominator is not fixed, the variance appears at first to be more complicated. The situation may be studied in the following manner:

Let N' be the population number in the classes that are being considered and n' the corresponding sample number. We will have

$$N' < N \quad ; \quad n' \leq n .$$

Then it may be shown that in random samples in which both n' and n are fixed, $p_{n'}$ follows the usual binomial distribution about the corresponding p.

What is happening can be indicated by appealing to an example. Suppose a population consists of the 5 elements A B C D E, where D and E are of no interest. Then, N' = 3 with N = 5. Samples of 3 are taken. The possible samples may be grouped according to the value of n'. ADE, BDE, and CDE give n' = 1. ABD, ABE, ACD, ACE, BCD, and BCE yield n' = 2. ABC gives n' = 3. By averaging over the ten samples, or over <u>any</u> group with fixed n', it is easy to see that an unbiased estimate of say A/(A + B + C) , will be obtained.

Hence,

$$E(p_{n'}/n,\ n') = \alpha/N' \tag{25}$$

where α = Numbers in the classes in the population corresponding to the classes in the sample used in forming the numerator for calculating $p_{n'}$. Further, for the variance, we have

$$V(p_{n'}/n,n') = \frac{N'-n'}{N'-1} \quad p\ q/n' \tag{26}$$

With these results we can now apply all the previous developements of this chapter. When normal theory is applied the confidence limits become

$$p = p_{n'} \pm t(\alpha)\sqrt{\frac{N'-n'}{N'-1} \quad \frac{p\ q}{n'}} \tag{27}$$

Now, we note two points:

1) While N is known, N' in general is not known. Quite often it is clear that n'/N' is negligible. In that case, we use

$$p = p_{n'} \pm t\sqrt{pq/n'} \,. \tag{27a}$$

2) If it seems advisable to make a finite population correction, we may assume that N'/N is estimated by n'/n. Then we can use

$$p = p_{n'} \pm t\sqrt{\frac{N-n}{N-1} \quad \frac{p\ q}{n'}} \,. \tag{27b}$$

Notice that n' still appears as the divisor for p q in (27b).

REFERENCES

(6) Bartlett, M. S.

"Subsampling for Attributes" Supplemental Journal of the Royal Statistical Society. Vol. IV, No. 1, 1937.

(7) Clopper, C. J. and Pearson, E. S.

Biometrika 26:404 (1934)

(8)Simon, L. E.

"An Engineer's Manual of Statistical Methods" John Wiley & Sons, New York, 1941.

(9)Snedecor, G. W.

"Statistical Methods" 4th Edition, Iowa State College Press, Ames, Iowa, 1946.

(10)Statistical Research Group
Columbia University

"Selected Techniques of Statistical Analysis" Edited by C. Eisenhart, M. W. Hastay, and W. Allen Wallis, McGraw-Hill, New York, 1947.

STRATIFIED RANDOM SAMPLING

5.1 Description. This type of sampling follows the general procedure of simple random sampling, but takes a preliminary step. The population of size N is first divided into sub-populations of sizes $N_1, N_2 \ldots N_k$. These sub-populations are called strata. Examples of such division are the use of counties within a state, or the separation of the labor force into factory, farm, mine, professional, and clerical groups. When the strata have been determined, a simple random sample is then taken from each stratum independently. The sample sizes within the strata are then $n_1, n_2, \ldots n_k$.

Stratification is a common procedure in sampling. The reasons for its general usage are

(1) If a heterogeneous population is divided into homogeneous strata, the accuracy of the sample can be increased, as will be shown later.

(2) The administrative considerations relating to the survey:

(a) The location of the field offices of the agency conducting the survey may require a division of the area by civil or political units.

(b) Publication policy often requires that data be available for sub-areas of the population.

(c) Action to be taken on the basis of the survey results may not apply uniformly to the whole area.

5.2 Theory for Stratified Random Sampling. The notation is as follows. Let $\bar{y}_{nj}$ be the sample mean and $\bar{y}_{pj}$ be the population mean in the j th stratum. Then,

$$\bar{y}_p = \frac{1}{N} \sum_{j=1}^{k} N_j \bar{y}_{pj} \tag{28}$$

For the estimate of $\bar{y}_p$, we take

$$\bar{y}_n = \frac{1}{N} \sum_{j=1}^{k} N_j \bar{y}_{nj} \tag{29}$$

In (29), we note that the eouation assumes knowledge of the N_j. Thus, more information is required for stratified sampling than for the simple case of an undivided population.

Next, we state that $E(\bar{y}_n) = \bar{y}_p$. This result can be readily obtained by application of Theorem 1a in each stratum.

Theorem 6: With $\bar{y}_n$ defined as in (29),

$$V(\bar{y}_n) = \frac{1}{N^2} \sum_{j=1}^{k} N_j (N_j - n_j) \sigma_j^2 / n_j , \qquad (30)$$

where

$$\sigma_j^2 = \sum_{i=1}^{N_j} \frac{(y_{ij} - \bar{y}_{pj})^2}{N_j - 1} = \text{population variance within the j th stratum.}$$

Proof: From the definitions of $\bar{y}_p$ and $\bar{y}_n$ in (28) and (29) we obtain

$$\bar{y}_n - \bar{y}_p = \frac{1}{N} \Sigma N_j (\bar{y}_{nj} - \bar{y}_{pj}) .$$

Then,

$$V(\bar{y}_n) = E (\bar{y}_n - \bar{y}_p)^2$$

$$= \frac{1}{N^2} \Sigma N_j^2 \left[E (\bar{y}_{nj} - \bar{y}_{pj})^2 + E \text{ (cross-product terms)} \right] . \qquad (31)$$

Since a simple random sample has been taken within each stratum, previous results can be applied. By Theorem 2, we have

$$E(\bar{y}_{nj} - \bar{y}_{pj})^2 = \frac{N_j - n_j}{N_j} \sigma_j^2 / nj . \qquad (32)$$

The sample taken within a stratum is independent of the sample taken within any other stratum, therefore,

$$E(\bar{y}_{nj} - \bar{y}_{pj}) (\bar{y}_{nm} - \bar{y}_{pm}) = 0 \text{ for } j \neq m. \qquad (33)$$

Inserting the results of (32) and (33) in (31) we obtain

$$V(\bar{y}_n) = \frac{1}{N^2} \Sigma N_j (N_j - n_j) \sigma_j^2 / n_j, \text{ the result as stated in (30).}$$

When nj/Nj is negligible, (30) may be reduced to $V(\bar{y}_n) = \frac{1}{N^2} \Sigma \; (N_j^2 \; \sigma_j^2)/n_j$ (34)

For estimating $V(\bar{y}_n)$, we do not know σ_j^2, but we can use the unbiased estimate $V(\bar{y}_n) = \frac{1}{N^2} \Sigma \; N_j \; (N_j - n_j) \; s_j^2/n_j$ (35)

where $s_j^2 = \sum_{i=1}^{n} (y_{ij} - \bar{y}_{nj})^2 \; / \; (n_j - 1)$ = estimated variance within the j th stratum.

(Refer Theorem 3).

5.3 Optimum Allocation. We now examine the problem of allocating the sample to the respective strata: that is, the choice of $n_1, n_2 \; . \; . \; . \; n_k$. From formula (30), the Variance V of the estimated mean $\bar{y}_n$ is seen to be a function of the n_j. Similarly the cost C of taking the sample will also be a function of the n_j. The principle which is used in selecting the n_j is to minimize V for fixed C. Sometimes C is minimized for a specified V; it will be found that this gives the same allocation as the minimizing of V for fixed C.

5.4 Cost functions. The form of the cost function depends on the type of survey. While investigation of cost functions has been rather meager up to the present time, the following type of function may serve as an example, which might be a satisfactory approximation for some kinds of surveys.

$$C = a + \sum_j b_j \sqrt{n} + \sum_j c_j \; n_j$$

This function has three constituents.

a = general overhead cost of the survey.

$b_j\sqrt{n_j}$ = travel cost within the j th stratum.

$c_j \; n_j$ = costs that are proportional to the sample size within the j th stratum (this includes the cost of enumeration).

Note that travel costs have been assumed proportional to the square root of the size of sample. This approximation is based on work by Mahalanobis (11) and Jessen (12), No general discussion of the optimum allocation for this cost function will be given. Two simple cases will be considered. First, we suppose that $b_j = 0$ and $c_j = c$, a constant. Then the cost function becomes

$$C = a + c\,(n_1 + n_2 + \dots + n_k) = a + c\,\Sigma\, n_j\ . \qquad (37)$$

Now $\Sigma\, n_j = n$, so we observe that C is proportional to n, the total sample size, since the cost per schedule is the same in all strata.

For the second case, we consider that the total cost is proportional to $\Sigma\, c_j\, n_j$, i.e., c_j, the cost per schedule, varies from stratum to stratum. Then we have

$$C = c_1\, n_1 + c_2\, n_2 + \dots c_k\, n_k. \qquad (38)$$

Cases I and II are presented below in Theorem 7 and 8, respectively.

5.5 <u>Theorem 7:</u> (Refer J. Neyman, Journal of the Royal Statistical Society, 97 (1939) 558-606). In stratified random sampling, $V(\bar{y}_n)$ is smallest for a fixed total size of sample if the sample is distributed with n_j proportional to $N_j\ \sigma_j$.

<u>Proof</u>: Using the Lagrangian multiplier we have

$$V(\bar{y}_n) + \lambda\, C = \frac{1}{N^2} \sum_{j=1}^{k} N_j \left(\frac{N_j}{n_j} - 1\right) \sigma_j^2 + \lambda\, (\Sigma\, n_j)$$

Differentiating with respect to n_j we obtain

$$\frac{-N_j^2\ \sigma_j^2}{N^2\ n_j^2} + \lambda = 0$$

The solution for n_j gives

$n_j = \dfrac{N_j\ \sigma_j}{N\sqrt{\lambda}}$ or n_j is proportional to $N_j\ \sigma_j$. By summing this result for n_j in both members we can simplify the result since

$$\Sigma\, n_j = n = \frac{\Sigma\, N_j\ \sigma_j}{N\sqrt{\lambda}}$$

Substituting for λ we find the actual value of n_j to be

$$n_j = n \frac{N_j \sigma_j}{\Sigma N_j \sigma_j} \qquad (39)$$

This result, due to Neyman, is very useful whenever the cost of taking the survey (apart from the fixed overhead) is proportional (or almost so) to the size of sample. Note that n_j depends on the product of the size of stratum and the standard deviation of the stratum. Other things being equal, a larger sample is needed in a variable stratum. In practice the values of σ_j will not be known when the sample is planned. Usable estimates of them can often be made either from general knowledge or previous experience with the population.

5.6 The Minimum Variance, CaseI: Now let us re-write the variance from (30) as

$$V(\bar{y}_n) = \frac{1}{N^2} \Sigma \left(\frac{N_j^2}{n_j} - N_j\right) \sigma_j^2$$

$$= \frac{1}{N^2} \Sigma \frac{N_j^2 \sigma_j^2}{n_j} - \frac{1}{N^2} \Sigma N_j \sigma_j^2 \qquad (40)$$

In (40), we substitute the results of Theorem 7, i.e., the value of n_j as given by (39). This yields for the minimum variance, Case I,

$$V(\bar{y}_n) \text{ min.} = \frac{1}{N^2} \frac{(\Sigma N_j \sigma_j)^2}{n} - \frac{1}{N^2} \Sigma N_j \sigma_j^2 . \qquad (41)$$

5.7 Corollary 1 to Theorem 7: If the finite population correction is negligible, the second term in the right member of (41)is small relative to the first. This gives

$$V(\bar{y}_n) \text{ min.} = \frac{(\Sigma N_j \sigma_j)^2}{N^2 n} \qquad (42)$$

Hence, the minimum standard error can be expressed as

$$s(\bar{y}_n) \text{ min.} = \frac{1}{\sqrt{n}} \; \frac{\Sigma N_j \sigma_j}{N} \tag{42a}$$

5.8 <u>Corollary 2 to Theorem 7, Proportional Sampling</u>. If $\sigma_j = \sigma$, a constant, that is, we have homogeneous variance for all the strata, then the optimum allocation occurs when n_j is proportional to N_j. For under this condition (39) reduces to

$$n_j/N_j = \frac{n}{\Sigma N_j} = n/N = \text{a constant.} \tag{43}$$

This type of sampling is called <u>proportional sampling</u>. With proportional sampling the calculation of the estimate is particularly simple, since

$$\bar{y}_n = \frac{1}{N} (\Sigma N_j \bar{y}_{nj}) = \frac{1}{n} (\Sigma n_j \bar{y}_{nj})$$

which is simply the sample total divided by the sample size. Thus, no weighting is required. Such samples are described as <u>self-weighting</u>.

5.9 <u>Theorem 8</u>: <u>Case II of Optimum Allocation</u>: Under the assumptions of (38), above, i.e., cost proportional to $c_j n_j$, the variance $V(\bar{y}_n)$ is a minimum for a given total cost if n_j is proportional to $N_j \sigma_j/\sqrt{c_j}$.

<u>Proof</u>: This is parallel to Case 1, Sec. 5.5. The quantity to be minimized is

$$V(\bar{y}_n) + \lambda C = \frac{1}{N^2} \Sigma N_j (N_j/n_j - 1) \sigma_j^2 + \lambda (\Sigma c_j n_j) \tag{44}$$

Differentiating and equating the result to zero we find

$$(-N_j^2/n_j^2) \sigma_j^2 + \lambda c_j = 0 .$$

Then $n_j \sqrt{\lambda} = N_j \sigma_j/\sqrt{c_j}$.

Summing again in both members and substituting the result obtained for

λ, we obtain

$$n_j = \frac{n(N_j \ \sigma_j/\sqrt{c_j})}{\Sigma(N_j \ \sigma_j/\sqrt{c_j})} \tag{45}$$

From the result for Case II, i.e., the variance is a minimum when n_j is proportional to $N_j \ \sigma_j/\sqrt{c_j}$, we deduce a simple statement of procedure for stratified sampling with the cost conditions assumed:

In a given stratum, take more samples

a. If the stratum is larger

b. If the stratum is more variable

c. If enumeration is cheaper in the stratum.

5.10 <u>Stratified Random Sampling from "Binomial Type" Populations</u>: We recall the discussion and theory presented in Sec. 4.1 to 4.7. The whole population falls into 2 classes. It is desired to estimate the percentage or proportion in each of the classes. In stratified sampling from this type of population we wish to divide the population so that the sub-populations, or strata, are homogeneous. For example, the partitioning should put most or all of the "yes" answers in one group of strata and the "no" answers in another group of strata.

The estimation proceeds as follows: We suppose n_j sampled in the j th stratum, and observe that g_j of the n_j fall in Class I. Then for the estimated population proportion in Class I, we have

$$p_n = \sum_{j=1}^{k} \frac{N_j}{N} \ \frac{g_j}{n_j} \tag{46}$$

In order to estimate the variance we apply Theorem 6 and then Theorem 4. We had

$V(\bar{y}_n) = \frac{1}{N^2} \Sigma \ N_j(N_j-n_j) \ \sigma_j^2/n_j$, By Theorem 4, $\sigma_j^2 = \frac{N_j}{N_j - 1} p_j \ q_j$.

Hence, with p_n defined as in (46),

$$V(p_n) = \frac{1}{N^2} \Sigma \frac{N_j (N_j - n_j)}{n_j} \quad \frac{N_j}{N_j - 1} \; p_j \, q_j \tag{47}$$

When the finite population correction can be ignored, we obtain

$$V(p_n) = \frac{1}{N^2} \; \Sigma \; N_j^2 \, p_j \, q_j/n_j \tag{47a}$$

To obtain a sample estimate of this variance, the observed values are substituted for the p_j and q_j of (47).

The optimum allocation for sampling from a "binomial type" population is as follows: Case I: With Σn_j = constant, n_j is proportional to $N_j \; \sigma_j$. Thus

$$n_j = N_j \frac{\sqrt{N_j}}{\sqrt{N_j - 1}} \quad \sqrt{p_j \; q_j} \tag{48}$$

Case II: Here $\Sigma \; c_j \; n_j$ = constant, and n_j is proportional to $N_j \; \sigma_j/\sqrt{c_j}$. Then we have

$$n_j = N_j \frac{\sqrt{N_j}}{\sqrt{N_j - 1}} \quad \frac{\sqrt{p_j \; q_j}}{\sqrt{c_j}} \tag{49}$$

Note: The results of this section can be extended to the "multinomial situation," refer Sec. 4.8.

5.11 Relative Accuracy of Stratified Random and Simple Random Samples. If intelligently used, stratification will nearly always result in a smaller variance of the estimated mean than is given by a comparable simple random sample. However, it is not true that any stratified sample gives a smaller variance than the comparable simple random sample: if the values of the n_j are far from optimum, stratified sampling may have a higher variance. The principal result is summarized in the following theorem. In this theorem the finite population correction (f.p.c.) is ignored, i.e., terms in $1/N_j$, n_j/N_j.

Theorem 9. If $n_j \propto N_j \sigma_j$ (i.e., the allocation is optimum in the sense of Neyman) then for samples of given total size n, the variance of the mean, $\bar{y}_n$, for V opt. $\leq$ V ran.

Proof: Some preliminary notes are needed. When the f.p.c. is ignored, the formula for the variance of the estimated mean from a stratified sample is

$$\text{V strat.} = \frac{1}{N^2} \Sigma \frac{N_j^2 \sigma_j^2}{n_j} \tag{50}$$

$$\text{If} \quad n_j = \frac{nN_j \sigma_j}{\Sigma N_j \sigma_j} \quad \text{(optimum allocation)}$$

this reduces to

$$\text{V opt.} = \frac{(\Sigma N_j \sigma_j)^2}{nN^2} \tag{51}$$

as previously noted, see (30), (34), and (42). Further, if $n_j = n \frac{N_j}{N}$ (i.e., sampling is proportional) the variance becomes

$$\text{V prop.} = \frac{\Sigma N_j \sigma_j^2}{nN} \tag{52}$$

Now

$$\text{V prop.} - \text{V opt.} = \frac{1}{nN} \left\{ \Sigma N_j \sigma_j^2 - \frac{(\Sigma N_j \sigma_j)^2}{N} \right\} \tag{53}$$

$$= \frac{1}{nN} \Sigma N_j (\sigma_j - \bar{\sigma})^2 \tag{53a}$$

where $\bar{\sigma} = (\Sigma N_j \sigma_j)/N$

This result shows that V opt. will always be smaller than V prop. The size of the difference depends on the amount of variation in the σ_j.

We now proceed to the main proof. For the simple random sample

$$\text{V ran.} = \frac{\sigma^2}{n}$$

where σ^2 is the variance of the whole population. But from an algebraic identity,

$$(N-1)\ \sigma^2 = \Sigma\ (N_j - 1)\ \sigma^2 + \Sigma\ N_j\ (\bar{y}_{pj} - \bar{y}_p)^2 \tag{54}$$

and since terms in $1/N_j$ are negligible, this may be written

$$N\ \sigma^2 = \Sigma\ N_j\ \sigma_j^2 + \Sigma\ N_j\ (\bar{y}_{pj} - \bar{y}_p)^2 \tag{54a}$$

Hence,

$$\text{V ran.} = \frac{\sigma^2}{n} = \frac{\Sigma\ N_j\ \sigma_j^2}{nN} + \frac{\Sigma\ N_j\ (\bar{y}_{pj} - \bar{y}_p)^2}{nN}$$

$$= \text{V prop.} + \frac{\Sigma\ N_j\ (\bar{y}_{pj} - \bar{y}_p)^2}{nN} \tag{55}$$

$$= \text{V opt.} + \frac{\Sigma\ N_j\ (\sigma_j - \bar{\sigma})^2}{nN} + \frac{\Sigma\ N_j\ (\bar{y}_{pj} - \bar{y}_p)^2}{nN} \tag{56}$$

This proves the theorem. It shows that the increase in accuracy from optimum allocation arises from two factors: (1) elimination of differences among the strata means, last term in the right member of (56), and (2) gain from optimum allocation over proportional allocation (middle term on the right). This second factor is to be expected, since a simple random sample allocates the n_j roughly proportionally.

Note: If the f.p.c. cannot be ignored, the result of Theorem 9 becomes

$$\text{V opt.} \leqslant \text{V ran.},$$

Provided that

$$\Sigma\, N_j\, (\bar{y}_{pj} - \bar{y}_p)^2 \geqslant \frac{\Sigma\, (N-N_j)\, \sigma_j^2}{N} \qquad (57)$$

This provisional condition is likely to be satisfied in almost all applications.

5.12 An Example to Illustrate Theorem 9: In Table 1 we present data from a complete census of Jefferson County, Iowa. The population consisted of 2,010 farms. Here we show the data for average corn acres per farm. Thus, the sampling unit is taken as one farm and the item on which the stratification is based is size of farm. Seven size groupings were established.

TABLE 1

AVERAGE CORN ACRES PER FARM BY SIZE OF FARM
JEFFERSON COUNTY, IOWA

Stratum No.	Farm Size Acres	N_j	Corn Acres $\bar{y}_{pj}$	Stratum Total $N_j\bar{y}_{pj}$	σ_j	Prop. Sampling n_j	$N_j\sigma_j$	Optimum Allocation n_j	$N_j\sigma_j^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	0-40	394	5.4	2127	8.3	20	3270	10	27141.0
2	41-80	461	16.3	7492	13.3	23	6131	18	81542.3
3	81-120	391	24.3	9515	15.1	19	5904	17	89150.4
4	121-160	334	34.5	11524	19.8	17	6613	19	130937.4
5	161-200	169	42.1	7110	24.5	8	4140	12	101430.0
6	201-240	113	50.1	5651	26.0	6	2938	9	76388.0
7	241---	148	63.8	9438	35.2	7	5210	15	183392.0
TOTAL		2,010	$\bar{y}_p$=26.3	52,857		100	34,206	100	689981.1

The original data are shown in columns (1) - (6). For a total sample size of 100 farms, column (7) shows the sample sizes in the respective strata for proportional sampling; column (9) gives the same data for sampling with optimum allocation. Since the sampling rate, 100/2010, is about 5 percent, the f.p.c. will be ignored throughout.

We proceed to calculate the variances of the estimated mean for three types of sampling. The variances are exact, since the complete population is known.

Simple Random Sampling:

The variance of the sample mean, $\bar{y}_n$, is V ran. = $\frac{\sigma^2}{n}$. In order to obtain σ^2, we may apply

$$N \sigma^2 = \Sigma N_j \sigma_j^2 + \Sigma N_j (\bar{y}_{pj} - \bar{y}_p)^2 \qquad (54a)$$

The first term on the right is given by the sum in column (10), Table 1. The second term on the right is given by summing the cross-products for columns (4) and (5), Table 1, thus, $\Sigma\left[5.4\ (2127) + \ldots + 63.8\ (9438)\right]$, and subtracting a correction term $(52857)^2/2010$, which gives 557,007.1. Summing the two terms, 689,981.1 + 557,007.1 = 1,246,988.2 = N σ^2, and dividing this result by Nn, we obtain V ran. = 6.20. The standard error is then S.E. $(\bar{y}_n)$ ran. = $\sqrt{6.20}$ = 2.49, and the coefficient of variation, C.V., is about 9.5%.

Proportional Allocation:

Using (52), we obtain for the variance of $\bar{y}_n$ with proportional sampling,

$$V \text{ prop.} = \frac{689981.1}{nN} = 3.43.$$

Then $\quad$ S.E. $(\bar{y}_n)$ prop. = $\sqrt{3.43}$ = 1.85

C. V. = 7.0%

Optimum Allocation:

Finally, the variance of $\bar{y}_n$ for optimum allocation may be obtained by using (51).

$$V \text{ opt.} = \frac{(34206)^2}{nN^2} = 2.90$$

S.E.$(\bar{y}_n)$ opt. = 1.70

C. V. = 6.5%

The comparison of sample size required to obtain the same accuracy by the several methods is a useful measure of efficiency. For comparing proportional with optimum allocation of the sample, we take $n = \frac{3.43}{2.90} \times 100 = 118$. Thus, about a 20% larger sample is required with proportional sampling to obtain the same accuracy as given by a sample of 100 under optimum allocation. The comparison of simple random sampling with optimum allocation gives $n = \frac{6.20}{2.90} \times 100 = 214$ as the size of sample required to obtain the same accuracy as a sample of 100 under optimum allocation. This result, 214, is slightly biased because we have ignored the f.p.c; the bias favors V opt. because the size of the f.p.c. increases as n increases.

5.13 Description of a Sample Survey: Since considerable background in stratified sampling has been given, we now discuss an actual sampling problem. A detailed description of this study is given by Deming & Simmons, Journal of the American Statistical Association, March, 1946, Vol. 41, p. 16-33. The survey, which used mailed questionnaires, was conducted in March 1945 for the Office of Price Administration (OPA). The population consisted of a list of 140,000 tire dealers on record with the OPA.

The information to be obtained by the survey was (1) the number of new truck and bus tires, and (2) the number of new passenger car tires, on hand by the dealers. The previous information, which was available for designing a sample, came from a fairly adequate census taken in September 1944 and a sample taken in December 1944. Both the census and the sample were taken principally by mail, and apparently the circumstances were such that the dealers replied readily by mail.

In setting up a stratification, a problem is met that is common to most surveys. There are two main items to be estimated--new truck and bus tires and new passenger car tires--and a stratification that is good for one of these may not be effective for the other. In this situation, one may either concentrate on the most important item, or try to reach some compromise that will be reasonably effective for both items. Deming and Simmons chose the latter approach. From a study of the previous data, they found (1) that many dealers (e.g., in service stations) had only car tires on hand (2) that dealers who had truck and bus tires tended also to have car tires, and that the number of car tires was roughly proportional to the number of truck and bus tires. This means that a stratification of this group by truck and bus tires would be fairly effective for car tiros. Also, they found (3) that some dealers primarily handle used tiros. Those data led to the following classification of the population.

TABLE 2.

STRATIFICATION OF TIRE DEALERS FOR MARCH 1945 OPA SURVEY

Group Designation	Size of Group	Description of group Dealers holding
A	27000	New truck & bus tires, except those defined as "used tire" dealers, group C.
B	40000	No new truck & bus tires, except those defined as "used tire" dealers, group C.
C	18000	Used tires > 40, and < 40 new pass. or truck or bus tires.
D	2000	Large numbers of tires, i.e., Mfrs. outlets
E	2000	(Newly authorized dealers)*
F	24000	(Non-respondents of Sept. 1944 survey)*
G	29000	(Respondents sending blank returns in the September survey)*

*It is to be noted that many in Group F may be out of business and that in Group G there may be many who have no tires on hand. The type of stock held by group E is not known.

The second stage of the classification comprised a further division of Groups A and B. The 27,000 in A were stratified by the number of new truck and bus tires on hand with classes 1-9, 10-19, 20-29, etc. The 40,000 in B were separated according to the number of new car tires on hand with classes of 0, 1-9, 10-19, etc.

The next problem was the allocation of the sample number or size, i.e., the n_j, to each stratum. The N_j were known and, since this was to be a mailed survey, the cost would be proportional to the N_j. Therefore, optimum allocation would be obtained by making n_j proportional to $N_j \sigma_j$. Again, a question arises. With two principal items of information to be obtained, for which item shall the allocation be made optimum--truck and bus tires, or car tires? The item selected was new car tires, and it appears to have been a good decision.

The information on the relevant σ's was obtained from the September and December surveys. The values as given by the December survey are shown in the following table:

TABLE 3.

STANDARD DEVIATIONS OF THE STRATA - OPA TIRE DEALERS SURVEY

	New Car Tires			
Size in Number of Tires on Hand	Mean $\bar{y}_{pj}$ **	Std. Dev. σ_j		Ratio $\sigma_j/\bar{y}_{pj}$
Group A *				
1-9	14.8	18.2		1.23
10-19	21.0	26.3		1.25
20-29	34.2 **	40.6		1.19
30-39	34.2	28.2		.82
			avg.	1.25
Group B ***				
0	1.0	3.6		3.6
1-9	6.7	8.2		1.22
10-19	13.0	9.9		.76
20-29	24.7	11.4		.46
30-39	32.0	12.4		.39
			avg.	.75

*Group sizes are based on holdings of new truck and bus tires.

**Group means are calculated from holdings of new car tires.

***Group sizes in B are based on holdings of new car tires.

From these data on the means and standard deviations in the strata, two general assumptions were made. For Group A, Deming and Simmons took $\sigma_j = 2\ \bar{y}_{pj}$, and for Group B, they took $\sigma_j = \bar{y}_{pj}$. These were conservative assumptions, though a greater variation in the survey to be taken in March was anticipated.

Now, we consider the problem of determining the size of sample for this survey. The accuracy to be obtained was specified. The coefficient of variation for total number of new tires on hand to be attained by the survey was set at 1.5%, or .015.

Let $n_j = \sigma_j\ N_j/k$ where k is an unknown constant to be determined,

then, $V(\bar{y}_n) = \frac{1}{N^2} \quad \frac{\Sigma\ N_j^2\ \sigma_j^2}{n_j}$ (omitting the f.p.c.).

Substituting for n_j, we obtain

$$V\ (\bar{y}_n) = \frac{k}{N^2}\ \Sigma\ N_j\ \sigma_j\ .$$

In this survey the estimate wanted was the <u>total</u> number of new tires on hand. We write this estimate as $Tn = N\ \bar{y}_n$. Hence, $V(Tn) = k\ \Sigma\ N_j\ \sigma_j$. At this stage we introduce from the preceding paragraph the assumptions on the σ's for Group A and Group B, and write

$$V\ (Tn) = k\left(\sum_A 2\ N_j\ \bar{y}_{pj} + \sum_B N_j\ \bar{y}_{pj}\right)$$ where the summations

are over the strata in Groups A and B, respectively.

$$= k\ (2\ T_A + T_B).$$

In this form, T_A and T_B indicate the population totals of number of tires in the groups. From the last result, we write the coefficient of variation of T_n as

$$C.V.\ (T_n) = \frac{\sqrt{k}\ \sqrt{2\ T_A + T_B}}{T_A + T_B}\quad ,$$

Before proceeding further it was necessary to estimate the number of new tires expected to be found on hand in the March survey. Such estimates were based on the numbers found on hand in the September and December surveys. Deming and Simmons estimated

$$T_A \text{ at } 1.6 \times 10^6$$

$$\text{and} \quad T_B \text{ at } 0.2 \times 10^6 \text{ .}$$

With the C.V.(T_n) already set at .015, we can now solve for k. Therefore,

$$k = \frac{(.015)^2 \; (1.8)^2 \times 10^{12}}{3.4 \times 10^6} = 214.$$

However, k was actually taken as 200 in order to simplify further calculations. This value of k required a sample size in Groups A and B of about 13% which strictly requires the use of the f.p.c., although it was omitted.

The allocation of the sample to the strata is now straightforward. In Group A we have n_j/N_j = the fraction to be sampled within a stratum = σ_j/k = $2\ \bar{y}_{pj}/200$. From this relation we obtain the percent sampled in the strata of Group A = $\bar{y}_{pj}$. Similarly, the percent sampled in the strata of Group B = $\bar{y}_{pj}/2$. An estimate of the $\bar{y}_{pj}$ for each of the strata in Groups A and B that would be found in the March survey then finally determined the strata sampling rates. In general, these values, $\bar{y}_{pj}$, were estimated from the December survey results. Table 4 below shows the sampling rates obtained for the strata in Group A.

TABLE 4.

SAMPLING RATES IN GROUP A - OPA TIRE DEALERS SURVEY

Size	N_j	Estimated $\bar{y}_{pj}$ for March	Sampling Rate
1-9	19,850	15*	1 in 6
10-19	3,250	22**	1 in 5
20-29	1,613	30	1 in 3
30-39		35	
40-49	894	45	1 in 2
50-59		55	
60+	1,662	? (100% taken)	1 in 1
	27,269		

*(December value - 14.8) **(December value - 21)

The method employed for taking a random sample of 3,300 out of the 19,850, 600 out of the 3,250, etc. was as follows. The members of each stratum were available on cards showing addresses. A random card was chosen as a starting point and all succeeding members of the sample were taken systematically at the designated sampling rate. Thus, in the first size group in Table 4, every 6th card was chosen thereafter. This method of sampling is known as systematic sampling and will be discussed later. In computing sampling errors, the authors assumed that their samples were equivalent to simple random samples within strata. Their comment on this point is that the sampling error of their sample is probably either equal to or slightly lower than the result given by the use of stratified random sampling formulae.

The remaining strata, i.e., Groups C through G, were handled as follows:

Group C - "Dealers holding more than 40 used tires": They were stratified by number of used tires, 40-49, 50-59, etc. Then a 25% sample, or 1 in 4, was taken in each stratum.

Group D - Manufacturers outlets: $\bar{y}_p$ = 75 for this group on previous survey. A 100% sample was taken.

Group E - Newly authorized dealers: One thousand new dealers were authorized between September and December. Hence, the size of the group was estimated as 2,000 for March, A 10% sample was taken.

Group F - Non-responses in September 1944: This group comprised 24,000 dealers. For the December survey a 4% sample (n = 997) had been taken from this group. The sample was classified into these categories:

(1)	Out of business	217
(2)	Unidentified	310
(3)	Located and schedule returned	470

The (3)rd category showed 11.9 new tires on hand per dealer in December. This indicated that Group F as a whole held many new tires. Determination of the sampling rate for Group F then followed this reasoning: $\bar{y}_p \geqslant 11.9\ (470/997) = 6$ from which σ was estimated as $3\ \bar{y}_p$ or approximately 18. By using the relation $n_j/N_j = \sigma_j/k = 18/200 = 9\%$ was obtained as the sample size in Group F. This value was deliberately cut to 5%, because of the difficulty of actually securing the sample from this group, i.e., greater cost.

Group G - Dealers sending in blank returns: This group was assumed to have few new tires. A 3% sample taken in December showed only 2.3 new tires per dealer. The comparison of this value with the first two strata of Group B, which had similar means, indicated that $\sigma = 2\ \bar{y}_p$ might be a reasonable assumption. Again, the application of $n_j/N_j \doteq \sigma_j/k$ gave $\frac{2(2.3)}{200} = 2.3\%$ sample. It was decided to take a 3% sample of this group again for the March survey.

Summary: The results of the March survey indicated that the desired precision had been attained. The example illustrates how sampling theory is combined with data from previous surveys to plan a new survey efficiently.

5.14 Estimation from a Sample of the Gain Due to Stratification: The formulae in Section 5.11 enable us to estimate the gain in accuracy due to stratification when a complete census of the population has been made. A similar estimate can be obtained when a stratified random sample has been taken. This estimate gives an appraisal of the utility of the stratification that was adopted in the survey. We will ignore finite population corrections in this section.

The data available from the stratified sample are the values of N_j, n_j, $\bar{y}_{nj}$, s_j^2 (estimate of the within-stratum variance σ_j^2). With the f.p.c. ignored, the estimated variance of the mean of the stratified sample is

$$\text{Estd. V strat.} = \frac{1}{N^2} \Sigma \frac{N_j^2 \ s_j^2}{n_j} = \Sigma \ \frac{W_j^2 \ s_j^2}{n_j} \tag{58}$$

where $W_j = N_j/N$.

We wish to compare this with an estimate of the variance of the mean that would have been obtained from a simple random sample. Now

$$\text{V ran.} = \frac{1}{n} \left[\frac{\Sigma \ (N_j - 1) \ \sigma_j^2 + \Sigma \ N_j \ (\bar{y}_{pj} - \bar{y}_p)^2}{(N - 1)} \right] \tag{59}$$

Since terms in $1/N_j$ are negligible, this may be simplified to

$$\text{V ran.} = \frac{1}{n} \left[\Sigma \ W_j \ \sigma_j^2 + \Sigma \ W_j \ (\bar{y}_{pj} - \bar{y}_p)^2 \right] \tag{60}$$

From the results for the stratified sample, there is no difficulty in obtaining an estimate of the first term inside the bracket. The second term requires investigation, since $\bar{y}_{pj}$ and $\bar{y}_p$ are not known.

Now

$\bar{y}_{nj} = \bar{y}_{pj} + \bar{e}_{nj}$, where $\bar{e}_{nj}$ is an error of sampling with $V\ (\bar{e}_{nj}) = \dfrac{\sigma_j^2}{n_j}$

Hence

$$E\ (\bar{y}_{nj}^2) = \bar{y}_{pj}^2 + \frac{\sigma_j^2}{n_j} \tag{61}$$

Thus

$$E\ \Sigma\ (W_j\ \bar{y}_{nj}^2) = \Sigma\ W_j\ \bar{y}_{pj}^2 + \Sigma\ \frac{W_j\ \sigma_j^2}{n_j} \tag{62}$$

Also

$$\Sigma\ W_j\ \bar{y}_{nj} = \Sigma\ W_j\ \bar{y}_{pj} + \Sigma\ W_j\ \bar{e}_{nj} \tag{63}$$

Hence

$$E \left\{ \Sigma\ W_j\ \bar{y}_{nj} \right\}^2 = (\Sigma\ W_j\ \bar{y}_{pj})^2 + \Sigma\ \frac{W_j^2\ \sigma_j^2}{n_j} \tag{64}$$

Subtract (64) from (62).

$$E\left[\Sigma\ W_j\ \bar{y}^2_{nj} - (\Sigma\ W_j\ \bar{y}_{nj})^2\right] = \Sigma\ W_j\ \bar{y}^2_{pj} - (\Sigma\ W_j\ \bar{y}_{pj})^2 + \Sigma\ \frac{W_j\ \sigma^2_j}{n_j} - \Sigma\ \frac{W^2_j\ \sigma^2_j}{n_j} \qquad (65)$$

It follows that an unbiased estimate of

$$\Sigma\ W_j\ (\bar{y}_{pj} - \bar{y}_p)^2$$

is given by

$$Q = \Sigma\ W_j\ \bar{y}^2_{nj} - (\Sigma\ W_j\ \bar{y}_{nj})^2 - \frac{\Sigma\ W_j\ s^2_j}{n_j} + \Sigma\ \frac{W^2_j\ s^2_j}{n_j} \qquad (66)$$

Finally

$$\text{Estd. V ran.} = \frac{1}{n}\left[\Sigma\ \ W_j\ s^2_j + Q\right] \qquad (67)$$

In order to illustrate the computations, we present a numerical example. The data are taken from the CPA Tire Dealers Survey as reported by Deming and Simmons (refer Sec. 5.13).

TABLE 5.

DATA AND CALCULATIONS FOR ESTIMATING GAIN DUE TO STRATIFICATION

GROUP A - OPA TIRE DEALERS SURVEY

Size of Stratum	N_j	n_j	$\bar{y}_{nj}$	s^2_j	W_j	$W^2_j\ s^2_j/n_j$	$W_j s^2_j/n_j$	$W_j\bar{y}_{nj}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1-9	19850	3000	4.1	34.8	.8032	.00748	.00932	3.29312
10-19	3250	600	13.0	92.2	.1315	.00266	.02021	1.70950
20-29	1007	340	25.0	174.2	.0407	.00085	.02085	1.01750
30-39	606	230	38.2	320.4	.0245	.00084	.03413	.93590
		4170			1.0000	.01183	.08451	6.95602

From the data in Table 5, we find

$$\text{V strat.} = \Sigma\ W^2_j\ s^2_j/n_j = .01183 \qquad (68)$$

Now,

$$V \text{ ran.} = \frac{1}{n}\left[\Sigma\, W_j\, s_j^2 + \Sigma\, W_j\, (\bar{y}_{pj} - \bar{y}_p)^2\right]$$

$$= \frac{1}{n}\left[55.02 + \Sigma\, W_j\, (\bar{y}_{pj} - \bar{y}_p)^2\right] \; . \qquad (69)$$

The second term in the brackets in (69) we estimate by applying (66).

$$Q = \Sigma\, W_j\, \bar{y}_{nj}^2 - (\Sigma\, W_j\, \bar{y}_{nj})^2 - \Sigma\, W_j\, s_j^2/n_j + \Sigma\, W_j^2\, s_j^2/n_j$$

$$= 96.91 \quad - \quad (6.95602)^2 - \quad .08451 \quad + \quad .01183 = 48.45 \qquad (70)$$

Then,

$$V \text{ ran.} = \frac{1}{4170}\,(55.02 + 48.45) = .02481 \quad , \qquad (71)$$

whereas V strat. was .01183.

The reduction from V ran. exceeds 50%, since the ratio of the variances is .01183/.02481 = .477.

<u>Simplification when σ_j^2 is constant and sampling is proportional</u>. In this case, which often arises in sampling field experiments, the results simplify considerably.

We have

$$\frac{n_j}{n} = \frac{N_j}{N} = W_j \text{ in all strata}$$

$$\sigma_j^2 = \text{constant which we write as} = \sigma_w^2$$

This is estimated by the <u>pooled</u> mean square within strata, s_w^2 . Then we have

$$\text{Estd. : } V \text{ strat.} = \Sigma\, \frac{W_j^2\, s_w^2}{n_j} = \frac{s_w^2}{n} \; . \qquad (72)$$

$$\text{Estd. } V \text{ ran.} = \frac{1}{n}\left[s_w^2 + Q\right] \qquad (73)$$

from (58), (66), and (67). The quantity Q now becomes

$$Q = \frac{1}{n}\;\Sigma\; n_j\, (\bar{y}_{nj} - \bar{y}_n)^2 - \frac{k}{n}\; s_w^2 + \frac{s_w^2}{n} \; . \qquad (74)$$

Hence,

$$\text{Estd. V ran.} = \frac{1}{n^2}\left[\Sigma\, n_j\,(\bar{y}_{nj} - \bar{y}_n)^2 + (n - k + 1)\, s_w^2\right] \qquad (75)$$

This quantity is easily calculated from an analysis of variance of the sample data into "among strata" and "~~written~~ within strata".

Analysis of variance for the stratified sample.

	d.f.	M.S.
Among strata	$(k - 1)$	$B = \Sigma\, n_j\,(\bar{y}_{nj} - \bar{y}_n)^2/(k-1)$
Within strata	$(n - k)$	$W = s_w^2$

From this the formula (75) may be written

$$\text{Estd. V ran.} = \frac{1}{n^2}\left[(k-1)\,B + (n-k+1)\,W\right] \qquad (76)$$

while

$$\text{Estd. V strat.} = \frac{1}{n}\,W \qquad (77)$$

Example: In sampling a field experiment for estimating number of wireworms on each plot, the plots were divided into halves and three random samples of soil were taken with a small boring tool in each half. (The sample was 9" square to a depth of 5"). There were 25 plots in the experiment. The analysis of variance of numbers of wireworms was as follows.

	d.f.	M.S.
Between strata (half-plots)	25	90.76 = B
Within strata	100	38.44 = W

Note that the conditions in the example are slightly different from those in the theory presented above. Each plot represents a separate population, divided into 2 strata. Thus k = 2 and n = 6. The analysis of variance gives the combined results for 25 stratified samples of this type.

$$\text{Est. V strat.} = \frac{38.44}{6} = 6.41$$

$$\text{Est. V ran.} = \frac{1}{36}\left[B + 5W\right] = \frac{1}{36}\left[90.76 + 5(38.44)\right] = 7.86$$

$$\text{R.E.} = 7.86/6.41 = 1.23.$$

Thus, stratification into halves increased the accuracy of the experiment by slightly under one-fourth.

5.15 Confidence Limits and Sample Size for Stratified Random Sampling: The variance is more complicated with stratification than with simple random sampling. (Refer Sec. 3.1 ff.) Functionally, we may express this variance in general as

$$V(\bar{y}_n) \text{ strat.} = f\ (\sigma_j,\ N_j,\ n_j).$$

After the determination of the strata, the first step is to allocate the sample to the strata, or to determine the ratios n_j/n. When this has been done, we may write the variance as a specific function

$$V(\bar{y}_n) = g\ (\sigma_j,\ N_j,\ n).$$

At this stage we note again that either the σ's must be known or good estimates of them must be available. Then the confidence limits are $\bar{y}_n \pm t(\alpha)\sqrt{V\ (\bar{y}_n)}$. To determine sample size, we equate $t(\alpha)\sqrt{V\ (\bar{y}_n)}$ to d, the specified confidence limit, and solve for n.

As an illustration of the above procedure, consider Case I of optimum allocation with $c_j = c =$ a constant. In (41) the minimum variance was expressed as

$$V\ (\bar{y}_n) \text{ min.} = \frac{1}{N^2}\left[\frac{(\Sigma\ N_j\ \sigma_j)^2}{n} - \Sigma\ N_j\ \sigma_j^2\right] . \qquad (41)$$

Therefore

$$\frac{t(\alpha)}{N}\sqrt{\frac{(\Sigma\ N_j\ \sigma_j)^2}{n} - \Sigma\ N_j\ \sigma_j^2} = d \qquad (78)$$

From (78) we obtain the solution for n as

$$n = \frac{(\Sigma\ N_j\ \sigma_j)^2}{\frac{N^2\ d^2}{t^2(\alpha)} + \Sigma\ N_j\ \sigma_j^2} \tag{79}$$

As a first approximation, the finite population correction is neglected. This gives

$$n_o = \frac{t^2(\alpha)\ (\Sigma\ N_j\ \sigma_j)^2}{N^2\ d^2} \tag{79a}$$

When n_o/N is not negligible, n is calculated directly from (79).

An interesting corollary can be derived from (51). Suppose $\sigma_j = \sigma =$ a constant. Then we have

$$n = \frac{\sigma^2\ N^2}{\frac{N^2\ d^2}{t^2(\alpha)} + N\ \sigma^2} = \frac{t^2(\alpha)\ \sigma^2/d^2}{1 + \frac{t^2(\alpha)\ \sigma^2}{N\ d^2}} \tag{80}$$

This result has the same form as was derived for simple random sampling. Thus,

$$n_o = t^2(\alpha)\sigma^2/d^2$$

and

$$n = \frac{n_o}{1 + n_o/N} \ .$$

The assumption that σ_j is constant is not unreasonable for some types of field crops or soil samplings. But the assumption is less plausible in human sampling, e.g., business and economic inquiries, where the σ_j are usually quite variable.

If proportional sampling is to be employed, the sample will be allocated according to the size of the strata, i.e.,

$$\frac{n_j}{N_j} = \frac{n}{N} \ ,$$

and then

$$n_j = \frac{N_j \, n}{N} \quad .$$

As shown in Theorem 6 the variance for a stratified sample when we do not have optimum allocation is given by (30). Hence, we may write

$$V(\bar{y}_n) \text{ prop.} = \frac{1}{N^2} \; \Sigma \, N_j(N_j - n_j) \; \sigma_j^2/n_j \; . \tag{81}$$

Substituting $\frac{N_j \, n}{N}$ for n_j in the formula for estimating V prop. we obtain

$$\frac{t(\alpha)}{N} \sqrt{\frac{N \, \Sigma \, N_j \, \sigma_j^2}{n} - \Sigma \, N_j \, \sigma_j^2} \;\; = \;\; d. \tag{82}$$

If n is solved for in equation (82) the result is

$$n \;=\; \frac{N \, \Sigma \, N_j \, \sigma_j^2}{\frac{d^2 \, N^2}{t^2(\alpha)} + \Sigma \, N_j \, \sigma_j^2} \;=\; \frac{\frac{t^2(\alpha) \; \Sigma \, N_j \, \sigma_j^2}{d^2 N}}{1 + \frac{1}{N^2}\left(\frac{t^2(\alpha) \Sigma \, N_j \, \sigma_j^2}{d^2}\right)} \tag{83}$$

Similarly, by ignoring the finite population correction factor, a first approximation becomes, from (52),

$$n_o \;=\; \frac{t^2(\alpha) \; \Sigma \, N_j \, \sigma_j^2}{d^2 \; N} \tag{83a}$$

If n_o/N is not negligible, n must be calculated from (83).

5.16 <u>Proximity as a Basis for Stratification</u>: In Section 5.1 one of the advantages presented for stratified sampling was the possibility of securing increased accuracy from the sample by dividing a heterogeneous population into homogeneous groups. Succeeding sections have shown how this is obtained. The question arises, "What criteria should be employed in stratifying a given population which is to be sampled?"

So far as possible, the criteria should be such that each stratum is homogeneous with respect to the items that are to be obtained in the survey.

Sometimes the most appropriate criteria are rather obvious from the nature of the survey; in other cases investigations are conducted in order to compare the effectiveness of different criteria. Frequently a compromise must be adopted, since the criterion that gives a good stratification for some items in the survey is poor for other items that seem equally important. Discussions of bases for stratification for economic items have been given by Stephan (13) and Hagood and Bernert (14), and for farm items by King and McCarty (15).

One principle that frequently holds is that adjacent sampling units are more alike than sampling units that are far apart. Hence, proximity of the units, or a geographical division of the population is used as a basis for stratification.

To indicate the results given by this procedure we shall consider several examples. The comparison of geographical stratification with simple random sampling may be made by calculating the relative efficiency. Here the relative efficiency of the stratified to the simple random sample is defined as the inverse ratio of their variances; that is, the variance of the mean from the random sample is divided by the variance of the mean from the stratified sample. Thus,

$$\text{R.E.} = \frac{\text{V random}}{\text{V strat.}} \qquad (84)$$

In (84) equal sized samples of n are assumed for both methods, simple random and stratified random sampling. When the finite population correction is negligible, (84) also gives the relative sizes of sample that must be taken to give the same variance for the estimated mean. This can be shown as follows: Suppose that the random sample is increased in size fron n to rn. Then, the variance of the mean of the random sample becomes σ^2/rn or $\frac{\text{V random}}{r}$. Now, if r is chosen so that $\frac{\text{V random}}{r} = \text{V strat.}$, we obtain

$$r = \frac{\text{V random}}{\text{V strat.}} = \text{R.E.} \qquad (85)$$

As the first example, we consider a problem in the counting of forestry nursery seedlings, refer F. A. Johnson (16). The seedlings were grown in long narrow beds. Sampling units were narrow strips across the beds. The number of seedlings in each sampling unit was determined by counting. Each bed was divided into about 20 strata. The pertinent results for comparing the sampling methods are given in Table 6.

TABLE 6.

Type of Seedling	R.E. or r	
	Bed #1	Bed #2
Silver Maple	1.29	1.49
American Elm	2.79	1.32
White Spruce	1.16	1.88
White Pine	1.15	-----

Table 7 shows results obtained for a number of typical farm economic items. In these investigations different sizes of strata were compared: townships, four-township blocks, counties, and type-of-farming areas within a state. A mean relative efficiency was calculated by averaging the individual relative efficiencies for each item.

TABLE 7.

State	No. of Items	Twp.	4-Twp.	County*	Type of farming area	State
Iowa - 1938	18	115	-	100	96	91
Iowa - 1939	19	121	-	100	97	91
Florida - 1942						
Citrus fruit area	14	144	119	100	-	
Truck farming area	15	111	-	100	-	
California - 1942	17	113	-	100	97	

*Average relative efficiencies were converted to a relative basis in each case by taking the county value as 100. Refer: Jessen (12) and Jessen & Houseman (17).

In both examples the increases in accuracy from geographic stratification are moderate rather than large. This appears to be typical of results with geographic stratification.

5.17 Effects of Errors in the Strata Totals: It frequently happens in practice that for some desirable type of stratification the strata totals N_j are not known exactly, being perhaps derived from a population count that is out of date, or from another sample. Definite statements about the consequences of basing a stratification upon erroneous weights cannot be made without considering particular cases. A few conclusions of a general nature can, however, be drawn.

For simplicity, finite population corrections will be ignored and the cost per unit is assumed the same in all strata. If the N_j were known, n_j would be chosen equal to $nN_j \sigma_j / \Sigma N_j \sigma_j$. The sample estimate of the population mean would be $\Sigma N_j \bar{y}_{nj}/N$, which may be written $\Sigma W_j \bar{y}_{nj}$. Its variance simplifies to

$$\frac{(\Sigma W_j \sigma_j)^2}{n} \tag{86}$$

Instead of the true stratum proportions W_j, we have estimates w_j. The sample estimated mean is $\Sigma w_j \bar{y}_{nj}$. The first point to note is that this estimate is biased. Its mean value in repeated sampling is $\Sigma w_j \bar{y}_{pj}$, while the true population mean is $\Sigma W_j \bar{y}_{pj}$. The bias amounts to $\Sigma (w_j - W_j) \bar{y}_{pj}$. Consequently, the error variance of this estimate contains two components: the variance about its own mean and the square of the bias. If optimum allocation is used (with, of course, the N_j replaced by their estimates) the first component is $(\Sigma w_j \sigma_j)^2/n$. The total variance is

$$\frac{(\Sigma w_j \sigma_j)^2}{n} + \left\{ \Sigma (w_j - W_j) \bar{y}_{pj} \right\}^2 \tag{87}$$

A more general form of this expression was given by Stephan (13). He points out that the first term in (87) will usually be about the same size as (86) - they are exactly the same if the variance is the same in all strata. The loss of accuracy from incorrect weights thus depends mainly on the size of the bias, which in individual cases might either be small or large. Further, for any given set of erroneous weights, the loss varies with the size of sample taken. This is so because the 'bias' component of the total variance is independent of the size of sample. With increasing sample size, a stage is reached where the 'bias' term predominates, and where the stratification would be less accurate than simple random sampling.

The preceding discussion does not help much in considering whether to stratify in a survey where the weights are known to be in error, because the size of the bias term cannot be predicted. Occasionally a standard error can be attached to the estimate of each N_j, from knowledge of the process by which these were estimated. If the estimates of the N_j are independent, and independent of the $\bar{y}_{nj}$, the average value of the bias component of the total variance is roughly, refer Cochran (18),

$$\Sigma \ (\bar{y}_{pj} - \bar{y}_p)^2 \ V \ (N_j)/N^2 \qquad (88)$$

where V (N_j) is the variance of our estimate of N_j. This quantity measures the expected increase in variance due to errors in the N_j.

King, McCarty and McPeek (19) applied this formula in research directed towards the estimation of yield per acre, protein and test weight in the wheat belt. They discuss the advisability of stratification by districts within each state. The total acreages N_j for each district were themselves estimated by a sample survey, so that some knowledge of the V (N_j) was available.

5.18 <u>Case Where the Strata Cannot be Identified in Advance</u>: In certain common types of survey it is not possible to tell accurately to what stratum a sampling unit belongs until the data have been secured from the unit. For example, in an election poll it may be useful to stratify according to the individual's vote at the last election. This will not be known until the individual has been contacted. A similar situation arises in a greater or less degree when stratification is by factors such as income, occupation, religious affiliation, ownership of telephone, etc. Of course, in such cases it is also likely that the strata sizes N_j may not be known exactly: we will, however, assume for the present discussion that reasonably good estimates of the N_j are available.

One procedure that can be used is to take a simple random sample of size n. Then classify the units into the strata on the basis of the information obtained about them. If $\bar{y}_{nj}$ is the mean of these units that fall in the j th stratum, use as an estimate

$$\bar{y}_w = \Sigma N_j \bar{y}_{nj}/N . \tag{89}$$

In other words we use the <u>true</u> strata sizes as weights to obtain a weighted mean, instead of taking the unweighted mean of the sample as our estimate.

If the sample is reasonably large, this technique is almost as accurate as <u>proportional</u> stratified sampling. Let m_j be the number in the sample that fall in the j th stratum, where m_j will vary from sample to sample. For samples in which the m_j are fixed,

$$V(\bar{y}_w) = \frac{1}{N^2} \Sigma N_j^2 \frac{\sigma_j^2}{m_j} \tag{90}$$

where the f.p.c. is ignored. The average value of this quantity in repeated sampling must now be calculated. This requires a little care, since it could happen that one or more of the m_j were zero. If this occurred, we should have to combine two or more strata before making the estimate. This would give a less accurate stratification and a higher variance for $\bar{y}_w$. However, with increasing <u>n</u> it may be shown that the probability that any m_j is zero is so small that the contribution to the variance from this source is negligible.

If the case where n_j is zero is ignored, Stephan (20) has shown that to terms of order n^{-2}.

$$E\left(\frac{1}{m_j}\right) = \frac{1}{nW_j} - \frac{1}{n^2 W_j} + \frac{1}{n^2 W_j^2} \qquad (91)$$

where $W_j = N_j/N$. Hence, substituting in (90),

$$E\left\{V(\bar{y}_w)\right\} = \frac{1}{n} \Sigma W_j \sigma_j^2 + 0 \ (n^{-2}) \qquad (92)$$

The leading term is the variance obtained with proportional stratified sampling (Sec. 5.14).

5.19 Quota Sampling: Another method that is used for this problem is to decide in advance the n_j values that are wanted from each stratum and to instruct the enumerator to continue sampling until the necessary "quota" has been obtained in each stratum. In the later stages of sampling, this may require considerable work on the part of the enumerator since most of the units that are contacted may fall in strata where the quota has already been met. If the enumerator chose the units initially at random, rejecting those that in later stages were not needed, this method would be equivalent to ordinary stratified sampling. The extra field work required to fill every quota might be very substantial.

As this method is used in practice by a number of agencies, the enumerator does not select units initially at random. Instead, he may use any information that will enable the quotas to be filled quickly (e.g., such as that people earning high incomes are not likely to live in slums). The object is to gain the advantages of stratification without the high field costs that might be incurred in an attempt to select units initially at random. The amount of latitude permitted to the enumerators varies from case to case. Unfortunately little is known about the accuracy of such "quota" methods as used in practice, relative to that given by more objective approaches.

5.20 The Problem of Non-Response: In many types of survey, there are certain units in the sample from which the desired information cannot be obtained at the first attempt. With human populations, this group may be persons who are not at home, or do not reply to a mail questionnaire. In crop surveys certain fields in the sample may not be ripe when the sampler reaches them. This 'non-response' group constitutes an important practical problem. To obtain information from it may require several attempts and be costly. To ignore it may result in a sample that has a bias of unknown dimensions. An ingenious application to this problem of the idea of stratified sampling has been made by Hansen and Hurwitz (21).

The population is envisaged as containing two strata. One, of size N_1 contains units that provide the information at the first try. The second, of size N_2, is the non-response stratum. The basic idea is that the second stratum should be sampled at a lower rate than the first, since the cost per unit is higher in that stratum. There is, however, the complication that neither the values of N_1 and N_2, nor even the units that fall in the two strata, is known in advance.

The first step, in the simplest case, is to take a random sample of n units. Of these let n_1 be the number that provide the data sought, and n_2 the number in the non-response group. By repeated efforts, the data are later obtained from a random sample of r_2 out of the n_2. If

$$n_2 = kr_2 \tag{93}$$

the quantity k is the ratio of the sampling rate in the first stratum to that in the second. The values of n (initial size of sample) and k are chosen so as to give a specified accuracy for the lowest cost.

The cost of taking the sample is

$$C = c_o n + c_1 n_1 + c_2 r_2 , \tag{94}$$

where the c's are costs per unit: c_o is the cost of making the first attempt, while c_1 and c_2 are the costs of getting and processing the data in the two strata respectively. Since the values of n_1 and n_2 will not be known until the first attempt is made, the expected cost must be used in planning the sample. The expected values of n_1 and r_2 are respectively $W_1 n$ and $W_2 n/k$, where $W_1 = N_1/N$. Thus expected cost is

$$c_o n + c_1 W_1 n + c_2 W_2 n/k . \tag{95}$$

Let $\bar{y}_{1n}$, $\bar{y}_{2r}$ be the sample means in the two strata, respectively, where the suffices n, r are used as a reminder that the sample in the first stratum is of size n_1, while that in the second is of size r_2. As an estimate of the population total, Hansen and Hurwitz take

$$y_s = \frac{N}{n} \left\{ n_1 \bar{y}_{1n} + n_2 \bar{y}_{2r} \right\} \tag{96}$$

Note that the second stratum receives a weight n_2, although the sample is only of size r_2. This is done in order to obtain an unbiased estimate.

The calculation of the variance of this estimate is not as straightforward as it might seem at first sight. For while n may be regarded as fixed, n_1 and n_2 and consequently r_2 vary from sample to sample as well as $\bar{y}_{1n}$ and $\bar{y}_{2r}$. In fact, n_1 and n_2 follow binomial distributions with probabilities N_1/N, N_2/N, respectively. We will suppose that k is fixed from sample to sample; i.e., it has been decided beforehand to what extent the second stratum will be under-sampled.

The easiest method of finding the variance is to introduce the quantity $\bar{y}_{2n}$, that is, the mean of the whole sample of size n_2 from the second stratum. We may introduce this quantity by expressing (96) as follows.

$$y_s = \frac{N}{n} \left\{ n_1 \bar{y}_{1n} + n_2 \bar{y}_{2n} \right\} + \frac{Nn_2}{n} (\bar{y}_{2r} - \bar{y}_{2n}) \qquad (97)$$

The first quantity is simply N times the mean of a random sample of n from the whole population. Its variance is therefore

$$\frac{N(N-n)}{n} \sigma^2$$

where σ^2 is the variance of the whole population. Further, when we find the variance of y_s, there will be no contribution from cross-products between the first and second terms. For if we average

$$\bar{y}_{2n}(\bar{y}_{2r} - \bar{y}_{2n})$$

over all random samples of size r_2 that can be drawn from a fixed sample of size n_2, the average will be zero. Consequently,

$$V(y_s) = \frac{N(N-n)}{n} \sigma^2 + \frac{N^2}{n^2} E \left\{ n_2^2 (\bar{y}_{2r} - \bar{y}_{2n})^2 \right\} \qquad (98)$$

Consider the second term. If $\bar{y}_{2p}$ is the population mean of the 'non-response' stratum, we have

$$(\bar{y}_{2r} - \bar{y}_{2p}) = (\bar{y}_{2r} - \bar{y}_{2n}) + (\bar{y}_{2n} - \bar{y}_{2p}) \qquad (99)$$

so that

$$E (\bar{y}_{2r} - \bar{y}_{2p})^2 = E (\bar{y}_{2r} - \bar{y}_{2n})^2 + E (\bar{y}_{2n} - \bar{y}_{2p})^2 \qquad (100)$$

there being no contribution from cross-product terms for the same reason as before. Now $\bar{y}_{2r}$ is the mean of a random sample of size r_2 from the second stratum, and $\bar{y}_{2n}$ is the mean of a random sample of size n_2 from the same stratum. Hence, <u>for fixed</u> n_2 <u>and</u> r_2 ,

$$\frac{(N_2 - r_2)}{N_2} \quad \frac{\sigma_2^2}{r_2} = E(\bar{y}_{2r} - \bar{y}_{2n})^2 + \frac{(N_2 - n_2)}{N_2} \quad \frac{\sigma_2^2}{n_2} \;, \qquad (101)$$

where σ_2^2 is the variance within the 'non-response' stratum. This gives

$$E\ (\bar{y}_{2r} - \bar{y}_{2n})^2 = \sigma_2^2 \left(\frac{1}{r_2} - \frac{1}{n_2} \right) = \sigma_2^2 \ \frac{(n_2 - r_2)}{n_2\ r_2} = \sigma_2^2 \ \frac{(k-1)}{n_2} \qquad (102)$$

from (93) and (101). Substitute in (98). Then

$$\begin{aligned} V(y_s) &= \frac{N(N-n)}{n}\ \sigma^2 + \frac{N^2}{n^2}\ (k-1)\ \sigma_2^2\ E\ (n_2) \\ &= \frac{N(N-n)}{n}\ \sigma^2 + \frac{N^2}{n^2}\ (k-1)\ \sigma_2^2\ \frac{nN_2}{N} \\ &= \frac{N(N-n)}{n}\ \sigma^2 + \frac{N\ N_2\ (k-1)}{n}\ \sigma_2^2 \ . \end{aligned} \qquad (103)$$

The first component is the variance that would be obtained if all n_2 in the non-response group were sampled: the second is the increase from sampling only r_2 of the n_2. The quantities n and k are then chosen to minimize (95) for a pre-assigned value of (103).

The solutions are:

$$k = \sqrt{\frac{c_2\ (\sigma^2 - W_2\ \sigma_2^2)}{\sigma_2^2\ (c_o + c_1\ W_1)}} \qquad (104)$$

$$n = \frac{N\ \left[\sigma^2 + W_2\ (k-1)\ \sigma_2^2\right]}{\frac{V}{N} + \sigma^2} \qquad (105)$$

where V is the value assigned to $V(y_s)$, the variance of the estimated population total. These formulae are identical with those given by Hansen and Hurwitz, though they appear on inspection to be slightly different. The difference arises because these authors use divisors N and N_2 respectively when defining σ^2 and σ_2^2, whereas we have used $(N-1)$ and (N_2-1).

The solutions depend on the unknown W_1 and W_2. If fairly close estimates of these can be made from earlier experience, the estimates may be used in place of the unknowns. Even if nothing is known in advance about W_1 and W_2, the authors develop an alternative method that gives in most cases a solution close to the optimum. Extensions to stratified sampling and to ratio estimation are also presented.

5.21 First example: This example is taken from the paper by Hansen and Hurwitz. They suppose that the first sample is taken by mail, and that the response rate is 50 percent. Further, the variance within the non-response group is the same as that within the whole population (this is unlikely to be exactly true in practice, but might serve as a first approximation). If these assumptions are made and if the f.p.c. is ignored, the variance of the estimated mean, from equation (103), simplifies to

$$N^2 = \sigma^2 (k + 1)/2n.$$

Thus all samples for which $(k + 1)/n$ have the same value will provide equal accuracy. As a standard of comparison, they choose an initial sample of size 1,000, in which all 500 non-respondents are later visited: that is, $n = 1{,}000$: $k = 1$. To obtain equal accuracy with other samples, we must have

$$k = \left(\frac{n}{500} - 1 \right) .$$

Such samples are shown in Table 8 for initial mailings of 2,000 and 5,000 schedules.

The cost in dollars was assumed to be of the form

$$C = (n + 4n_1 + 45r_2)/10.$$

These costs were obtained by assuming that the cost is 10 cents per questionnaire mailed, that the processing of a completed questionnaire costs 40 cents and that it costs $4.10 to carry through a field interview. The costs of the three samples described above are shown in Table 8.

TABLE 8.

SAMPLES OF DIFFERENT SIZES THAT LEAD TO SAME PRECISION OF RESULTS, THROUGH JOINT USE OF MAIL AND ENUMERATION METHODS ASSUMING A 50 PERCENT RESPONSE RATE

n	n_1	n_2	r_2	Schedules Tabulated	Cost
(1)	(2)	(3)	(4)	(5)	(6)
1,000	500	500	500	1,000	$2,550
2,000	1,000	1,000	333	1,333	2,099
5,000	2,500	2,500	278	2,778	2,751

n = Number of questionnaires mailed out

n_1 = Number of mail respondents

n_2 = Number of non-respondents to mail canvass

r_2 = Number of field interviews among the non-respondents

The middle sample is the cheapest: in the first sample there is too much sampling of the non-respondents, while in the third sample there is too little.

In this way we could determine the most economical sample by trying various combinations of n and k. Alternatively, by substitution in (104), we find that the optimum k value is $\sqrt{7.5}$, or 2.739. This gives $n = 500\ (3.739)$, or 1,870. Consequently, the optimum sample is such that 1,870 schedules are mailed initially. Of the 935 that are not returned, we enumerate by visitation 935/2.739, or 341. The cost will be found to be $2,096. It is evident that the middle of the three samples in Table 8 was very close to the optimum.

5.22 Second example: This is intended mainly to illustrate the type of bias that arises quite commonly in samples taken by mail: it is not an application of the Hansen-Hurwitz approach. The data come from an experimental sampling of fruit orchards in North Carolina, conducted in 1946. A list was available showing the number of fruit trees for each grower having more than 100 trees. The object of the sample was to obtain information about the number of peach trees and their production of peaches. (More accurately, the object was to devise and study methods for estimating such data by sampling).

A schedule was mailed to each member of the population. There was less than a 10 percent response. A second and a third mailing were sent out: these together raised the response to 41 percent. The returns to the three responses are summarized in Table 9. The principal points of interest are: (i) the steady decline in the number of fruit trees per grower in the successive responses, these being 456 at the first request, 382 at the second, 340 at the third, and 290 for the non-respondents. The larger operators tend to respond more easily: (ii) Both the second and third requests were substantially more successful than the first.

After the third request, a visitation survey, which will not be described in detail, was taken from the non-respondents. This survey was stratified according to the number of fruit trees per county in the non-respondent group and to the location of the county.

TABLE 9.

RESPONSE TO THREE REQUESTS OF A MAILED INQUIRY SENT TO GROWERS IN NORTH CAROLINA HAVING 100 OR MORE FRUIT TREES*

	No. of Growers	No. of Fruit Trees	Average No. of Fruit Trees per grower
Growers on the mailing list to whom schedules were sent.	3,241	1,064,899	329
Schedules returned unclaimed.	125	39,442	315
Remainder of Population	3,116	1,025,457	329
Response to first request	300	136,859	456
Response to second request	543	207,662	382
Response to third request	434	147,387	340
Total Response	1,277	491,908	385
Percent Total Response	41%	48%	
Total Non-Respondents	1,839	533,549	290
Percent Total Non-Respondents	59%	52%	

*Six counties of concentrated peach production were dealt with separately, i.e., by a complete enumeration.

REFERENCES

(11) Mahalanobis, P. C. "A Sample Survey of the Acreage Under Jute in Bengal" Sankhya, 4. pp. 511-530, 1940.

(12) Jessen, R. J. "Statistical Investigation of a Sample Survey for Obtaining Farm Facts" Iowa Agr. Exp. Sta. Res. Bull. 304, 1942.

(13) Stephan, F. F. "Stratification in Representative Sampling" Journal of Marketing, 6, pp. 38-46, 1941.

(14) Hagood, M. J. and Bernert, E. H. "Component Indices as a Basis for Stratification in Sampling" Journal American Statistical Association, 40, pp. 330-341, 1945.

(15) King, A. J. and McCarty, D. E. "Application of Sampling to Agricultural Statistics, with Emphasis on Stratified Samples" Journal of Marketing, 5, pp. 462-474, 1941.

(16) Johnson, F. A. "A Statistical Study of Sampling Methods for Tree Nursery Inventories" Journal of Forestry, 41, pp. 674-679, 1943.

(17) Jessen, R. J. and Houseman, E. E. "Statistical Investigations of Farm Sample Surveys taken in Iowa, Florida, and California" Iowa Agr. Exp. Sta. Res. Bull. 329, 1944.

(18) Cochran, W. G. "The Use of the Analysis of Variance in Enumeration by Sampling" Journal American Statistical Association, 34, pp. 492-510, 1939.

(19) King, A. J., McCarty, D. E., and McPeek, M. "An Objective Method of Sampling Wheat Fields to Estimate Production and Quality of Wheat" U. S. Dept. of Agr. Tech. Bull. 814, 1942.

(20) Stephan, F. F. "The Expected Value and Variance of the Reciprocal and Other Negative Powers of a Positive Bernoullian Variate" Ann. Math. Stat., 16, pp. 50-61, 1945.

(21) Hansen, M. H. and Hurwitz, W. N. "The Problem of Non-Response in Sample Surveys" Journal Amer. Stat. Asso., 41, pp. 517-529, 1946.

Systematic Sampling

6.1 We now consider a method of sampling, quite commonly used, that differs markedly from random sampling. Suppose that there are N = nk units in the population and that these are numbered. To select a sample of n units, we take a unit at random from the first k units; and every kth subsequent unit. For instance, if k is 15 and if the first unit drawn is number 13, the subsequent units are numbers 28, 43, 58, and so on. The selection of the first unit determines the whole sample. This type of sample will be called an "every kth" systematic sample.

The apparent advantages of this method over simple random sampling are:

(i) It is easier to draw and often easier to administer without mistakes. The saving in time of drawing may be quite large if slight departures are made from the strict " every kth" rule. For instance, if the units are described on cards which have <u>not</u> been numbered but which are all of the same size and lie in a file drawer, a card can be drawn out, say every inch along the file, as measured by a ruler. This operation is very speedy, whereas strict random sampling would be rather slow.

(ii) Intuitively it seems likely to be more accurate than random sampling. In effect, it stratifies the population into n strata, namely the first k units, the second k units and so on. We might therefore expect the systematic sample to be about as accurate as a stratified random sample with <u>one</u> unit per stratum. The difference is that with the systematic sample the units all occur at the same relative position in the stratum, while with the stratified random sample, position in the stratum is determined separately by randomization within each stratum. The systematic sample is spread more evenly over the population, and this fact has sometimes made it considerably more accurate than stratified random sampling.

In practice, one variant of the systematic sample is to choose each unit at or near the center of the stratum: the idea being that it will represent the stratum better than if it occurs near one end. No thorough investigation of the efficacy of this type of sampling appears to have been made, and attention will be confined to the case where the first unit in the sample is drawn at random from the first k in the population. The sampling theory was first developed by W. G. and L. H. Madow (22). It is rather more complex than might have been expected.

6.2 Sampling theory: For simplicity in presenting the theory, we assume that N is exactly equal to nk, where n is the size of sample to be taken and k is an integer. In practice N will be of the form (nk + r), where r is less than k. This will disturb slightly the results stated below in Theorems 10 and 11, which are not exactly true. The disturbance is probably negligible if n exceeds 50.

Theorem 10. The sample mean $\bar{y}_n$ is an unbiased estimate of the population mean $\bar{y}_p$.

Proof: This is rather obvious. Let the observations in the population be $y_1, y_2, \ldots y_{nk}$, and let

$$m_i = \left\{ y_i + y_{i+k} \ldots . + y_{i+(n-1)k} \right\} /n \quad . \qquad (106)$$

If y_i is the observation chosen when we draw the random number between 1 and k in order to start the sample, then m_i is the corresponding sample mean. Since every i between 1 and k is equally likely to be selected

$$E(m_i) = \left\{ m_1 + m_2 + \ldots + m_k \right\} /k \ .$$

From (106) this is clearly equal to $\bar{y}_p$.

<u>Variance of the estimate</u>. The variance may be expressed in a number of different ways. One form, due to the Madows (22) is given in Theorem 11.

<u>Theorem 11</u>. The variance of the mean of the systematic sample is

$$V\ (\bar{y}_n) = \frac{\sigma^2}{n} \left\{ \frac{N-1}{N} + \frac{2}{n} \sum_{d=1}^{n-1} (n-d) \rho'_{kd} \right\}$$

where ρ'_{kd} is the non-circular serial correlation coefficient for lag kd, defined by the equation

$$k(n-d)\ \sigma^2 \rho'_{kd} = \sum_{i=1}^{k(n-d)} (y_i - \bar{y}_p)\ (y_{i+kd} - \bar{y}_p)\ . \tag{107}$$

<u>Proof</u>: By definition,

$$V\ (\bar{y}_n) = E\ (\bar{y}_n - \bar{y}_p)^2 = \frac{1}{k} \sum_{i=1}^{k} (m_i - \bar{y}_p)^2 = \frac{1}{n^2 k} \sum_{i=1}^{k} (nm_i - n\bar{y}_p)^2$$

But $(nm_i - n\bar{y}_p) = (y_i - \bar{y}_p) + (y_{i+k} - \bar{y}_p) + \ldots + (y_{i+(n-1)k} - \bar{y}_p)$

When this is squared and added over all <u>k</u> values of <u>i</u>, the squared terms amount to

$$\sum_{i=1}^{N} (y_i - y_p)^2 = (N-1)\ \sigma^2\ .$$

The cross product terms will be seen to involve every pair of observations that differ by a multiple of <u>k</u>. These may be grouped according to the multiple of <u>k</u>. Thus there are k(n-1) products from observations that are k units apart: k(n-2) products from observations that are 2k units apart, and so on. Consequently

$$V(\bar{y}_n) = \frac{1}{n^2 k} \left\{ (N-1)\ \sigma^2 + 2 \sum_{i=1}^{k(n-1)} (y_i - \bar{y}_p)\ (y_{i+k} - \bar{y}_p) \right.$$

$$\left. + 2 \sum_{i=1}^{k(n-2)} (y_i - \bar{y}_p)\ (y_{i+2k} - \bar{y}_p) + \ldots + 2 \sum_{i=1}^{k} (y_i - \bar{y}_p)\ (y_{i+(n-1)k} - \bar{y}_p) \right\}$$

When we introduce the serial correlation coefficients as defined in (107), this becomes

$$V(\bar{y}_n) = \frac{1}{n^2 k} \left\{ (N-1)\,\sigma^2 + 2k \sum_{d=1}^{n-1} (n-d)\, \rho'_{kd}\, \sigma^2 \right\}$$

$$= \frac{\sigma^2}{n} \left\{ \frac{N-1}{N} + \frac{2}{n} \sum_{d=1}^{n-1} (n-d)\, \rho'_{kd} \right\} . \qquad (108)$$

Note: For a random sample of size n, the corresponding result would be

$$V(\bar{y}_n) = \frac{\sigma^2 (N-n)}{Nn} = \frac{\sigma^2 (k-1)}{n \,.\, k} .$$

Formula (108) shows that if the serial correlation coefficients are positive, the systematic sample is less accurate than the random sample. The formula also suggests that if the serial correlation coefficients are negative and sufficiently large, the systematic sample is likely to be more accurate. Since it is difficult to visualize what values the serial coefficients will take in a particular population, no simple general conclusions about the efficacy of systematic sampling can be drawn from the formula.

Theorem 12: This gives an alternative form for $V(\bar{y}_n)$ which is more suitable for comparisons with stratified samples.

$$V(\bar{y}_n) = \frac{N-n}{N} \frac{\sigma_w^2}{n} \left\{ 1 + \frac{2}{n} \sum_{d=1}^{n-1} (n-d)\, \rho_{(kd)w} \right\} \qquad (109)$$

where σ_w^2 is the "within-stratum" average variance, defined by

$$n(k-1)\, \sigma_w^2 = \sum_{i=1}^{n} (y_i - \bar{y}_{pi})^2$$

$\bar{y}_{pi}$ being the mean of the stratum to which y_i belongs. Further, $\rho_{(kd)w}$ is the "within-stratum" serial correlation coefficient for lag kd, defined by

$$(k-1)\ (n-d)\ \sigma_w^2\ \rho_{(kd)w} = \sum_{i=1}^{k(n-d)} (y_i - \bar{y}_{pi})\ (y_{i+kd} - \bar{y}_{p,\ i+kd}) .$$

Proof: This is similar to that of Theorem 11. Since

$$n\bar{y}_p = \bar{y}_{pi} + \bar{y}_{p,i+k} + \ldots + \bar{y}_{p,\ i+(n-1)k} ,$$

we have

$$(nm_i - n\bar{y}_p) = (y_i - \bar{y}_{pi}) + (y_{i+k} - \bar{y}_{p,\ i+k}) + \ldots \text{ (to } n \text{ terms)}.$$

The rest of the proof follows exactly the same method as in Theorem 11, and will be omitted.

Note: For a stratified random sample with one unit per stratum, the corresponding result is

$$V\ (\bar{y}_n) = \frac{(N-n)}{N}\ \frac{\sigma_w^2}{n}$$

Comparison with (109) shows that the systematic and stratified random samples will have equal accuracy if the lag correlations within strata are zero for all pairs of units that are a multiple of k apart.

6.3 Further comparison of systematic with random samples: As has been indicated, there are no simple general results about the accuracy of systematic samples relative to random and stratified random samples. Comparisons can be made for specific populations either by the preceding variance formulae or by direct methods. Several are given by the Madows (22). Two will be described briefly.

Linear trend: If the population consists solely of a linear trend, we may assume that $y_i = i$. Since

$$\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6} \ ; \quad \sum_{i=1}^{N} i = \frac{N(N+1)}{2} ,$$

the population variance σ^2 is given by

$$(N-1)\ \sigma^2 = \frac{N(N+1)\ (2N+1)}{6} - \frac{N^2(N+1)^2}{4N} = \frac{N(N^2-1)}{12} . \quad (110)$$

Hence the variance of the mean of a random sample of size n is

$$V_{ran} = \frac{(N-n)}{N} . \frac{\sigma^2}{n} = \frac{n(k-1)}{nk} \quad \frac{nk(N+1)}{12n} = \frac{(k-1)\ (N+1)}{12}$$

To find the variance within strata σ_w^2, we need only replace N by k in (110). This gives

$$V_{strat} = \frac{(N-n)}{N} . \frac{\sigma_w^2}{n} = \frac{n(k-1)}{nk} . \frac{k(k+1)}{12n} = \frac{(k^2-1)}{12n}$$

The variance for the systematic sample may easily be found directly. It is clear that the mean of the second systematic sample exceeds that of the first by 1, while the mean of the third exceeds that of the second by 1, and so on. Thus the means may be represented by the numbers 1, 2, 3, ... k. Hence

$$\Sigma\ (\bar{y}_n - \bar{y}_p)^2 = \frac{k(k^2-1)}{12} ,$$

by a further application of (110), with k for N. This gives

$$V_{sys} = \frac{(k^2-1)}{12} .$$

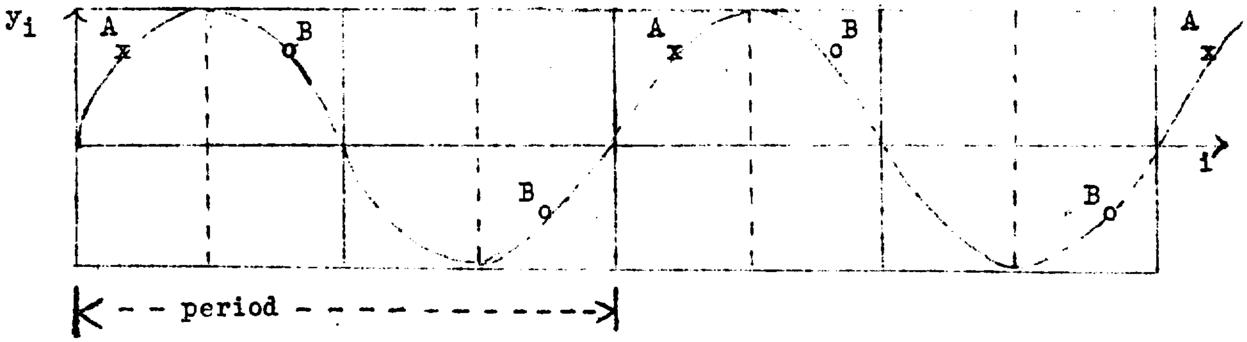
This result may be checked by applying the general formula (109) to this population. It will be found that $\rho_{(kd)w} = 1$, for all d.

From the formulae we deduce that

$$V_{strat} < V_{sys} < V_{ran} .$$

Thus for removing the effect of an unknown linear trend, the systematic sample is much more effective than the random sample, but less effective than the stratified random sample.

Periodic trend: If the population consists of a periodic trend, e.g., a simple sine curve, the effectiveness of the systematic sample depends on the value of k. This may be seen pictorially.



In this representation, the height of the curve is the observation y_i. The sample points A represent the case least favorable to the systematic sample. In this case k is equal to the period of the sine curve. Every observation within the systematic sample is exactly the same, so that the sample is no more accurate than an single observation taken at random from the population. This holds whenever k is any integral multiple of the period.

The most favorable case (sample B) occurs when k is an odd multiple of the half-period. Every systematic sample has a mean exactly equal to the true population mean, since successive deviations above and below the middle line cancel. The sampling variance of the mean is therefore zero. Between these two cases the sample has various degrees of effectiveness, depending on the relation between k and the period.

Natural populations: A few comparisons have also been made from natural populations. For instance, Johnson (16) studied 13 populations in which the observations were the numbers of seedlings in successive feet in a forest nursery bed. In seven beds containing seedbed stock of high variability, the variance of the mean of the systematic sample was only about half that for the stratified random sample: both were much more accurate than the simple random sample. The results for these beds appear in Table 10. In the remaining six beds, which had more homogeneous transplant stock, the systematic and the stratified sample were about equal in accuracy, both being again superior to the simple random sample. For estimating the areas under different types of cover (e.g., grass, woodland) from a map, Osborne (23) found the systematic sample twice to four times as accurate as the stratified sample. In these investigations the stratified sample had a stratum size <u>2k,</u> with 2 samples per stratum so as to permit estimation of the sampling error. The results would probably remain substantially the same if the stratum size were reduced to <u>k</u>. It may be anticipated that for populations where y_i shows 'continuous' variation--in the sense that observations near one another are likely to give similar results-the systematic sample will often be more effective than stratified random sampling. A theoretical investigation on this point has been made by Cochran (24). A useful elementary discussion of systematic samples, with application to part of Johnson's data, has been given by L. H. Madow (25).

TABLE 10.

VARIANCES OF SAMPLE MEAN NUMBERS OF SEEDLINGS
(F. A. Johnson's Data)

	Bed	V_{ran}	V_{strat}	V_{sys}	Estimate of V_{sys} by Method	
					(1)	(2)
Silver Maple	1	2.62	2.01	0.91	2.8	2.5
	2	3.26	2.19	0.74	3.6	2.9
American Elm	1	25.7	9.2	4.8	28.4	12.6
	2	20.8	15.8	15.5	22.6	18.6
White Spruce	1	13.4	11.9	5.5	17.2	11.2
	2	9.0	4.8	2.0	11.6	6.4
White Pine	1	19.4	16.8	8.2	21.0	21.9

6.4 *Estimation of the variance from a single sample*: Given the results of a single random sample, we can calculate an unbiased estimate of the variance of the sample mean, the estimate being unbiased *whatever the form of the population*. This useful property does not hold for the systematic sample. This may be seen by means of the 'sine curve' example. Let

$$y_i = m + a\sin (\pi\, i/2)$$

where $k = 4$ and $i = 1, 2, \ldots 4n$. The successive observations are

$$m + a,\ m,\ m - a,\ m,\ m + a,\ m,\ m - a,\ m,\ \ldots .$$

If $i = 1$ is chosen, *all* members of the systematic sample have the value $(m + a)$. For the other three possible choices of i, all members have the values *m*, $(m - a)$, or *m* respectively. Thus from a *single* sample we have no means of finding out or estimating the value of *a*, since we observe only $(m + a)$, *m*, or $(m - a)$. But the true sampling variance of the mean of the systematic sample is $a^2/2$.

Consequently, no reliable estimate of the standard error can be attached to a systematic sample, in the sense that this can be done for a random sample. What is usually done in practice is to make some assumption about the nature of the population, and to use a variance formula that will be reasonably unbiased if the assumption happens to be correct. For instance, if it is believed that the observations are ordered essentially at random, the variance formula for a random sample might be used. If it is believed that there will be differences among strata, but no serial correlation within strata, an estimate such as

$$\left(\frac{(N-n)}{Nn}\right) \sum_{i=1}^{n-1} (y_i - y_{i+k})^2/2(n-1) \qquad (111)$$

might be used. This estimate is likely to be positively biased, since it contains strata differences: it might not be far in error if differences between <u>neighboring</u> strata were small. To deal with the case where serial correlation was present, Osborne (23) used a more complex formula which seemed to work well for the type of natural population with which he was dealing. A type of formula appropriate to a population with an exponential correlogram has been given by Cochran (24), and an interesting general study of the problem by Matérn (26). All such methods are, of course, hazardous, and should be supplemented by detailed study, whenever possible, of the properties of the type of population that is being sampled.

The application of two formulae of this type to Johnson's data is shown in Table 10 (righ hand columns). Method (1) is the method given in formula (111), based on successive differences. It considerably overestimates the variance for the <u>stratified</u> sample and is scarcely within sight of the true variance for the systematic sample. Method 2 uses the estimate

$$\frac{(N-n)}{Nn} \sum_{i=1}^{n-2} (y_i - 2y_{i+k} + y_{i+2k})^2/6(n-2) .$$

This would be appropriate if the population contained a linear trend plus random deviations. However, it also fails in this case, where the population contained a quasi-continuous variation of a more complex type.

An alternative approach that is being investigated by Yates is to take supplementary observations along with the systematic sample. The extra observations will be used to obtain more information about the nature of the population and so to provide a more reliable estimate of variance. Results have not yet been published, though the method shows promise.

REFERENCES

(22) Madow, W. G. and L. H. "On the Theory of Systematic Sampling" Ann. Math. Stat. 15, pp. 1-24, 1944.

(16) Johnson, F. A. "A Statistical Study of Sampling Methods for Tree Nursery Inventories" Journal of Forestry, 41 pp. 674-679, 1943.

(23) Osborne, J. G. "Sampling Errors of Systematic and Random Surveys of Cover-type Areas" Jour. Amer. Stat. Asso., 37, pp. 256-264, 1942.

(24) Cochran, W. G. "Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations" Ann. Math. Stat., 17, pp. 164-177, 1946.

(25) Madow, L. H. "Systematic Sampling and Its Relation to Other Sampling Designs" Jour. Amer. Stat. Asso., 41, pp. 204-217, 1946.

(26) Matérn, B. "Methods of Estimating the Accuracy of Line and Sample Plot Surveys" Medd. fr. Statens Skogsforsknings Institut, Band 36, p. 1, 1947.

TYPE OF SAMPLING UNIT

7.1 Sometimes the population can be divided into units in various ways. For example, we might regard a city as composed either of a number of city blocks, or of a number of households, or of a number of persons. Similarly, in soil sampling, the tool with which the sample of soil is extracted can be constructed of various sizes and shapes, each of which determines a different subdivision of a field into units. A change in the type of sampling unit will usually affect both the cost of taking the sample and the accuracy. The determination of the optimum type of unit is therefore of importance from the point of view of reducing costs.

The optimum unit is that which gives the desired variance for the sample estimate at minimum cost. In order to compare two different units, we must find the size of sample needed with each unit, and the cost of taking this size of sample for each unit. It is quite often found that when a given percentage of the population is sampled, a large unit provides a less accurate estimate than a small unit. However, the sample tends to cost less with the large unit. The situation is not always so: Hansen and Hurwitz (27) have pointed out that for the estimation of the sex ratio, a household is roughly twice as accurate as a person (for a given percent sampled), because of the common presence of husband and wife in the same household.

7.2 A simple example: Johnson's data (28) for white pine seedlings provide a simple example. There were six rows in the bed (or population) and the rows were 434 feet long. The object in sampling is to estimate the total number of seedlings in the bed. Clearly there are many ways in which the bed can be divided into sampling units. The relevant data for four types of units are shown below.

TABLE 11.

DATA FOR FOUR TYPES OF SAMPLING UNITS

	Type of Unit			
	One foot row	Two-feet row	One foot bed	Two-feet bed
N_i = number of units in pop.	2,604	1,302	434	217
σ_i^2 = pop. variance per unit	2.537	6.746	23.094	68.558
Number of feet of row that can be counted in 15 mins.	44	62	78	108

The units were (i) one foot of a single row (ii) two feet of a single row. In both these cases it was assumed that the sample would be stratified by rows (one-sixth of the sample being taken from each row) so that the variances represent variances within rows. (iii) One foot of the complete width of the bed and (ii) two feet of the complete width of the bed. For these cases it was assumed that simple random samples would be taken.

Since the principal cost is that of locating and counting the units, costs were estimated by a time study (last row of Table 11). A larger bulk of sample can be counted in 15 minutes with the larger units, since less time is spent in moving from one unit to another.

The item to be estimated is the population total number of seedlings. In studies of this type, a population total is more convenient to discuss than a population mean, since the mean per s.u. for a two-feet bed unit is quite a different quantity from the mean per s.u. for a one foot row unit, whereas the population total has the same meaning for all units. If the f.p.c. is ignored, the variance of the estimated population total is

$$N_i^2 \frac{\sigma_i^2}{n_i}$$

where i = 1, 2, 3, 4 stands for the type of unit, n_i for the number of units in the sample and N_i for that in the population. We want this variance to be the same for all units. Thus if the smallest unit is chosen as a standard, the values of the other n_i that give the same accuracy as the smallest unit satisfy the equation

$$n_i = n_1 \left(\frac{N_i}{N_1} \right)^2 \frac{\sigma_i^2}{\sigma_1^2} \quad .$$

For example, the value of n_2 comparable to n_1 in this respect is

$$n_2 = n_1 . (\frac{1}{4}) . \frac{6.746}{2.537} = .665 \; n_1 \; .$$

These data are shown in Table 12, first line.

TABLE 12.

COMPARABLE SAMPLE SIZES AND COSTS

	Type of Unit			
	One foot row	Two feet row	One foot bed	Two feet bed
Comparable values of n_i	n_1	.665 n_1	.253 n_1	.188 n_1
Comparable sample sizes (in one-foot row units)	n_1	1.330 n_1	1.518 n_1	2.256 n_1
Comparable costs	c_1	.944 c_1	.856 c_1	.919 c_1
Relative net efficiency	100	106	117	109

The next step is to find the comparable sample sizes in terms of single feet of row, since the cost data are expressed in these terms. For n_2 we multiply the previous line by 2, because the unit contains two feet of row. These data appear in the second line of Table 12. It will be observed that as the size of the unit increases, the size of sample required to obtain equal accuracy also increases: in fact with the two-feet bed unit the sample must be 2 1/2 times as large as with the one-foot row unit.

The cost of taking n_1 of the smallest units may be expressed as

$$c_1 = n_1/44,$$

since this is the time required in 15-minute intervals. Similarly the cost with the second unit is

$$\frac{1.330\ n_1}{62} = 1.330 \times \frac{44}{62}\ c_1 = .944\ c_1 ,$$

as shown in the Table. All the larger units cost somewhat less than the smallest unit. If we define net efficiency as inversely proportional to cost, the relative net efficiencies are as given in the last line of the Table. From these data the one-foot bed width appears to be the best type of unit of those compared.

Note 1. For examples of this kind the comparable costs may be obtained directly without going through the intermediate steps. If z_i is the relative size of the i th unit to the smallest unit, the reader may verify that the costs for equal accuracy are proportional to

$$\sigma_i^2\ C_i\ /\ z_i$$

where C_i is the cost of taking a given bulk of sample with the i th unit. Thus to compare costs with the first and third units, we compare

$$\frac{2.537}{1 \times 44} = .0577 \quad \text{and} \quad \frac{23.094}{6 \times 78} = .0493 ,$$

since the one-foot bed is six times as large as the one-foot row.

Note 2. The previous example might be criticised on the grounds that whatever unit was chosen, the sample taken in practice would either be a stratified random sample or an 'every k th' systematic sample, whereas the comparisons assumed no stratification along the length of the bed. When comparing different types of unit, it is advisable to make the comparisons for the kind of sampling that is to be used: or if this has not been decided, for the kinds that are under consideration. A change in the method of sampling may change the relative costs of the different types of unit. A highly effective stratification, for instance, tends to make comparisons more favorable to the larger units, though the influence of stratification is not always in this direction. Some data on stratification as affecting the relative efficiency of large and small units are given for farm sampling by Jessen (17). In the same way, comparisons of type of unit will depend on the method of estimation that is used (see Section 9).

7.3 Comparisons from Sample Data: In the previous example the variances of the various sampling units were obtained from a complete census. When only sample data are taken, a slight change in the procedure is sometimes necessary. To illustrate, we consider a farm sample taken in North Carolina in 1942 in order to estimate farm employment. For details see reference (29). In effect, the method of drawing the sample was to locate points at random on the map, and to choose as sampling unit the three farms that were nearest to each point. Thus the sampling unit comprises a group of three neighboring farms. This method of selecting farms gives a large farm a greater chance of being included in the sample than a small farm, so that the average farm size in the sample tends to be biased upwards. Any effects of bias will be ignored in the present discussion.

The sample was stratified, the stratum being a group of townships that were similar in density of farm population and in ratio of cropland to farmland. Some data for the sample taken in May are shown below.

TABLE 13.

SIZES OF POPULATION AND SAMPLE

	Population	Sample
No. of strata	587	572
No. of sampling units	72,849	1,397
No. of farms	217,976	4,166

It will be noted that a few strata were not sampled: further, the number of farms per unit was very slightly under 3 (this discrepancy will be ignored). The sample was about 1.9 percent of the population.

From this sample we can compare the cluster of three farms that was actually used with the single farm. We shall not go into the cost aspects of the comparison, the purpose being to show how to estimate comparable variances. The first step is to compute an analysis of variance of the sample data, shown below for the number of paid workers.

TABLE 14.

ANALYSIS OF VARIANCE (NUMBER OF PAID WORKERS)

	d.f.	mean square
Between units within strata	825	6.218
Between farms within units	2,768	2.918
Total: Between farms within strata	3,593	3.676

This analysis is computed on a single-farm basis.

We wish to compare the accuracy of the population total number of paid workers as estimated by (i) a sample of n individual farms, (ii) a sample of n/3 clusters of 3 farms each. Each sample will be stratified into the strata that were used.

For (i) the variance of the estimated state total (ignoring f.p.c.) is $N^2\ \sigma_1^2/n$, where N is the number of farms in the state and

σ_1^2 is the variance between farms within strata. To estimate σ_1^2, it might at first be thought that we could use the mean square between farms within strata as found in the sample: that is, the mean square 3.676 as given in the last line of Table 14. However, the sample taken was not a random sample of farms within strata, but a random sample of groups of three farms. This fact causes the estimate to be biased.

An unbiased estimate of σ_1^2 may be obtained by making an analysis of variance, similar to Table 14, for the complete population.

TABLE 15.

ANALYSIS OF VARIANCE FOR THE COMPLETE POPULATION

	d.f.	Estimated mean square
Between units within strata	72,262	6.218
Between farms within units	145,127	2.918
Total: Between farms within strata	217,389	4.015

The degrees of freedom are obtained from the data in Table 13. The argument is that if we <u>had</u> analyzed the complete population, the mean square in the last line of the table would be the exact value for the variance between farms within strata. We do not know the population values for the mean squares between units within strata or between farms within units. But the figure 6.218 obtained from the sample is an unbiased estimate of the former, and the figure 2.918 is an unbiased estimate of the latter. Hence an unbiased estimate of the mean square σ_1^2 between farms within strata is

$$\frac{72,262 \text{ x } 6.218 + 145,127 \text{ x } 2.918}{217,369} = 4.015.$$

If σ_2^2 is the variance within strata for the 'three-farm' unit, the variance of the estimated state total will be

$$\left(\frac{N}{3}\right)^2 \quad \frac{\sigma_2^2}{n/3} = \frac{N^2}{n} \quad \frac{\sigma_2^2}{3} \quad ,$$

because the population contains only N/3 of these clusters, and the sample size is n/3. The figure 6.218 in the analysis of variance is an unbiased estimate of $\sigma_2^2/3$, since the mean square between the cluster totals has already been divided by 3 to transfer it to a single-farm basis. Consequently, for the same total size of sample, the comparable variances for the two units are

4.015 (single farm) and 6.218 (group of 3 farms).

Thus the sample size must be about 50 percent larger with the cluster unit than with single farms. Consideration of costs would make the result more favorable to the larger unit.

7.4 A Variance Function: Attempts have been made by various authors, notably Jessen (12) and Mahalanobis (5), to develop a general law which shows how the sampling error changes with the sige of unit. Suppose that the smallest unit is called an element, and that the large unit contains $\underline{M}$ neighboring elements. It has been found in several agricultural surveys that the variance $\underline{W}$ between elements within the large unit is related to $\underline{M}$ by means of the formula

$$W = AM^g , \qquad g > 0 , \tag{112}$$

where $\underline{A}$ and $\underline{g}$ are constants that do not depend on $\underline{M}$. In this representation $\underline{W}$ increases steadily as the size of the large unit increases, the curve being concave upwards. A curve of this type might be expected when there are forces that exert a similar influence on elements that are close together. Thus climate, soil type, topography, access to markets, and so on tend to make neighboring farms have similar features.

Note that the formula applies to the variance within the large unit and not to the sampling error for the large unit, the latter being derived from the variance *among* large units. We can derive a corresponding relation for the sampling error. Suppose that the population contains N elements, i.e., N/M large units. The following analysis of variance holds for the variation among elements in the population.

	d.f.	Mean square
Between large units	$\frac{N}{M} - 1$	B
Between elements within large units	$\frac{N}{M}\ (M - 1)$	W
Between elements in the population	$(N - 1)$	T

From this it follows that

$$\frac{(N-M)B}{M} = (N-1)\ T - \frac{N(M-1)W}{M}$$

Obviously the quantity T does not depend on M. Hence B is expressed as a function of M and of the three constants A, g, and T by the relation

$$B = \left\{M(N-1)\ T - N(M-1)\ AM^{g}\right\}/(N-M). \qquad (113)$$

The constants A, g, and M are estimated from the data. For this purpose we require (i) an estimate of the variance among elements in the complete population, so as to obtain T (ii) an estimate of the variance between elements within large units for at least *two* values of M, so as to obtain A and g. If the relation holds, we can then predict the value of B, and hence the sampling variance with the large unit, for any value of M.

Hendricks (30) has pointed out that the complete population might be regarded as a single large sampling unit containing N elements. If formula (113) holds, we may therefore put $T = AN^{g}$.

Substitution in (113) gives

$$B = AMN \left\{(N-1)N^{g-1} - (M-1)M^{g-1}\right\} / (N-M) . \qquad (114)$$

The formula now depends on only two constants, A and g. It can therefore be estimated from the variance among elements in the population, plus the variance within the large unit for one value of M. It may happen, however, that while (112) holds for small values of M, it fails to hold for the very large value $M = N$. In this event the more general formula (113) for B should be used. For applications of (114) to agricultural data, see Hendricks (30) and McVay (33).

7.5 A Cost Function: In connection with surveys where the elements are farms, and the larger units, or clusters, are groups of neighboring farms, Jessen (12) has developed a function that expresses the cost of taking the sample in terms of M. The discussion below presents a simplified form of this cost function.

We suppose that the sample contains n large units, each with M elements. Two components of cost are distinguished. The component c_1Mn consists of costs that vary directly with the total number of elements (farms): thus c_1 contains the cost of an interview and the cost of travel from farm to farm within the large unit.

The second component, $c_2\sqrt{n}$, measures the cost of travel between the areas. By tests on a map it was found that this cost, for a fixed population, varies with the square root of the number of sampling units. Total cost is therefore of the form

$$C\ (M,n) = c_1Mn + c_2\sqrt{n} . \qquad (115)$$

The best choice of M and n is that which minimizes the cost for a specified value of the variance of the estimate. If we are estimating the mean per farm for some item, the variance, ignoring f.p.c., will be B/Mn, since there are Mn farms in the sample and B is the variance between the units on a single-farm basis. Simple random sampling is assumed. Taking the more general form (113) for B, we have

$$V(M,n) \;=\; \frac{B}{Mn} \;=\left\{(N-1)\;T - N(M-1)AM^{g-1}\right\}/n(N-M). \qquad (116)$$

Since N is assumed very large this reduces to

$$V(M,n) \;=\; \left\{T - (M-1)AM^{g-1}\right\}/n. \qquad (117)$$

The algebraic solution is a little complex, though its application in a particular problem presents no great difficulty. We shall consider one aspect of the solution that leads to some interesting conclusions. We have to minimize

$$C \;+\; \lambda\, V$$

for a specified value of V. Since $\partial V/\partial n = -V/n$, the equations on differentiation with respect to n and M are

$$c_1 M \;+\; \frac{1}{2}\;c_2\,n^{-\frac{1}{2}} = \lambda\, V/n\;. \qquad (118)$$

$$c_1 n \qquad\qquad = -\lambda\,\partial V/\partial M\;. \qquad (119)$$

Dividing (119) by (118) so as to eliminate λ, we find that

$$\frac{M}{V}\;\frac{\partial V}{\partial M} \;=\; -\;\frac{1}{1 \;+\; \dfrac{c_2}{2c_1 M\sqrt{n}}} \qquad (120)$$

Now if equation (115) for the cost is solved as a quadratic in $\sqrt{n}$, it will be found, after some manipulation, that

$$\frac{2c_1 M\sqrt{n}}{c_2} = \left\{1 + \frac{4\ C\ c_1 M}{c_2^2}\right\}^{+\frac{1}{2}} - 1$$

Substituting in (112) we find

$$\frac{M}{V}\ \frac{\partial V}{\partial M} = \left\{1 + \frac{4\ C\ c_1 M}{c_2^2}\right\}^{-\frac{1}{2}} - 1 \qquad (121)$$

The important point about this equation is that it does not involve **n**, as may be verified from the form of V, equation (117). It is an equation from which we can solve directly for M. Further, the left hand side does not involve any of the cost factors. The right hand side involves M only in the combination $4\ C\ c_1\ M/c_2^2$. Hence if the variance function is unchanged but the cost factors vary, M will respond to these variations in such a way that the quantity $4\ C\ c_1\ M/c_2^2$ remains constant.

Now c_1 increases if the length of interview increases, while c_2 decreases if travel becomes cheaper, or if the farms in a given area become more dense. These facts lead to the conclusion that the optimum size of sampling unit becomes smaller if (i) the length of interview increases (ii) travel becomes cheaper (iii) the elements (farms) become more dense or (iv) the total amount of money used (C) increases. The conclusions are, of course, a consequence of the type of cost function that has been used and would require re-examination for a different type of cost function.

7.6 Cases where the Large Units Vary in Size: This happens in numerous surveys. A household, for example, contains differing numbers of individuals while an area of land, as used in farm surveys, will contain differing numbers of farms. If several specific sizes of unit are being compared, and if the variance has been estimated directly for each size, the methods of section 7.2 may be applied without change. The construction of a variance function requires a more elaborate analysis of variance to take account of variations in the M's. See (12) and (29).

The best method of estimating a population total also requires consideration. Suppose that the i th sampling unit has M_i elements and that the item total for the unit is y_i. The method considered thus far for estimating the population total is to calculate the mean per s.u., $\Sigma\ y_i/n$, and multiply by the number of s.u.'s in the population. If y_i is roughly proportional to M_i, as will often happen, this estimate may be rather poor, since its variance will depend on the variation in the M_i. An alternative is to calculate the mean per <u>element</u> $\Sigma\ y_i/\Sigma\ M_i$, and multiply by the number of elements in the population. This is frequently more accurate than the estimate based on the mean per s.u. . The sampling variance of this type of estimate is not covered by the formulae given previously in these notes, since both $\Sigma\ y_i$ and $\Sigma\ M_i$ will vary from sample to sample, so that the estimate involves the ratio of two random variates. Sampling variances for ratio estimates are given in Section 9.

7.7 <u>Possible Bias with Small Units</u>: It sometimes is found that small units give biased estimates, the bias arising from uncertainty about the boundaries of the unit. For example, Homeyer and Black (31) found that in sampling for the yield of oats, units 2' x 2' gave yields about 8 percent higher than units 3' x 3'. They express the opinion that the results for the larger unit are probably also biased upwards, because samplers tend to place boundary plants inside the unit when there is doubt. Sukhatme (32) gives similar comparisons in sampling for wheat and paddy.

REFERENCES

(27) Hansen, M. H. and Hurwitz, W. N. "Relative Efficiencies of Various Sampling Units in Population Inquiries" Jour. Amer. Stat. Asso., 37, pp. 89-94, 1942.

(28) Johnson, F. A. " A Statistical Study of Sampling Methods for Tree Nursery Inventories" Iowa State College M. S. Thesis, 1941.

(12) Jessen, R. J. "Statistical Investigation of a Sample Survey for Obtaining Farm Facts" Iowa Agr. Exp. Sta. Res. Bull. 304, 1942.

(17) Jessen, R. J. and Houseman, E. E. "Statistical Investigations of Farm Sample Surveys Taken in Iowa, Florida and California" Iowa Agr. Exp. Sta. Res. Bull. 329, 1944.

(29) Finkner, A. L., Morgan, J. J., and Monroe, R. J. "Methods of Estimating Farm Employment from Sample Data in North Carolina" N. C. Agr. Exp. Sta. Tech. Bull. 75, 1943.

(5) Mahalanobis, P. C. "On Large Scale Sample Surveys" Phil.Trans. Roy. Soc. London, B, 231, 1944.

(30) Hendricks, W. A. "The Relative Efficiencies of Groups of Farms as Sampling Units" Jour. Amer. Stat. Asso., 39, pp. 367-376, 1944.

(31) Homeyer, P. G. and Black, C. A. "Sampling Replicated Field Experiments on Oats for Yield Determinations" Soil Sci. Proc., 11, pp. 341-344, 1946.

(32) Sukhatme, P. V. "The Problem of Plot-size in Large-scale Yield Surveys" Jour. Amer. Stat. Asso., 42, pp. 297-310, 1947.

(33) McVay, F. E. "Sampling Methods Applied to Estimating Numbers of Commercial Orchards in a Commercial Peach Area" Jour. Amer. Stat. Asso., 42, pp. 533-540, 1947.

SUBSAMPLING

8.1 Suppose that the population is divided into N large sampling units, and that each of these contains M smaller units, which we will now call *sub-units*. If sub-units within the same unit give closely similar results, it may seem a waste of time to enumerate all M sub-units. Consequently, it is a common practice to enumerate only m of the M sub-units in each unit. In the presentation of the initial theory, these m will be assumed chosen at random from the M. This technique is called *subsampling*, since the sampling unit is not completely enumerated, but is itself sampled. For instance, in estimating the production of wheat in an area by sampling the standing crop when it is ripe, the field might be the sampling unit. It would not be feasible for a travelling crew to cut and thresh the whole of every wheat field that came into the sample. Instead, small areas of wheat (sub-units) are cut from each field. Studies have indicated that it is not economical to cut more than a small part of each field, so that in this case m/M is likely to be quite small. Similarly, in sampling the inhabitants of a town, a block may be the sampling unit, and a few persons or households selected from each block that comes into the sample.

Note. From another point of view, subsampling is the same thing as incomplete stratification. For we might regard the sub-unit as the sampling unit, and the unit as the stratum. The sampling technique is then such that only certain of the strata are sampled.

8.2 *Elementary theory*: We assume that the observation y_{ij} from the j th sub-unit of the i th unit is of the form

$$y_{ij} = \mu + b_i + w_{ij} \qquad (121)$$

where μ represents the population mean, b_i varies from unit to unit with mean zero and variance σ_b^2, and w_{ij} varies from sub-unit to sub-unit with mean zero and variance σ_w^2. All values of b_i, w_{ij}

are assumed mutually independent, and the number N of units in the population is assumed infinite. The units are chosen at random from the population, and the sub-units at random from the units.

From (121) it follows that the sample mean per sub-unit when m sub-units are taken from each of n units is

$$\bar{y}_{nm} = \mu + (b_1 + b_2 + \ldots + b_n)/n + (w_{11} + w_{12} + \ldots + w_{nm})/nm \qquad (122)$$

Hence,

$$V(\bar{y}_{nm}) = E(\bar{y}_{nm} - \mu)^2 = \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nm} \qquad (123)$$

Note that an increase in m diminishes only the contribution from the variance within units: an increase in n diminishes both components of the variance. For an estimate of the population total, we use $NM\bar{y}_{nm}$: the variance is then multiplied by $(NM)^2$.

8.3 Estimation of the variance: When a sample of this type has been taken, we may compute the following analysis of variance, on a sub-unit basis.

TABLE 16.

ANALYSIS OF VARIANCE WITH SUBSAMPLING

	d.f.	Mean square	Estimate of
Between sampling units	(n-1)	$B = m\,\Sigma(\bar{y}_{i.}-\bar{y}_{nm})^2/(n-1)$	$\sigma_w^2 + m\,\sigma_b^2$
Within units between sub-units	n(m-1)	$W = \Sigma(y_{ij}-\bar{y}_{i.})^2\,/n(m-1)$	σ_w^2

where $\bar{y}_{i.}$ is the mean of the m observations from the i th unit. It may be shown by algebra that the expected values of B and W are as shown in the right hand column of the Table,

Consequently from (123), an unbiased estimate of the variance of the sample mean $\bar{y}_{nm}$ is simply B/nm. The value of W is not required.

8.4 *Prediction of the variance for other subsampling rates:* From the analysis of variance in Table 16, we can also predict the variance of the sample mean for sampling and subsampling rates different from those actually used. This information may be useful in the planning of future samples on the same type of population.

Suppose that in the initial sample there were *m* sub-units sampled per unit and *n* units. We wish to estimate the variance of the sample mean under the supposition that these numbers were changed to *m'* and *n'* respectively. By (123), this variance is

$$V(\bar{y}_{n'm'}) = \frac{\sigma_b^2}{n'} + \frac{\sigma_w^2}{n'm'} \qquad (124)$$

From Table 16, unbiased estimates of σ_b^2 and σ_w^2 are

$$s_b^2 = (B-W)/m \quad ; \quad s_w^2 = W \quad .$$

Hence the estimated variance of the sample mean is

$$\frac{s_b^2}{n'} + \frac{s_w^2}{n'm'} = \frac{1}{n'}\left[\frac{B}{m} + W\left(\frac{1}{m'} - \frac{1}{m}\right)\right] \qquad (125)$$

Example: King and Jebe (34) report the following analysis of variance in sampling wheat fields in North Dakota, 1938. Two small samples were taken from each field, and the fields were stratified by districts.

TABLE 17.
ANALYSIS OF VARIANCE OF WHEAT YIELDS (BUSHEL PER ACRE) *

	d.f.	Mean squares
Between fields within districts	217	180
Within fields between subsampling units.	222	38

* Since the analysis presented by King and Jebe refers to a field *mean*, the mean squares have been multiplied by 2 to place it on a sub-unit basis.

The fields were not chosen at random, but by following routes designed so as to give good coverage of the area. Consequently, the mean square between fields may be a slight overestimate of the figure that would be obtained from a random sample of fields. For purposes of illustration, it will be assumed that the technique may be applied here. Further, effects of variation in field size are ignored.

We will consider how the variance of the sample mean is affected by (i) doubling the number of fields, with 2 subsamples per field; (ii) keeping the number of fields unchanged, but taking 4 subsamples per field; (iii) keeping the number of fields unchanged, but completely harvesting the fields.

If there are n fields in the original sample, the variance of the sample mean is 180/2n, or 90/n. By substitution in (125) the reader may verify that the corresponding variances for cases (i) and (ii) are

$$V_i = \frac{45}{n} \quad : \quad V_{ii} = \frac{80.5}{n} .$$

To solve case (iii), we have to assume that complete harvesting would be equivalent to taking all possible sub-units out of every field in the sample. Since the size of the sub-unit was very small compared to the size of a field, this implies that $m' = \infty$. The formula gives

$$V_{iii} = \frac{71}{n} .$$

The results illustrate the point that when there is an substantial variance between units, the variance of the sample mean cannot be decreased rapidly by increasing the subsampling rate: it is necessary to sample more units.

8.5. Application to field experiments: This type of theory may be applied to field experiments in cases where plot yields are