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, 1949
# Table of Contents

1. **Introduction**
   1.1. Advantages of the sampling method ........................................ 1
   1.2. General procedure in sampling .............................................. 1
   1.3. Scope of the course .......................................................... 5
   1.4. General principles ............................................................ 5
   1.5. Errors in sample surveys .................................................... 6
   References ..................................................................................... 7

2. **Basic Theory**
   2.1. General description .............................................................. 8
   2.2. Mean and variance ............................................................... 8
   2.3. Simple random sampling ....................................................... 8
   2.4. Expectation of \( \bar{y}_n \) ..................................................... 9
   2.5. Expectation of \( (\bar{y}_n)^2 \) ................................................... 9
   2.6. Variance of the mean of a random sample ............................... 10
   2.7. Estimation of the variance from sample data ........................... 11

3. **Confidence Limits and Estimation of Sample Size with Simple Random Sampling**
   3.1. Confidence limits ............................................................... 12
   3.2. Size of sample needed ........................................................... 12
   3.3. Case II: Variance guessed in advance ..................................... 12
   3.4. Example of Case I ............................................................... 14
   3.5. Case II: Variance estimated by preliminary sample .................. 14

4. **Sampling from "Binomial Type" Populations**
   4.1. General description ............................................................. 18
   4.2. Variance of binomial type population .................................... 18
   4.3. Variance of the sample proportion from a simple random sample .. 18
   4.4. Estimation of the variance of the sample proportion ............... 19
   4.5. Confidence limits for the sample proportion ........................... 19
   4.6. Confidence limits when normal theory does not apply ............... 20
   4.7. Estimation of sample size required ....................................... 21
   4.8. Extension to more than 2 classes in the population ................. 21
   4.9. Case II: Ratios involving all classes .................................... 22
   4.10. Case II: Ratios involving only some classes ......................... 22
   References .................................................................................... 24

5. **Stratified Random Sampling**
   5.1. General description ............................................................ 25
   5.2. Theory for stratified random sampling ................................... 25
   5.3. Optimum allocation .............................................................. 27
   5.4. Cost functions ....................................................................... 27
   5.5. Neyman's proof of optimum allocation ................................... 28
   5.6. The minimum variance ......................................................... 29
5.7. Simplification when f.p.c. is negligible ........................................... 29
5.8. Proportional sampling. ................................................................. 30
5.9. Optimum allocation when costs vary among strata. ......................... 30
5.10. Stratified random sampling from "binomial type" populations. ........ 31
5.11. Relative accuracy of stratified random and simple random sample. .... 33
5.12. Example illustrating relative accuracies ........................................... 35
5.13. Description of a sample survey ................................................... 37
5.14. Estimation from a sample of the gain due to stratification ............. 43
5.15. Confidence limits and sample size for stratified random sampling .... 48
5.16. Proximity as a basis for stratification ......................................... 50
5.17. Effects of errors in strata totals ............................................... 55
5.18. Case where the strata cannot be identified in advance ................. 54
5.19. Quota sampling ............................................................................. 56
5.20. The problem of non-response. ......................................................... 57
5.21. Example I ......................................................................................... 61
5.22. Example II ......................................................................................... 63
References .............................................................................................. 64-65

6. SYSTEMATIC SAMPLING. ................................................................. 66

6.1. Description and advantages ............................................................. 66
6.2. Sampling theory. ............................................................................... 67
6.3. Further comparison of systematic with random samples ................. 70
6.4. Estimation of the variance from a single sample. .......................... 74
References .............................................................................................. 76

7. TYPE OF SAMPLING UNIT. ............................................................... 77

7.1. General principle for the choice of unit ......................................... 77
7.2. Example. ........................................................................................... 77
7.3. Comparisons from sample data ....................................................... 81
7.4. A variance function. ......................................................................... 84
7.5. A cost function. ................................................................................ 86
7.6. Cases where the large units vary in size ........................................ 88
7.7. Possible bias with small units .......................................................... 89
References .............................................................................................. 90

8. SUB-SAMPLING ..................................................................................... 91

8.1. General description. .......................................................................... 91
8.2. Elementary theory. ............................................................................ 91
8.3. Estimation of the variance ............................................................... 92
8.4. Prediction of the variance for other sub-sampling rates ................... 93
8.5. Application to field experiments ...................................................... 94
8.6. Alternative mathematical model ...................................................... 97
8.7. The finite population correction ....................................................... 99
8.8. Estimation of the variance when the finite population correction is required ......................................................... 101
8.9. Stratified sampling of the units ....................................................... 103
8.10. Sub-sub-sampling ............................................................................ 104
8.11. Sub-sampling when units are unequal in size .................................. 106
8.12. Numerical example ........................................... 109
8.13. Selection with arbitrary probabilities ...................... 111
8.14. Application in practice ........................................ 113
8.15. Extension to stratified sampling, .............................. 115
        References ............................................. 117

9. OTHER METHODS OF ESTIMATION OF A POPULATION TOTAL ....... 118

9.1. Introduction .................................................. 118
9.2. The ratio estimate ............................................ 118
9.3. Estimation from a sample ..................................... 120
9.4. Comparison of the ratio estimate and the mean per sampling
       unit estimate ............................................ 121
9.5. Other applications of the ratio estimate ...................... 122
9.6. Ratio estimates in stratified sampling ........................ 124
9.7. Optimum allocation with a ratio estimate ..................... 126
9.8. The linear regression estimate ............................... 131
9.9. Variance of the estimate ..................................... 131
9.10. Comparison of regression estimate with ratio estimate and
       mean per sampling unit estimate .......................... 134
        References ............................................. 137

10. DOUBLE SAMPLING .................................................. 138

10.1. General description .......................................... 138
10.2. Case where the x variate is used for stratification .......... 139
10.3. Case where the x variate is used for regression ............. 143
        References ............................................. 146

11. ADDITIONAL NOTES ................................................ 147

11.1. Extension of the general principle ........................... 147
11.2. Aron sampling ............................................... 150
11.3. Control of human errors ...................................... 151
11.4. Descriptions of actual surveys ............................... 152
        References ............................................. 153
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. Jobe and A. L. Finkner, Resident Collaborators, Agricultural Estimates, Bureau of Agricultural Economics. Mr. Jobe 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 Beavis 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 purpose 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 comprise 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


Mahalanobis, P. C.

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:

\[
\bar{Y}_p = \frac{y_1 + y_2 + \ldots + y_N}{N} \tag{1}
\]

\[
\sigma^2 = \frac{\sum (y_i - \bar{Y}_p)^2}{N-1} \tag{2}
\]

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 \( \binom{N}{n} \) 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 \( \binom{N}{n} \) 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:

\[
\begin{align*}
\text{ABC} & \quad \text{ABD} & \quad \text{ABE} & \quad \text{ACD} & \quad \text{ADE} \\
\text{ACE} & \quad \text{BCD} & \quad \text{BCE} & \quad \text{BDE} & \quad \text{CDE}
\end{align*}
\]
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 \( E(\bar{y}_n) \) as the average over all the \( \frac{N!}{(N-n)!n!} \) 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.

**Theorem 2a:** \( E(\bar{y}_n) = \bar{y}_p \) (3)

\[
E(\bar{y}_n) = \frac{1}{n} E \left( y_1 + y_2 + \ldots + y_N \right)
\]

Since every unit appears in an equal number of samples, \( E(y_1 + \ldots + y_n) \) must be some multiple of \( E(y_1 + \ldots + y_N) \). 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 **Theorem 2b:**

\[
E \left( \bar{y}_n^2 \right) = \frac{1}{Nn} \frac{N-n}{N-1} \frac{N}{n} \sum_{i=1}^{N} y_i^2 + \frac{N(n-1)}{n(n-1)} \bar{y}_p^2
\] (4)

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

**Proof:**

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

\[
= \frac{y_1^2 + y_2^2 + \ldots + y_n^2}{n^2} + \frac{2}{n^2} \left( y_1 y_2 + \ldots + y_{n-1} y_n \right)
\]

By symmetry,

\[
E \left( y_1^2 \ldots + y_n^2 \right) = \frac{n}{N} \left( y_1^2 \ldots + y_N^2 \right)
\]
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,
\[
E (\overline{y}_n^2) = \frac{1}{nN} \sum_{i=1}^{N} y_i^2 + \frac{2(n-1)}{nN(N-1)} (y_1 y_2 + \ldots + y_{N-1} y_N).
\]

But
\[
2(y_1 y_2 + \ldots + y_{N-1} y_N) = y_1 (y_2 + \ldots + y_N) + y_2 (y_1 + y_3 + \ldots + y_N) + \ldots + y_N (y_1 + y_2 + \ldots + y_{N-1})
\]

\[
= y_1 (N\overline{y}_p - y_1) + y_2 (N\overline{y}_p - y_2) + \ldots + y_N (N\overline{y}_p - \overline{y}_N)
\]

\[
= N\overline{y}_p (y_1 + y_2 + \ldots + y_N) - y_1^2 - y_2^2 - \ldots - y_N^2
\]

\[
= N^2 \overline{y}_p^2 - \frac{N}{1} \overline{y}_1^2.
\]

Introducing this last reduction in \( E (\overline{y}_n^2) \), we obtain
\[
E (\overline{y}_n^2) = \frac{1}{nN} \left( 1 - \frac{n-1}{N-1} \right) \sum_{i=1}^{N} y_i^2 + \frac{N(n-1)}{nN(N-1)} \overline{y}_p^2
\]

\[
= \frac{1}{nN} \left( \frac{N-n}{N-1} \right) \sum_{i=1}^{N} y_i^2 + \frac{N(n-1)}{nN(N-1)} \overline{y}_p^2
\]

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

\[
E (\overline{y}_n - \overline{y}_p)^2 = \frac{N-n}{N} \frac{\sigma^2}{n}
\]

Proof: Expand the above, obtaining
\[
E (\overline{y}_n^2) - 2E \overline{y}_n \overline{y}_p + \overline{y}_p^2
\]

\[
= E (\overline{y}_n^2) - \overline{y}_p^2, \text{ by Theorem 1a.}
\]

Substitution from Theorem 1b gives
\[
\frac{1}{nN} \left( \frac{N-n}{N-1} \right) \sum_{i=1}^{N} y_i^2 + \left( \frac{N(n-1)}{nN(N-1)} - 1 \right) \overline{y}_p^2
\]

\[
= \frac{1}{nN} \left( \frac{N-n}{N-1} \right) \sum_{i=1}^{N} y_i^2 - \frac{N-n}{nN(N-1)} \overline{y}_p^2
\]
\[ -11 - \]

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

\[ = \frac{N-n}{N} \frac{\sigma^2}{n} \]

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), \( \frac{\sigma^2}{\bar{y}_n} \) depends primarily on \( n \), and not on \( \frac{n}{N} \). For instance, if \( \sigma^2 \) 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 \( \sigma^2 \) from the sample data.

\[ s^2 = \frac{\sum (y_i - \bar{y}_n)^2}{n-1} \]

is an unbiased estimate of \( \sigma^2 \). (6)

Proof:

\[ E(s^2) = \frac{1}{n-1} E \left( \sum y_i^2 - N \bar{y}_n^2 \right) \]

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

Combining the first two terms in brackets, this reduces to

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

\[ = \sigma^2 \]

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

\[ s_{\bar{y}_n} = \frac{\sqrt{(N-n)} \frac{s}{\sqrt{n}}}{\sqrt{N}} \]

(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

\[
\bar{y}_p = \bar{y}_n \pm t(\alpha, n-1) \sqrt{\frac{N-n}{N}} \frac{s}{\sqrt{n}}
\]

where \( t(\alpha, n-1) \) is the value of \( t \) corresponding to a significance level \( \alpha \), 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 \( \alpha \) (e.g., .05, .10, .20) and a margin of error \( d \) allowable in the sample mean. That is, we want

\[
P \left\{ \left| \bar{y}_n - \bar{y}_p \right| \geq d \right\} = \alpha
\]

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 \( \alpha \) 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 \( \sigma \) in the population. In Case 1, \( \sigma \) is estimated from previous sampling of a similar population, or simply by intelligent guesswork. Since the estimated \( \sigma \) is likely to be itself in error, we cannot expect more than a rough estimate of \( n \). If \( \sigma \) were to be correct, the value of \( d \) would be given by
\[ d = \left| \bar{y}_n - \bar{y}_p \right| = t(\alpha, \infty) \sqrt{\frac{n-n}{N}} \cdot \frac{\sigma}{\sqrt{n}} \]  

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

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

or

\[ n = \frac{N\sigma^2 t^2(\alpha, \infty) / d^2}{N + \frac{\sigma^2 t^2(\alpha, \infty)}{d^2}} = \frac{\sigma^2 t^2(\alpha, \infty) / d^2}{1 + \frac{1}{N} \cdot \sigma^2 t^2(\alpha, \infty) / d^2} \]  

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

\[ n_0 = \sigma^2 t^2 / d^2. \]

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

\[ n = \frac{n_0}{1 + \frac{n_0}{N}} \]

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 (6) 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 \( \sigma \) 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 \( d = (19) \times (0.1) \),

Then,

\[ n = \frac{25}{1 + 95/430} = 78. \]

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 \( \sigma \) may turn out to be wrong, and even if \( \sigma \) is correct, the \( \sigma \) that is found when the sample is taken will differ from \( \sigma \). 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 \( \sigma \) 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 \( \sigma \), 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)

\[
t(\alpha, n_1 - 1)s/\sqrt{\bar{y}_1}.
\]

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

\[
s^2 t^2(\alpha, n_1 - 1)/d^2
\]

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

\[
P\left( \left| \bar{y}_n - \bar{y}_p \right| \geq d \right) \leq \alpha.
\]

Sketch of proof. The proof assumes that the observations, \( \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n \), are normally distributed about \( \bar{y}_p \). Throughout the proof, \( d, \alpha \) and \( n_1 \) are assumed to be fixed quantities. The total sample size \( n \) is not fixed, but is a random variable, since its value depends on the value of \( \alpha \) that turns up in the first sample. Nevertheless, for fixed \( \alpha, n \) is fixed, and the quantity
\[ \sqrt{n} (\bar{y}_n - \bar{y}_p) \]
is normally distributed with mean zero and variance \( \sigma^2 \). Hence, this quantity follows the normal distribution whether \( \sigma \) 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

\[ P \left( \left| \frac{\sqrt{n} (\bar{y}_n - \bar{y}_p)}{s} \right| \geq t(\alpha, n_1 - 1) \right) = \alpha \quad (15) \]

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

\[ \sqrt{n} \geq s t(\alpha, n_1 - 1)/d, \text{ or } \sqrt{n}/s \geq t(\alpha, n_1 - 1)/d, \quad (16) \]

so that

\[ \left| \frac{\sqrt{n} (\bar{y}_n - \bar{y}_p)}{s} \right| \geq \left| \frac{t(\bar{y}_n - \bar{y}_p)/d}{s} \right| \]

Hence, from (15)

\[ P \left( \left| \frac{t(\bar{y}_n - \bar{y}_p)/d}{s} \right| > t \right) \leq \alpha \]

I.e.,

\[ P \left( \left| \frac{\bar{y}_n - \bar{y}_p}{s} \right| \geq d \right) \leq \alpha . \]

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 \( d = 10, \alpha = 0.05 \). From previous information, \( \sigma \) is guessed as about 50 (though this guess may be seriously in error). With this value of \( \sigma \), 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

\[
n \geq t^2 s^2 / d^2 = (4.040) (1,938) / 100 = 78.3
\]

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

\[
\frac{t^2 s^2}{d^2} = \frac{1 + \frac{1}{N} \cdot \frac{t^2 s^2}{d^2}}{N}
\]
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,

\[
\bar{y}_n = \frac{\text{Number in Sample in A}}{n} = p_n,
\]

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:

\[
s^2 = \frac{N}{N-1} pq \quad \text{where} \quad q = 1-p.
\]  \hspace{1cm} (17)

Proof: By definition,

\[
s^2 = \frac{1}{N-1} \left( \sum y_i^2 - N \bar{y}_p^2 \right) \quad (i = 1, 2 \ldots N).
\]

\[
= \frac{1}{N-1} \left( Np - N \bar{y}_p^2 \right) = \frac{N}{N-1} pq \quad \text{from the coding}
\]

and definition of \( p \) given in (4.1).

4.3 Theorem 5. Variance of the sample proportion from a simple random sample is

\[
\frac{N-n}{N-1} \frac{p \cdot q}{n} \hspace{1cm} (18)
\]

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,

\[ V(\hat{y}_n) = \frac{N-n}{N} \frac{1}{n} \frac{N}{N-1} p \ q \]

\[ = \frac{N-n}{N-1} \frac{p \ q}{n} \]

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

\[ V(p_n) = \frac{N-n}{(N-1)n} \ p_n \ q_n \quad \text{where} \quad q_n = 1 - p_n \]  \hspace{1cm} (19)

It is to be noted, however, that \( E(p_n q_n) = \frac{N(n-1)}{n(N-1)} \ p q \)

Therefore, an unbiased estimate of

\[ V(p_n) = \frac{N-n}{(N-1)n} \frac{n(N-1)}{N(n-1)} \ p_n \ q_n = \frac{N-n}{N(n-1)} \ p_n \ q_n \]  \hspace{1cm} (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

\[ p = p_n \pm t(\alpha) \frac{\sqrt{N-n}}{\sqrt{N-1}} \frac{\sqrt{p \ q}}{\sqrt{n}} \]  \hspace{1cm} (21)

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

\[ p_n \pm t(\alpha) \frac{\sqrt{N-n}}{\sqrt{N-1}} \frac{\sqrt{P_n \ q_n}}{\sqrt{n}} \quad \text{where} \quad t(\alpha) \ \text{is taken with} \ \sigma \ \text{degrees of freedom} \]  \hspace{1cm} (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 K. 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

$$p = \frac{p_n + k \pm k/\sqrt{1 + 2p_n q_n/k}}{1 + 2k}$$

where $k = t^2(\alpha)/2n$ (23) and $q_n = 1 - p_n$.

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
--- | ---
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

\[
n = \frac{t^2 \frac{p_n q_n}{d^2}}{1 - \frac{1}{n} \left(1 - \frac{t^2 \frac{p_n q_n}{d^2}}{d^2}\right)}
\]  

(24)

When the finite population correction can be ignored the solution becomes simply \( n = t^2 \frac{p_n q_n}{d^2} \). 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:
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

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

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

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,

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

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

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

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 any group with fixed $n'$, it is easy to see that an unbiased estimate of say $A/(A + B + C)$, will be obtained.

Hence,

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

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

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

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

$$d = p_{n'} \pm t(\alpha) \sqrt{\frac{N' - n'}{N' - 1} \frac{p \cdot q}{n'}}$$ (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
\[ p = p_n' \pm t_{\sqrt{pq/n'}}. \]  

(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

\[ p = p_n' \pm t_{\sqrt{\frac{N-n}{N-1}} \frac{pq}{n'}}. \]  

(27b)

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

REFERENCES

(6) Bartlett, M. S.


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

Biometrika 26:404 (1934)

(8) Simon, L. E.


(9) Snedecor, G. W.


(10) Statistical Research Group

Columbia University

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,

\[
\bar{y}_p = \frac{1}{N} \sum_{j=1}^{k} N_j \bar{y}_{pj}
\]  
(28)

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

\[
\bar{y}_n = \frac{1}{N} \sum_{j=1}^{k} N_j \bar{y}_{nj}
\]  
(29)
In (29), we note that the equation 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),

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

(30)

where

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

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

$$\bar{y}_n - \bar{y}_p = \frac{1}{N} \sum_{j=1}^{k} N_j \left( \bar{y}_{nj} - \bar{y}_{pj} \right).$$

Then,

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

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

(31)

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

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

(32)

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

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

(33)

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

$$V(\bar{y}_n) = \frac{1}{N^2} \sum_{j=1}^{k} N_j \frac{(N_j-n_j)}{N_j-1} \sigma_j^2/n_j,$$ the result as stated in (30).
When \( n_j/N_j \) is negligible, (30) may be reduced to \( V(\bar{y}_n) = \frac{1}{n^2} \sum_j (N_j^2 c_j^2)/n_j \) \( (34) \)

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

where \( s_j^2 = \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{nj})^2 / (n_j - 1) \) = 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 \( n_1, n_2, \ldots, n_k \).

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.

\[ C = a + \sum_j b_j \sqrt{n_j} + \sum_j c_j n_j \]

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

\[
C = a + c \left( n_1 + n_2 + \ldots + n_k \right) = a + c \sum n_j.
\] (37)

Now \( \sum n_j = n, \) 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 \( \sum c_j n_j, \) i.e., \( c_j, \) the cost per schedule, varies from stratum to stratum. Then we have

\[
C = c_1 n_1 + c_2 n_2 + \ldots + c_k n_k.
\] (38)

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

5.5 Theorem 7: (Refer J. Neyman, Journal of the Royal Statistical Society, 97 (1936) 558–606). In stratified random sampling, \( \nu(\bar{y}_n) \) is smallest for a fixed total size of sample if the sample is distributed with \( n_j \) proportional to \( N_j \sigma_j. \)

Proof: Using the Lagrangian multiplier we have

\[
\nu(\bar{y}_n) + \lambda C = \frac{1}{N^2} \sum_{j=1}^{k} N_j \left( \frac{N_j}{n_j} - 1 \right) \sigma_j^2 + \lambda \left( \sum n_j \right)
\]

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

\[
n_j = \frac{N_j \sigma_j}{N \sqrt{\lambda}} \quad \text{or} \quad n_j \text{ is proportional to } N_j \sigma_j.
\]

By summing this result for \( n_j \) in both members we can simplify the result since

\[
\sum n_j = n = \frac{\sum N_j \sigma_j}{N \sqrt{\lambda}}
\]
Substituting for \( \lambda \) we find the actual value of \( n_j \) to be

\[
n_j = n \frac{N_j \sigma_j}{\sum N_j \sigma_j^2}
\]  

(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 \( \sigma_j \) 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, Case I: Now let us re-write the variance from (30) as

\[
V(\bar{y}_n) = \frac{1}{N^2} \sum \frac{N_j^2}{n_j} - N_j \sigma_j^2
\]

\[
= \frac{1}{N^2} \sum \frac{N_j^2 \sigma_j^2}{n_j} - \frac{1}{N^2} \sum N_j \sigma_j^2
\]  

(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,

\[
V(\bar{y}_n)_{\text{min.}} = \frac{1}{N^2} \frac{(\sum N_j \sigma_j)^2}{n} - \frac{1}{N^2} \sum N_j \sigma_j^2.
\]  

(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

\[
V(\bar{y}_n)_{\text{min.}} = \frac{(\sum N_j \sigma_j)^2}{N^2 n}
\]  

(42).
Hence, the minimum standard error can be expressed as

$$s(\bar{y}_n)_{\text{min}} = \frac{1}{\sqrt{N}} \sum_j \frac{N_j \sigma_j}{N}$$  \hspace{1cm} (42a)

5.8 Corollary 2 to Theorem 7. Proportional Sampling. 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

$$\frac{n_j}{N_j} = \frac{n}{\sum N_j} = \frac{n}{N} = \text{a constant.}$$  \hspace{1cm} (43)

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

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

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

5.9 Theorem 8: Case II of Optimum Allocation: Under the assumptions of (36), 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 \( N_j \frac{\sigma_j}{\sqrt{c_j}} \).

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

$$V(\bar{y}_n) + \lambda C = \frac{1}{N^2} \sum N_j (n_j/n_j - 1) c_j^2 + \lambda \left( \sum c_j n_j \right)$$  \hspace{1cm} (44)

Differentiating and equating the result to zero we find

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

Then \( n_j \sqrt{\lambda} = \frac{N_j}{\sigma_j} \).

Summing again in both members and substituting the result obtained for
\begin{equation}
\lambda_n \text{ we obtain}
\begin{align*}
n_j &= \frac{n(N_j \sigma_j/\sqrt{\sigma_j})}{\Sigma(N_j \sigma_j/\sqrt{\sigma_j})} \\
\end{align*}
\tag{45}
\end{equation}

From the result for Case II, i.e., the variance is a minimum when \( n_j \) is proportional to \( N_j \sigma_j/\sqrt{\sigma_j} \), 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 Stratified Random Sampling from "Binomial Type" Populations:

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 \( \varepsilon_j \) of the \( n_j \) fall in Class I. Then for the estimated population proportion in Class I, we have

\begin{equation}
\hat{p}_n = \sum_{j=1}^{k} \frac{N_j}{N} \frac{\varepsilon_j}{n_j} \\
\end{equation}
\tag{46}

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

We had

\begin{equation}
\text{Var}(\hat{p}_n) = \frac{1}{N^2} \Sigma N_j (N_j-n_j) \sigma_j^2/n_j. \text{ By Theorem 4, } \sigma_j^2 = \frac{N_j}{N_j-1} p_j q_j. \\
\end{equation}
Hence, with \( p_n \) defined as in (46),

\[
V(p_n) = \frac{1}{N_j^2} \sum \frac{N_j (N_j - n_j)}{n_j} \frac{N_j}{N_j - 1} p_j q_j
\]  \hspace{1cm} (47)

When the finite population correction can be ignored, we obtain

\[
V(p_n) = \frac{1}{N_j^2} \sum N_j^2 p_j q_j / n_j
\]  \hspace{1cm} (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 \( \Sigma N_j = \) constant, \( n_j \) is proportional to \( N_j \sigma_j \).

Thus

\[
n_j = N_j \frac{\sqrt{N_j}}{\sqrt{N_j - 1}} \frac{\sqrt{p_j q_j}}{\sqrt{c_j}}
\]  \hspace{1cm} (48)

Case II: Here \( \Sigma c_j n_j = \) constant, and \( n_j \) is proportional to \( N_j \sigma_j / c_j \).

Then we have

\[
n_j = N_j \frac{\sqrt{N_j}}{\sqrt{N_j - 1}} \frac{\sqrt{p_j q_j}}{\sqrt{c_j}}
\]  \hspace{1cm} (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

$$V_{strat.} = \frac{1}{N^2} \Sigma \frac{N_j^2 \sigma_j^2}{n_j}$$

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

this reduces to

$$V_{opt.} = \frac{(\Sigma N_j \sigma_j)^2}{nN^2}$$

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

$$V_{prop.} = \frac{\Sigma n_j \sigma_j^2}{nN}$$

Now

$$V_{prop.} - V_{opt.} = \frac{1}{nN} \left\{ \Sigma n_j \sigma_j^2 - \frac{(\Sigma N_j \sigma_j)^2}{N} \right\}$$

$$= \frac{1}{nN} \Sigma n_j (\sigma_j - \bar{\sigma})^2$$

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

- - -
This result shows that \( V_{\text{opt.}} \) will always be smaller than \( V_{\text{prop.}} \). The size of the difference depends on the amount of variation in the \( \sigma_j \).

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

\[
V_{\text{ran.}} = \frac{\sigma^2}{n}
\]

where \( \sigma^2 \) is the variance of the whole population. But from an algebraic identity,

\[
(N-1) \sigma^2 = \sum (N_j - 1) \sigma_j^2 + \sum N_j (\bar{y}_{pj} - \bar{y}_p)^2 \tag{54}
\]

and since terms in \( 1/N_j \) are negligible, this may be written

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

Hence,

\[
V_{\text{ran.}} = \frac{\sigma^2}{n} = \frac{\sum N_j \sigma_j^2}{nnN} + \frac{\sum N_j (\bar{y}_{pj} - \bar{y}_p)^2}{nN}
\]

\[
= V_{\text{prop.}} + \frac{\sum N_j (\bar{y}_{pj} - \bar{y}_p)^2}{nN} \tag{55}
\]

\[
= V_{\text{opt.}} + \frac{\sum N_j (\sigma_j - \bar{y})^2}{nN} + \frac{\sum N_j (\bar{y}_{pj} - \bar{y}_p)^2}{nN} \tag{56}
\]

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

\[
V_{\text{opt.}} \leq V_{\text{ran.}}
\]
Provided that

\[ \sum N_j (\bar{y}_{pj} - \bar{y}_p)^2 \geq \frac{\sum (N-N_j) \sigma_j^2}{N} \]  

(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**

<table>
<thead>
<tr>
<th>Stratum No.</th>
<th>Farm Size</th>
<th>N_j Acres</th>
<th>Corn Acres</th>
<th>Corn Stratum Total</th>
<th>Prop. Sampling</th>
<th>Optimum Allocation N_j \sigma_j^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>1</td>
<td>0-40</td>
<td>394</td>
<td>5.4</td>
<td>2127</td>
<td>8.3</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>41-80</td>
<td>461</td>
<td>16.3</td>
<td>7492</td>
<td>13.3</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>81-120</td>
<td>391</td>
<td>24.3</td>
<td>9515</td>
<td>15.1</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>121-160</td>
<td>334</td>
<td>34.5</td>
<td>11524</td>
<td>19.8</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>161-200</td>
<td>169</td>
<td>42.1</td>
<td>7110</td>
<td>24.5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>201-240</td>
<td>113</td>
<td>50.1</td>
<td>5651</td>
<td>26.0</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>241-</td>
<td>148</td>
<td>63.8</td>
<td>9438</td>
<td>35.2</td>
<td>7</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>2,010 \bar{y}_p = 26.3</strong></td>
<td></td>
<td><strong>52,857</strong></td>
</tr>
</tbody>
</table>

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 \( V \, \text{ran.} = \frac{\sigma^2}{n} \). In order to obtain \( \sigma^2 \), we may apply

\[
N \sigma^2 = \sum N_j \sigma_j^2 + \sum N_j (\bar{y}_{pj} - \bar{y}_p)^2
\]

(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, \( \sum 5.4 (2127) + \ldots + 63.8 (9438) \), and subtracting a correction term \( (52887)^2/2010 \), which gives 557,007.1. Summing the two terms, 689,981.1 + 557,007.1 = 1,246,988.2 = \( N \sigma^2 \), and dividing this result by \( Nn \), we obtain \( V \, \text{ran.} = 6.20 \). The standard error is then \( \text{S.E.} (\bar{y}_n) \, \text{ran.} = \sqrt{6.20} = 2.49 \), 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 \( \text{S.E.} (\bar{y}_n) \, \text{prop.} = \sqrt{3.43} = 1.85 \)

C. V. = 7.0%

**Optimum Allocation:**

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

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

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

C. V. = 6.6%
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 \( n = \frac{3.43}{2.90} \times 100 = 118.3 \). 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 tires. Also, they found (3) that some dealers primarily handle used tires. These data led to the following classification of the population.

**TABLE 2.**

**STRATIFICATION OF TIRE DEALERS FOR MARCH 1945 CPA SURVEY**

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

*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 $\sigma$'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 - CPA TIRE DEALERS SURVEY

<table>
<thead>
<tr>
<th>Size in Number of Tires on Hand</th>
<th>Mean $\bar{Y}_{pj}^{**}$</th>
<th>Std. Dev. $\sigma_j$</th>
<th>Ratio $\sigma_j / \bar{Y}_{pj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group A</strong>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-9</td>
<td>14.8</td>
<td>18.2</td>
<td>1.23</td>
</tr>
<tr>
<td>10-19</td>
<td>21.0</td>
<td>26.3</td>
<td>1.25</td>
</tr>
<tr>
<td>20-29</td>
<td>34.2**</td>
<td>40.6</td>
<td>1.19</td>
</tr>
<tr>
<td>30-39</td>
<td>34.2</td>
<td>28.2</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>avg.</strong></td>
<td></td>
<td></td>
<td>1.25</td>
</tr>
<tr>
<td><strong>Group B</strong>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1.0</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>1-9</td>
<td>6.7</td>
<td>8.2</td>
<td>1.22</td>
</tr>
<tr>
<td>10-19</td>
<td>13.0</td>
<td>9.9</td>
<td>.76</td>
</tr>
<tr>
<td>20-29</td>
<td>24.7</td>
<td>11.4</td>
<td>.46</td>
</tr>
<tr>
<td>30-39</td>
<td>32.0</td>
<td>12.4</td>
<td>.39</td>
</tr>
<tr>
<td><strong>avg.</strong></td>
<td></td>
<td></td>
<td>.75</td>
</tr>
</tbody>
</table>

*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,
\[
V(\bar{y}_n) = \frac{1}{N^2} \sum \frac{n_j^2 \sigma_j^2}{n_j} \quad \text{(omitting the f.p.c.)}
\]
Substituting for \( n_j \), we obtain
\[
V(\bar{y}_n) = \frac{k}{N^2} \sum N_j \sigma_j \cdot
\]
In this survey the estimate wanted was the total number of new tires on hand. We write this estimate as \( T_n = N \bar{y}_n \). Hence, \( V(T_n) = k \sum N_j \sigma_j \). At this stage we introduce from the preceding paragraph the assumptions on the \( \sigma \)'s for Group A and Group B, and write
\[
V(T_n) = k \left( \sum_{A} 2 N_j \bar{y}_{pj} + \sum_{B} N_j \bar{y}_{pj} \right) \quad \text{where the summations are over the strata in Groups A and B, respectively.}
\]
\[= k \left( 2 T_A + T_B \right).\]
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
\[
C.V. (T_n) = \frac{\sqrt{k} \sqrt{2 T_A + T_B}}{T_A + T_B}.
\]
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. Denning and Simmons estimated

\[ T_A \text{ at } 1.6 \times 10^6 \]

and \[ T_B \text{ at } 0.2 \times 10^6 \].

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

\[ k = \frac{(0.015)^2 (1.8)^2 \times 10^{12}}{3.4 \times 10^6} = 214. \]

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 = \(\sigma_j/k = 2 \tilde{y}_{pj}/200\). From this relation we obtain the percent sampled in the strata of Group A = \(\tilde{y}_{pj}\). Similarly, the percent sampled in the strata of Group B = \(\tilde{y}_{pj}/2\). An estimate of the \(\tilde{y}_{pj}\) 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, \(\tilde{y}_{pj}\), 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 - CPA TIRE DEALERS SURVEY

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

*(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 3rd 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: \( \tilde{y}_p > 11.9 \times (470/997) = 6 \) from which \( \sigma \) was estimated as 3 \( \tilde{y}_p \) 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 \( \sigma = 2 \tilde{y}_p \) might be a reasonable assumption. Again, the application of \( n_j/N_j = \sigma_j/k \) gave \( 2(2.3)/200 = 2.3\% \) 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 \( \bar{N}_j \), \( n_j \), \( \bar{Y}_{nj} \), \( s_j^2 \) (estimate of the within-stratum variance \( c_j^2 \)). With the f.p.c. ignored, the estimated variance of the mean of the stratified sample is

\[
\text{Estd. } V \text{ strat.} = \frac{1}{N^2} \sum \frac{N_j^2 s_j^2}{n_j} = \sum \frac{W_j^2 s_j^2}{n_j}
\]

(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

\[
V \text{ ran.} = \frac{1}{n} \left[ \frac{\sum (N_j - 1) c_j^2 + \sum N_j (\bar{Y}_{pj} - \bar{Y}_p)^2}{(N - 1)} \right]
\]

(59)

Since terms in \( 1/N_j \) are negligible, this may be simplified to

\[
V \text{ ran.} = \frac{1}{n} \left[ \sum W_j c_j^2 + \sum W_j (\bar{Y}_{pj} - \bar{Y}_p)^2 \right]
\]

(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

\[ \bar{Y}_{nj} = \bar{Y}_{pj} + \epsilon_{nj} \]

where \( \epsilon_{nj} \) is an error of sampling with \( V (\epsilon_{nj}) = \frac{\sigma_j^2}{n_j} \)

Hence

\[ E (\bar{Y}_{nj}^2) = \bar{Y}_{pj}^2 + \frac{\sigma_j^2}{n_j} \]

(61)

Thus

\[ E \sum (W_j \bar{Y}_{nj}^2) = \sum W_j \bar{Y}_{pj}^2 + \sum \frac{W_j \sigma_j^2}{n_j} \]

(62)

Also

\[ \sum W_j \bar{Y}_{nj} = \sum W_j \bar{Y}_{pj} + \sum W_j \epsilon_{nj} \]

(63)

Hence

\[ E \left\{ \sum W_j \bar{Y}_{nj} \right\}^2 = (\sum W_j \bar{Y}_{pj})^2 + \sum \frac{W_j^2 \sigma_j^2}{n_j} \]

(64)
Subtract (64) from (62).

\[
\mathbb{E} \left[ \sum W_j \bar{y}_{nj}^2 - (\sum W_j \bar{y}_{nj})^2 \right] = \sum W_j \bar{y}_{pj}^2 - (\sum W_j \bar{y}_{pj})^2 + \sum \frac{W_j \sigma_j^2}{n_j} - \sum \frac{W_j^2 s_j^2}{n_j} \tag{65}
\]

It follows that an unbiased estimate of

\[
\sum W_j (\bar{y}_{pj} - \bar{y}_p)^2
\]

is given by

\[
Q = \sum W_j \bar{y}_{nj}^2 - (\sum W_j \bar{y}_{nj})^2 - \sum \frac{W_j s_j^2}{n_j} + \sum \frac{W_j^2 s_j^2}{n_j} \tag{66}
\]

Finally

\[
\text{Zstd. V ran.} = \frac{1}{n} \left[ \sum W_j s_j^2 + Q \right] \tag{67}
\]

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 - CPA TIRE DEALERS SURVEY**

<table>
<thead>
<tr>
<th>Size of Stratum</th>
<th>N_j</th>
<th>n_j</th>
<th>\bar{y}_{nj}</th>
<th>s_j^2</th>
<th>W_j</th>
<th>W_j s_j^2 / n_j</th>
<th>W_j s_j^2 / n_j</th>
<th>W_j \bar{y}_{nj}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-9</td>
<td>19850</td>
<td>3000</td>
<td>4.1</td>
<td>34.8</td>
<td>.8032</td>
<td>.00748</td>
<td>.00932</td>
<td>3.28312</td>
</tr>
<tr>
<td>10-19</td>
<td>3250</td>
<td>600</td>
<td>13.0</td>
<td>92.2</td>
<td>.1315</td>
<td>.02666</td>
<td>.02021</td>
<td>1.70950</td>
</tr>
<tr>
<td>20-29</td>
<td>1007</td>
<td>340</td>
<td>25.0</td>
<td>174.2</td>
<td>.0407</td>
<td>.00085</td>
<td>.02085</td>
<td>1.01750</td>
</tr>
<tr>
<td>30-39</td>
<td>606</td>
<td>230</td>
<td>38.2</td>
<td>320.4</td>
<td>.0245</td>
<td>.00084</td>
<td>.03415</td>
<td>.93590</td>
</tr>
<tr>
<td></td>
<td>4170</td>
<td></td>
<td></td>
<td></td>
<td>1.0000</td>
<td>.01183</td>
<td>.08451</td>
<td>6.95602</td>
</tr>
</tbody>
</table>

From the data in Table 5, we find

\[
V_{\text{strat.}} = \sum W_j^2 s_j^2 / n_j = .01183 \tag{68}
\]
Now,

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

\[
= \frac{1}{n} \left[ 55.02 + \sum W_j \left( \overline{y}_{pj} - \overline{y}_p \right)^2 \right] .
\]  

(69)

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

\[
Q = \sum W_j \overline{y}_{nj}^2 - \left( \sum W_j \overline{y}_{nj} \right)^2
= \sum W_j s_j^2/n_j + \sum W_j \overline{y}_{pj} \overline{y}_p
\]

\[
= 96.91 - (6.95602)^2 - .08451 + .01183 = 48.45
\]

(70)

Then,

\[
V\text{ ran.} = \frac{1}{41.70} (55.02 + 48.45) = .02481 ,
\]

(71)

whereas \( V\text{ strat.} = .01183 . \)

The reduction from \( V\text{ ran.} \) exceeds 50\%, since the ratio of the variances is \( .01183/.02481 = .477 . \)

\textbf{Simplification when \( \sigma_j^2 \) is constant and sampling is proportional.} 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 \textbf{pooled} mean square within strata, \( s_w^2 \). Then we have

\[
\text{Estd.} : V\text{ strat.} = \sum \frac{W_j s_w^2}{n_j} = \frac{s_w^2}{n} .
\]

(72)

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

(73)

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

\[
Q = \frac{1}{n} \sum n_j \left( \overline{y}_{nj} - \overline{y}_n \right)^2 - \frac{k}{n} s_w^2 + \frac{s_w^2}{n} .
\]

(74)
Hence,
\[
\text{Estd. V ran.} = \frac{1}{n^2} \left[ \sum n_j (\bar{y}_{nj} - \bar{y}_n)^2 + (n - k + 1) s_w^2 \right]
\]  \hspace{0.5cm} (75)

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

**Analysis of variance for the stratified sample.**

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>M.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among strata</td>
<td>(k - 1)</td>
<td>(B = \sum n_j (\bar{y}_{nj} - \bar{y}_n)^2/(k-1))</td>
</tr>
<tr>
<td>Within strata</td>
<td>(n - k)</td>
<td>(W = s_w^2)</td>
</tr>
</tbody>
</table>

From this the formula (75) may be written
\[
\text{Estd. V ran.} = \frac{1}{n^2} \left[ (k-1) B + (n-k + 1) W \right]
\]  \hspace{0.5cm} (76)

while
\[
\text{Estd. V strat.} = \frac{1}{n} W
\]  \hspace{0.5cm} (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,

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>M.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between strata (half-plots)</td>
<td>25</td>
<td>90.76 = B</td>
</tr>
<tr>
<td>Within strata</td>
<td>100</td>
<td>38.44 = W</td>
</tr>
</tbody>
</table>

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.

\[
\text{Estd. V strat.} = \frac{38.44}{6} = 6.41
\]
Est. V ran. = $\frac{1}{36} \left[ B + 5W \right] = \frac{1}{36} \left[ 90.76 + 5(38.44) \right] = 7.86$

R.E. = $\frac{7.86}{6.41} = 1.23$.

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 $\sigma$'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

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

Therefore

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

(78)
From (79) we obtain the solution for \( n \) as

\[
    n = \frac{(\Sigma N_j \sigma_j)^2}{N^2 d^2 + \Sigma N_j \sigma_j^2}
\]  

(79)

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

This gives

\[
    n_o = \frac{t^2(\alpha) (\Sigma N_j \sigma_j)^2}{N^2 d^2}
\]  

(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

\[
    n = \frac{\sigma^2 N^2}{N^2 d^2 + N \sigma^2} = \frac{t^2(\alpha) \sigma^2/d^2}{1 + \frac{t^2(\alpha) \sigma^2}{N d^2}}
\]  

(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 \( \sigma_j \) 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 \( \sigma_j \) 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

\[ n_j = \frac{N_j n}{N} \]  

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

\[ V(\bar{y}_n)_{\text{prop.}} = \frac{1}{N^2} \sum N_j (N_j - n_j) \sigma_j^2/n_j \]  

(81)

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

\[ \frac{t(\alpha)}{N} \sqrt{\frac{\sum N_j \sigma_j^2}{n} - \sum N_j \sigma_j^2} = d \]  

(82)

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

\[ n = \frac{\sum N_j \sigma_j^2}{\frac{d^2 N^2}{t^2(\alpha)} + \sum N_j \sigma_j^2} = \frac{\frac{t^2(\alpha) \sum N_j \sigma_j^2}{d^2 N}}{1 + \frac{1}{N^2} \left( \frac{\frac{t^2(\alpha) \sum N_j \sigma_j^2}{d^2}}{d^2} \right)} \]  

(83)

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

\[ n_o = \frac{t^2(\alpha) \sum N_j \sigma_j^2}{d^2 N} \]  

(83a)

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

5.16 Proximity as a Basis for Stratification: 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, 

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

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 from \( n \) to \( n_r \). Then, the variance of the mean of the random sample becomes \( \sigma^2/n_r \) or \( \frac{V_{\text{random}}}{r} \). Now, if \( r \) is chosen so that \( \frac{V_{\text{random}}}{r} = V_{\text{strat.}} \), we obtain

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

<table>
<thead>
<tr>
<th>Type of Seedling</th>
<th>R.E. or r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bed #1</td>
</tr>
<tr>
<td>Silver Maple</td>
<td>1.29</td>
</tr>
<tr>
<td>American Elm</td>
<td>2.79</td>
</tr>
<tr>
<td>White Spruce</td>
<td>1.16</td>
</tr>
<tr>
<td>White Pine</td>
<td>1.15</td>
</tr>
</tbody>
</table>

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.**

<table>
<thead>
<tr>
<th>State</th>
<th>No. of Items</th>
<th>Twp.</th>
<th>4-Twp.</th>
<th>County</th>
<th>Type of farming area</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iowa - 1938</td>
<td>18</td>
<td>115</td>
<td>-</td>
<td>100</td>
<td>96</td>
<td>91</td>
</tr>
<tr>
<td>Iowa - 1939</td>
<td>19</td>
<td>121</td>
<td>-</td>
<td>100</td>
<td>97</td>
<td>91</td>
</tr>
<tr>
<td>Florida - 1942</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citrus fruit area</td>
<td>14</td>
<td>144</td>
<td>119</td>
<td>100</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Truck farming area</td>
<td>15</td>
<td>111</td>
<td>-</td>
<td>100</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>California - 1942</td>
<td>17</td>
<td>113</td>
<td>-</td>
<td>100</td>
<td>97</td>
<td></td>
</tr>
</tbody>
</table>

*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

\[
\frac{(\Sigma W_j \sigma_j)^2}{n} \quad (86)
\]

Instead of the true stratum proportions \( W_j \), we have estimates \( \bar{w}_j \). The sample estimated mean is \( \Sigma \bar{w}_j \bar{y}_{nj} \). The first point to note is that this estimate is biased. Its mean value in repeated sampling is \( \Sigma \bar{w}_j \bar{y}_{pj} \), while the true population mean is \( \Sigma W_j \bar{y}_{pj} \). The bias amounts to

\[
\Sigma (\bar{w}_j - W_j) \bar{y}_{pj} \].

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 \( (\Sigma \bar{w}_j \sigma_j)^2 / n \). The total variance is

\[
\frac{(\Sigma \bar{w}_j \sigma_j)^2}{n} + \left\{ \Sigma (\bar{w}_j - W_j) \bar{y}_{pj} \right\}^2 \quad (87)
\]

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),

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

(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 di-
rected 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 Case Where the Strata Cannot be Identified in Advance: 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

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

In other words we use the true 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 proportional 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,

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

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

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

$$\mathbb{E} \left\{ V \left( \overline{\gamma}_W \right) \right\} = \frac{1}{n} \Sigma W_j \sigma_j^2 + 0 \ (n^{-2})$$ (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

\[
  n_2 = kr_2
\]  

(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

\[ C = c_0 n + c_1 n_1 + c_2 r_2, \]  

(94)

where the \( c_j \)'s are costs per unit: \( c_0 \) 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

\[ c_0 n + c_1 W_1 n + c_2 W_2 n/k. \]  

(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

\[ \bar{y}_s = \frac{N}{n} \left\{ n_1 \bar{y}_{1n} + n_2 \bar{y}_{2r} \right\} \]  

(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 \( W_1/N \), \( W_2/N \), 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.

\[
y_s = \frac{N}{n} \left\{ n_1 \bar{y}_{1n} + n_2 \bar{y}_{2n} \right\} + \frac{N n_2}{n} \left( \bar{y}_{2r} - \bar{y}_{2n} \right) \tag{97}
\]

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 \( \sigma^2 \) 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

\[
\bar{y}_{2n} \left( \bar{y}_{2r} - \bar{y}_{2n} \right)
\]

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,

\[
\text{V}(y_s) = \frac{N(N-n)}{n} \sigma^2 + \frac{N^2}{n^2} \mathbb{E} \left\{ n_2^2 \left( \bar{y}_{2r} - \bar{y}_{2n} \right)^2 \right\} \tag{98}
\]

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

\[
(\bar{y}_{2r} - \bar{y}_{2p}) = (\bar{y}_{2r} - \bar{y}_{2n}) + (\bar{y}_{2n} - \bar{y}_{2p}) \tag{99}
\]

so that

\[
\mathbb{E} (\bar{y}_{2r} - \bar{y}_{2p})^2 = \mathbb{E} (\bar{y}_{2r} - \bar{y}_{2n})^2 + \mathbb{E} (\bar{y}_{2n} - \bar{y}_{2p})^2 \tag{100}
\]

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, for fixed $n_2$ and $r_2$,

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

where $\sigma_2^2$ is the variance within the 'non-response' stratum. This gives

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

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

$$V(y_\ell) = \frac{N(N-n)}{n} \sigma_2^2 + \frac{N^2}{n^2} (k-1) \sigma_2^2 \frac{n N_2}{N}$$

$$= \frac{N(N-n)}{n} \sigma_2^2 + \frac{N N_2 (k-1)}{n} \sigma_2^2 \quad (103)$$

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:

$$k = \sqrt{\frac{c_2 (\sigma^2 - W_2 \sigma_2^2)}{\sigma_2^2 (c_0 + c_1 W_1)}} \quad (104)$$

$$n = \frac{NV}{N} + \sigma^2 \quad (105)$$
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 $\sigma^2$ and $\sigma_2^2$, 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

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

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

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

<table>
<thead>
<tr>
<th>n</th>
<th>n_1</th>
<th>n_2</th>
<th>r_2</th>
<th>Schedules Tabulated</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
</tr>
<tr>
<td>1,000</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>1,000</td>
<td>$2,550</td>
</tr>
<tr>
<td>2,000</td>
<td>1,000</td>
<td>1,000</td>
<td>333</td>
<td>1,333</td>
<td>2,099</td>
</tr>
<tr>
<td>5,000</td>
<td>2,500</td>
<td>2,500</td>
<td>278</td>
<td>2,778</td>
<td>2,751</td>
</tr>
</tbody>
</table>

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 not the 935 that are 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*

<table>
<thead>
<tr>
<th></th>
<th>No. of Growers</th>
<th>No. of Fruit Trees</th>
<th>Average No. of Fruit Trees per grower</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growers on the mailing list to whom schedules were sent.</td>
<td>3,241</td>
<td>1,064,899</td>
<td>329</td>
</tr>
<tr>
<td>Schedules returned unclaimed.</td>
<td>125</td>
<td>39,442</td>
<td>315</td>
</tr>
<tr>
<td>Remainder of Population</td>
<td>3,116</td>
<td>1,025,457</td>
<td>329</td>
</tr>
<tr>
<td>Response to first request</td>
<td>300</td>
<td>136,659</td>
<td>456</td>
</tr>
<tr>
<td>Response to second request</td>
<td>543</td>
<td>207,662</td>
<td>382</td>
</tr>
<tr>
<td>Response to third request</td>
<td>434</td>
<td>147,387</td>
<td>340</td>
</tr>
<tr>
<td>Total Response</td>
<td>1,277</td>
<td>491,908</td>
<td>365</td>
</tr>
<tr>
<td>Percent Total Response</td>
<td>41%</td>
<td>46%</td>
<td></td>
</tr>
<tr>
<td>Total Non-Respondents</td>
<td>1,839</td>
<td>533,549</td>
<td>290</td>
</tr>
<tr>
<td>Percent Total Non-Respondents</td>
<td>59%</td>
<td>52%</td>
<td></td>
</tr>
</tbody>
</table>

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

### REFERENCES


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 \( k \)th subsequent unit. For instance, if \( k \) is 15 and if the first unit drawn is number 15, 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 \( k \)th" 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 \( k \)th" rule. For instance, if the units are described on cards which have not 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 one 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. 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

\[
m_i = \left\{ y_1 + y_{1+k} + \cdots + y_{1+(n-1)k} \right\} / n.
\]

(106)

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

\[
E(m_i) = \left\{ m_1 + m_2 + \cdots + m_k \right\} / k,
\]
From (106) this is clearly equal to \( \bar{y}_p \).

**Variance of the estimate.** The variance may be expressed in a number of different ways. One form, due to the Nalons (22) is given in Theorem 11.

**Theorem 11.** The variance of the mean of the systematic sample is

\[
\text{Var} (\bar{y}_n) = \frac{\sigma^2}{n} \left\{ \frac{N-1}{N} + \frac{2}{n} \sum_{d=1}^{n-1} (n-d) C'_{kd} \right\}
\]

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

\[
k(n-d) \sigma^2 C'_{kd} = \sum_{i=1}^{k} (y_i - \bar{y}_p) (y_{i+kd} - \bar{y}_p).
\]

**Proof:** By definition,

\[
\text{Var} (\bar{y}_n) = E (\bar{y}_n - \bar{y}_p)^2 = \frac{1}{k} \sum_{i=1}^{k} (m_i - \bar{y}_p)^2 = \frac{1}{k} \sum_{i=1}^{k} (nm_i - n\bar{y}_p)^2
\]

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 \( k \) values of \( i \), the squared terms amount to

\[
\sum_{i=1}^{N} (y_i - \bar{y}_p)^2 = (N-1) \sigma^2.
\]

The cross product terms will be seen to involve every pair of observations that differ by a multiple of \( k \). These may be grouped according to the multiple of \( k \). 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

\[
\text{Var} (\bar{y}_n) = \frac{1}{n^2k} \left\{ \frac{(N-1)}{\sigma^2} + 2 \sum_{i=1}^{k} (y_i - \bar{y}_p) (y_{i+k} - \bar{y}_p) + \sum_{i=1}^{k} (y_i - \bar{y}_p) (y_{i+2k} - \bar{y}_p) + \ldots + 2 \sum_{i=1}^{k} (y_i - \bar{y}_p) (y_{i+(n-1)k} - \bar{y}_p) \right\}
\]
When we introduce the serial correlation coefficients as defined in (107), this becomes
\[
V(\bar{y}_n) = \frac{1}{n^2 k} \left\{ (N-1) \sigma^2 + 2k \sum_{d=1}^{n-1} (n-d) \rho_{kd}^2 \right\} \\
= \frac{\sigma^2}{n} \left\{ \frac{N-1}{N} + \frac{2}{n} \sum_{d=1}^{n-1} (n-d) \rho_{kd}^2 \right\} \quad (108)
\]

**Note:** For a random sample of size \( n \), the corresponding result would be
\[
V(\bar{y}_n) = \frac{\sigma^2 (N-n)}{Nn} = \frac{\sigma^2 (k-1)}{n \cdot k}
\]

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 efficiency 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.

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

where \( \sigma_w^2 \) 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
\( \log \text{kd}, \text{defined by} \)

\[
(k-1)(n-d) \sigma^2_w \rho(kd)_w = \sum_{i=1}^{k(n-d)} (y_i - \bar{y}_{pl}) (y_{i+kd} - \bar{y}_p, i + kd).
\]

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

\[
ny_p = \bar{y}_{pl} + \bar{y}_{p,1+k} + \ldots + \bar{y}_{p, i+(n-1)k}
\]

we have

\[
(nm_i - ny_p) = (y_i - \bar{y}_{pl}) + (y_{i+k} - \bar{y}_p, i+k) + \ldots \text{(to n terms)}.
\]

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

\[
\text{Var}(\bar{y}_n) = \frac{(N-n)}{N} \frac{\sigma_w^2}{n}
\]

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 \( y_1 = 1 \). 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 $\sigma^2$ is given by

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

(110)

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

$$
\bar{y}_{\text{ran}} = \frac{(N-n)}{N} \cdot \frac{\sigma^2}{n} = \frac{n(k-1)}{nk} \cdot \frac{nk(N+1)}{12n} = \frac{(k-1)(N+1)}{12n}
$$

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

$$
\bar{y}_{\text{strat}} = \frac{(N-n)}{N} \cdot \frac{\sigma^2_w}{n} = \frac{n(k-1)}{nk} \cdot \frac{k(k+1)}{12n} = \frac{(k^2-1)}{12n}
$$

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

$$
\bar{y}_{\text{sys}} = \frac{(k^2-1)}{12}
$$

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

$$
\bar{y}_{\text{strat}} < \bar{y}_{\text{sys}} < \bar{y}_{\text{ran}}.
$$

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, 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
foot 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 \( 2k \), 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 \( k \). 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.
Kadow (25).
### Table 10.

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

|       | Bed | \(V_{\text{ran}}\) | \(V_{\text{strat}}\) | \(V_{\text{sys}}\) | Estimate of \(V_{\text{sys}}\) 
|-------|-----|-------------------|-------------------|-------------------|-------------------------------
|       | (1) | (2)               | (3)               | (4)               | (5)                           |
| 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              | 5.4               | 5.5               | 22.6                          | 15.8                          |
| White Spruce | 1   | 13.4              | 11.9              | 5.5               | 17.2                          | 12.2                          |
|        | 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 true sampling variance of the mean of the systematic sample is \( s^2 / 2 \).
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

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

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 neighboring 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 Mátórn (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 (right hand columns). Method (1) is the method given in formula (111), based on successive differences. It considerably overestimates the variance for the stratified sample and is scarcely within sight of the true variance for the systematic sample. Method 2 uses the estimate
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


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

<table>
<thead>
<tr>
<th>Type of Unit</th>
<th>One foot row</th>
<th>Two-feet row</th>
<th>One foot bed</th>
<th>Two-feet bed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_1$ = number of units in pop.</td>
<td>2,604</td>
<td>1,302</td>
<td>434</td>
<td>217</td>
</tr>
<tr>
<td>$c_1^2$ = pop. variance per unit</td>
<td>2,537</td>
<td>6.746</td>
<td>23,094</td>
<td>68,558</td>
</tr>
<tr>
<td>Number of feet of row that can be counted in 15 mins.</td>
<td>44</td>
<td>62</td>
<td>78</td>
<td>108</td>
</tr>
</tbody>
</table>

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

$$\frac{2}{N_1} \frac{c_1^2}{n_1}$$
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

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

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

\[
    n_2 = n_1 \left( \frac{1}{4} \right) \frac{6.746}{2.537} = .665 \, n_1.
\]

These data are shown in Table 12, first line.

**Table 12.**

**Comparative Sample Sizes and Costs**

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

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 \( r_1 \) of the smallest units may be expressed as

\[
c_1 = r_1 / 44, \nonumber
\]

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

\[
\frac{1.330 \times r_1}{62} = 1.330 \times \frac{44}{62} \quad c_2 = 0.944 \quad c_1, \nonumber
\]

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

\[
\sigma_i^2 \frac{c_i}{z_i},
\]

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

\[
\frac{2.537}{1 \times 44} = .0577 \quad \text{and} \quad \frac{22.094}{6 \times 78} = .0493, \nonumber
\]

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 consensus. 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 1943 in order to estimate farm employment. For details see reference (20). 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
TABLE 13.

SIZES OF POPULATION AND SAMPLE

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of strata</td>
<td>587</td>
<td>572</td>
</tr>
<tr>
<td>No. of sampling units</td>
<td>72,849</td>
<td>1,397</td>
</tr>
<tr>
<td>No. of farms</td>
<td>217,976</td>
<td>4,166</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between units within strata</td>
<td>825</td>
<td>6.218</td>
</tr>
<tr>
<td>Between farms within units</td>
<td>2,768</td>
<td>2.918</td>
</tr>
<tr>
<td>Total: Between farms within strata</td>
<td>3,593</td>
<td>3,676</td>
</tr>
</tbody>
</table>

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 \( N^2 \sigma^2 / n \), where \( N \) is the number of farms in the state and
$\sigma_1^2$ is the variance between farms within strata. To estimate $\sigma_1^2$, 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 $\sigma_1^2$ 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**

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>Estimated mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between units within strata</td>
<td>72,262</td>
<td>6.218</td>
</tr>
<tr>
<td>Between farms within units</td>
<td>145,127</td>
<td>2.918</td>
</tr>
<tr>
<td>Total: Between farms within strata</td>
<td>217,389</td>
<td>4.015</td>
</tr>
</tbody>
</table>

The degrees of freedom are obtained from the data in Table 13. The argument is that if we had 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 $\sigma_1^2$ between farms within strata is

$$\frac{72,262 \times 6.218 + 145,127 \times 2.918}{217,389} = 4.015.$$ 

If $\sigma_2^2$ is the variance within strata for the 'three-farm' unit, the variance of the estimated state total will be
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 $\sigma^2/3$, 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 (6), to develop a general law which shows how the sampling error changes with the size of unit. Suppose that the smallest unit is called an element, and that the large unit contains $M$ neighboring elements. It has been found in several agricultural surveys that the variance $\bar{W}$ between elements within the large unit is related to $M$ by means of the formula

$$W = AM^g, \quad g > 0,$$

(112)

where $A$ and $g$ are constants that do not depend on $M$. In this representation $\bar{W}$ 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.

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>Mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between large units</td>
<td>( \frac{N}{M} - 1 )</td>
<td>( B )</td>
</tr>
<tr>
<td>Between elements within large units</td>
<td>( \frac{N}{M} (N - 1) )</td>
<td>( W )</td>
</tr>
<tr>
<td>Between elements in the population ( (N - 1) )</td>
<td>( (N - 1) )</td>
<td>( T )</td>
</tr>
</tbody>
</table>

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 \( A, \theta, \) and \( T \) by the relation

\[
B = \left\{ \frac{M(N-1)T - N(M-1)\theta \kappa}{(N-M)} \right\} / (N-M).
\] (113)

The constants \( A, \theta, \) and \( \kappa \) 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 \( \theta \). 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 \( T = A\theta \kappa \).
Substitution in (113) gives

\[ B = AN \left\{ \frac{(N-1)N^{g-1}}{M} - \frac{(M-1)M^{g-1}}{N} \right\} / (N-M). \]  

(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 \( N \) 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

\[ C(M,n) = c_1Mn + c_2\sqrt{n}. \]  

(115)
The best choice of $\bar{x}$ and $\bar{y}$ 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 $B/n$, since there are $kn$ 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 $\sigma^2$, we have

$$V(M,n) = \frac{\sigma^2}{kn} = \left\{ \frac{(N-1)}{N} T - N(N-1)AM^{N-1} \right\} / n(N-M).$$  (116)

Since $N$ is assumed very large this reduces to

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

$$c_1M + \frac{1}{2} c_2 n - \frac{1}{2} = \lambda V/n.$$  (118)

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

Dividing (119) by (118) so as to eliminate $\lambda$, we find that

$$\frac{n}{V} \frac{\partial V}{\partial M} = -\frac{1}{1 + \frac{c_2}{2c_1M\sqrt{n}}}.$$  (120)

Now if equation (115) for the cost is solved as a quadratic in $\sqrt{n}$, it will be found, after some manipulation, that
\[ \frac{2c_1 k \sqrt{n}}{c_2} = \left\{ 1 + \frac{4 C c_1 M}{c_2^2} \right\} + \frac{1}{2} - 1 \]

Substituting in (42a) we find

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

The important point about this equation is that it does not involve \( M \), 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 ith sampling unit has \( y_i \) 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 element \( \Sigma y_i / \Sigma M_i \), 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 Possible Biases with Small Units: 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


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 comes 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 \( m/k \) 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

\[
y_{ij} = \mu + b_i + w_{ij}
\]

(121)

where \( \mu \) represents the population mean, \( b_i \) varies from unit to unit with mean zero and variance \( \sigma_b^2 \), and \( w_{ij} \) varies from sub-unit to sub-unit with mean zero and variance \( \sigma_w^2 \). 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

\[
\bar{y}_{nm} = \mu + (b_1 + b_2 + \ldots + b_n)/n + (w_{11} + w_{12} + \ldots + w_{nm})/mn
\]

Hence,

\[
V(\bar{y}_{nm}) = E(\bar{y}_{nm} - \mu)^2 = \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nm}
\]  \( \text{(122)} \)

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**

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>Mean square</th>
<th>Estimate of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between sampling units</td>
<td>( n-1 )</td>
<td>( B = m \Sigma (\bar{y}<em>{1i} - \bar{y}</em>{nm})^2 / (n-1) )</td>
<td>( \sigma_b^2 + m \sigma_w^2 )</td>
</tr>
<tr>
<td>Within units between sub-units</td>
<td>( n(m-1) )</td>
<td>( W = \Sigma (y_{1ij} - \bar{y}_{1i})^2 / n(m-1) )</td>
<td>( \sigma_w^2 )</td>
</tr>
</tbody>
</table>

where \( \bar{y}_{1i} \) 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 \( m \) units. We wish to estimate the variance of the sample mean under the supposition that these numbers were changed to \( m' \) and \( m' \) respectively. By (123), this variance is

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

From Table 16, unbiased estimates of \( \sigma_b^2 \) and \( \sigma_w^2 \) are

\[
s_b^2 = \frac{(B-W)}{m}; \quad s_w^2 = W .
\]

Hence the estimated variance of the sample mean is

\[
\frac{s_b^2}{n'} + \frac{s_w^2}{n'm'} = \frac{1}{n'} \left[ \frac{B}{m} + W \left( \frac{1}{m'} - \frac{1}{m} \right) \right] \tag{126}
\]

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>
<thead>
<tr>
<th>TABLE 17.</th>
<th>ANALYSIS OF VARIANCE OF WHEAT YIELDS (BUSHEL PER ACRE)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d.f.</td>
</tr>
<tr>
<td>Between fields within districts</td>
<td>217</td>
</tr>
<tr>
<td>Within fields between subsampling units.</td>
<td>222</td>
</tr>
</tbody>
</table>

*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 \( \frac{180}{2n} \), or \( \frac{90}{n} \). By substitution in (125) the reader may verify that the corresponding variances for cases (i) and (ii) are

\[
V_i = \frac{45}{n} : V_{ii} = \frac{80.5}{n}.
\]

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
obtained by taking samples from the plots. In fact, the earliest applications of the theory were to the sampling of field experiments. The mathematical model becomes a little more complex because of the treatment and block effects, but the essential analysis is the same. For a randomized blocks experiment, let $y_{ijk}$ be the yield from the $k$ th subsampling unit in the $j$ th replicate of the $i$ th treatment. Then

$$y_{ijk} = \mu + t_i + r_j + e_{ijk},$$

where $\mu$ represents the general mean, $t_i$ the effect of the $i$ th treatment, $r_j$ that of the block or replicate, and $e_{ijk}$ the residual.

The last component is separated into two parts: a part $b_{ij}$ depending only on the plot to which the sub-unit belongs, and a part $w_{ijk}$ varying from sub-unit to sub-unit within the plot. If the analysis of variance (on a sub-unit basis) is computed, it can be shown algebraically that the following expectations hold.

$$E(\text{Experimental error mean square}) = \sigma^2_w + m \sigma^2_b,$$

$$E(\text{mean square between sub-units within plots}) = \sigma^2_w,$$

where $m$ is the number of sub-units taken per plot. Further, if there are $n$ replicates, the experimental error variance of a treatment mean (i.e., a mean over $mn$ sub-units) is

$$V(\bar{y}_{i..}) = \frac{\sigma^2_w + mn \sigma^2_b}{mn} = \frac{\sigma^2_b}{n} + \frac{\sigma^2_w}{mn} \quad (126)$$

In such cases the cost function will frequently be of the form

$$C = c_1 n + c_2 mn, \quad (127)$$

where $c_1$ is the component of cost proportional to the amount of replication but independent of the amount of sampling, while $c_2$ is
a component proportional to the number of subsamples taken. If the cost is minimized for a specified value of the variance (126), we find

$$m = \frac{c_w}{c_b} \sqrt{\frac{c_1}{c_2}}$$  \hspace{1cm} (128)$$

This equation determines the optimum amount of sampling per plot. The accompanying number of replications is then calculated from equation (126).

For a more complete development of this theory and its application to the sampling of cereal experiments, see Yates and Zacopanay (35). One common device is to use stratification (often called 'local control') within each plot. For instance, if eight samples are taken from each plot, the plot is divided into quarters, two samples being taken from each quarter. The theory is unchanged except that the mean square between sub-units within plots is replaced by the mean square between sub-units within strata (quarters).

One further point deserves mention. In order to perform F and t tests of the treatment effects in the experiment, we need an estimate of the experimental error variance ($\sigma_w^2 + m \sigma_b^2$), but we do not need an estimate of the subsampling error $\sigma_w^2$. Consequently, so far as the drawing of conclusions from the experiment is concerned, we can take only one subsample per plot, or we can use a method such as systematic sampling which does not provide an unbiased estimate of the sampling error. On the other hand, if we wish to use the results to learn something about the optimum amount of sampling in future experiments, an estimate of $\sigma_w^2$ is required for the use of formula (125). Thus in the exploratory stages of sampling, it is advisable to ensure that an unbiased estimate of $\sigma_w^2$ will be available. When
the optimum method of sampling has been learned, this requirement can
be dropped. For example, if expensive chemical determinations are to
be made on the samples, all the samples from a plot may be bulked and
only a single determination made for each plot.

8.6 Alternative mathematical model: We return to the considera-
tion of subsampling in sample surveys. One consequence of the mathe-
matical model, as will be seen from Table 16, page 92, is that the
true variance between sampling units is always at least as large as
the variance between sub-units within units. The model does not
allow for the possibility that the variance within units might be
larger than that between units. Situations may arise in which this
is so. One example previously mentioned is that of the sex-ratio,
when the unit is a household and the sub-unit a person, (27). The
mean-square between persons within a household is substantially larger
than the mean square between households. This happens because there
is a negative correlation between the sexes of members of the same
households, owing to the fact that many households contain both hus-
band and wife. Although it is less likely to do so, the same effect
may arise in field experiments if there is competition between plants
within a plot.

The extension of our model to this case has been given by Yates
and Faucpanay (35). Alternatively, Hansen and Hurwitz (27) suggest
the use of the intra-class correlation coefficient. Instead of (121)
we have

\[ Y_{ij} = \mu + w_{ij} \]  \hspace{1cm} (129)

The quantities \( w_{ij} \) all have mean zero and variance \( \sigma^2 \). Any pair of
sub-units \( w_{ij}, w_{ik} \), that are in the same unit are correlated, with
correlation coefficient \( \rho \), while elements in different units are
uncorrelated.

With this model, the expectations of the mean squares in the analysis of variance on a sub-unit basis may be shown to be as in Table 16, which corresponds to Table 16 for the previous model.

**Table 18.**

**Analysis of Variance with Sub-sampling: Alternative Model**

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>Mean square</th>
<th>Estimate of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between sampling units</td>
<td>(n-1)</td>
<td>$E = mE_{n}(\bar{y}_{1} - \bar{y})^2/(n-1)$</td>
<td>$\sigma^2 {1/(m-1) + 1}$</td>
</tr>
<tr>
<td>Within units between sub-units</td>
<td>n(m-1)</td>
<td>$W = \Sigma (y_{1j} - \bar{y}_{1})^2/n(m-1)$</td>
<td>$\sigma^2 (1 - \rho)$</td>
</tr>
</tbody>
</table>

If $\rho$ is positive, $E$ will have a larger expectation than $W$ and the results obtained are exactly the same as those obtained with the previous model. The case where $E$ is expected to be less than $W$ is covered by negative values of $\rho$. Note, however, that $\rho$ cannot be less than $-1/(m-1)$, for such values would give $E$ a negative expected value.

This property of the intraclass correlation coefficient is well known.

For certain applications, it is known that some pairs of sub-units within the same unit will be correlated, but others will not. Thus for full generality we would require a model in which $\rho_{ijk}$ is the correlation between the $j$th and $k$th sub-units within the $i$th unit. The only effect of this elaboration is to replace $\rho$ in Table 18 by $\bar{\rho}$, the simple average of all these correlation coefficients. For the case of the sex ratio presented by Hansen and Hurwitz, suppose that the typical household consists of husband, wife, and two children, and let $w_{ij}$ be 1 for a male and 0 for a female. Six possible pairs can be formed from the members of the household. The correlation between the sex of husband and wife is $-1$, but there will
be no correlation between the sexes of the five other pairs (excluding rare cases such as identical twins). Consequently \( \bar{\rho} = -1/6 \). It follows that the variance between sampling units would be expected to be about 4/7 of that within units.

8.7 The finite population correction: Thus far it has been assumed that \( n/N \) is small; this should be remembered when using previous results. We now suppose that the population contains \( N \) units, each with \( M \) sub-units, while the sample has \( n \) units, each with \( m \) sub-units. By definition, the true variance of the sample mean is

\[
V(\bar{y}_{nm}) = E(\bar{y}_{nm} - \bar{y}_{NM})^2.
\]

First, it is necessary to re-define \( \sigma_w^2 \) and \( \sigma_b^2 \), so that they refer to a finite population. Consider the following analysis of variance for the complete population:

**TABLE 19.**

**ANALYSIS OF VARIANCE FOR THE COMPLETE POPULATION (SUB-UNIT BASIS)**

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>Mean square</th>
<th>Defined as equal to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between units</td>
<td>( N(M-1) )</td>
<td>( \sum_{i=1}^{N} \sum_{j=1}^{M} (\bar{y}<em>{ij} - \bar{y}<em>i) \sum</em>{j=1}^{M} (\bar{y}</em>{ij} - \bar{y}_i)^2 / (N-1) )</td>
<td>( \sigma_w^2 + M \sigma_b^2 )</td>
</tr>
<tr>
<td>Within units between sub-units</td>
<td>( N(M-1) )</td>
<td>( \sum_{i=1}^{N} \sum_{j=1}^{M} (\bar{y}_{ij} - \bar{y}_i)^2 / (NM-1) )</td>
<td>( \sigma_w^2 )</td>
</tr>
</tbody>
</table>

where \( \bar{y}_i \) denotes the mean of the \( i \)th unit. We define \( \sigma_w^2 \) and \( \sigma_b^2 \) so that the equations given in the two lines of the analysis are valid. With these definitions, as will be seen later, the expected values of \( B \) and \( W \) remain as given in Table 16.

**Theorem 13:**

\[
V(\bar{y}_{nm}) = E(\bar{y}_{nm} - \bar{y}_{NM})^2 = \frac{(N-n)}{N} \sigma_b^2 + \frac{(MN-mm)}{MN} \sigma_w^2.
\]
Proof: Write

$$\bar{y}_{nm} - \bar{y}_{nM} = (\bar{y}_{nm} - \bar{y}_{nM}) + (\bar{y}_{nM} - \bar{y}_{nM}) ,$$

(131)

where $$\bar{y}_{nM}$$ denotes the mean that would be obtained if the $$n$$ units in the sample were all enumerated completely. If we square both sides and take the average over all sets of samples that could be drawn, there will be no contribution from the cross-product term on the right, since for any fixed set of $$n$$ units,

$$E(\bar{y}_{nm}) = \bar{y}_{nM} .$$

Consider the first term on the right. At present we restrict attention to a fixed set of $$n$$ units. If each of these units is regarded as a stratum, the sample from these units is a proportionally stratified sample, since $$m$$ are taken out of every $$M$$. Consequently we can apply the formula in Theorem 6, page 26, for the variance of the mean of a stratified sample. This gives

$$E(\bar{y}_{nm} - \bar{y}_{nM})^2 = \frac{1}{(nM)^2} \sum_{j=1}^{n} \frac{M(M-m)}{m} \frac{\sigma^2}{w_j}$$

where $$\sigma^2_{w_j}$$ is the variance within the $$j$$th unit. This may be rewritten

$$\frac{(n-m)}{M} \frac{1}{mn} \frac{-\sigma^2}{\sigma_{wn}}$$

(132)

where $$\sigma^2_{wn}$$ is the average variance within these $$n$$ units. If we further average over all possible sets of $$n$$, it is clear that the average of $$\sigma^2_{wn}$$ is $$\sigma^2_w$$. Hence

$$E(\bar{y}_{nm} - \bar{y}_{nM})^2 = \frac{(n-m)}{M} \frac{1}{mn} \sigma^2_w .$$

(133)

The contribution from the second term on the right of equation (131) presents no difficulty, since $$\bar{y}_{nM}$$ is the mean of a simple random
sample of \(n\) units, each completely enumerated. Consequently,

\[
E(\bar{y}_{nnM} - \bar{y}_{NNM})^2 = \frac{(N-n)}{nN} \left\{ \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{M} \right\}
\]

(134)

since by the definition of \(\sigma_b^2\) in Table 19, the variance of the mean of a unit is \((\sigma_b^2 + \sigma_w^2/M)\).

From (133) and (134), we obtain finally

\[
E(\bar{y}_{nm} - \bar{y}_{NNM})^2 = \frac{(M-m)}{M} \frac{\sigma_w^2}{mn} + \frac{(N-n)}{nN} \left\{ \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{M} \right\}
\]

\[
= \frac{(N-n)}{N} \frac{\sigma_b^2}{n} + \frac{(MN - mn)}{MN} - \frac{\sigma_w^2}{mn} .
\]

**Note:** As \(N\) becomes large, the formula for the variance reduces to

\[
\frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{mn},
\]

which is the same as our earlier formula (123). This implies that the earlier formula requires \(n/N\) to be small, but does not require \(m/M\) to be small.

8.8 Estimation of the variance when the f.p.c. is required:

The previous formula \(E/nnm\) for the estimated variance also needs revision to take account of the f.p.c. For this estimate we use as before the sample analysis of variance, as given in Table 16, page 92. However, \(\sigma_w^2\) and \(\sigma_b^2\) are now as defined in Table 19, page 99.

Since the \(n\) units are chosen at random out of the \(N\) and since each \(m\) is chosen at random out of \(M\), it is easy to see that in repeated sampling the expectation of \(\bar{W}\), the sample mean square within units, is equal to \(\sigma_w^2\), the corresponding population mean square.

The mean value of \(E\) is less obvious. We have

\[
E = m \sum (\bar{y}_i - \bar{y}_{nm})^2/(n-1).
\]
Write
\[ \bar{y}_{i.} = \bar{y}_{1M} + \bar{e}_{i.}, \]

where \( \bar{y}_{1M} \) is the mean of all \( M \) sub-units in the \( i \)th unit. Then
\[
E(\bar{e}_{i.}^2) = \frac{\bar{y}_{1M} - \bar{y}_{nm}}{M} \cdot \frac{\sigma_{ir}^2}{m},
\]

since this is the variance of the mean of a random sample of \( m \) sub-units out of \( M \). Similarly write
\[ \bar{y}_{nm} = \bar{y}_{nM} + \bar{e}_{nm}. \]

In equation (133) we proved that
\[
E(\bar{y}_{nm} - \bar{y}_{nM})^2 = E(\bar{e}_{nm}^2) = \frac{M-m}{M} \cdot \frac{\sigma_{wn}^2}{mn}.
\]

Now,
\[
\sum_{i=1}^{n} (\bar{y}_{i} - \bar{y}_{nm})^2 = \sum_{i=1}^{n} \left[ (\bar{y}_{1M} - \bar{y}_{nM}) + \bar{e}_{i} - \bar{e}_{nm} \right]^2.
\]

When we expand and take the expectation for a fixed set of \( n \) units, we have
\[
E \left\{ \sum_{i=1}^{n} (\bar{y}_{i} - \bar{y}_{nm})^2 \right\} = \sum_{i=1}^{n} (\bar{y}_{1M} - \bar{y}_{nM})^2 + \frac{(M-m)}{M} \sum \sigma_{ir}^2 - \sigma_{wn}^2.
\]

Since \( \sigma_{wn}^2 = \sum \sigma_{ir}^2 \), we obtain, on division by \( (n-1) \),
\[
E \left\{ \frac{(Y_{i} - Y_{nm})^2}{(n-1)} \right\} = \frac{(\bar{y}_{1M} - \bar{y}_{nM})^2}{(n-1)} + \frac{(M-m)}{M} \frac{\sigma_{wn}^2}{(n-1)}.
\]

By Theorem 3, page 11, the first term on the right is an unbiased estimate of the population variance between the true means of the units. Take the mean over all possible selections of \( n \) out of \( M \).

This gives
\[
E (B/m) = \sum_{i=1}^{N} \frac{(\bar{y}_{1M} - \bar{y}_{nM})^2}{(N-1)} + \frac{(M-m)}{M} \frac{\sigma_{wn}^2}{(n-1)}.
\]

\[
= \sigma_{b}^2 + \frac{\sigma_{w}^2}{M} + \left( \frac{1}{m} - \frac{1}{M} \right) \frac{\sigma_{wn}^2}{M}.
\]
\[ \sigma^2_b + \frac{\sigma^2_w}{m} \]

Hence

\[ E(B) = \sigma^2_w + m \sigma^2_b. \]

(135)

The result is the same as for the case where no f.p.c. is required.

**Theorem 14:** An unbiased estimate of \( \nu(\tau_{nm}) \), taking into account the f.p.c., is

\[ \frac{1}{mn} \left\{ \frac{(N-n)}{N} B + \frac{(M-m)}{M} \frac{n}{N} W \right\} \]

(136)

This follows at once from the preceding results. An unbiased estimate of \( \sigma^2_b \) is \( (B-W)/m \). On substituting in formula (130) for \( \nu(\tau_{nm}) \) and collecting the two terms in \( W \), we obtain the result, which has been given by Yates (36).

It may be noted that if \( m = M \), the formula reduces to that applicable to simple random sampling of the units, since in this case units in the sample are being enumerated completely. If \( n=N \), the formula becomes that for proportional stratified sampling, since every unit is being sampled, so that the units serve as strata. If \( n/N \) is negligible, the formula reduces to that given earlier in this section.

8.9 **Stratified sampling of the units:** Subsampling may be combined with any type of sampling of the units. Similarly, the subsampling itself may employ stratification, or systematic sampling. We shall not enter into these elaborations. The formulae for subsampling with stratified sampling of the units will, however, be given, since this combination is common in practice.

Let the suffix \( j \) refer to the stratum. The population variances \( \sigma^2_{bj} \) and \( \sigma^2_{w_j} \) will in general be defined separately for each stratum, since they may vary from stratum to stratum. The definition of
Table 19 will be used in each stratum. The jth stratum contains $N_j$ units, each with $M_j$ sub-units, while the sample from the stratum has $n_j$ units and $m_j$ sub-units in each unit. The estimated population mean per sub-unit is

$$\frac{\sum_j M_j N_j \overline{y}_{nmj}}{\sum_j M_j N_j}$$

Its variance is

$$\frac{\sum_j (M_j N_j)^2 \text{Var}(\overline{y}_{nmj})}{(\sum_j N_j)^2} = \sum_j (M_j N_j)^2 \left[ \frac{(N_j - n_j)}{N_j} \frac{\sigma_{b_j}^2}{n_j} + \frac{(M_j N_j - m_j n_j)}{M_j N_j} \right]$$

$$\frac{\sigma_{w_j}^2}{m_j n_j} \left/ \left( \sum_j M_j N_j \right)^2 \right. \right)$$

from formula (130). Unbiased sample estimates can be obtained from (136). The results simplify considerably if the variances and sampling rates are the same in all strata.

8.10 Sub-subsampling: It is sometimes advisable to carry the process of subsampling a stage further by sampling the sub-units instead of enumerating them completely. For instance, in certain surveys to estimate crop production in India (22), the village is a convenient sampling unit. Within a village, only certain of the fields growing the crop in question are selected, so that the field is a sub-unit. When a field is selected, only certain parts of it are cut for the determination of yield per acre; thus the sub-unit itself is sampled. If physical or chemical analyses were being made on the crop, an additional subsampling might be used, since these determinations are often made on only a part of the sample.
cut from a field.

Results for the elementary theory will be given briefly. The population contains \( N \) units, each with \( M \) sub-units, each of which has \( P \) sub-sub-units. The corresponding numbers for the sample are \( n, m, \) and \( p \) respectively. The model is

\[
y_{ijk} = \mu_i + b_i + w_{ij} + z_{ijk},
\]

The variance of the sample mean per sub-sub-unit is

\[
\text{var}(\bar{y}_{nmp}) = \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nm} + \frac{\sigma_z^2}{nmp}.
\]

The sample analysis of variance (on a sub-sub-unit basis) is as follows.

**TABLE 20.**

**ANALYSIS OF VARIANCE FOR SUB-SUBSAMPLING**

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>Mean square</th>
<th>Estimate of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between units</td>
<td>( n-1 )</td>
<td>B ( \sigma_b^2 + p \sigma_w^2 + mp \sigma_z^2 )</td>
<td></td>
</tr>
<tr>
<td>Between sub-units * within units</td>
<td>( n(m-1) )</td>
<td>W ( \sigma_z^2 + p \sigma_w^2 )</td>
<td></td>
</tr>
<tr>
<td>Between sub-sub-units * within sub-units</td>
<td>( nm(p-1) )</td>
<td>Z ( \sigma_z^2 )</td>
<td></td>
</tr>
</tbody>
</table>

Consequently, an unbiased estimate of the variance of the sample mean is \( B/nmp \). As before, an unbiased estimate can also be obtained of the variance for values of \( n, m, \) and \( p \) different from those used.

To obtain the finite population corrections, we define the basic variances by an analysis of variance for the complete population: e.g., the mean square between units in the complete
population is defined to be \((\sigma_z^2 + P \sigma_w^2 + M P \sigma_b^2)\) and so on. By methods similar to those in Section 8.8, we then find

\[
V(\overline{y}_{mp}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_b^2 + \left(\frac{1}{mn} - \frac{1}{MN}\right)\sigma_w^2 + \left(\frac{1}{mnp} - \frac{1}{MNP}\right)\sigma_z^2, \tag{139}
\]

of which an unbiased estimate is

\[
\frac{1}{mnp} \left[ \frac{(N-n)}{N} B + \frac{(M-m)}{M} \cdot \frac{n}{N} W + \frac{(P-p)}{P} \cdot \frac{n}{N} \cdot \frac{m}{M} \cdot Z \right] \tag{140}
\]

The extension of these formulae to further subsampling is obvious.

8.11 Subsampling when the units are unequal in size: Thus far it has been assumed throughout this section that every unit contains the same number \(M\) of sub-units. In practice this is often not the case. In a national farm survey, for example, the county may be the unit, while the sub-unit is a farm or a group of farms. The numbers of these sub-units in a county may vary considerably. When \(M\) changes from unit to unit, the situation is more complex. The development of methods of sampling and their variance formulae for this case is due mainly to Hansen and Hurwitz (37).

Suppose that the \(i\) th unit has \(M_i\) sub-units. For simplicity, we assume \(n = 1\); i.e., only a single unit is chosen from the population. If this unit is the \(i\) th unit, let \(n_i\) sub-units be sampled at random from it. The mean of the observations from these \(n_i\) sub-units is denoted by \(\overline{y}_{is}\) (s for sample), while the true mean of the unit is \(\overline{y}_{ip}\) (p for population). The mean of the whole population, \(\overline{y}_p\) is the quantity to be estimated.

It seems natural to use the sample mean \(\overline{y}_{is}\) as an estimate of the population mean \(\overline{y}_p\). This estimate is, however, biased. In repeated sampling from the same unit, the average of \(\overline{y}_{is}\) will be
\( \bar{y}_{ip} \) But if we give every unit an equal chance of being the unit selected, the average of \( \bar{y}_{ip} \) in repeated sampling will be

\[
\sum_{i=1}^{N} \frac{\bar{y}_{ip}}{N} = \bar{y}_u \quad \text{(unweighted mean)}
\]

whereas the true population mean is

\[
\bar{y}_p = \Sigma M_i \bar{y}_{ip}/M, \text{ where } M = \Sigma M_i.
\]

To find the sampling error variance of this estimate, write

\[
(\bar{y}_{is} - \bar{y}_p) = (\bar{y}_{is} - \bar{y}_{ip}) + (\bar{y}_{ip} - \bar{y}_u) + (\bar{y}_u - \bar{y}_p)
\]

If we square and take the expectation over all possible samples, all cross-product terms vanish, and we obtain

\[
V(\bar{y}_{is}) = \frac{1}{N} \sum_{i=1}^{N} \frac{(M_i - m_i)}{M_i} \sigma^2_i + \frac{1}{N} \Sigma (\bar{y}_{ip} - \bar{y}_u)^2
\]

Within units \hspace{1cm} Between units

\[+ (\bar{y}_u - \bar{y}_p)^2 \]

(141)

Bias.

The variance comprises three components: one arising from variation within units, one from that between the true means of the units, and one from the bias. The quantity \( \sigma^2_i \) in the first term is the variance within the \( i \)th unit, defined in the usual way.

The values of the \( m_i \) have not been specified. With sufficient information, they could be chosen so as to minimize expected cost.

In practice, the most common choices are to have either all \( m_i \) equal, or to have \( m_i \) proportional to \( M_i \), that is, to subsample a fixed proportion of whatever unit is selected. Note that the choice of the \( m_i \) affects only the first of the three components
of the variance, that arising from variation within units.

An unbiased estimate can be made. If we multiply \( \bar{y}_{is} \) by \( M_i \), we obtain an unbiased estimate of the total of the \( i \)th unit. A further multiplication by \( N/M \) provides an unbiased estimate of the population mean per sub-unit. We may call this estimate \( \bar{y}_t \), since it is derived from an estimate of the unit total. Now

\[
(y_t - \bar{y}_p) = \frac{NM_i \bar{y}_{is}}{M} - \frac{\Sigma M_i \bar{y}_{ip}}{M}
\]

\[
= \frac{NM_i (\bar{y}_{is} - \bar{y}_{ip})}{M} + \frac{N T_i - \Sigma T_i}{M}
\]

where we write \( T_i \) for the true unit total, \( M_i \bar{y}_{ip} \). It follows that

\[
V(\bar{y}_t) = \frac{N}{M^2} \Sigma_{i=1}^{N} M_i(M_i - n_i) \frac{\sigma_i^2}{m_i} + \frac{N}{M^2} \Sigma_{i=1}^{N} (T_i - \bar{T})^2
\]

(142)

where \( \bar{T} = \Sigma T_i/N \), is the unweighted mean of the unit totals.

The 'between-unit' part of this variance (second term on the right) arises from the variation among the unit totals. Consequently, this component is affected by variations in the \( M_i \) as well as by variations in the sub-unit means \( \bar{y}_{ip} \), unless it happens that the two are correlated in such a way that their product is rather constant. Frequently, this component is so large that \( \bar{y}_t \) has a much larger variance than the biased estimate based on the mean per sub-unit. Thus, neither estimate is fully satisfactory.

In this situation, Hansen and Hurwitz (37) propose that the units be selected, not with equal probability \( 1/N \), but with probabilities \( M_i/M \) proportional to their sizes. In order to do this, we cumulate the \( M_i \) and select a random number between 1 and \( M \). The unit in which this number falls in the cumulated totals is the unit chosen.
The effect is to make the sample mean \( \bar{\bar{y}}_i \) an unbiased estimate of \( \bar{y}_p \). For, in repeated sampling, the \( i \)th unit will appear with frequency \( M_i / N \), so that

\[
E(\bar{\bar{y}}_i) = \sum_{i=1}^{N} \frac{M_i}{N} \bar{y}_{ip} = \bar{y}_p.
\]

For the sampling variance, we have

\[
(\bar{\bar{y}}_i - \bar{y}_p) = (\bar{\bar{y}}_i - \bar{y}_{ip}) + (\bar{y}_{ip} - \bar{y}_p).\]

When we compute the average of the squares in repeated sampling, the \( i \)th unit is again weighted by \( M_i / N \), so that

\[
V(\bar{\bar{y}}_i) = E((\bar{\bar{y}}_i - \bar{y}_p)^2) = \sum_{i=1}^{N} \frac{M_i}{N} \frac{(M_i - m_i)}{M_i} \frac{\sigma_i^2}{m_i} \]

\[+ \sum \frac{M_i}{N} (\bar{y}_{ip} - \bar{y}_p)^2 \quad (143)\]

For many populations, this variance is found to be smaller than that of either of the two preceding methods, though this result need not always happen.

8.12 Numerical example: It may be instructive to apply these results to a small population, artificially constructed. The data are as follows.

**TABLE 21.**

**ARTIFICIAL POPULATION WITH UNITS OF UNEQUAL SIZES**

<table>
<thead>
<tr>
<th>Unit</th>
<th>( \bar{y}_{ij} )</th>
<th>( M_i )</th>
<th>( T_i )</th>
<th>( \sigma_i^2 )</th>
<th>( \bar{y}_{ip} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>'0,1'</td>
<td>2</td>
<td>1</td>
<td>.500</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>'1,2,2,3'</td>
<td>4</td>
<td>8</td>
<td>.667</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>'3,3,4,4,5,5'</td>
<td>6</td>
<td>24</td>
<td>.800</td>
<td>4.0</td>
</tr>
</tbody>
</table>
There are three units, with respectively 2, 4 and 6 sub-units. The reader may verify the figures for $T_1$, $\sigma^2$, and $\bar{y}_1$. The population mean $\bar{y}_p$ is $33/12$, or 2.75. The unweighted mean of the $\bar{y}_{1p}$ is 2.167, so that the bias in the first method is -.583. Its square, the contribution to the variance, is .340. One unit is to be selected, and two sub-units sampled from it. We consider four methods.

**Method I.**

Selection: unit with equal probability, two sub-units from it.
Estimate: $\bar{y}_{is}$ (biased).

**Method II.**

Selection: unit with equal probability, $\frac{1}{2} \frac{M_i}{M}$ sub-units from it,
Estimate: $\bar{y}_{is}$ (biased).

**Method III.**

Selection: unit with equal probability, two sub-units from it.
Estimate: $NM_i \bar{y}_{is} / M_i$ (unbiased).

**Method IV.**

Selection: unit with probability $M_i / M$, two sub-units from it.
Estimate: $\bar{y}_{is}$ (unbiased).

Method II (proportional subsampling) does not guarantee a sample size of two (it may be 1, 2, or 3). The average sample size is, however, two.

By application of the sampling error formulae (141), (142), and (143), the reader may verify the following computations:

**TABLE 22.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Contribution to variance from</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within Units</td>
<td>Between Units</td>
</tr>
<tr>
<td>I</td>
<td>.145</td>
<td>2.056</td>
</tr>
<tr>
<td>II</td>
<td>.183</td>
<td>2.056</td>
</tr>
<tr>
<td>III</td>
<td>.256</td>
<td>5.792</td>
</tr>
<tr>
<td>IV</td>
<td>.189</td>
<td>1.813</td>
</tr>
</tbody>
</table>
Though the example is artificial, the results are rather typical. Method IV gives the smallest variance because it has the smallest contribution from 'between units'. Method I (equal size of subsample) is slightly better than Method II (proportional subsampling). Method III, though unbiased, is very inferior.

The total possible number of samples is quite small (it is 22 for Methods I, III, and IV). It is a useful exercise to verify the total variances in Table 22 by constructing the estimates from every possible sample.

In the applications of these results, it is more usually desired to estimate a population total than a mean per sub-unit. For the estimated population total, we need only multiply the previous estimates by $\bar{x}$; their variances become multiplied by $N^2$.

8.13 Selection with arbitrary probabilities: It may happen that the sizes $N_1$ of the units are known only roughly. In the sampling of towns, where the unit is a block and the sub-unit a household, the number of households per block is usually obtained from city maps, but such maps may be out of date or in error. To meet this situation, Hansen and Hurwitz (37) have investigated the theory when the units are selected with probabilities proportional to an estimate of size. Their results also apply to any arbitrary assignment of the probabilities. We consider the estimation of a population total, the population being as in previous sections. Let $P_i$ be the probability assigned to the $i$th unit, where the $P_i$ are any set of numbers that are all greater than zero and add to unity.

First assign a sampling rate $\bar{x}$ to the population, e.g., 1 percent or 5 percent. If the $i$th unit is chosen, we take a sample of size $n_i$ from it, and use as the estimate of the population total $n_i \bar{Y}_i / \bar{x}$; in other words, the sample total, divided by the sampling
rate. The mean value of this estimate in repeated sampling is

\[ \sum_{i=1}^{N} \frac{F_i m_i \bar{y}_{ip}}{t} \]

If this is to equal the true population total, \( \sum m_i \bar{y}_{ip} \), we must have

\[ m_i = \frac{t M_i}{F_i} \] (144)

This means that whatever probabilities are assigned, an unbiased estimate is obtained if \( m_i \) is chosen as in (144). Note that the formula requires knowledge of the true size \( M_i \) for the unit that is selected (though not for any other units). If this is not known in advance, it is counted during the survey. Such counting is usually known as pre-listing.

The variance of the estimate is easily obtained. The error of estimate is

\[ \frac{m_i}{t} \bar{y}_{is} - \bar{y}_{p} = \frac{N}{F_i} \bar{y}_{is} - \frac{m_i}{F_i} \bar{y}_{p} = \frac{M_i}{F_i} (\bar{y}_{is} - \bar{y}_{ip}) + \frac{M_i}{F_i} \bar{y}_{ip} - \bar{y}_{p} \]

Each square receives a weight \( F_i \). Thus

\[ v(m_i \bar{y}_{is}/t) = \sum_{i=1}^{N} \frac{N_i}{F_i} (M_i - m_i) \frac{c_i^2}{m_i} + \sum_{i=1}^{N} \frac{P_i}{F_i} \left( \frac{m_i \bar{y}_{ip}}{P_i} - \bar{y}_{p} \right)^2 \] (145)

If \( P_i = M_i/M \) it will be found that this reduces (apart from the factor \( M^2 \)) to (143) for the variance when probabilities are proportionate to sizes. If \( P_i = 1/N \), (initial probabilities equal), it reduces to formula (142) for the unbiased estimate when probabilities are equal and the subsampling is proportionate.

One comment should be made about the "between units" contribution to the variance (last term on the right of (145)). Unless \( P_i = M_i/M \), i.e., probabilities are proportional to sizes, this term is affected by variations in the \( M_i \) as well as by variations in the unit means \( \bar{y}_{ip} \). This means that if the \( P_i \) are based on
estimated sizes, we do not quite eliminate the effect of variations
in the true sizes from the error variance. If the estimates of size
are good, however, the inflation of the variance from this source is
likely to be small.

8.14 Application in practice: Once the idea is grasped, sampl-
ing with probability proportional to estimated size is not difficult
to apply in practice. Suppose, for instance, that the population
contains six city blocks, of which one is to be selected. The sub-
unit is the household, and we want to sample 5 percent of the popu-
lation so that \( t = 1/20 \). The expected numbers of households (e.n.o.h.)
in the six blocks are as shown below.

<table>
<thead>
<tr>
<th>Block</th>
<th>E.n.o.h.</th>
<th>Cumulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>57</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>82</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>121</td>
</tr>
</tbody>
</table>

We draw a random number between 1 and 121: let this be 96. The
block selected is no. 5, which covers the range from 83 to 105 in
the cumulation.

The sampler visits block no. 5, and counts all the households
in it. Suppose he finds that there are actually 31 households. The
number that he has to enumerate can be found in either of two simple
ways. Since \( P_i = 23/121 \), and \( t = 1/20 \), we may apply (144) and
obtain

\[
M_i = \frac{1.23 \times 121}{20.23} = 8, \text{ to the nearest integer.}
\]

Alternatively, we may note that the sampling rate for the
block chosen, that is, \( m_i/M_i \), is equal to \( t/P_i \). This is known
before the block is pre-listed. Thus the enumerator can be told
in advance the rate at which the block is to be subsampled. This
method is useful when an 'every k th' systematic sample is to be
used for the subsampling. In the present case $t/P_i$ is equal to
121/20.23, or about 1 in 4. After numbering the 31 households which
he finds, the enumerator could choose a random number between 1 and 4,
say 2, and visits the households numbered 2, 6, 10, 14, 18, 22, 26,
and 30 on his list. The reader will notice that we do not choose
$m_i$ so as to satisfy (144) exactly, because of the restriction that $m_i$
must be an integer. The disturbance from this cause will usually be
negligible.

Sometimes no estimates of the $M_i$ are available before the sample
is taken. The best procedure in this case depends on several factors,
of which two are (i) how much it costs to obtain estimates of the $M_i$
and (ii) how much the $M_i$ actually vary. If the cost is high, it may
be best to draw the $M_i$ with equal probability and use the biased esti-
mates of the population mean or total. An interesting case of this
problem is described by Jessen et al (38). They were sampling blocks
in Greek towns, and in some towns had no usable estimates of the
numbers of households in the blocks. They considered three procedures:
(i) drawing the blocks with equal probabilities, (ii) making a rapid
preliminary cruise of the town in order to tie together small blocks
so as to build artificial blocks that appeared to have roughly the
same numbers of households. Also, blocks which obviously had no
households could be eliminated in the process of cruising. The object
is, of course, to diminish the variations in the $M_i$. Blocks would
then be chosen with equal probability. (iii) Cruising the town
slowly enough to permit estimates to be made of the number of house-
holds in each block. Blocks were then chosen with probability pro-
portional to estimated sizes.
8.15 Extension to stratified sampling: The case which we have been discussing is not very practical, in that only one sampling unit is chosen from the population. As these methods are applied in practice, the population is divided into strata, one unit being chosen from each stratum. The formulae for the sampling error variances are built up from the preceding formulae, which will of course apply to a single stratum. The suffix \( j \) denotes the stratum. The following notation is analogous to that previously used.

- \( M_{ij} \): number of sub-units in \( i \) th unit of \( j \) th stratum.
- \( N_j \): total number of sub-units in \( j \) th stratum.
- \( m_{ij} \): number of sub-units sampled in \( i \) th unit of \( j \) th stratum.
- \( N_j \): number of units in \( j \) th stratum.
- \( \sigma_{ij}^2 \): variance within \( i \) th unit of \( j \) th stratum.
- \( \bar{y}_{ijs} \): sample mean in \( i \) th unit of \( j \) th stratum.
- \( \bar{y}_{ijp} \): true mean of \( i \) th unit of \( j \) th stratum.
- \( \bar{y}_{jp} \): true mean of \( j \) th stratum.
- \( \bar{y}_{ju} \): unweighted mean of \( \bar{y}_{ijp} \) within \( j \) th stratum.

We quote the error variances for three procedures for estimating the population total.

I. Units chosen with equal probability \( 1/N_j \) within strata. The estimate is

\[
\bar{y} = \sum_{j} \frac{M_j}{N_j} \bar{y}_{ijs} \quad \text{(biased)}.
\]

II. Units chosen with probability proportional to relative size within strata. \( P_{ij} = M_{ij}/M_j \). The estimate is as in I (unbiased).
\[
V = \sum_j N_j \left[ \frac{N_j}{\sum_{i=1}^{m_{ij}} (M_{ij} - m_{ij})} \frac{c_{ij}}{m_{ij}} + \frac{N_j}{\sum_{i=1}^{m_{ij}} (\bar{y}_{ijp} - \bar{y}_{jp})^2} \right].
\] (147)

III. Sampling rate \(t_j\) in the \(j\) th stratum. Units chosen with arbitrary probabilities \(P_{ij}\). (Adding to 1 within each stratum).

\(m_{ij}\) taken as \(t_j M_{ij} / P_{ij}\). The estimate is

\[
\frac{\sum_j m_{ij} \bar{y}_{ijp}}{t_j} \quad \text{(unbiased)}.
\]

\[
V = \sum_j N_j \left[ \frac{N_j}{\sum_{i=1}^{M_{ij}} \frac{M_{ij} - m_{ij}}{P_{ij}}} \frac{c_{ij}}{m_{ij}} + P_{ij} \left( \frac{N_j \bar{y}_{ijp}}{P_{ij}} \right)^2 \right].
\] (148)

When probabilities are proportional to actual or estimated size, the restriction that only one unit be taken per stratum is not trivial. If more than one unit is chosen per stratum, it is impossible to keep the probability proportional to size unless sampling is done with replacement or by some equivalent device. The simplest method is to use an 'every k th' systematic sample. For instance, suppose that in the example of Section 8.14 we wished to sample 2 of the six city blocks. Since the e.n.o.h. in the population is 121, we could take \(k = 60\), and choose a random number between 1 and 60, say 43. The blocks chosen are those that contain households 43 and 103, i.e., blocks 3 and 5. However, we have in effect divided the population into two strata and taken one unit from each. Consequently, with these methods it is not possible to compute an unbiased sample estimate of the error variance. If pairs of strata can be formed such that there is not much difference between the members of each pair, an estimate that is serviceable may be made from the differences between the two sample means in each pair.
REFERENCES


OTHER METHODS OF ESTIMATION OF A POPULATION TOTAL

9.1 Rather naturally, persons engaged in sampling have favored methods of estimation that can be computed easily and rapidly. Since questionnaires often contain a large number of questions, there is a great advantage in methods of estimation that require little more than simple addition, which can be performed on an IBM tabulator. The potentialities of complex methods of estimation have been little explored. The gain in accuracy from a superior method of estimation may, however, be secured fairly cheaply, since only the final computations are affected, and there are likely to be cases, with certain important estimates, where quite elaborate calculations would be justified if a substantial increase in accuracy resulted. Two methods of estimation which require more calculation than the mean per s.u. estimate, but which usually result in increased accuracy if applicable, are the ratio and the linear regression methods. In these methods, an auxiliary variate \( x \), correlated with \( y \), must be obtained for each unit in the sample. In addition, the population total \( X_p \) of \( x \) must be known. In practice, \( x \) is often the value of \( y \) on some previous occasion when a complete census was taken. The aim in both methods is to obtain increased accuracy by taking advantage of the correlation between \( y \) and \( x \). We consider first simple random sampling.

9.2 The ratio estimate: For the population total, this estimate, which is simple to compute, is:

\[
Y_R = \frac{Y_s}{X_s} X_p
\]  

(148)

where \( Y_s, X_s \) are the sample totals of \( y \) and \( x \). The comparable estimate based on the mean per s.u. is, of course, \( N \bar{Y}_n \), or \( N Y_s/n \).
Theorem 15: The variance of \( Y_R \) in large samples is given approximately as

\[
\sigma(Y_R) = \left[ \frac{N(N-n)}{n} \right] \left[ \frac{(N-1)}{N} \sigma_y^2 + \frac{R_p^2}{R_p} \rho \sigma_y \sigma_x \right]
\]

(149)

where \( R_p = \frac{\bar{y}_p}{\bar{x}_p} \) is the population ratio of \( y \) to \( x \) and \( \rho \) is the correlation coefficient between \( y \) and \( x \). Formula (149) can be shown to be algebraically identical with

\[
N(Y_R) = \left( \frac{N(N-1)}{n} \right) \left( \frac{N-n}{N} \right) \left( \frac{\sigma_y^2}{\bar{y}_p^2} + \frac{\sigma_x^2}{\bar{x}_p^2} - \frac{2 \text{cov}(y, x)}{\bar{y}_p \bar{x}_p} \right).
\]

(150)

Sketch of proof: The result holds as an asymptotic approximation when \( n \) is large and \( n/N \) not too large. A rigorous proof requires fairly advanced mathematics. The following argument is not rigorous in that it does not justify the discarding of certain terms in the analysis.

\[
Y_R = \frac{\bar{y}_n}{\bar{x}_n} = \frac{\bar{y}_p + \Delta y}{\bar{x}_p + \Delta x}
\]

\[
= \frac{\bar{y}_p}{\bar{x}_p} \frac{N \bar{x}_p}{N} \left( 1 + \frac{\Delta y}{\bar{y}_p} \right) \left( 1 + \frac{\Delta x}{\bar{x}_p} \right)^{-1}
\]

\[
= \frac{N \bar{y}_p}{\bar{x}_p} \left( 1 + \frac{\Delta y}{\bar{y}_p} \right) \left( 1 + \frac{\Delta x}{\bar{x}_p} \right)^{-1}
\]

\[
= \frac{N \bar{y}_p}{\bar{x}_p} \left( 1 + \frac{\Delta y}{\bar{y}_p} - \frac{\Delta x}{\bar{x}_p} \right)
\]

approximately, this being the first term in a Taylor series expansion. Therefore,

\[
E(Y_R) = N \bar{y}_p \text{ since } E(\Delta y) = E(\Delta x) = 0.
\]
Now
\[ Y_R - \bar{Y}_R = N \bar{\bar{y}}_p \left( \frac{\Delta y}{\bar{y}_p} - \frac{\Delta x}{\bar{x}_p} \right) \]

where
\[
(\Delta y) = (\bar{y}_n - \bar{y}_p) \frac{2}{n} \\
\mathbb{E} (\Delta y)^2 = \frac{N-n}{N} \frac{\sigma_y}{n}, \text{ etc.}
\]

Therefore, the variance of \( Y_R \) is approximately

\[
V(Y_R) = \mathbb{E}[Y_R - \bar{Y}_R]^2 = \left( \frac{N^2 \bar{y}_p^2}{n} \right) \left( \frac{N-n}{N} \right) \left( \frac{\sigma_y^2}{\bar{y}_p^2} + \frac{\sigma_x^2}{\bar{x}_p^2} - 2 \frac{\mathbb{Cov}(y, x)}{\bar{y}_p \bar{x}_p} \right)
\]

9.3 Estimation from a sample: In the estimation of the variance of \( Y_R \) from a sample, \( \sigma_y^2 \) is estimated by \( s_y^2 = \frac{\sum(y - \bar{y})^2}{n-1} \), \( \sigma_x^2 \) is estimated by

\[
s_x^2 = \frac{\sum(x - \bar{x})^2}{n-1}
\]

and the covariance of \( y \) and \( x \) is estimated by

\[
s_{yx} = \frac{\sum(y - \bar{y})(x - \bar{x})}{n-1}
\]

Therefore, the estimated variance of \( Y_R \) becomes

\[
V(Y_R) = \left( \frac{N^2 \bar{y}_n^2}{n} \right) \left( \frac{N-n}{N} \right) \left( \frac{s_y^2}{\bar{y}_n^2} + \frac{s_x^2}{\bar{x}_n^2} - 2 \frac{s_{yx}}{\bar{y}_n \bar{x}_n} \right)
\]

Alternative forms that are sometimes easier to compute may be developed. From (151),

\[
V(Y_R) = \frac{N(N-n)}{n} \left( s_y^2 + \frac{\bar{y}_n^2 s_x^2}{\bar{x}_n^2} - 2 \frac{\bar{y}_n}{\bar{x}_n} s_{yx} \right)
\]

We may write \( R_s = \frac{\bar{y}_n}{\bar{x}_n} \), or \( Y_s/X_s \), for the sample ratio. Further, it is easy to verify that (151) is the same as

\[
V(Y_R) = \frac{N(N-n)}{n(n-1)} \left[ \sum y_i^2 + R_s^2 \sum x_i^2 - 2 R_s \sum y_i x_i \right]
\]

(153)

where the sums are uncorrected sums of squares or products over the sample. This is often a convenient form for calculation. We may
also write

\[ V(Y_R) = \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} (y_i - \bar{R}_s x_i)^2 \]  

(154)

9.4 Comparison of the ratio estimate and the 'mean per s.u.'
estimate: The variance of the population total, as estimated by the
mean per s.u., is

\[ V(N \bar{y}_n) = N(N-n) \left( \frac{\sigma_y^2}{n} \right) \]  

(155)

Hence from (149), \( V(Y_R) \) is less than \( V(N \bar{y}_n) \) if the following in-
equality is true:

\[ \sigma_y^2 + \frac{R^2}{n} \sigma_x^2 - 2 R \rho \sigma_y \sigma_x < \sigma_y^2 \]

or

\[ 2 \rho_{yx} \sigma_y \sigma_x > \sigma_x^2 \left( \frac{\bar{y}_p}{\bar{x}_p} \right) \]

\[ \rho_{yx} > \frac{1}{2} \frac{\sigma_x}{\bar{y}_p} \frac{\sigma_y}{\bar{x}_p} \]

\[ \rho_{yx} > \frac{1}{2} \frac{\text{coefficient of variation of } x}{\text{coefficient of variation of } y} \]

In general, if the coefficient of variation of \( x \) is greater than
twice the coefficient of variation of \( y \), then the ratio method will
be less efficient than the mean per s.u. method. If \( x \) is the value
of \( y \) on some previous occasion, the two coefficients of variation
may be about equal. In this case, the ratio estimate is superior
if \( \rho \) exceeds \( \frac{1}{2} \).

**Theorem 16:** The ratio estimate is a "best unbiased linear
estimate" if two conditions are satisfied (Cochrane, (39)):

(i) the relation between \( y \) and \( x \) is a straight line through
    the origin, and

(ii) the variance of \( y \) about this line is proportional to \( x \).
Proof: We assume \( N \) infinite. The mathematical model is

\[
y = \beta x + e,
\]

where \( e \) is a random variable with mean zero and variance \( \mu_x \).

Hence,

\[
\bar{y}_p = \beta \bar{x}_p + \bar{e}_p
\]

\[
= \beta \bar{x}_p \quad (\text{since } \bar{e}_p = 0)
\]

By Markoff's theorem (40), the best linear unbiased estimate of \( \bar{y}_p \) (i.e., \( \beta \bar{x}_p \)) is \( b \bar{x}_p \) where \( b \) is the least squares estimate of \( \beta \). This is

\[
b = \frac{\sum w y x}{\sum w x^2}
\]

where

\[
w = \frac{1}{\sigma^2} = \frac{1}{\lambda x}
\]

Hence,

\[
b = \frac{\sum y}{\sum x} = \frac{\sum y}{\sum x}
\]

In exploratory work, a graph plotting the sample values of \( y \) against \( x \) is therefore useful in considering whether the ratio estimate is likely to be the best available.

Where conditions (i) and (ii) are not satisfied, the distribution of the ratio estimate in small samples has not yet been expressed in convenient terms, despite numerous attempts. Unless condition (i) holds, the estimate is biased, Basel (46), though the bias is usually negligible relative to the sampling error. In large samples, the distribution tends to normality with the approximate variance expressed in formula (149). Unfortunately, no simple rule seems to be available for giving the limits of error in the approximate formula.

9.5 Other applications of the ratio estimate: The previous discussion was concerned with the ratio estimate as a means of
estimating the population total of \( y \). Often the purpose of a sample is to estimate ratios. For instance, if the unit is a household, we might wish to estimate the sex-ratio, or the fraction of the population with ages between 5 and 10 years. As an estimate, we can use 
\[ R_s = \frac{Y_s}{X_s} \]
where \( Y_s \) is, say, the number of persons in the sample who are between the ages of 5 and 10, and \( X_s \) is the total number of persons in the sample. Note that both \( X_s \) and \( Y_s \) will vary from household to household, so that the previous analysis and formulae apply to this ratio. Of course, we want the factor \( X_p \) from the estimate, and divide the variance formulae (150) - (154) by \( X_p^2 \), since for such cases we want the variance of \( R_s \) itself.

Ratio estimates are often useful when we have a large unit which contains a varying number of sub-units. For example, the unit might consist of all farms whose farmsteads lie in some area, these areas being delineated on a map so as to cover the population. The sample contains \( n \) areas. To estimate total farm income, we could take the arithmetic mean of the total farm incomes in the different areas, and multiply by \( N \), the population number of areas. If the numbers of farms per area vary greatly, total farm income per area may also do so, with the result that this estimate has a high variance. If the total number of farms in the population is known, an alternative estimate is to divide the total farm income in the sample by the total number of farms in the sample, and multiply by the total number of farms in the population. This estimate is a ratio estimate, since it is of the form \( (Y_s/X_s)X_p \), where \( x \), the number of farms, is a random variable from area to area. Consequently, in computing the variance of the mean per farm estimate, we must use the formulae applicable to a ratio estimate. This fact has sometimes been overlooked.
9.6 Ratio estimates in stratified sampling: There are several ways in which a ratio estimate of a population total \( Y_p \) can be made. One is to make a separate ratio estimate of the total of each stratum and add these totals. If \( Y_{sj} \), \( X_{sj} \), are the sample totals in the \( j \) th stratum and \( X_{pj} \) is the stratum total for \( x \), this estimate

\[
Y_{Rs} = \sum_j \frac{Y_{sj}}{X_{sj}} X_{pj} = \sum_j \left( \frac{\bar{Y}_{nj}}{\bar{X}_{nj}} N_j \bar{X}_{pj} \right)
\]

(156)

It is clear that no assumption is made that the true ratio remains fixed from stratum to stratum: the estimate postulates, however, a knowledge of the separate \( X_{pj} \).

Since sampling is independent in the different strata, the variance is found simply by summation of terms as given in formula (150). This gives

\[
V(Y_{Rs}) = \sum_j \frac{N_j^2}{n_j} \frac{X_{pj}^2}{\bar{X}_{pj}^2} \left[ \frac{\sigma_{yj}^2}{\bar{Y}_{pj}^2} + \frac{\sigma_{xj}^2}{\bar{X}_{pj}^2} - 2 \text{cov} x_j y_j \right]
\]

which may be written as

\[
V(Y_{Rs}) = \sum_j \frac{N_j (N_j - n_j)}{n_j} \left[ \frac{\sigma_{yj}^2}{\bar{Y}_{pj}^2} + \frac{R_{pj}^2 \sigma_{xj}^2}{\bar{X}_{pj}^2} - 2 R_{pj} \sigma_{yj} \sigma_{xj} \right]
\]

(157)

where \( R_{pj} = \frac{\bar{Y}_{pj}}{\bar{X}_{pj}} \) is the true ratio for the stratum.

An alternative estimate, derived from a single combined ratio has been used by Hansen, Hurwitz, and Murrell (41). This is

\[
Y_{Rc} = \frac{\sum_j \frac{N_j}{n_j} Y_{sj}}{\sum_j \frac{N_j}{n_j} X_{sj}} \quad X_{p} = \left( \frac{\sum_j \frac{N_j}{n_j} \bar{Y}_{nj}}{\sum_j \frac{N_j}{n_j} \bar{X}_{nj}} \right) N \bar{X}_{p}
\]

(158)
Write \( \bar{u}_n = \Sigma N_j \bar{y}_{nj} / N \); \( \bar{v}_n = \Sigma N_j \bar{x}_{nj} / N \).

Then \( \mathbb{E}(\bar{u}_n) = \bar{y}_p \); \( \mathbb{E}(\bar{v}_n) = \bar{x}_p \).

If we apply the argument in Theorem 15, page 119, we find

\[
V\left( \frac{\bar{u}_n}{\bar{v}_n} \right) = \frac{N^2 - 2}{\bar{v}_p^2} \left\{ \frac{\text{V}(\bar{u}_n)}{\bar{y}_p^2} + \frac{\text{V}(\bar{v}_n)}{\bar{x}_p^2} - \frac{2 \text{cov}(\bar{u}_n, \bar{v}_n)}{\bar{y}_p \bar{v}_p} \right\}
\]

But, \( \text{V}(\bar{u}_n) = \frac{1}{N^2} \Sigma N_j(N_j - n_j) \frac{\sigma_j^2}{n_j} \),

with corresponding results for \( \text{V}(\bar{v}_n) \) and the covariance.

Hence,

\[
\text{V}(Y_{Re}) = \frac{\bar{y}_p^2}{\bar{v}_p^2} \Sigma \left[ \frac{N_j(N_j - n_j)}{n_j} \right] \left[ \frac{\sigma_j^2}{\bar{y}_p^2} + \frac{\sigma_j^2}{\bar{x}_p^2} - \frac{2 \text{cov}(y, x)}{\bar{y}_p \bar{x}_p} \right]
\]

Formula (159) can be shown to be algebraically identical with

\[
\text{V}(Y_{Re}) = \Sigma \frac{N_j(N_j - n_j)}{n_j} \left[ \sigma_j^2 + \frac{R_p^2 \sigma_j^2}{\bar{x}_p^2} - 2 R_p \rho_j \sigma_j \sigma_x \right]
\]

Formula (160) differs from (157) only in that the single ratio

\[ R_p = \frac{Y_p}{X_p} \]

replaces \( R_{yj} \). To compare (160) with (157) we can write

\[
\text{V}(Y_{Re}) = \text{V}(Y_{Rs}) + \Sigma \frac{N_j(N_j - n_j)}{n_j} \left[ (R_p^2 - R_{yj}^2) \sigma_x^2 - 2(R_p - R_{yj}) \rho_j \sigma_y \sigma_x \right]
\]

\[
= \text{V}(Y_{Rs}) + \Sigma \frac{N_j(N_j - n_j)}{n_j} \left[ (R_{yj} - R_p)^2 \sigma_x^2 + 2(R_{yj} - R_p) \rho_j \sigma_y \sigma_x \right]
\]

\[
- R_{yj} \sigma_x^2
\]

- 125 -
The last term on the right is usually small. (It vanishes if within each stratum the relation between $y$ and $x$ is a straight line through the origin). It follows that unless $R_{pj}$ is constant from stratum to stratum, the use of a separate ratio estimate in each stratum is likely to be more accurate. The advantage appears to be small unless the variation in $R_{pj}$ is marked.

For sample estimates of (157) and (160) we substitute sample estimates of $R_{pj}$ and $R_p$ in the appropriate places. The sample mean squares $s_{yj}^2$ and $s_{xj}^2$ are substituted for the corresponding variances, and the sample covariance for the term $\rho_{yj} s_{yj} s_{xj}$. It will be noted in general, the sample mean square and covariance must be calculated separately for each stratum.

Example: For illustration, we use the data from the Jefferson County, Iowa, study discussed on p. 35. For this example $y$ refers to acres in corn and $x$ to acres in the farm. The population is divided into two strata (instead of seven as in the original example), the first stratum containing farms of size up to 160 acres. We assume a sample of 100 farms. When stratified sampling is used, we assume 70 farms taken from stratum 1 and 30 from stratum 2, this being roughly optimum allocation in the sense of Neyman. The necessary data are given in Table 23.

<table>
<thead>
<tr>
<th>Strata</th>
<th>Size (farm acres)</th>
<th>Nj</th>
<th>$s_{yj}^2$</th>
<th>$s_{yj} s_{xj}$</th>
<th>$s_{xj}^2$</th>
<th>$R_{pj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-160</td>
<td>1580</td>
<td>312</td>
<td>494</td>
<td>2055</td>
<td>.2351</td>
</tr>
<tr>
<td>2</td>
<td>over 160</td>
<td>430</td>
<td>922</td>
<td>858</td>
<td>7619</td>
<td>.2242</td>
</tr>
<tr>
<td>For complete pop.</td>
<td>2010</td>
<td>620</td>
<td>1453</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strata</th>
<th>$\bar{y}_{pj}$</th>
<th>$\bar{x}_{pj}$</th>
<th>nj</th>
<th>$q_j = s_{yj}^2/n_j$</th>
<th>$V_j$</th>
<th>$V'_{j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.51</td>
<td>82.56</td>
<td>70</td>
<td>.008288</td>
<td>194</td>
<td>193</td>
</tr>
<tr>
<td>2</td>
<td>51.63</td>
<td>244.85</td>
<td>30</td>
<td>.001525</td>
<td>287</td>
<td>287</td>
</tr>
<tr>
<td>For c.p.</td>
<td>26.30</td>
<td>117.28</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We consider five methods of estimating the population mean corn acres per farm. The f.p.c. will be ignored.

(i) Simple random sample: mean per farm estimate.

\[ V_1 = \frac{\sigma_y^2}{n} = \frac{620}{100} = 6.20. \]

(ii) Simple random sample: ratio estimate.

\[ V_2 = \frac{1}{n} \left[ \frac{\sigma_y^2 + \frac{R_x \sigma_x^2}{p} - 2 R_{xy} \sigma_y}{\sigma_y} \right] \]

\[ = \frac{1}{100} \left[ 620 + (.2242)^2 (7619) - 2(.2242)(1453) \right] \]

\[ = 3.51 \]

(iii) Stratified random sample: mean per farm estimate.

\[ V_3 = \frac{1}{N^2} \sum \frac{N_j^2}{n_j} \sigma_{yj}^2 = \sum Q_j \sigma_{yj}^2 = 4.16. \]

(iv) Stratified random sample: ratio estimate using a separate ratio in each stratum.

\[ V_4 = \sum Q_j \left[ \frac{\sigma_{yj}^2 + \frac{R_{xj} \sigma_{xj}^2}{p_j} - 2 R_{yj} \sigma_{yj}}{\sigma_{yj}} \right] = \sum Q_j V_j = 3.07. \]

(v) Stratified random sampling: Ratio estimate using a combined ratio.

\[ V_5 = \sum Q_j \left[ \frac{\sigma_{yj}^2 + \frac{R_x \sigma_x^2}{p} - 2 R_{xj} \sigma_{yj}}{\sigma_{yj}} \right] = \sum Q_j V_j = 3.09. \]

The relative information obtained by the various methods can be summarized as follows:

<table>
<thead>
<tr>
<th>Sampling Method</th>
<th>Method of Estimation</th>
<th>R. I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Simple random</td>
<td>Mean per s.u.</td>
<td>100</td>
</tr>
<tr>
<td>(ii) Simple random</td>
<td>Ratio</td>
<td>177</td>
</tr>
<tr>
<td>(iii) Stratified random</td>
<td>Mean per s.u.</td>
<td>149</td>
</tr>
<tr>
<td>(iv) Stratified random</td>
<td>Separate ratio</td>
<td>202</td>
</tr>
<tr>
<td>(v) Stratified random</td>
<td>Combined ratio</td>
<td>201</td>
</tr>
</tbody>
</table>

The results bring out an interesting point that is of rather
general application. Stratification by size of farm accomplishes the same purpose as the use of a ratio estimate on farm size, namely to eliminate the effect of variations in farm size from the sampling error. For instance, the gain from a ratio estimate is 77 percent when simple random sampling is used, but is only 35 percent (202 against 149) when stratified sampling is used. In fact, in the original example on p. 35, where seven strata were used, the variance of the mean per farm estimate was seen to be 2.90, which is lower than any of the variances above. With seven strata, there is no further gain from the use of a ratio estimate over a mean per farm estimate.

Consequently, in the design of samples one often may choose whether to introduce some factor into the stratification, or to utilize it in the method of estimation, or perhaps to use it in both ways. The best decision will depend on the circumstances. Relevant points are: (i) some factors, e.g., geographical location, are more easily introduced into the stratification than into the method of estimation; (ii) the issue depends on the relation between \( y \) and \( x \). All simple methods of estimation work most effectively with a linear relation. With a complex or discontinuous relation, stratification may be more effective, since if there are enough strata, stratification will eliminate the effects of almost any kind of relation between \( y \) and \( x \).

9.7 Optimum Allocation with a ratio estimate: The optimum allocation of the \( n_j \) may be different when a ratio estimate is used than when a mean per s.u. is used. In discussing this point, we shall use formula (157) on the assumption that in practice, it will differ little from (160). The quantity \( \sigma^2 = \sigma^2_{y_j} + R^2_{pj} \sigma^2_{x_j} - 2 R_{pj} \rho_j \sigma_{y_j} \sigma_{x_j} \) is the variance within the \( j \)th stratum of the variate \( d = (y - R_{pj}x) \). This variance will be denoted by \( \sigma^2_{dj} \). If (157) is minimized subject to a total cost of the form \( \sum c_j n_j \), it is found that the \( n_j \) must be
chosen proportional to \( \frac{N_j \sigma_{dj}}{\sqrt{c_j}} \), whereas with a mean per s.u. estimate, \( n_j \) is chosen proportional to \( \frac{N_j \sigma_{yj}}{\sqrt{c_j}} \).

In the case where the ratio estimate is a best unbiased linear estimate, \( \sigma_{dj} \) will be proportional to \( \sqrt{x} \). The \( n_j \) would then be made proportional to \( \frac{N_j \sqrt{\bar{x}_{pj}}}{\sqrt{c_j}} \). In other cases the variance of \( d \) may be more nearly proportional to \( x^2 \). This leads to the allocation of \( n_j \) proportional to \( \frac{N_j \bar{x}_{pj}}{\sqrt{c_j}} \), that is, to the stratum total of \( x \), divided by the square root of the unit cost. An example of the latter case is discussed by Hansen, Hurwitz, and Gurney (41) for a sample designed to estimate retail store sales.

**Example:** The different methods of allocation can be compared using data collected in a complete enumeration of 256 commercial peach orchards in the Sandhills area of North Carolina in June 1946. The purpose of this survey was to determine the most efficient sampling procedure for estimating commercial peach production in this area. Information was obtained on the number of peach trees per orchard and estimated total peach production. The high correlation between these two variables suggested the use of a ratio estimate. For this illustration, the area was divided geographically into three strata. The number of peach trees in an orchard is denoted by \( x \) and the expected production in bushels of peaches by \( y \). Only the first ratio estimate \( y_{Rs} \) (based on a separate ratio in each stratum) will be used since the principle is the same for both types of stratified ratio estimates. Four different methods of allocation will be compared: (i) \( n_j \) proportional to \( N_j \), (ii) \( n_j \) proportional to \( N_j \sigma_{yj} \), (iii) \( n_j \) proportional
to \( N_j \sqrt{x_{pj}} \), and, (iv) \( n_j \) proportional to \( N_j \sqrt{x_{pj}} \). A sample size of 100 will be considered. The data needed for these comparisons are summarized in Table 24.

**TABLE 24.**

**DATA FROM THE NORTH CAROLINA PEACH SURVEY**

<table>
<thead>
<tr>
<th>Strata</th>
<th>( \sigma_{xj}^2 )</th>
<th>( \sigma_{yj}^2 )</th>
<th>( \sigma_{yj}^2 )</th>
<th>( x_{j} )</th>
<th>( y_{j} )</th>
<th>( \bar{x}_{j} )</th>
<th>( \bar{y}_{j} )</th>
<th>( R_{j} )</th>
<th>( V_{j} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5186</td>
<td>6462</td>
<td>8699</td>
<td>72.01</td>
<td>93.27</td>
<td>53.80</td>
<td>69.48</td>
<td>1.29133</td>
<td>658</td>
</tr>
<tr>
<td>2</td>
<td>2387</td>
<td>3100</td>
<td>4614</td>
<td>48.65</td>
<td>67.93</td>
<td>31.07</td>
<td>45.64</td>
<td>1.40475</td>
<td>573</td>
</tr>
<tr>
<td>3</td>
<td>4877</td>
<td>4817</td>
<td>7311</td>
<td>69.33</td>
<td>85.51</td>
<td>56.97</td>
<td>66.39</td>
<td>1.16547</td>
<td>2706</td>
</tr>
<tr>
<td>Total</td>
<td>3898</td>
<td>4434</td>
<td>6409</td>
<td>62.43</td>
<td>80.06</td>
<td>44.45</td>
<td>56.47</td>
<td>1.27053</td>
<td>1433</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strata</th>
<th>( N_j )</th>
<th>( n_j )</th>
<th>( N_j \sqrt{\bar{x}_{pj}} )</th>
<th>( \bar{x}_{pj} )</th>
<th>( \bar{y}_{pj} )</th>
<th>( \bar{z}_{pj} )</th>
<th>( \bar{z}_{pj} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47</td>
<td>18</td>
<td>4384</td>
<td>22</td>
<td>7.33</td>
<td>344.5</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>113</td>
<td>46</td>
<td>5016</td>
<td>40</td>
<td>5.57</td>
<td>657.3</td>
<td>39</td>
</tr>
<tr>
<td>3</td>
<td>91</td>
<td>36</td>
<td>7781</td>
<td>38</td>
<td>7.55</td>
<td>687.1</td>
<td>41</td>
</tr>
<tr>
<td>Total</td>
<td>256</td>
<td>100</td>
<td>20181</td>
<td>100</td>
<td>20.45</td>
<td>1698.9</td>
<td>100</td>
</tr>
</tbody>
</table>

The upper part of the table shows the basic data. The lower part gives the calculations needed to obtain the four different types of allocation. The actual values of the \( n_j \) for each type appear in the columns headed (i) - (iv) respectively.

From (157),

\[
V(Y_{Re}) = \sum_{j} N_j \left( n_j - n_j \right) V_j, \quad \text{where} \quad V_j = \sigma_{yj}^2 + R_{pj}^2 \sigma_{xj}^2 - 2 R_{pj} \sigma_{yxj}.
\]

Note that the quantities \( V_j \) are the same for all four allocations; they are given at the extreme right of the top half of Table 24.

The variances and relative information for the different methods are shown in Table 25.
### TABLE 25.

**Comparison of Four Methods of Allocation**

<table>
<thead>
<tr>
<th>Method of Allocation</th>
<th>Variance</th>
<th>Relative Information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strata 1</td>
<td>Strata 2</td>
</tr>
<tr>
<td>nj proportional to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nj</td>
<td>49,824</td>
<td>105,033</td>
</tr>
<tr>
<td>Nj (\sigma_{yj}^2)</td>
<td>35,144</td>
<td>131,847</td>
</tr>
<tr>
<td>Nj (\sqrt{x_{pj}})</td>
<td>41,750</td>
<td>136,964</td>
</tr>
<tr>
<td>Nj (\bar{z}_{pj})</td>
<td>35,144</td>
<td>181,710</td>
</tr>
</tbody>
</table>

There is not a great deal to choose between the different allocations, as would be expected since the \(n_j\) do not differ greatly in the four methods. Method (iv), in which allocation is proportional to the total number of peach trees in the stratum, is the best.

#### 9.8 The linear regression estimate

We assume that the sample is a simple random sample. To utilize this estimate, we first compute from the sample the least squares regression coefficient \(b\) of \(y\) on \(x\), where

\[
b = \frac{\sum (y - \bar{y}_n) (x - \bar{x}_n)}{\sum (x - \bar{x}_n)^2}.
\]

The estimate of the population mean of \(y\) is then taken as

\[
\bar{y}_{Lr} = \bar{y}_n + b(\bar{x}_p - \bar{x}_n)^2
\]

The sample arithmetic mean \(\bar{y}_n\) is adjusted for the difference between the mean value of \(x\) in the population and that in the sample. The estimate requires a knowledge of the total number \(N\) of units and of the population total of \(x\).

#### 9.9 Variance of the estimate

To develop the elementary theory, we assume that \(N\) is infinite and that

\[
y = \alpha + \beta (x - \bar{x}_p) + e,
\]

(162)
where \( e \) is a random variable with mean zero for any \( x \) and constant variance \( \sigma_e^2 \).

It follows from (162) that \( \bar{y}_p = \alpha \). Further algebraic consequences of (162) are

\[
\bar{y}_n = \alpha + \beta (\bar{x}_n - \bar{x}_p) + \bar{e}_n,
\]

\[
b = \beta + \frac{\sum e (x - \bar{x}_n)}{\sum (x - \bar{x}_n)^2},
\]

consequently, the error of estimate, \( (\bar{y}_{LR} - \bar{y}_p) \), will be found to be

\[
\bar{e}_n + (\bar{x}_p - \bar{x}_n) \frac{\sum e (x - \bar{x}_n)}{\sum (x - \bar{x}_n)^2}.
\]

If the \( x \)'s are regarded as fixed from sample to sample, this is a linear function of the \( e \)'s. Since the mean value of \( e \) is zero, we conclude that the regression estimate is unbiased. From the formula for the variance of a linear function, the variance of the estimate works out as

\[
V(\bar{y}_{LR}) = \sigma_e^2 \left[ \frac{1}{n} + \frac{(\bar{x}_p - \bar{x}_n)^2}{\sum (x - \bar{x}_n)^2} \right],
\]

\[
= \frac{c_y^2 (1 - \rho^2)}{n} \left[ 1 + \frac{n(\bar{x}_p - \bar{x}_n)^2}{\sum (x - \bar{x}_n)^2} \right],
\]

where \( \rho \) is the correlation coefficient between \( y \) and \( x \).

The sample estimate of this variance is obtained by substituting for \( c_y^2 (1 - \rho^2) \) the mean square of the deviations of \( y \) from the sample regression on \( x \) (following the usual regression rule, we assign \( n-2 \) degrees of freedom to the sum of squares of deviations).

It will be observed in (167) that the variance depends on the set of \( x \) values that happen to turn up in the sample. This fact does not hinder the practical use of the formula, since all the \( x \) values
that appear in (167) are known when the sample has been drawn. For comparison with other estimates, however, the average variance of the regression estimate under random sampling is needed. From (167) this clearly depends on the form of the frequency distribution of the \(x\)'s. The mean value of (167) may be expanded in a series of inverse powers of \(n\), the sample size. Retaining the two leading terms we obtain

\[
\overline{V}(\overline{y}_{lr}) = \frac{c_y^2 (1 - \rho^2)}{n} \left[ 1 + \frac{1}{n} + \frac{3 + 2\gamma_1^2}{n^2} \right]
\]  

(168)

where \(\gamma_1\) is Fisher's (42) measure of relative skewness \((\gamma_1^2 = k_3^2/k_2^3)\).

If the \(x\)'s were normally distributed, \(\gamma_1\) would be zero, and the exact value for the term in brackets would be \((n-2)/(n-3)\).

If \(n\) is reasonably large, we may regard the factor in brackets as unity. This gives

\[
\overline{V}(\overline{y}_{lr}) = \frac{c_y^2 (1 - \rho^2)}{n}
\]  

(169)

The preceding theory is rather restricted in its scope, since it assumes (i) that the true regression is linear, (ii) that the deviations from the regression have a constant variance, and, (iii) that \(N\) is infinite. With regard to (i) and (ii), it may be shown that if \(n\) is large enough so that terms in \(1/n\) are negligible, formula (169) still holds even if the true regression is not linear and the residual variance depends on \(x\). (Cochran, (39)). For small values of \(\gamma_1\), the preceding theory would require some modification.

When the finite size of population is taken into account, the regression estimate is slightly biased, though the bias is unimportant so far as practical use is concerned. The effect on the variance is approximately to multiply it by the usual factor \((N-n)/N\).

The preceding discussion referred entirely to the estimation of the population mean. To estimate the population total, we multiply

\[
\]
the estimate of the mean by N and its variance by $n^2$.

9.10 Comparison with the ratio estimate and the mean per s.u.: 

For these comparisons we assume the sample size n sufficiently large so that formula (169) may be used, and that the approximate formula for the variance of the ratio also is valid. The three comparable variances are:

\[ \text{V}(\bar{y}_{Lr}) = \frac{(N-n)}{N} \frac{\sigma_y^2 (1-\rho^2)}{n}, \quad \text{(regression)} \]

\[ \text{V}(\bar{y}_R) = \frac{(N-n)}{Nn} \left( \frac{\sigma_y^2}{\rho^2} - 2 \frac{\rho}{\rho^2} \frac{\sigma_y \sigma_x}{\rho^2} + \frac{\rho^2}{\rho^2} \sigma_x^2 \right), \quad \text{(ratio)} \]

\[ \text{V}(\bar{y}_n) = \frac{(N-n)}{Nn} \sigma_y^2, \quad \text{(mean per s.u.)} \]

It is obvious that the variance of the regression estimate is smaller than that of the mean per s.u. unless $\rho = 0$, in which case the two variances are equal.

Further, the variance of the regression estimate is less than that of the ratio estimate if

\[ -\frac{\sigma_y^2 \rho^2}{\sigma_x^2} \leq -2 \frac{\rho}{\rho^2} \frac{\sigma_y \sigma_x}{\rho^2} + \frac{\rho^2}{\rho^2} \sigma_x^2, \]

where we have written $\rho \sigma_y \sigma_x$ for $\sigma_{yx}$. This is equivalent to

\[ 0 \leq (\rho \sigma_y - R_p \sigma_x)^2. \]

Therefore the regression estimate is more accurate than the ratio estimate unless:

\[ \rho = R_p \frac{\sigma_x}{\sigma_y} = \text{coefficient of variation of } x \text{ coefficient of variation of } y \]

(170)

in which case the two have equal variances. Equation (170) holds whenever the relation between y and x is a straight line through the origin, so that in this event, the regression and ratio estimates are equally
accurate. It is interesting to note that the regression estimate is as good as the ratio estimate, even when the latter is a best unbiased estimate.

The regression estimate is more laborious to compute, principally owing to the work in calculating $b$. If there is an appreciable saving in time, an inefficient estimate of $b$ can often be used instead of the least squares estimate. If the estimate of $b$ has an efficiency $E$, ($E < 1$), the fractional increase in the variance of the regression estimate of $\hat{\beta}^2$ is about $\frac{(1-E)}{nE}$. With large $n$, even a highly inefficient estimate of $b$ causes only a trivial increase in the variance.

A simple method for obtaining an estimate of $b$ has been proposed by Hendricks and was used by Finkner, Morgen, and Monroe (29). Under this system, the sampling units are separated into two approximately equal groups on the basis of size. Averages are then computed for each group for both $x$ and $y$. The estimate of $b$ then becomes

$$b = \frac{\bar{y}_1 - \bar{y}_s}{\bar{x}_1 - \bar{x}_s}$$

where $\bar{y}_1$ and $\bar{x}_1$ are the respective means of the group containing the larger sampling units and $\bar{y}_s$ and $\bar{x}_s$ are the means of the group containing the smaller sampling units.

It should be remembered that with the least squares estimate of $b$, one can obtain an unbiased sample estimate of $\sigma^2_y (1 - \rho^2)$ very quickly, whereas with other estimates of $b$, the 'short cut' calculation of the sample residual mean square does not apply.

**Example:** The accuracy of the regression, ratio, and mean per s.u. estimate from a simple random sample can be compared using data collected in the complete enumeration of commercial peach orchards described on page 129. In this example, $y$ is the estimated peach
production of an orchard and \( x \) the number of peach trees in the orchard. The relevant data are \( \sigma_y^2 = 6409, \sigma_{yx} = 4434, \sigma_x^2 = 3893, \)
\( R_p = 1.270, \bar{r} = .887, n = 100, N = 256. \)

\[
V(\bar{y}_{Lr}) = \frac{N-n}{N} \frac{\sigma_y^2 (1-R^2)}{n} \left[ 1 + \frac{1}{n-3} \right]
\]
\[
= \left( \frac{256-100}{256} \right) \frac{6409 \left( 1 - .787 \right)}{100} \left[ 1 + \frac{1}{97} \right]
\]
\[
= 8.40
\]

\[
V(\bar{y}_{R}) = \frac{N-n}{N} \frac{1}{n} \left[ \sigma_y^2 + R_p^2 \sigma_x^2 - 2 R_p \sigma_{yx} \right]
\]
\[
= \left( \frac{256-100}{256} \right) \frac{1}{100} \left[ 6409 + (1.613) (3893) - (2) (1.270) (4434) \right]
\]
\[
= 8.74
\]

\[
V(\bar{y}_{n}) = \frac{N-n}{N} \frac{\sigma_y^2}{n}
\]
\[
= \left( \frac{256-100}{256} \right) \left( \frac{6409}{100} \right)
\]
\[
= 39.05
\]

There is little to choose between the regression and ratio estimates, as might be expected from the nature of the variables. Both techniques are greatly superior to the mean per s.u.

The relative efficiencies of the three methods of estimation are

<table>
<thead>
<tr>
<th>Method</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean per s.u.</td>
<td>100%</td>
</tr>
<tr>
<td>Ratio</td>
<td>447%</td>
</tr>
<tr>
<td>Regression</td>
<td>465%</td>
</tr>
</tbody>
</table>
REFERENCES


DOUBLE SAMPLING

10.1 As we have seen, the use of ratio or regression estimates requires a knowledge of the true population mean of the auxiliary variable \( x \). Similarly, if it is desired to stratify the population according to the values of \( x \), a knowledge of the number of units in the population that have \( x \) values between specified limits is needed. This demands detailed information about the frequency distribution of \( x \) in the population. Quite often such information is lacking, or is known only roughly, for \( x \) variables that we would like to use in this way.

It may happen that \( x \) can be measured relatively cheaply by a sample. In this case, even though the purpose of a survey is to estimate a number of \( y \) variates, it may pay to devote part of the funds to a large preliminary sample in which \( x \) alone is measured. From this sample we can make a good estimate of the population mean of \( x \), if a ratio or regression estimate is envisaged. Alternatively, we can make good estimates of the population numbers \( N_j \) in strata based on the distribution of \( x \). Of course, by devoting funds to a special sample for \( x \), we must cut down the size of the main survey on \( y \). Consequently, the technique will increase accuracy only if the gain in accuracy from ratio or regression estimates or from stratification more than offsets the loss due to the reduction in size of the main sample.

A simple application has been given by Watson (43). The problem was to estimate the mean leaf area of the leaves on a plant. The determination of the area of a leaf by planimeter is rather tedious. However, there is a close correlation between leaf area and leaf weight, and it is very easy to determine the mean weight per leaf for a number of leaves. The procedure is therefore to weigh all the leaves on the plant (so that in this case the large sample is the complete population). A small sample of leaves is then selected for the determination of leaf
areas. These are later adjusted by means of the regression of area on weight. A similar application which uses eye estimates in timber cruising has been mentioned by Cochran (44), and applications to the estimation of forage yields by Wilm et al (45).

10.2 Case where the \( x \) variate is used for stratification: The theory for this case was first given by Neyman (46). The following discussion covers much the same ground, though in considerably less detail.

We wish to stratify the population into a number of classes according to the value of \( x \). Let \( W_j = N_j/N \) be the true (though unknown) proportion of the population that falls in the \( j \) th stratum. The first sample is a random sample of size \( L \), and \( w_j = L_j/L \) is the proportion of \( x \) values found in the \( j \) th stratum. Thus \( w_j \) is an estimate of \( W_j \), and the \( w_j \) follow the usual multinomial distribution. (We assume that the true number \( N_j \) in any stratum is so large that it may be considered infinite).

The second sample is a stratified random sample in which \( y \) is measured; \( n_j \) units are drawn from the \( j \) th stratum. As usual, the variance within the \( j \) th stratum is denoted by \( c_j^2 \). In the simplest case, the cost of the two samples will be of the form

\[
C = c_1L + c_2n, \tag{171}
\]

where \( c_2 \) is presumed large relative to \( c_1 \).

The problem is to choose \( L \) and the \( n_j \) (and consequently \( n \)) so as to minimize the variance of the estimate for a given cost. We must then verify whether the minimum variance is smaller than can be attained by the use of a single random sample in which \( y \) alone is measured.

The first step is to set up the estimate and determine its variance. The true population mean is
\[ \sum_j W_j \bar{y}_{pj} . \]

As estimate we use
\[ \sum_j w_j \bar{y}_{sj} . \]

Note that \( w_j \) and the sample means \( \bar{y}_{sj} \) are both subject to error. The problem is one of stratification where the strata totals are not known exactly. Write
\[ v_j = W_j + u_j ; \quad \bar{y}_{sj} = \bar{y}_{pj} + e_j . \]

Then the error of estimate may be expressed as
\[ \sum_j (v_j \bar{y}_{sj} - W_j \bar{y}_{pj}) = \sum_j (W_j e_j + u_j \bar{y}_{pj} + u_j e_j) \quad (172) \]

Since \( u_j \) and \( e_j \) are independently distributed, and since each has mean value zero, it follows from (172) that the estimate is unbiased.

The variance is a little troublesome. When we square (172) and take the expectation, there will be contributions from squared terms, and from cross-product terms between different strata. Consider first the squared terms. These are
\[ \mathbb{E} \left[ \sum_j \left( W_j e_j + u_j \bar{y}_{pj} + u_j e_j \right)^2 \right] \]
\[ = \sum_j \left[ \frac{W_j^2 \mathbb{E} (e_j^2) + \bar{y}_{pj}^2 \mathbb{E} (u_j^2) + \mathbb{E} (u_j^2) \mathbb{E} (e_j^2) }{n_j} \right] , \]
all other terms vanishing when the expectation is taken. This gives
\[ \sum_j \left[ \frac{W_j^2 \sigma_j^2}{n_j} + \frac{-2 \bar{y}_{pj} \sigma_j (1-W_j)}{L} + \frac{W_j(1-W_j)}{L} \cdot \sigma_j^2 \right] \quad (173) \]

Now consider cross-product terms between different strata. If \( j \) and \( k \) refer to two strata, there is no contribution from terms of the form \( \sigma_j \sigma_k \), since sampling is independent in different strata.

The only contribution is that from terms in \( u_j u_k \). For the multinomial distribution,

\[
E(u_j u_k) = -w_j w_k / L,
\]

so that the cross-products contribute

\[
\sum_{j \geq k} -2 \bar{v}_{pj} \bar{v}_{pk} w_j w_k / L.
\]  

(174)

If the middle term in (173) is combined with (174), the reader may verify that these together amount to

\[
\sum w_j (\bar{v}_{pj} - \bar{v}_p)^2 / L.
\]

Hence, the final form of the variance is

\[
V = \sum_j \left[ \left( \frac{w_j}{L} + \frac{w_j (1-w_j)}{L} \right) + \frac{\sigma_j^2}{n_j} + \frac{w_j (\bar{v}_{pj} - \bar{v}_p)^2}{L} \right].
\]

(175)

The term free from \( L \) is the familiar expression for the variance when the stratum sizes are known exactly. The effects of the errors in the sample are therefore to increase the within-stratum contribution to the variance and to introduce a between-stratum component.

A considerable amount of information about the population is required in order to use this result. Estimates are needed both of the within-stratum variances and of the effectiveness of stratification.

The values of the \( n_j \) and \( L \) that lead to the minimum variance are rather complicated. It is clear that \( n_j \) should be proportional to

\[
\sigma_j \sqrt{\frac{w_j^2 + \frac{w_j (1-w_j)}{L}}{}}.
\]

Since the second term inside the root will usually be small compared with the first, Neyman suggests taking \( n_j \) proportional to \( w_j \sigma_j \), as a
first approximation. Thus

\[ n_j = n \frac{(W_j \sigma_j)}{\sum (W_j \sigma_j)} . \]

If this value is substituted into (175), with the term in \( W_j (1-W_j) \)

ignored, we obtain

\[ V' = \frac{(\sum W_j \sigma_j)^2}{n} + \sum \frac{W_j (\bar{y}_{pj} - \bar{y}_p)^2}{L} \]  

(176)

\[ = \frac{a}{n} + \frac{b}{L} \quad \text{(say)} . \]  

(177)

If this approximate form of the variance is minimized by choice

of \( n \) and \( L \) for a given cost of the form (171), it is easily found that

\[ \frac{n}{L} = \left[ \frac{ac_1}{bc_2} \right]^{-\frac{1}{2}} . \]  

(178)

This equation with (171) serves to determine \( n \) and \( L \).

Example: This example is artificial, but will give some idea of

the calculations involved. We use the Jefferson, Iowa, data previously

considered (page 126). The \( x \) variate, farm size, is to be used to

divide the population into two strata: farms up to 160 acres and farms

over 160 acres. Assume that it costs 10 times as much to sample for

corn acres \( (y) \) as for farm size \( (x) \), and let the cost be of the form

\[ C = 100 = 0.1L + n . \]  

(179)

This means that if double sampling is not used \( (L = 0) \), we can afford
to take a sample of 100 farms to estimate corn acres.

The relevant data for the population are:

<table>
<thead>
<tr>
<th>Strata</th>
<th>( W_j )</th>
<th>( \sigma_j^2 )</th>
<th>( \sigma_j )</th>
<th>( \bar{y}_{pj} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.786</td>
<td>312</td>
<td>17.7</td>
<td>19.404</td>
</tr>
<tr>
<td>2</td>
<td>0.214</td>
<td>222</td>
<td>30.4</td>
<td>51.626</td>
</tr>
<tr>
<td>Complete</td>
<td>620</td>
<td>26.297</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We find

\[ a = (\Sigma W_j \sigma_j)^2 = 417, \]

\[ b = \Sigma W_j \left( \bar{x}_p - \frac{\bar{x}}{\bar{y}} \right)^2 = 175 \]

so that

\[ \frac{n}{L} = \sqrt{\frac{417}{175}} \cdot \frac{1}{\frac{1}{10}} = 4.98 . \]

From the cost equation (179) we obtain

\[ L = \frac{100}{.588} = 170 \quad ; \quad n = 170 \times 4.98 = 83 . \]

At this point we may verify that the neglected term in \( W_j(1-W_j) \) in (175) is in fact negligible. From (177) we then have

\[ V_{\text{min}} = \frac{417}{83} + \frac{175}{170} = 5.02 + 1.03 = 6.05 \]

For a random sample of size 100, with no double sampling, we would have

\[ V = \frac{520}{100} = 6.20 \]

From this it appears that there would be only a trifling gain from double sampling.

10.3 Case where the \( x \) variate is used for regression: In most of the applications that have appeared in the literature, the \( x \) variate has been used to make a regression rather than a ratio estimate. For this reason the regression case will be discussed. We assume that the population is infinite and that

\[ y = \alpha + \beta (x - \bar{x}_p) + e \quad (180) \]

where \( e \) has mean zero and variance \( \sigma_e^2 = \sigma_y^2 (1 - \rho^2) \). In the first (large) sample, of size \( L \), we measure only \( x \); in the second, of size \( n \), we measure both \( x \) and \( y \). The estimate of \( \bar{y}_p \) is

\[ \bar{y}_{ds} = \bar{y}_n + b(\bar{x}_L - \bar{x}_n) \quad (181) \]
where \( b \) is the least squares regression coefficient of \( y \) on \( x \), computed from the small sample.

As an algebraic consequence of (180) it will be found that the error of estimate

\[
y_{ds} - \bar{y}_p = \bar{e}_n + (\bar{y}_L - \bar{y}_n) \frac{\Sigma e (x - \bar{x}_n)}{\Sigma (x - \bar{x}_n)^2} + \beta (\bar{x}_L - \bar{x}_p)
\]

If we consider the \( x \) values fixed in both the small and the large sample, the last term on the right remains fixed and might be regarded as a bias. For fixed \( x \)'s, the variance

\[
E(y_{ds} - \bar{y}_p^2) = \sigma_y^2 (1 - \rho^2) \left[ \frac{1}{n} + \frac{(\bar{x}_L - \bar{x}_n)^2}{\Sigma (x - \bar{x}_n)^2} \right] + \beta^2 (\bar{x}_L - \bar{x}_p)^2 . \tag{182}
\]

As is typical of regression formulae, the variance depends on the sets of \( x \) values that happen to turn up. For comparison with other sampling methods, we would like an average variance in repeated sampling from the same population. This average presents some difficulty.

An average can be obtained if we assume (i) that the large sample is drawn at random, (ii) that the small sample is a random sample drawn from the large sample and, (iii) that the \( x \)'s are normally distributed.

The value of the average is

\[
\bar{E} (y_{ds} - \bar{y}_p^2) = \sigma_y^2 (1 - \rho^2) \left[ \frac{1}{n} + \left( \frac{1}{n} - \frac{1}{L} \right) \frac{1}{(n-3)} \right] + \frac{\beta^2 \sigma_x^2}{L} , \tag{183}
\]

which may be re-written

\[
\frac{\sigma_y^2 (1 - \rho^2)}{n} \left[ 1 + \frac{(L-n)}{L} \frac{1}{(n-3)} \right] + \frac{\rho^2 \sigma_y^2}{L} . \tag{184}
\]

If the \( x \)'s are not normally distributed, the only term affected is that in \( 1/(n-3) \), as discussed previously on page 133. As regards assumption (ii), it is rather unlikely that the small sample would be drawn at random from the large sample. Instead, we would usually
draw the small sample so as to obtain a wide spread in the values of $x_i$ and so reduce the contribution from the sampling error of $b$. The effect would be to reduce (perhaps considerably) the term in $1/(n-3)$; the exact amount of the reduction would require further investigation.

The best method of estimating the variance from a sample (or rather the two samples) is likewise not too clear. Formula (182) is not usable as it stands. The sample mean square deviation $s_{y,x}^2$ from the regression is an unbiased estimate of $\sigma_y^2 (1 - \rho^2)$. But we do not know the value of $(\bar{x}_L - \bar{x}_P)^2$. It seems necessary to use the following hybrid of (182) and (184),

$$v = \sigma_y^2 (1 - \rho^2) \left[ \frac{1}{n} + \frac{(\bar{x}_L - \bar{x}_n)^2}{\Sigma (x - \bar{x}_n)^2} \right] + \frac{s_y^2}{L} \quad . \quad (185)$$

Since $s_{y,x}^2$ is an unbiased estimate of $\sigma_y^2 (1 - \rho^2)$ and since

$$s_y^2 = \Sigma (y - \bar{y}_n)^2 / (n-1)$$

is an unbiased estimate of $\sigma_y^2$, it follows that

$$\Sigma (s_y^2 - s_{y,x}^2) = \rho^2 \sigma_y^2$$

Hence, for a sample estimate of the variance we can use

$$s_{y,x}^2 \left[ \frac{1}{n} + \frac{(\bar{x}_L - \bar{x}_n)^2}{\Sigma (x - \bar{x}_n)^2} \right] + \frac{(s_y^2 - s_{y,x}^2)}{L} \quad . \quad (186)$$

If the f.p.c. is introduced, this formula becomes changed to

$$s_{y,x}^2 \left[ \left( \frac{1}{n} - \frac{1}{N} \right) + \frac{(\bar{x}_L - \bar{x}_n)^2}{\Sigma (x - \bar{x}_n)^2} \right] + (s_y^2 - s_{y,x}^2) \left( \frac{1}{L} - \frac{1}{N} \right) \quad . \quad (187)$$

A development of the theory has been given by Chandali Bose (47).

She notes that in some applications the small sample may be drawn
quite separately from the large sample. This changes the term in

\[
\left( \frac{1}{n} - \frac{1}{L} \right)
\]

in (183) to

\[
\left( \frac{1}{n} + \frac{1}{L} \right)
\]

with a corresponding change in (184).

REFERENCES


ADDITIONAL NOTES

These notes, which cover a few topics not discussed in preceding sections, are intended mainly to indicate further reading.

11.1 Extension of the general principle: The principle of maximum accuracy for given cost, or minimum cost for given accuracy, is not completely satisfactory. The principle assumes that in some way either the cost or the accuracy is fixed in advance. Now the specification of the desired degree of accuracy usually involves some arbitrariness. If a coefficient of variation of 1.5 percent is demanded, the sample will not be regarded as useless should the coefficient turn out to be 1.6 percent. The advance specification of a sum of money that must be spent on the sample is also open to criticism, for the accuracy obtained from this expenditure may be substantially more, or substantially less, than is needed for the use that is to be made of the estimates. Two attempts to utilize a more general principle, in which optimum cost and optimum accuracy are determined simultaneously, will be briefly described.

In order to apply the principle, one must be able to answer the question: how much is a given degree of accuracy worth? Any decisions that are based on an estimate from a sample will presumably be more fruitful if the estimate has a low error than if it has a high error. In certain cases we may be able to calculate, in monetary terms, the loss 1 (z) that will be incurred in a decision through an error of amount z in the estimate. Although the actual value of z is not predictable in advance, sampling theory may enable us to predict the frequency distribution p (z, n) of z, which for a specified method of sampling will depend on the size of sample n. Hence the expected loss for a given size of sample is
\[ L(n) = \int l(z) p(z,n) \, dz. \]

The purpose in taking the sample is to diminish this loss. If \( C(n) \) is the cost of a sample of size \( n \), clearly \( n \) should be chosen so as to minimize

\[ C(n) + L(n) \]

since this is the total cost involved in taking the sample and in making decisions from its results. Choice of \( n \) so as to minimize this quantity will determine both the optimum amount of money to be spent on sampling and the optimum accuracy. The idea is presented here only in its simplest form; it may be extended to cover a choice between different sampling methods.

In the application described by Blythe (48), the selling price of a lot of standing timber is \( SV \), where \( S \) is the price per unit volume, and \( V \) is the volume of timber in the lot. The number \( N \) of logs in the lot is counted, and the average volume per log is estimated from a sample of \( n \) logs. If \( \sigma \) is the standard deviation per log for the sampling method used, the standard deviation of the estimate of \( V \) will be \( N \sigma / \sqrt{n} \), (ignoring finite population correction).

Suppose that this estimate is made and paid for by the seller. The buyer provisionally accepts the estimate of the amount of timber which he has bought. Subsequently, however, he finds out the correct volume purchased, and the seller reimburses him if he has paid for more than was delivered. If he has paid for less than was delivered, the buyer does not mention the fact. In this situation the seller loses whenever he underestimates the volume, but does not gain when he overestimates it. The situation is artificial, but serves to illustrate the application of the principle to a case that does not require
complex mathematics. (This presentation is slightly different from that of Blythe).

When he underestimates the volume by an amount \( z \), the seller loses an amount \( Sz \). Thus we may take \( 1(z) \) as zero when \( z \) is negative and as \( Sz \) when \( z \) is positive, where \( z \) is the amount of underestimation. On the assumption of a normal distribution of sampling errors, \( p(z,n) \) is the normal distribution with mean zero and variance \( \frac{N^2 \sigma^2}{n} \). Hence

\[
L(n) = \frac{\sqrt{\frac{n}{2\pi}}} {N\sigma} \int_0^\infty e^{-\frac{n z^2}{2N^2 \sigma^2}} dz = \frac{S N \sigma}{\sqrt{2\pi} \sqrt{n}}
\]

If we suppose further that the cost of measuring the volume of a log is \( c \), the cost function \( C(n) \) is \( cn \). The quantity to be minimized is therefore

\[
cn + \frac{S N \sigma}{\sqrt{2\pi} \sqrt{n}}
\]

Differentiation with respect to \( n \) leads to the solution

\[
n = \left( \frac{S N \sigma}{2c \sqrt{2\pi}} \right)^\frac{2}{3}
\]

In the example due to Nordin (49), a manufacturer takes a sample in order to estimate the size of a market which he intends to enter. If the size is known accurately, the amount of fixed equipment and the production per unit period can be adjusted so as to maximize expected profit. Errors in the estimated size of market will result in choices of these two factors that fall short of the optimum, and lead to a smaller expected profit. The sample size \( n \) should therefore be such that the addition of an \((n+1)\)th unit to the sample increases the profit expectation by exactly the cost of the \((n+1)\)th unit.

In many cases it will be difficult to apply these ideas because no way can be found to translate the effect of a sampling error into
monetary terms. Moreover, an estimate may be used by different persons for quite diverse purposes. Nevertheless, the question of the standard of accuracy needed in sample estimates has received too little attention, and this type of research may point in a fruitful direction.

11.2 Area sampling: All sampling methods for which we have presented theory require something equivalent to a listing of the population, since this is needed to draw either random, stratified random or "every k th" systematic samples. For many types of population no listings are available, and this imposes a serious handicap to the use of theoretically sound methods. For the sampling of human populations, the method of area sampling represents a major achievement towards overcoming this difficulty. The sampling unit is a compact area of land, usually shown on a map. These areas are constructed so that they completely cover the map which shows the population that is to be sampled. In other words, a listing of the population into areas is deliberately made.

In the Master Sample of agriculture, designed primarily for farm surveys, every county in the United States has been divided in this way into areas. These average about 2 1/2 square miles in area and contain from 4 to 8 farms each on the average, though these numbers differ in different parts of the country and vary considerably for individual areas. The next step, which presents difficulties, is to devise rules such that each element in the population is clearly associated with one and only one area. For instance, if the population is a population of farms, we require a rule such that every farm in the population 'belongs' to one and only one area. If this rule is found, a random sample of areas provides a random 'cluster' sample of farms. The relevant theory is that given in chapters 7 (type of sampling unit) and 8 (subsampling). Further, since the number of farms per area cannot conveniently be kept constant, it is usually found that
ratio or regression estimates involving the number of farms are more accurate than a simple expansion by the ratio of the number of areas in the population to that in the sample.

In certain parts of the country, the rule that the farm is associated with the area on which the farmstead lies works fairly well. More complex rules are needed for cases where the farm has no farmstead, or where the farm consists of multiple tracts. A sample of areas can also be used, e.g., for surveys of rural housing, where the population becomes a population of houses rather than of farms, and changes in the rules are made accordingly. For a more detailed description, see King and Jessen (50).

For samples involving visitation of houses in large towns, the areal unit is usually a city block, which can be outlined on a city map. It is customary to stratify the blocks, and to utilize subsampling, only a certain fraction of the houses in a block that is selected being visited. The method is discussed by Hansen and Hauser (51); another useful reference is Hansen and Dening (52).

11.3 Control of human errors: Mahalanobis (53) has described a number of devices used in his sampling work in order to obtain information on the extent of human errors. One device is to have certain sampling units enumerated twice by different workers (or teams of workers), who do not know on which units this duplication is to occur. By means of a t test one can examine whether there is a consistent difference between the results for the two workers. A second device is the use of what Mahalanobis calls "interpenetrating samples". If for instance there are four strata and five teams, each team might be assigned to enumerate one-fifth of the units in each stratum. From the results the following analysis of variance can be computed.
Between strata  d.f.  3
Between teams  4
Interaction; strata x teams  12
Within teams between units  -

From this analysis the presence of consistent differences among teams, or of differences in individual strata, can be examined. Of course, if differences between teams exist, they enter into the real sampling error of the estimate: the sampling error as calculated from the standard formulae given in previous chapters would be an underestimate.

11.4 Description of actual surveys: The following references contain accounts (in whole or in part) of actual surveys, and are useful in studying the practical application of sampling techniques.


REFERENCES


