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 μ 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.

$$E(\text{Experimental error mean square}) = \sigma_w^2 + m\,\sigma_b^2,$$

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

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

$$V(\bar{y}_{i..}) = \frac{\sigma_w^2 + m\sigma_b^2}{mn} = \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{mn} \qquad (126)$$

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

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

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

$$m = \frac{\sigma_w}{\sigma_b} \sqrt{\frac{c_1}{c_2}} \qquad (128)$$

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 σ_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 σ_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 σ_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 consideration of subsampling in sample surveys. One consequence of the mathematical 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 husband 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 Zacopanay (35). Alternatively, Hansen and Hurwitz (27) suggest the use of the intra-class correlation coefficient. Instead of (121) we have

$$y_{ij} = \mu + w_{ij} \qquad (129)$$

The quantities w_{ij} all have mean zero and variance σ^2. Any pair of sub-units w_{ij}, w_{ik}, that are in the same unit are correlated, with correlation coefficient ω, 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 18, which corresponds to Table 16 for the previous model.

TABLE 18.

ANALYSIS OF VARIANCE WITH SUBSAMPLING: ALTERNATIVE MODEL

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

If ρ is positive, B 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 B is expected to be less than W is covered by negative values of ρ. Note, however, that ρ cannot be less than $-1/(m-1)$, for such values would give B 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 ρ_{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 ρ 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

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

First, it is necessary to re-define σ_w^2 and σ_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)

	d.f.	Mean square	Defined as equal to
Between units	(N-1)	$M \sum_{i=1}^{N} (\bar{y}_{iM} - \bar{y}_{NM})^2/(N-1)$	$= \sigma_w^2 + M\,\sigma_b^2$
Within units between sub-units	N(M-1)	$\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_{iM})^2/N(M-1)$	$= \sigma_w^2$

where $\bar{y}_{iM}$ denotes the mean of the i th unit. We define σ_w^2 and σ_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:

$$V(\bar{y}_{nm}) = E(\bar{y}_{nm} - \bar{y}_{NM})^2 = \frac{(N-n)}{N} \frac{\sigma_b^2}{n} + \frac{(MN-mn)}{MN} \frac{\sigma_w^2}{mn} . \quad (130)$$

Proof: Write

$$\bar{y}_{nm} - \bar{y}_{NM} = (\bar{y}_{nm} - \bar{y}_{nM}) + (\bar{y}_{nM} - \bar{y}_{NM}) , \tag{131}$$

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

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

where σ_{wj}^2 is the variance within the j th unit. This may be rewritten

$$\frac{(M-m)}{M} \cdot \frac{1}{mn} \bar{\sigma}_{wn}^2 \tag{132}$$

where $\bar{\sigma}_{wn}^2$ 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 $\bar{\sigma}_{wn}^2$ is σ_w^2. Hence

$$E(\bar{y}_{nm} - \bar{y}_{nM})^2 = \frac{(M-m)}{M} \cdot \frac{1}{mn} \sigma_w^2 . \tag{133}$$

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,

$$E(\bar{y}_{nM} - \bar{y}_{NM})^2 = \frac{(N-n)}{nN} \left\{ \sigma_b^2 + \frac{\sigma_w^2}{M} \right\} \qquad (134)$$

since by the definition of σ_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

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

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

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 B/nm 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, σ_w^2 and σ_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 W, the sample mean square within units, is equal to σ_w^2 , the corresponding population mean square.

The mean value of B is less obvious. We have

$$B = m \Sigma (\bar{y}_{i.} - \bar{y}_{nm})^2/(n-1).$$

Write

$$\bar{y}_{i.} = \bar{y}_{iM} + \bar{e}_{i.} ,$$

where $\bar{y}_{iM}$ is the mean of all $\underline{M}$ sub-units in the i th unit. Then

$$E(\bar{e}_{i.}^{\,2}) = \frac{M-m}{M} \cdot \frac{\sigma_{wi}^2}{m} ,$$

since this is the variance of the mean of a random sample of $\underline{m}$ sub-units out of $\underline{M}$. Similarly write

$$\bar{y}_{nm} = \bar{y}_{nM} + \bar{e}_{nm} .$$

In equation (132) we proved that

$$E(\bar{y}_{nm} - \bar{y}_{nM})^2 = E(\bar{e}_{nm}^{\,2}) = \frac{M-m}{M} \cdot \frac{\bar{\sigma}_{wn}^2}{mn} .$$

Now,

$$\sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{nm})^2 = \sum_{i=1}^{n} \left\{ (\bar{y}_{im} - \bar{y}_{nM}) + \bar{e}_i - \bar{e}_{nm} \right\}^2 .$$

When we expand and take the expectation for a fixed set of $\underline{n}$ units, we have

$$E \sum_{i=1}^{n} \left\{ (\bar{y}_i - \bar{y}_{nm})^2 \right\} = \sum_{i=1}^{n} (\bar{y}_{im} - \bar{y}_{nM})^2 + \frac{(M-m)}{Mm} \left\{ \Sigma\, \sigma_{wi}^2 - \bar{\sigma}_{wn}^2 \right\}$$

Since $n\,\bar{\sigma}_{wn}^2 = \Sigma\, \sigma_{wi}^2$, we obtain, on division by (n-1),

$$\frac{E \left\{ \Sigma\, (\bar{y}_i - \bar{y}_{nm})^2 \right\}}{(n-1)} = \frac{\Sigma\, (\bar{y}_{im} - \bar{y}_{nm})^2}{(n-1)} + \frac{(M-m)}{Mm}\, \bar{\sigma}_{wn}^2$$

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 $\underline{n}$ out of $\underline{N}$. This gives

$$E\,(B/m) = \sum_{i=1}^{N} \frac{(\bar{y}_{im} - \bar{y}_{NM})^2}{(N-1)} + \frac{(M-m)}{Mm}\, \sigma_w^2 .$$

$$= \sigma_b^2 + \frac{\sigma_w^2}{M} + \left(\frac{1}{m} - \frac{1}{M} \right) \sigma_w^2$$

$$= \sigma_b^2 + \frac{\sigma_w^2}{m}$$

Hence $$E(B) = \sigma_w^2 + m\,\sigma_b^2 . \qquad (135)$$

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

Theorem 14: An unbiased estimate of $V(\bar{y}_{nm})$, taking into account the f.p.c., is

$$\frac{1}{mn}\left\{\frac{(N-n)}{N}\,B + \frac{(M-m)}{M}\,\frac{n}{N}\,W\right\} \qquad (136)$$

This follows at once from the preceding results. An unbiased estimate of σ_b^2 is $(B-W)/m$. On substituting in formula (130) for $V(\bar{y}_{nm})$ 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 σ_{bj}^2 and σ_{wj}^2 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 j th 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

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

Its variance is

$$\frac{\sum_j (M_j N_j)^2 V(\bar{y}_{nmj})}{(\Sigma M_j N_j)^2} = \sum_j (M_j N_j)^2 \left[\frac{(N_j - n_j)}{N_j} \frac{\sigma_{bj}^2}{n_j} + \frac{(M_j N_j - m_j n_j)}{M_j N_j} \frac{\sigma_{wj}^2}{m_j n_j} \right] \Big/ (\Sigma M_j N_j)^2 , \qquad (137)$$

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 (32), 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

$$y_{ijk} = \mu + b_i + w_{ij} + z_{ijk} ,$$

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

$$V(\bar{y}_{nmp}) = \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nm} + \frac{\sigma_z^2}{nmp} . \qquad (138)$$

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

TABLE 20.

ANALYSIS OF VARIANCE FOR SUB-SUBSAMPLING

	d.f.	Mean square	Estimate of
Between units	(n-1)	B	$\sigma_z^2 + p\,\sigma_w^2 + mp\,\sigma_b^2$
Between sub-units within units	n(m-1)	W	$\sigma_z^2 + p\,\sigma_w^2$
Between sub-sub-units within sub-units	nm(p-1)	Z	σ_z^2

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

$$V(\bar{y}_{nmp}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_b^2 + \left(\frac{1}{mn} - \frac{1}{MN}\right)\sigma_w^2 + \left(\frac{1}{mnp} - \frac{1}{MNP}\right)\sigma_z^2 , \quad (139)$$

of which an unbiased estimate is

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

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 m_i sub-units be sampled at random from it. The mean of the observations from these m_i sub-units is denoted by $\bar{y}_{is}$ (s for sample), while the true mean of the unit is $\bar{y}_{ip}$ (p for population). The mean of the whole population, $\bar{y}_p$ is the quantity to be estimated.

It seems natural to use the sample mean $\bar{y}_{is}$ as an estimate of the population mean $\bar{y}_p$. This estimate is, however, biased. In repeated sampling from the same unit, the average of $\bar{y}_{is}$ 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

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

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

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

Within units — Between units

$$+ (\bar{y}_u - \bar{y}_p)^2 \qquad (141)$$

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 σ_i^2 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 <u>first</u> 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

$$(\bar{y}_t - \bar{y}_p) = \frac{NM_i\,\bar{y}_{is}}{M} - \frac{\Sigma\, M_i\,\bar{y}_{ip}}{M}$$

$$= \frac{NM_i\,(\bar{y}_{is} - \bar{y}_{ip})}{M} + \frac{N\,T_i - \Sigma\,T_i}{M} \quad ,$$

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

$$V(\bar{y}_t) = \frac{N}{M^2} \sum_{i=1}^{N} M_i(M_i - m_i)\,\frac{\sigma_i^2}{m_i} + \frac{N}{M^2} \sum_{i=1}^{N} (T_i - \bar{T})^2 \qquad (142)$$

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 $\bar{y}_{is}$ an unbiased estimate of $\bar{y}_p$. For, in repeated sampling, the i th unit will appear with frequency M_i/M, so that

$$E(\bar{y}_{is}) = \sum_{i=1}^{N} \frac{M_i}{M} \bar{y}_{ip} = \bar{y}_p .$$

For the sampling variance, we have

$$(\bar{y}_{is} - \bar{y}_p) = (\bar{y}_{is} - \bar{y}_{ip}) + (\bar{y}_{ip} - \bar{y}_p) .$$

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

$$V(\bar{y}_{is}) = E(\bar{y}_{is}-\bar{y}_p)^2 = \sum_{i=1}^{N} \frac{M_i}{M} \frac{(M_i-m_i)}{M_i} \frac{\sigma_i^2}{m_i}$$

$$+ \Sigma \frac{M_i}{M} (\bar{y}_{ip} - \bar{y}_p)^2 \qquad (143)$$

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

Unit	y_{ij}	M_i	T_i	σ_i^2	$\bar{y}_{ip}$
1	0,1	2	1	.500	0.5
2	1,2,2,3	4	8	.667	2.0
3	3,3,4,4,5,5	6	24	.800	4.0

There are three units, with respectively 2,4 and 6 sub-units. The reader may verify the figures for T_i, σ_i^2 and $\bar{y}_{ip}$. The population mean $\bar{y}_p$ is 33/12, or 2.75. The unweighted mean of the $\bar{y}_{ip}$ 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, $\frac{1}{2} M_i$ sub-units from it.
Estimate: $\bar{y}_{is}$ (biased).

Method III.

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

VARIANCES OF SAMPLE ESTIMATES

Method	Contribution to variance from Within Units	Between Units	Bias	Total Variance
I	.145	2.056	.340	2.541
II	.183	2.056	.340	2.579
III	.256	5.792	.000	6.048
IV	.189	1.813	.000	2.002

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 $\underline{M}$: their variances become multiplied by M^2.

8.13 Selection with arbitrary probabilities: It may happen that the sizes M_i 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 $\underline{t}$ to the population, e.g., 1 percent or 5 percent. If the i th unit is chosen, we take a sample of size m_i from it, and use as the estimate of the population total $m_i \bar{y}_{is}/t$; in other words, the sample total, divided by the sampling

rate. The mean value of this estimate in repeated sampling is

$$\sum_{i=1}^{N} P_i m_i \bar{y}_{ip}/t \quad .$$

If this is to equal the true population total, $\Sigma M_i \bar{y}_{ip}$, we must have

$$m_i = tM_i/P_i \quad . \tag{144}$$

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

$$\frac{m_i}{t}\bar{y}_{is} - M\bar{y}_p = \frac{M_i}{P_i}\bar{y}_{is} - M\bar{y}_p = \frac{M_i}{P_i}(\bar{y}_{is}-\bar{y}_{ip}) + \frac{M_i}{P_i}\bar{y}_{ip} - M\bar{y}_p$$

Each square receives a weight P_i. Thus

$$V(m_i\bar{y}_{is}/t) = \sum_{i=1}^{N} \frac{M_i}{P_i}(M_i-m_i)\frac{\sigma_i^2}{m_i} + \sum_{i=1}^{N} P_i\left(\frac{M_i\bar{y}_{ip}}{P_i} - M\bar{y}_p\right)^2 \tag{145}$$

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, sampling 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 subunit is the household, and we want to sample 5 percent of the population so that t = 1/20. The expected numbers of households (e.n.o.h.) in the six blocks are as shown below.

Block	E.n.o.h.	Cumulated
1	10	10
2	30	40
3	17	57
4	25	82
5	23	105
6	16	121

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 $t = 1/120$, we may apply (144) and obtain

$$m_i = \frac{1.31.121}{20.23} = 8, \text{ to the nearest integer.}$$

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 estimate 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 twon slowly enough to permit estimates to be made of the number of households in each block. Blocks were then chosen with probability proportional 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.
M_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.
σ^2_{ij}	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

$$\sum_j M_j \bar{y}_{ijs} \qquad \text{(biased)}.$$

$$V = \sum_j \frac{M_j^2}{N_j}\left[\sum_{i=1}^{N_j} \frac{(M_{ij} - m_{ij})}{M_{ij}} \frac{\sigma^2_{ij}}{m_{ij}} + \sum_{i=1}^{N_j} (\bar{y}_{ijp}-\bar{y}_{ju})^2 + (\bar{y}_{ju} - \bar{y}_{jp})^2\right] \qquad (146)$$

II. Units chosen with probability proportional to relative size within strata. $P_{ij} = M_{ij}/M_j$. The estimate is as in I (unbiased).

$$V = \sum_j M_j \left[\sum_{i=1}^{N_j} (M_{ij} - m_{ij}) \frac{\sigma_{ij}^2}{m_{ij}} + \sum_{i=1}^{N_j} M_{ij} (\bar{y}_{ijp} - \bar{y}_{jp})^2 \right] . \quad (147)$$

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

$$\sum_j m_{ij} \bar{y}_{ijs}/t_j \qquad \text{(unbiased).}$$

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

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

(34) King, A. J. and Jebe, E. H. "An Experiment in the Pre-harvest Sampling of Wheat Fields" Iowa Agr. Exp. Sta. Res. Bull. 273, 1940.

(35) Yates, F. and Zacopanay, I. "The Estimation of the Efficiency of Sampling with Special Reference to Sampling in Cereal Experiments" Jour. Agr. Sci., 25, pp. 545-577, 1935.

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

(36) Yates, F. "A Review of Recent Statistical Developments in Sampling and Sampling Surveys" Jour. Roy. Stat. Soc., 109, pp. 12-42, 1946.

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

(37) Hansen, M. H. and Hurwitz, W. N. " On the Theory of Sampling from Finite Populations" Ann. Math. Stat., 14, pp. 333-362, 1943.

(38) Jessen, R. J., Blythe, R. H., Kempthorne, O. and Deming, W. Edwards "On a Population Sample for Greece" Jour. Amer. Stat. Asso., 42, pp. 357-384, 1947.

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:

$$Y_R = \frac{Y_s}{X_s} X_p \qquad (148)$$

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

$$V(Y_R) = \left[\frac{N(N-n)}{n}\right]\left[\sigma_y^2 + R_p^2\ \sigma_x^2 - 2\,R_p\ \rho\,\sigma_y\,\sigma_x\right] \qquad (149)$$

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

$$N(Y_R) = \left(\frac{N^2\,\bar{y}_p^{\,2}}{n}\right)\left(\frac{N-n}{N}\right)\left(\frac{\sigma_y^2}{\bar{y}_p^2} + \frac{\sigma_x^2}{\bar{x}_p^2} - \frac{2\ \mathrm{cov}\ (y\ x)}{\bar{y}_p\ \bar{x}_p}\right). \qquad (150)$$

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.

$$Y_R = \frac{\bar{y}_n}{\bar{x}_n}\ N\,\bar{x}_p$$

$$= \left(\frac{\bar{y}_p + \Delta y}{\bar{x}_p + \Delta x}\right) N\,\bar{x}_p$$

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

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

$$= N\,\bar{y}_p \left(1 + \frac{\Delta y}{\bar{y}_p} - \frac{\Delta x}{\bar{x}_p}\right) \text{ approximately, this being}$$

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

$$\left[Y_R - E(Y_R)\right] = N \bar{y}_p \left(\frac{\Delta y}{\bar{y}_p} - \frac{\Delta x}{\bar{x}_p} \right)$$

where

$$(\Delta y) = (\bar{y}_n - \bar{y}_p)$$

$$E\,(\Delta y)^2 = \frac{N-n}{N} \frac{\sigma_y^2}{n} \text{, etc.}$$

Therefore, the variance of Y_R is approximately

$$V(Y_R) = E\left[Y_R - E(Y_R)\right]^2 = \left(\frac{N^2 \bar{y}_p^2}{n} \right) \left(\frac{N-n}{N} \right) \left(\frac{\sigma_y^2}{\bar{y}_p^2} + \frac{\sigma_x^2}{\bar{x}_p^2} - \frac{2\,\mathrm{cov}(y\,x)}{\bar{y}_p\,\bar{x}_p} \right)$$

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

$s_x^2 = \frac{\Sigma(x - \bar{x})^2}{n-1}$ and the covariance of $\underline{y}$ and $\underline{x}$ is estimated by

$s_{yx} = \frac{\Sigma(y-\bar{y})\,(x-\bar{x})}{n-1}$. Therefore, the estimated variance of Y_R becomes

$$V(Y_R) = \left(\frac{N^2 \bar{y}_n^2}{n} \right) \left(\frac{N-n}{N} \right) \left(\frac{s_y^2}{\bar{y}_n^2} + \frac{s_x^2}{\bar{x}_n^2} - \frac{2\,s_{yx}}{\bar{y}_n\,\bar{x}_n} \right) \quad (151)$$

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

$$V(Y_R) = \frac{N(N-n)}{n} \left(s_y^2 + \frac{\bar{y}_n^2\, s_x^2}{\bar{x}_n^2} - 2 \frac{\bar{y}_n}{\bar{x}_n}\, s_{yx} \right) \quad (152)$$

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

$$V(Y_R) = \frac{N(N-n)}{n(n-1)} \left[\Sigma\, y_i^2 + R_s^2\, \Sigma\, x_i^2 - 2\, R_s\, \Sigma\, y_i\, x_i \right] \quad (153)$$

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

$$V(Y_R) = \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} (y_i - R_s x_i)^2 \qquad (154)$$

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

$$V_{(N\bar{y}_n)} = N(N-n)\left(\frac{\sigma_y^2}{n}\right) \qquad (155)$$

Hence from (149), $V(Y_R)$ is less than $V_{(N\bar{y}_n)}$ if the following inequality is true:

$$\sigma_y^2 + R_p^2 \sigma_x^2 - 2 R_p \rho \sigma_y \sigma_x < \sigma_y^2$$

or

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

$$\rho_{yx} > \frac{1}{2} \frac{\frac{\sigma_x}{\bar{x}_p}}{\frac{\sigma_y}{\bar{y}_p}},$$

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

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 ρ exceeds $\frac{1}{2}$.

Theorem 16: The ratio estimate is a "best unbiased linear estimate" if two conditions are satisfied (Cochran, (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 λx. Hence,

$$\bar{y}_p = \beta \bar{x}_p + \bar{e}_p$$

$$= \beta \bar{x}_p \text{ (since } \bar{e}_p = 0)$$

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 β. This is

$$b = \frac{\Sigma w y x}{\Sigma w x^2} \quad \text{where}$$

$$w = \frac{1}{\sigma_e^2} = \frac{1}{\lambda x}$$

Hence,

$$b = \frac{\Sigma y}{\Sigma x} = \frac{\bar{y}_n}{\bar{x}_n}$$

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, Hasel (48), 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 $R_s = Y_s/X_s$, 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 omit 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

$$Y_{Rs} = \sum_j \frac{Y_{sj}}{X_{sj}} X_{pj} = \sum_j \left(\frac{\bar{y}_{nj}}{\bar{x}_{nj}} N_j \bar{x}_{pj} \right) \qquad (156)$$

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

$$V(Y_{Rs}) = \sum_j \frac{N_j}{n_j} \bar{y}^2_{pj} \left(\frac{N_j - n_j}{N_j} \right) \left[\frac{\sigma^2_{yj}}{\bar{y}^2_{pj}} + \frac{\sigma^2_{xj}}{\bar{x}^2_{pj}} - \frac{2 \operatorname{cov} x_j y_j}{\bar{y}_{pj} \bar{x}_{pj}} \right]$$

which may be written as

$$V(Y_{Rs}) = \sum_j \frac{N_j (N_j - n_j)}{n_j} \left[\sigma^2_{yj} + R^2_{pj} \sigma^2_{xj} - 2R_{pj} \rho_j \sigma_{yj} \sigma_{xj} \right] \qquad (157)$$

where $R_{pj} = \frac{Y_{pj}}{X_{pj}}$ is the true ratio for the stratum.

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

$$Y_{Rc} = \frac{\sum \frac{N_j}{n_j} Y_{sj}}{\sum \frac{N_j}{n_j} X_{sj}} X_p = \left(\frac{\sum N_j \bar{y}_{nj}}{\sum N_j \bar{x}_{nj}} \right) N \bar{x}_p \qquad (158)$$

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

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

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

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

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

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

Hence,

$$V(Y_{Rc}) = \bar{y}_p^2 \Sigma \left[\frac{N_j\ (N_j - n_j)}{n_j}\right] \left[\frac{\sigma_{yj}^2}{\bar{y}_p^2} + \frac{\sigma_{xj}^2}{\bar{x}_p^2} - \frac{2\ \mathrm{cov}\ (y\ x)}{\bar{y}_p\ \bar{x}_p}\right] \qquad (159)$$

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

$$V(Y_{Rc}) = \Sigma \frac{N_j\ (N_j-n_j)}{n_j} \left[\sigma_{yj}^2 + R_p^2\ \sigma_{xj}^2 - 2\ R_p\ \rho_j\ \sigma_{yj}\ \sigma_{xj}\right] \qquad (160)$$

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

$R_p = \frac{Y_p}{X_p}$ replaces R_{pj}. To compare (160) with (157) we can write

$$V(Y_{Rc}) = V(Y_{Rs}) + \Sigma \frac{N_j(N_j-n_j)}{n_j} \left[(R_{pj}^2 - R_p^2)\ \sigma_{xj}^2 - 2(R_{pj} - R_p)\rho_j\ \sigma_{yj}\ \sigma_{xj}\right]$$

$$= V(Y_{Rs}) + \Sigma \frac{N_j(N_j-n_j)}{n_j} \left[(R_{pj}-R_p)^2\ \sigma_{xj}^2 + 2(R_{pj}-R_p)\ (\rho_j\ \sigma_{yj}\ \sigma_{xj}\right.$$

$$\left. - R_{pj}\ \sigma_{xj}^2)\right]$$

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^2_{yj} and s^2_{xj} are substituted for the corresponding variances, and the sample covariance for the term $\rho_j \sigma_{yj} \sigma_{xj}$. It will be noted that 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 23.

DATA FROM JEFFERSON COUNTY, IOWA

Strata	Size (farm acres)	N_j	σ^2_{yj}	σ_{yxj}	σ^2_{xj}	R_{pj}
1	0-160	1580	312	494	2055	.2351
2	over 160	430	922	858	7357	.2019
For complete pop.		2010	620	1453	7619	.2242

Strata	$\bar{y}_{pj}$	$\bar{x}_{pj}$	n_j	$Q_j = W_j^2/n_j$	V_j	V'_j
1	19.51	82.56	70	.008828	194	193
2	51.63	244.85	30	.001525	887	907
For c.p.	26.30	117.28	100			

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.

$$V_2 = \frac{1}{n}\left[\sigma_y^2 + R_p^2 \sigma_x^2 - 2 R_p \sigma_{yx}\right]$$

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

$$= 3.51$$

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

$$V_3 = \frac{1}{N^2} \Sigma \frac{N_j^2}{n_j} \sigma_{yj}^2 = \Sigma Q_j \sigma_{yj}^2 = 4.16.$$

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

$$V_4 = \Sigma Q_j \left[\sigma_{yj}^2 + R_{pj}^2 \sigma_{xj}^2 - 2 R_{pj} \sigma_{yxj}\right] = \Sigma Q_j V_j = 3.07.$$

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

$$V_5 = \Sigma Q_j \left[\sigma_{yj}^2 + R_p^2 \sigma_{xj}^2 - 2 R_p \sigma_{yxj}\right] = \Sigma Q_j V_j' = 3.09.$$

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

	Sampling Method	Method of Estimation	R. I.
(i)	Simple random	Mean per s.u.	100
(ii)	Simple random	Ratio	177
(iii)	Stratified random	Mean per s.u.	149
(iv)	Stratified random	Separate ratio	202
(v)	Stratified random	Combined ratio	201

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 $\underline{y}$ and $\underline{x}$. 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 $\underline{y}$ and $\underline{x}$.

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 $(\sigma_{yj}^2 + R_{pj}^2 \sigma_{xj}^2 - 2 R_{pj} \rho_j \sigma_{yj} \sigma_{xj})$ is the variance within the j th stratum of the variate $d = (y - R_{pj}x)$. This variance will be denoted by σ_{dj}^2. If (157) is minimized subject to a total cost of the form $\Sigma c_j n_j$, 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, σ_{dj} will be proportional to $\sqrt{\bar{x}}$. 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 $\bar{x}^2$. 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 $N_j \sqrt{\bar{x}_{pj}}$, and, (iv) n_j proportional to $N_j \bar{x}_{pj}$. 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

Strata	σ^2_{xj}	σ_{yxj}	σ^2_{yj}	σ_{xj}	σ_{yj}	$\bar{x}_j$	$\bar{y}_j$	R_j	V_j	
1	5186	6462	8699	72.01	93.27	53.80	69.48	1.29133	658	
2	2367	3100	4614	48.65	67.93	31.07	43.64	1.40475	573	
3	4877	4817	7311	69.83	85.51	56.97	66.39	1.16547	2706	
Total	3898	4434	6409	62.43	80.06	44.45	56.47	1.27053	1433	
Strata	N_j	(i)	$N_j\sigma_{yj}$	(ii)	$\sqrt{\bar{x}_{pj}}$	$N_j\sqrt{\bar{x}_{pj}}$	(iii)	$N_j\bar{x}_{pj}$	(iv)	
1	47	18	4384	22	7.33	344.5	20	2529	22	
2	113	46	8016	40	5.57	657.3	39	3666	32	
3	91	36	7781	38	7.55	687.1	41	5184	46	
Total	256	100	20181	100	20.45	1688.9	100	11379	100	

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),

$$V(Y_{Rs}) = \sum_j \frac{N_j (N_j - n_j)}{n_j} V_j \text{ , where } V_j = \sigma^2_{yj} + R^2_{pj} \sigma^2_{xj} - 2 R_p \sigma_{yxj} .$$

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

Method of Allocation n_j proportional to		Variance Strata 1	2	3	Total	Relative Information
(i)	N_j	49,824	105,833	376,215	531,872	100
(ii)	$N_j \sigma_{yj}$	35,144	131,847	343,446	510,437	104
(iii)	$N_j \sqrt{\bar{x}_{pj}}$	41,750	136,964	300,312	479,026	111
(iv)	$N_j \bar{x}_{pj}$	35,144	181,710	240,888	457,742	116

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

$$b = \Sigma (y-\bar{y}_n)(x-\bar{x}_n)/\Sigma (x-\bar{x}_n)^2 .$$

The estimate of the population mean of y is then taken as

$$\bar{y}_{Lr} = \left\{\bar{y}_n + b(\bar{x}_p-\bar{x}_n)\right\} \qquad (161)$$

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

$$y = \alpha + \beta (x-\bar{x}_p) + e, \qquad (162)$$

where e is a random variable with mean zero for any x and constant variance σ_e^2 .

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

$$\bar{y}_n = \alpha + \beta\ (\bar{x}_n - \bar{x}_p) + \bar{e}_n\ , \tag{163}$$

$$b = \beta + \frac{\Sigma\ e\ (x - \bar{x}_n)}{\Sigma\ (x - \bar{x}_n)^2}\ , \tag{164}$$

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

$$\bar{e}_n + (\bar{x}_p - \bar{x}_n)\ \frac{\Sigma\ e\ (x - \bar{x}_n)}{\Sigma\ (x - \bar{x}_n)^2} \tag{165}$$

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

$$V(\bar{y}_{Lr}) = \sigma_e^2 \left[\frac{1}{n} + \frac{(\bar{x}_p - \bar{x}_n)^2}{\Sigma\ (x - \bar{x}_n)^2} \right] \tag{166}$$

$$= \frac{\sigma_y^2\ (1 - \rho^2)}{n} \left[1 + \frac{n(\bar{x}_p - \bar{x}_n)^2}{\Sigma\ (x - \bar{x}_n)^2} \right] \tag{167}$$

where ρ is the correlation coefficient between y and x.

The sample estimate of this variance is obtained by substituting for $\sigma_y^2\ (1 - \rho^2)$ 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

$$\bar{V}(\bar{y}_{Lr}) = \frac{\sigma_y^2 (1 - \rho^2)}{n} \left[1 + \frac{1}{n} + \frac{3 + 2\gamma_1^2}{n^2} \right] \qquad (168)$$

where γ_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, γ_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

$$V(\bar{y}_{Lr}) = \frac{\sigma_y^2 (1 - \rho^2)}{n} \qquad (169)$$

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 n, 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:

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

$$V(\bar{y}_R) = \frac{(N-n)}{Nn} (\sigma_y^2 - 2 R_p \sigma_{yx} + R_p^2 \sigma_x^2) , \qquad \text{(ratio)}$$

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

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

$$- \sigma_y^2 \rho^2 \leq - 2 R_p \rho \sigma_y \sigma_x + R_p^2 \sigma_x^2 ,$$

where we have written $\rho \sigma_y \sigma_x$ for σ_{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:

$$\rho = R_p \frac{\sigma_x}{\sigma_y} = \frac{\text{coefficient of variation of x}}{\text{coefficient of variation of y}} \qquad (170)$$

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 $\underline{b}$. If there is an appreciable saving in time, an inefficient estimate of $\underline{b}$ can often be used instead of the least squares estimate. If the estimate of $\underline{b}$ has an efficiency $\underline{E}$, $(E < 1)$, the fractional increase in the variance of the regression estimate of $N \bar{y}_p$ is about $\frac{(1-E)}{nE}$. With large $\underline{n}$, even a highly inefficient estimate of $\underline{b}$ causes only a trivial increase in the variance.

A simple method for obtaining an estimate of $\underline{b}$ has been proposed by Hendricks and was used by Finkner, Morgan, 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

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

where $\bar{y}_l$ and $\bar{x}_l$ 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 $\underline{b}$, one can obtain an unbiased sample estimate of $\sigma_y^2 (1 - \rho^2)$ very quickly, whereas with other estimates of $\underline{b}$, 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 $\sigma_y^2 = 6409$, $\sigma_{yx} = 4434$, $\sigma_x^2 = 3898$, $R_p = 1.270$, $\rho^2 = .887$, $n = 100$, $N = 256$.

$$V(\bar{y}_{Lr}) = \frac{N-n}{N} \frac{\sigma_y^2 (1-\rho^2)}{n} \left[1 + \frac{1}{n-3}\right]$$

$$= \left(\frac{256-100}{256}\right) \frac{6409\ (1 - .787)}{100} \left[1 + \frac{1}{97}\right]$$

$$= 8.40$$

$$V(\bar{y}_R) = \frac{N-n}{N} \frac{1}{n} \left[\sigma_y^2 + R_p^2 \sigma_x^2 - 2 R_p \sigma_{yx}\right]$$

$$= \frac{(256-100)}{256} \frac{1}{100} \left[6409 + (1.613)\ (3898) - (2)\ (1.270)\ (4434)\right]$$

$$= 8.74$$

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

$$= \left(\frac{256-100}{256}\right)\left(\frac{6409}{100}\right)$$

$$= 39.05$$

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

Mean per s.u.	-	100%
Ratio	-	447%
Regression	-	465%

REFERENCES

(39) Cochran, W. G. "Sampling Theory When the Sampling Units Are of Unequal Sizes" Jour. Amer. Stat. Asso., 37, pp. 199-212, 1942.

(40) David, F. N. and Neyman, J. "Extention of the Markoff Theorem on Least Squares" Stat. Res. Mem. 2, 105, 1938.

(41) Hansen, M. H., Hurwitz, W. N. and Gurney, M. "Problems and Methods of a Sample Survey of Business" Jour. Amer. Stat. Asso., 41, pp. 173-189, 1946.

(42) Fisher, R. A. "Statistical Methods for Research Workers" Edinburgh, Oliver and Boyd. Section 14.

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

(48) Hasel, A. A. "Estimation of Volume in Timber Stands by Strip Sampling" Ann. Math. Stat. 13, pp. 179-206, 1942.

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 $\underline{x}$. Similarly, if it is desired to stratify the population according to the values of $\underline{x}$, a knowledge of the number of units in the population that have $\underline{x}$ values between specified limits is needed. This demands detailed information about the frequency distribution of $\underline{x}$ in the population. Quite often such information is lacking, or is known only roughly, for $\underline{x}$ variables that we would like to use in this way.

It may happen that $\underline{x}$ can be measured relatively cheaply by a sample. In this case, even though the purpose of a survey is to estimate a number of $\underline{y}$ variates, it may pay to devote part of the funds to a large preliminary sample in which $\underline{x}$ alone is measured. From this sample we can make a good estimate of the population mean of $\underline{x}$, 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 $\underline{x}$. Of course, by devoting funds to a special sample for $\underline{x}$, we must cut down the size of the main survey on $\underline{y}$. 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 $\underline{x}$. 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 $\underline{x}$ 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 $\underline{y}$ is measured: n_j units are drawn from the j th stratum. As usual, the variance within the j th stratum is denoted by σ_j^2. In the simplest case, the cost of the two samples will be of the form

$$C = c_1 L + c_2 n \ , \qquad (171)$$

where c_2 is presumed large relative to c_1.

The problem is to choose $\underline{L}$ and the $\underline{n}_j$ (and consequently $\underline{n}$) 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

$$\sum_j W_j \bar{y}_{pj} \cdot$$

As estimate we use

$$\sum_j w_j \bar{y}_{sj} \cdot$$

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

$$w_j = W_j + u_j \quad ; \quad \bar{y}_{sj} = \bar{y}_{pj} + e_j \cdot$$

Then the error of estimate may be expressed as

$$\sum_j (w_j \bar{y}_{sj} - W_j \bar{y}_{pj}) = \sum_j (W_j e_j + u_j \bar{y}_{pj} + u_j e_j) \qquad (172)$$

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

$$E\left[\sum_j (W_j e_j + u_j \bar{y}_{pj} + u_j e_j)^2\right]$$

$$= \sum_j \left[W_j^2 E(e_j^2) + \bar{y}_{pj}^2 E(u_j^2) + E(u_j^2) E(e_j^2)\right] ,$$

all other terms vanishing when the expectation is taken. This gives

$$\sum_j \left[\frac{W_j^2 \sigma_j^2}{n_j} + \frac{\bar{y}_{pj}^2 W_j (1-W_j)}{L} + \frac{W_j(1-W_j)}{L} \cdot \frac{\sigma_j^2}{n_j}\right] \qquad (173)$$

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 $e_j e_k$, 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

$$\sum_{k>j} -2\bar{y}_{pj}\, \bar{y}_{pk}\, W_j\, W_k / L . \qquad (174)$$

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

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

Hence, the final form of the variance is

$$V = \sum_j \left[\left\{ W_j^2 + \frac{W_j\,(1-W_j)}{L} \right\} \frac{\sigma_j^2}{n_j} + \frac{W_j\,(\bar{y}_{pj} - \bar{y}_p)^2}{L} \right] . \qquad (175)$$

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 large 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

$$\sigma_j \sqrt{W_j^2 + \frac{W_j(1-W_j)}{L}} .$$

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

$$n_j = n\ (W_j\ \sigma_j)\ /\ \Sigma\ (W_j\ \sigma_j)\ .$$

If this value is substituted into (175), with the term in $W_j(1-W_j)$ ignored, we obtain

$$V' = \frac{(\Sigma\ W_j\ \sigma_j)^2}{n} + \Sigma\ \frac{W_j(\bar{y}_{pj}-\bar{y}_p)^2}{L} \qquad (176)$$

$$= \frac{a}{n} + \frac{b}{L} \quad \text{(say)}. \qquad (177)$$

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

$$\frac{n}{L} = \left[\frac{ac_1}{bc_2}\right]^{\frac{1}{2}} \qquad (178)$$

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

$$C = 100 = 0.1L + n. \qquad (179)$$

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:

Strata	W_j	σ_j^2	σ_j	$\bar{y}_{pj}$
1	.786	312	17.7	19.404
2	.214	922	30.4	51.626
Complete pop.		620		26.297

We find $a = (\Sigma W_j \sigma_j)^2 = 417,$

$$b = \Sigma W_j (\bar{y}_{pj} - \bar{y}_p)^2 = 175$$

so that $$\frac{n}{L} = \sqrt{\frac{417}{175} \cdot \frac{1}{10}} = .488 .$$

From the cost equation (179) we obtain

$$L = \frac{100}{.588} = 170 \; ; \; n = 170 \times .488 = 83 .$$

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

$$V_{min} = \frac{417}{83} + \frac{175}{170} = 5.02 + 1.03 = 6.05$$

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

$$V = \frac{620}{100} = 6.20$$

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 $\underline{x}$ 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

$$y = \alpha + \beta (x - \bar{x}_p) + e \qquad (180)$$

where $\underline{e}$ 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 $\underline{x}$: in the second, of size $\underline{n}$, we measure both $\underline{x}$ and $\underline{y}$. The estimate of $\bar{y}_p$ is

$$y_{ds} = \bar{y}_n + b(\bar{x}_L - \bar{x}_n) \qquad (181)$$

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

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

If we consider the $\underline{x}$ 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

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

As is typical of regression formulae, the variance depends on the sets of $\underline{x}$ 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

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

which may be re-written

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

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 $\underline{x}$, and so reduce the contribution from the sampling error of $\underline{b}$. 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 $s^2_{y.x}$ from the regression is an unbiased estimate of $\sigma^2_y (1-\rho^2)$. 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).

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

Since $s^2_{y.x}$ is an unbiased estimate of $\sigma^2_y (1-\rho^2)$ and since

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

is an unbiased estimate of σ^2_y, it follows that

$$E (s^2_y - s^2_{y.x}) = \rho^2 \sigma^2_y \quad .$$

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

$$s^2_{y.x} \left[\frac{1}{n} + \frac{(\bar{x}_L - \bar{x}_n)^2}{\Sigma (x-\bar{x}_n)^2} \right] + \frac{(s^2_y - s^2_{y.x})}{L} \qquad (186)$$

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

$$s^2_{y.x} \left[\left(\frac{1}{n} - \frac{1}{N} \right) + \frac{(\bar{x}_L - \bar{x}_n)^2}{\Sigma (x-\bar{x}_n)^2} \right] + (s^2_y - s^2_{y.x} \left(\frac{1}{L} - \frac{1}{N} \right) . \qquad (187)$$

A development of the theory has been given by Chaneli 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

(43) Watson, D. J. "The Estimation of Leaf Areas" Jour. Agr. Sci., 27, p. 474, 1937.

(44) Cochran, W. G. "The Use of the Analysis of Variance in Enumeration by Sampling" Jour. Amer. Stat. Asso., 37, pp. 199-212, 1939.

(45) Wilm, H. G., Costello, D. F. and Klipple, G. E. "Estimating Forage Yield by the Double-Sampling Method" Jour. Amer. Soc. Agron., 36, pp. 194-203, 1944.

(46) Neyman, J. "Contribution to the Theory of Sampling Human Populations" Jour. Amer. Stat. Asso., 33, pp. 101-116, 1938.

(47) Bose, Chameli "Note on the Sampling Error in the Method of Double-Sampling" Sankhya, 6, p. 330, 1943.

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 $l(z)$ 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 $\underline{n}$, 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 $\underline{n}$ 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 σ 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 $l(z)$ 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 $N^2 \sigma^2/n$. Hence

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

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

$$cn + \frac{S N \sigma}{\sqrt{2\pi\, n}}$$

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 dute 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 Deming (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.

	d.f.
Between strata	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.

(54) Sampling Staff, Bureau of the Census. A Chapter in Population Sampling. U. S. Government Printing Office, Washington, D. C. 1947. A detailed account of a sample which covered a number of large cities, the principal object being to investigate overcrowding. A stratified sample of city blocks was used, with subsampling of the blocks.

(38) Jessen, R. J., Blythe, R. H., Kempthorne, O., and Deming, W. Edwards. On a Population Sample for Greece. Jour. Amer. Stat. Asso. 42, pp. 357-384, 1947. A population sample extending over a whole country where no previous sampling work had been done.

(55) Yates, F. and Finney, D. J. Statistical Problems in Field Sampling for Wire Worms. Ann. Appl. Biol., 29, pp. 156-167, 1942. An extensive sample of farm fields.

(56) Cornell, F. G. A Stratified Sample of a Small Finite Population. Jour. Amer. Stat. Asso., 42, pp. 523-532, 1947. A mail sample of universities and colleges in the U. S., in order to obtain estimates of total enrollments.

(57) Blankenship, A. B. (Editor) How to Conduct Consumer and Opinion Research. Harper and Bros., New York, 1946. An excellent source for information about many of the commercial uses of sampling.

(58) Hendricks, W. A. Mathematics of Sampling. Va. Agr. Exp. Sta. Special Tech. Bull., 1948. This contains a series of lecture notes covering approximately the same ground as the present notes, and is highly recommended as supplementary reading.

REFERENCES

(48) Blythe, R. H. "The Economics of Sample Size Applied to the Scaling of Sawlogs" Biom. Bull., 1, pp. 67-70, 1945.

(49) Nordin, J. A. "Determining Sample Size" Jour. Amer. Stat. Asso., 39, pp. 497-506, 1944.

(50) King, A. J. and Jessen, R. J. "The Master Sample of Agriculture" Jour. Amer. Stat. Asso., 40, pp. 38-56, 1945.

(51) Hansen, M. H. and Hauser, P. M. "Area Sampling - Some Principles of Sample Design" Public Opinion Quart., pp. 183-193, Summer 1945.

(52) Hansen, M. H. and Dening, W. Edwards. "On Some Census Aids to Sampling" Jour. Amer. Stat. Asso., 38, pp. 353-357, 1943.

(53) Mahalanobis, P. C. "Recent Experiments in Statistical Sampling in the Indian Statistical Institute" Jour. Roy. Stat. Soc., 109, pp. 325-378, 1946.