pp 17-20

Estimation of Sampling Errors in Multi-Stage Surveys

by

J. Durbin

The Johns Hopkins University*

and

The London School of Economics and Political Science

1. Introduction.

Recent developments in high-speed data-processing equipment make it possible in principle to compute valid estimates of sampling error on a routine basis even for highly complex sample designs. Such estimates are needed not only so that standard errors may be attached to the quantities produced by the survey but also, and this point is often overlooked in theoretical discussions, to assist in the maintenance of effective cost control. Because of the close relation between costs and variances the magnitude of the sampling error provides the client with a useful indicator of the reliability of the information collected in relation to the cost of the survey. Furthermore, it is hard to see how he can make a rational choice between alternative sample designs and estimation methods for future surveys except on the basis of reliable estimates of sampling errors obtained from similar surveys in the past.

Yet of the large body of theory of variance estimation for surveys available in the literature very little has ever been used in practice. The reason is that the sheer volume of data collected in a large-scale survey precludes the use of any but the simplest formulae if computing time is to be kept within reasonable limits. This remains true even when the most advanced equipment is employed for processing the data. Because of this,

* This research was supported in part by the ~~Air Force Office of Scientific Research by Contract No. 49(638)-1302~~ Office of Naval Research awarded to the Department of ~~Statistics, The Johns Hopkins University, Baltimore, Maryland.~~

the survey worker who wishes to estimate his sampling errors is usually faced with the choice between an inefficient sample design for which an adequate theory is available or a design which seems to be obviously more efficient but for which approximations of unknown validity are required if variances are to be computed. There is thus a strong need for the development of efficient survey designs for which the error-estimation procedures appear feasible and worth-while to the practical worker and are at the same time adequate from a theoretical point of view.

In this paper we shall consider the problem of error estimation for a class of multi-stage designs from this standpoint. Although at the outset the approach is ~~admittedly~~ theoretical, the main effort in the work to be described lies in the attempt to develop simplifications intended to facilitate the use of the methods in practice. The designs considered are those in which first-stage units are stratified and the selections made with probability proportional to size within strata. It is well known that when the selection of first-stage units is carried out with replacement, simple procedures are available for the estimation of sampling errors. The replicated sample designs of Deming (1960) and the interpenetrating designs of Mahalanobis (1946) make use of this property. However, in spite of the theoretical advantages of with-replacement sampling few survey practitioners are prepared to use it in practice. They dislike having to include the same unit twice or more in the sample, as they are sometimes required to do by the theory. In my view they are right about this. The fact that repititions are possible in with-replacement sampling clearly demonstrates the inefficiency of the method.

A way out adopted by some is to carry out the sampling without replacement but to calculate the variance as if the sampling had been done with replacement. The objection to this is that the resulting estimate of variance is biased. It has been shown (Durbin, 1953) that for the common design of two first-stage units per stratum the bias is twice the reduction in variance achieved by using without-replacement sampling instead of with-replacement sampling. The degree of bias is not, therefore, negligible and is certainly worth removing if this can be done in a way that is feasible from a practical point of view. Some evidence on the amount of bias has been given by Des Raj (1964); more is given below in section 8 .

The main object of the work described in this paper was to develop methods of selection and variance-estimation for sampling without replacement which preserve as much as possible of the simplicity of with-replacement error calculation. We consider only samples of size two since this is by far the most important case in practice. Samples of size two permit the maximum degree of stratification consistent with the provision of a contribution to the estimate of error from every stratum. We consider two methods of selection. The first is described in section 2 and is exact in the sense of achieving specified probabilities of inclusion and at the same time giving unbiased estimates of sampling variances of linear estimates. Sections 4-6 discuss modifications designed to simplify both the selection and the variance-estimating procedures. Even with the maximum degree of simplication, however, this method may be found too complicated in some situations. An approximate method is therefore discussed in section 6 which leads to calculations scarcely any more

complicated than those for sampling with replacement. Finally, some numerical results obtained by applying the methods to British election statistics are presented.

2. Selection of two units with unequal probabilities without replacement (Method 1).

We suppose that there are N units (usually comprising a stratum of first-stage units) with sizes $x_1, \ldots, x_N$. Let $p_1, \ldots, p_N$ denote the selection probabilities defined by $p_i = x_i / \sum x_j$. Clearly $\sum p_i = 1$. The requirements we shall impose on the selection procedure are threefold:

1. The probability of inclusion of a unit should be strictly proportional to its size.
2. The calculations needed to apply the method must be simple.
3. The joint probability of inclusion of each pair of units should be calculable. (This is needed for variance estimation).

Between them these requirements eliminate many well-known methods. For instance, requirement 1 eliminates that due to Rao, Hartley and Cochran (1962). Requirement 2 eliminates methods based on iterative calculations such as that proposed by Yates and Grundy (1953). Requirement 3 eliminates systematic selection down the cumulated size/scale, as considered by Hartley and Rao (1962).

The proposed method (Method 1) is as follows. Choose the first unit with probability p_i and the second with probability proportional to

$$(1) \qquad p_j \left(\frac{1}{1 - 2p_i} + \frac{1}{1 - 2p_j} \right) \qquad (j \neq i) .$$

Then the total probability π_i of inclusion of the i^{th} unit is $2p_i (i = 1, \ldots, N)$.

To prove this, let the conditional probability of choosing the j^{th} unit second given that the i^{th} unit is chosen first be denoted by $p_{j\cdot i}$. Then

$$p_{j\cdot i} = \lambda_i p_j \left(\frac{1}{1 - 2p_i} + \frac{1}{1 - 2p_j}\right) .$$

Since $\sum_{j\neq i} p_{j\cdot i} = 1$ we have,

$$\lambda_i^{-1} = \sum_{j\neq i} p_j \left(\frac{1}{1 - 2p_i} + \frac{1}{1 - 2p_j}\right)$$

$$= \frac{1 - p_i}{1 - 2p_i} + \sum_{k=1}^{N} \frac{p_k}{1 - 2p_k} - \frac{p_i}{1 - 2p_i}$$

$$= 1 + \sum_{k=1}^{N} \frac{p_k}{1 - 2p_k} .$$

Thus λ_i does not depend on i and its suffix may be dropped. The probability of getting the i^{th} unit first and the j^{th} unit second is therefore

$$p_i p_{j\cdot i} = \lambda p_i p_j \left(\frac{1}{1 - 2p_i} + \frac{1}{1 - 2p_j}\right) .$$

Since this is symmetric in i and j it is also the probability of getting the j^{th} unit first and the i^{th} unit second. The total probability of getting the i^{th} unit second therefore equals the total probability of getting the i^{th} unit first which is p_i by the method of selection. The probability of inclusion of the i^{th} unit is therefore $\pi_i = 2p_i$. The joint probability of choosing the i^{th} and j^{th} units together is

$$(2) \qquad \pi_{ij} = 2\lambda p_i p_j \left(\frac{1}{1 - 2p_i} + \frac{1}{1 - 2p_j} \right) \quad (j \neq i = 1,\ldots,N) \ ,$$

where

$$(3) \qquad \lambda = \left(1 + \sum_{k=1}^{N} \frac{p_k}{1 - 2p_k} \right)^{-1} \ .$$

The method can obviously be applied whenever $\max p_i < \frac{1}{2}$. If $\max p_i = \frac{1}{2}$ the largest unit should always be included, the second unit being chosen from the remainder with probability proportional to p_i. Now the condition for selection with probability proportional to p_i to be possible is that $\max p_i \leq \frac{1}{2}$. This follows since the probability of inclusion of the i^{th} unit must be $2p_i$ and this cannot be greater than one. Thus Method 1 can be used whenever sampling with probability proportional to p_i without replacement is possible.

It turns out that Method 1 gives the same joint probabilities of inclusion as a method due to Brewer (1963), although it was developed in ignorance of Brewer's work. In Brewer's method a unit is first selected with probability proportional to $p_i(1 - p_i)/(1 - 2p_i)$ and a second unit with probability proportional to $p_j (j \neq i)$. The joint probability of including the i^{th} and j^{th} units in the example is therefore proportional to

$$\frac{p_i(1 - p_i)}{1 - 2p_i} \cdot \frac{p_j}{1 - p_i} + \frac{p_j(1 - p_j)}{1 - 2p_j} \cdot \frac{p_i}{1 - p_j} \ ,$$

which gives the same π_{ij} as (2). I am indebted to Dr. J.N.K. Rao for calling my attention to this point. Method 1 has an important advantage over Brewer's method, however, since unlike Brewer's method it permits the use of the grouping device considered in the next section; this can result in substantial computational savings.

As was indicated in Durbin (1965), the method can sometimes be extended to permit the selection of samples of sizes greater than two. Thus for samples of size three, one takes selection probabilities p_k'' for the third unit proportional to $p_k' [(1-2p_j')^{-1} + (1-2p_k')^{-1}]$ $(k \neq i,j)$ where $p_j' = \lambda^{-1} p_j \{(1-2p_i)^{-1} + (1-2p_j)^{-1}\}$ $(j \neq i)$, the first two units selected being the i^{th} and the j^{th}. For samples of size four one could take probabilities proportional to $p_l''\{(1-2p_k'')^{-1} + (1-2p_l'')^{-1}\}$ and so on. Unfortunately the method does not always work since the selection probabilities may under certain circumstances become negative. In practice this is unimportant since samples of size two are adequate for variance estimation.

It is worth noting that Method 1 can be employed with the Lahiri selection procedure which avoids the listing of sampling units. Lahiri's procedure is as follows. Denoting the sizes of the units by $x_1, \ldots, x_N$ let x_o be a number such that $x_o \geq x_i$ for all i. Let i be a random digit from 1 to N and let u be a random number in the range $0 \leq u \leq x_o$. If $u \leq x_i$ accept the i^{th} unit. If $u \geq x_i$ reject the i^{th} unit, choose another random pair (i,u) and repeat the operation. Continue until a unit has been accepted. This ensures that a unit has been selected with probability proportional to x_i. The same device is now used for the selection of the second unit using $x_j' = x_j\{(X-2x_i)^{-1} + (X-2x_j)^{-1}\}$ $(j \neq i)$, where $X = \sum_1^N x_i$, in place of x_j and $x_o' \geq \max x_j'$ in place of x_o. How useful the technique is in this context is arguable, since if variance estimates are to be calculated we need to know the value of λ in order to compute π_{ij} from (2). Instead of using Lahiri's method and computing λ from (3) it might well be simpler to note that $p_{j \cdot i} = \lambda x_j'$ and to obtain λ from the condition $\sum_{j \neq i} p_{j \cdot i} = 1$.

3. Estimation of variance.

Consider first stratified single-stage sampling and suppose we want to estimate the variance of the linear estimator

$$t = \sum_{h=1}^{k} (y_{hi} + y_{hj}) \quad , \tag{4}$$

where y_{hi} and y_{hj} are the contributions from the two units selected from the h^{th} stratum ($h = 1,\ldots,k$). The variance of this is

$$V(t) = \sum_{h=1}^{k} \sum_{i \neq j} (\pi_{hi}\,\pi_{hj} - \pi_{hij})(y_{hi} - y_{hj})^2 \quad , \tag{5}$$

where π_{hi} and π_{hj} are the probabilities of selection of the i^{th} and j^{th} units from the h^{th} stratum and π_{hij} is the probability that they are chosen together. An unbiased estimate is

$$\hat{V}(t) = \sum_{h=1}^{k} \left(\frac{\pi_{hi}\,\pi_{hj}}{\pi_{hij}} - 1 \right) (y_{hi} - y_{hj})^2 \quad . \tag{6}$$

Formulae (5) and (6) are due to Yates and Grundy (1953).

For the multi-stage case, which is our main concern in this paper, we consider the same estimator t given by (4) except that for given i and j, y_{hi} and y_{hj} are now random variables based on sampling at the second and successive stages. The unbiased estimate of variance is now

$$\hat{V}(t) = \sum_{h=1}^{k} \left(\frac{\pi_{hi}\pi_{hj}}{\pi_{hij}} - 1 \right) (y_{hi} - y_{hj})^2 + \sum_{h=1}^{k} \left(\pi_{hi}\, s_{hi}^2 + \pi_{hj}\, s_{hj}^2 \right) \quad , \tag{7}$$

where s_{hi}^2 and s_{hj}^2 are unbiased estimates of the variances of y_{hi} and y_{hj}

An additional note of warning should be mentioned in regard to formulae (6) and (7). The factor $\pi_{hi}\,\pi_{hj}\,\pi_{hij}^{-1} - 1$ is a random variable which fluctuates from sample to sample. It is possible for it to take large values and this introduces instability into the estimate of variance. In this respect these formulae compare unfavorably with the with-replacement formula where the coefficient of $(y_{hi} - y_{hj})^2$ is constant. Before using (6) and (7) in practice one should therefore include some provision for damping down any instability arising from this effect. This will be borne in mind in putting forward specific proposals in later sections of the paper. Meanwhile it is worth mentioning that an effective way of dealing with the problem is to replace $\pi_i\,\pi_j\,\pi_{ij}^{-1} - 1$ by one whenever it exceeds one. The bias resulting from this device is completely negligible in practice and the variability of the resulting estimate of variance should be less than that of the with-replacement estimate.

Of course, in practice very few estimators take the linear form (4). Most estimators in survey work are ratio estimators of the form

$$(10) \qquad r = \sum_{h=1}^{k} (y_{hi} + y_{hj}) \Big/ \sum_{h=1}^{k} (y'_{hi} + y'_{hj}) = t/t' \quad \text{say} \quad .$$

However, as is well known, good approximate variance formulae for r can be obtained by replacing the y_{hi} in formulae for the variance of t by $(y_{hi} - r\,y'_{hi})/t'$. Consequently, for the discussion of generalities such as concern us here we may base the treatment on the linear form (4).

4. Grouping of units within strata.

The work of sample selection and variance estimation can often be reduced very substantially by the grouping device now to be described. This possibility arises since Method 1 can be employed whenever max $p_i \leq \frac{1}{2}$. It is often the case that the size of the largest unit in a stratum is substantially less than half the total size of the stratum. The units are arranged in groups such that each group contains as few units as possible subject to the requirement that the max p_i within each group $\leq$ half the total of the p_i's in the group. Now select two units from the stratum with replacement with probability proportional to p_i. If two units from different groups are selected accept both. If both selections give units in the same group (not necessarily different), accept the first and reject the second, making a further selection from the group by Method 1 with p_i replaced by $p_i' = p_i / \sum' p_i$, where $\sum'$ denotes summation over the units in the group.

The probability of getting i^{th} unit first and the j^{th} unit second is now $p_i p_j$ if the units come from different groups and is

$$\lambda' p_i \left(\sum' p_i \right) p_j' \left(\frac{1}{1 - 2p_i'} + \frac{1}{1 - 2p_j'} \right) , \tag{11}$$

where

$$\lambda' = \left\{ 1 + \sum' \frac{p_k'}{1 - 2p_k'} \right\}^{-1} \tag{12}$$

if the units come from the same group. In both cases this is a symmetric function of i and j so the total probability of the i^{th} unit coming second is the same as the probability of the i^{th} unit coming first, namely p_i. Thus the probability of inclusion of the i^{th} unit is $\pi_i = 2p_i$ as before.

This device can obviously yield substantial economies in the selection of the sample when the maximum group sizes within strata are not too large in relation to total strata sizes. When units from different groups are chosen the selection is as simple as for sampling with replacement. The extra calculations necessary for the application of Method 1 only have to be carried out when the same group is chosen on both selections.

The economies in the calculation of variance estimates are, however, much greater. In the first place the factor $\pi_{hi}\,\pi_{hj}\,\pi_{hij}^{-1} - 1$ equals unity whenever the units are selected from different groups since $\pi_{hij} = 2p_{hi}\,p_{hj}$, where p_{hi} is the selection probability for the i^{th} unit in the h^{th} stratum. Secondly, we need only add a contribution for sampling at the second stage whenever the units come from the same group. (For simplicity we shall from now on speak of "the second stage" when we mean "the second and subsequent stages". The formulae we give apply to multi-stage designs with any number of stages). This can be seen from the following.

Let $y_{hi} = \mu_{hi} + d_{hi}$, where μ_{hi} is the conditional expectation for fixed i of y_{hi} due to sampling at the second-stage. The contribution to the overall variance V(t) arising from second-stage sampling in the h^{th} stratum is $E(d_{hi} + d_{hj})^2$. This equals $E(d_{hi} - d_{hj})^2$ since second-stage sampling within different first-stage units is done independently. We also have

$$E(y_{hi} - y_{hj})^2 = E(\mu_{hi} - \mu_{hj} + d_{hi} - d_{hj})^2$$

$$= E(\mu_{hi} - \mu_{hj})^2 + E(d_{hi} - d_{hj})^2 \quad .$$

Thus if the first-stage contribution to the estimate of variance is $(y_{hi}-y_{hj})^2$, as is the case where the i^{th} and j^{th} units come from different groups, this automatically includes the correct allowance for the component of variance due to sampling at the second stage. Thus no additional second-stage component need be added.

For the first-stage component when the units come from the same group we need the factor $\pi_{hi}\ \pi_{hj}\ \pi_{hij}^{-1} - 1$. From (11) we have

$$\pi_{hij} = 2\lambda' p_i p_j \left(\frac{1}{1 - 2p_i} + \frac{1}{1 - 2p_j} \right)$$

where we drop the suffix h from the p's. Thus

$$(1') \qquad \pi_{hi}\ \pi_{hj}\ \pi_{hij}^{-1} - 1 = \frac{2}{\lambda'\{(1 - 2p_i)^{-1} + (1 - 2p_j)^{-1}\}} - 1 ,$$

which is the same value that would have been obtained if the stratum had consisted entirely of this one group. It can easily be put in the form (8) with p replaced by p' and N replaced by the number of units in the group.

To obtain the appropriate second-stage component we see from (7) that what we require to estimate is $E\left\{ \sum_{h=1}^{k} (\pi_{hi}\ s_{hi}^2 + \pi_{hj}\ s_{hj}^2) \right\}$ $= \sum_{h=1}^{k} \sum_{i} \pi_{hi}^2\ \sigma_{hi}^2$ where σ_{hi}^2 is the variance of y_{hi} due to second-stage sampling. An unbiased estimator of this is obtained by including a term $\pi'_{hi}\ s_{hi}^2 + \pi'_{hj}\ s_{hj}^2$ from the h^{th} stratum only when the two units from the stratum came from the same group. By π'_{hi} we mean $2p_{hi} / \sum' p_{hi}$ where $\sum'$ indicates summation over units in the same group while as before s_{hi}^2 denotes

an unbiased estimator of the variance of y_{hi} due to sampling at the second stage. When the units come from different groups no contribution is included.

The estimator is unbiased since conditional on the two units coming from a particular group $E(\pi'_{hi}\, s_{hi}^2 + \pi'_{hj}\, s_{hj}^2) = \sum' \pi'^2_{hi}\, \sigma_{hi}^2$. The probability of getting two units from that group is $\left(\sum' p_{hi}\right)^2$. Thus the overall expectation is $\left(\sum' p_{hi}\right)^2 \sum' \pi'^2_{hi}\, \sigma_{hi}^2$ summed over all groups i.e. $\sum_i \pi_{hi}^2\, \sigma_{hi}^2$ where $\sum_i$ indicates summation over all units in the stratum.

The results of this section on variance estimation can be summarized in the following interesting and useful

RULE: For strata in which the selected units come from different groups compute the contribution to the variance as if the sampling had been carried out with replacement. For strata in which the units come from the same group compute the contribution to the variance as if each stratum consisted only of the group containing the units.

This rule should be applied to both first- and second-stage components. It obviously simplifies greatly the task of variance estimation. However, the same note of warning should be mentioned as was given at the end of the previous section about the importance of avoiding undue instability in the factors $\pi'_{hi}\, \pi'_{hj}\, \pi'^{-1}_{hij} - 1$. This factor should be replaced by unity whenever it exceeds unity.

5. Further simplification of variance formulae.

The grouping device in suitable cases takes us a good part of the way towards making variance estimation for without-replacement sampling comparable in simplicity with that for with-replacement sampling. We now consider what further progress can be made in this direction. There are two things to aim at, first the elimination of the irritating factors $\pi'_{hi}\ \pi'_{hj}\ \pi'^{-1}_{hij} - 1$ and secondly the simplification of the procedure for obtaining the second-stage component s^2_{hi}.

We have already noted that where $\pi'_{hi}\ \pi'_{hj}\ \pi'^{-1}_{hij} - 1$ exceeds unity it should be replaced by unity. We now propose that where it is less than unity it should be replaced by unity with probability $\pi'_{hi}\ \pi'_{hj}\ \pi'^{-1}_{hij} - 1$ and by zero otherwise. This will clearly not affect the expected value of the estimate.

The selection of the groups for which factors are replaced by unity should be done in the following way. Suppose there are q strata for which the units come from the same group and have factors less than one. Denote the factors $\pi'_{hi}\ \pi'_{hj}\ \pi'^{-1}_{hij} - 1$ by $a_1,\ldots,a_q$. Form the cumulative sums $\sum_1^s a_r$ $(s = 1,\ldots,q)$. Let u be a random number between 0 and 1. If any of the quantities $u, u+1,\ldots,u+q-1$ lie between $\sum_1^{s-1} a_r$ and $\sum_1^s$ then the s^{th} factor is replaced by one, otherwise it is replaced by zero. This method ensures that the probability that the s^{th} group has factor unity is a_s while keeping the number of unit factors fairly stable.

When the factor is unity the contribution to the first-stage component of variance is, of course, $(y_{hi} - y_{hj})^2$, which is the same as if the sampling had been done with replacement. As in the last section the

term $(y_{hi} - y_{hj})^2$ contains automatically the correct allowance for second-stage sampling. Thus we need only consider the inclusion of second-stage components from strata for which the factor is zero, i.e. the first-stage component is absent. A suitable unbiased estimator is therefore obtained by including the second-stage component $s_{hi}^2 + s_{hj}^2$ when the first-stage component $(y_{hi} - y_{hj})^2$ is absent and by excluding it when $(y_{hi} - y_{hj})^2$ is present.

Let us now consider in detail the construction of a satisfactory form for the second-stage component. It would clearly be a great advantage from the standpoint of the use of data-processing equipment if the second stage component took the same form as the first-stage component $(y_{hi} - y_{hj})^2$. This can be arranged by designing the sample so that within each of the two first-stage unitsfrom strata for which a second stage component of variance is required the sample is laid out in the form of two independent interpenetrating subsamples each of half the total size required. If the contributions of the two subsamples from the i^{th} unit to the overall estimate t are denoted by y_{hi1} and y_{hi2}, and similarly for the j^{th} unit, the appropriate second-stage component is $(y_{hi1} + y_{hj1} - y_{hi2} - y_{hj2})^2$. This evidently has the same form as the first-stage component $(y_{hi} - y_{hj})^2$ that would have been used if required. The only difference is that for the first-stage component we square the difference between units while for the second-stage component we square the difference across units.

Summing up the main points of this procedure we note that each stratum contributes a single component to the estimate of the sampling variance. This component is either a first-stage component or a second-stage component. If for a particular stratum the two units come from

different groups, or the factors $\pi'_{hi}\,\pi'_{jh}\,\pi'^{-1}_{hij} - 1$ has been replaced by one, the appropriate component is the first-stage component $(y_{hi} - y_{hj})^2$. If the two units come from the same group and the factor $\pi'_{hi}\,\pi'_{hj}\,\pi'^{-1}_{hij} - 1$ has been replaced by zero the appropriate component is the second-stage component $(y_{hi1} + y_{hj1} - y_{hi2} - y_{hj2})^2$. In no case does one need both a first-stage component and a second-stage component from the same stratum.

It is evident that the estimate of variance obtained in this way is no more complicated so far as data-processing is concerned than the with-replacement estimate - in fact it has an identical form. The price that is paid for the advantages of without-replacement sampling lies in the use of a more complicated procedure for sample selection . In this connection it should be emphasised that interpenetrating samples at the second stage are only needed for strata designated to provde a second-stage component of variance to the estimate. In all other cases, whether or not the first-stage units come from the same group, the sampling used at the second and later stages can take any form whatsoever provided only that it is free from bias and that the sampling is carried out independently in different first-stage units.

6. A useful approximate method (Method 2).

For situations in which Method 1 is regarded as too complicated the following is worthy of consideration. The method, which is almost as simple as sampling without replacement but is more efficient, is a slight variant of one ~~described by~~ due to Kish (1965 p. 229). It can be applied whenever units are not too dissimilar in size within strata .

It may also be regarded as a special case of Stenning's (1958) method in ~~which~~ the number of units per group of equi-sized units equals two

As far as possible, strata are arranged to have even numbers of single-stage units. Units are listed in order of size within strata and are marked off in adjacent pairs. Two selections are made with replacements with probability proportional to size. If the two selections result in two different units these are accepted. If they give the same unit this is accepted and the other number of the pair is taken. The case of strata with odd numbers of units is considered at the end of the section.

The method is approximate only to the extent that members of the same pair differ in size. If pairs consisted exactly of equi-sized units the method would give unbiased estimates, unbiased estimates of error and would be more efficient than sampling with replacement. Assuming this to be true, the factor $\pi_{hi}\,\pi_{hj}\,\pi_{hij}^{-1} - 1 = 1$ when the units selected belong to different pairs and $\ldots = 0$ when they belong to the same pair. This follows since $\pi_{hij} = 2p_{hi}\,p_{hj}$ when the units belong to different pairs and $\ldots = 4\,p_{hi}^{2}$ when they belong to the same pair. We therefore have a situation like that considered in the last section in which the first-stage component of variance $(y_{hi} - y_{hj})^2$ has a coefficient of one in some strata and zero in others. As shown there, when the coefficient is one the second-stage component of variance is automatically allowed for, and when the coefficient is zero an estimate of the second-stage component is all that is required from that stratum.

As before we therefore have the following rules for the sample layout and for variance estimation. Where the units come from different pairs the sampling at the second and later stages can take any form whatever that is free from bias provided it is carried out independently in the two units. The appropriate contribution to the variance is the first-stage component $(y_{hi} - y_{hj})^2$. Where the units belong to the same pair the

Of course, listing is only required either to obtain a substitute for a repetition or to ascertain whether two units belong to the same pair when calculating the variance. Thus it need not be carried out in advance in practice. Only when the two units are adjacent when ordered by size in listing needed to see whether they belong to the same pair. If the units are not adjacent this will usually be obvious on inspection.

-19-

second-stage sample is arranged in the form of pair of independent interpenetrating samples within each first-stage unit. The appropriate contribution to the variance is then $(y_{hi1} + y_{hj1} - y_{hi2} - y_{hj2})^2$ where y_{hi1}, y_{hi2} are the contributions to t of the two half-samples in the i^{th} unit. The computer program for variance estimation is therefore as simple as for sampling with replacement.

The method is so attractive that it is worth considering what modifications would be needed to make it exact. The simplest possibility would be to replace the unequal probabilities p_i and p_j in a pair by equal values $\frac{1}{2}(p_i + p_j)$, making suitable modifications of the sampling fractions at the later stages of sampling in order to preserve the overall probabilities of inclusion of the final-stage units. With this modification the method is exact. Alternatively, the initial selection probabilities could be modified to make the probabilities of inclusion of the first-stage units strictly proportional to the p_i. One difficulty with this is that the factors $\pi_{hi}\,\pi_{hj}\,\pi_{hij}^{-1} - 1$ are no longer exactly 1 or 0 . My own preference would be to regard Method 2 as a good approximate method which is definitely preferable to sampling with replacement and to use Method 1 if an exact method is required.

It remains to indicate suitable modifications for dealing with strata containing odd numbers of units. Arrange in pairs as before except for the smallest three units, say the i, j, k^{th}. If two distinct units of the three are chosen accept them. If i is repeated take i and j, if j is repeated take j and k and if k is repeated take k and i . Whichever pair is chosen give it a 1/3 chance of contributing a first-stage component to the variance of the form $(y_{hi} - y_{hj})^2$ and a 2/3 chance of

contributing a second-stage component of the form $(y_{hi1} + y_{hj1} - y_{hi2} - y_{hj2})^2$. Assuming that the three units are equi-sized and that members of pairs are equi-sized this gives an unbiased estimate of variance.

7. Reduction of number of degrees of freedom.

In spite of the simplifications of the previous sections further measures may be needed. The reason is that for a design based on k strata, and therefore with 2k first-stage units, essentially 2k different tabulations are required for each variable whose variance is to be computed. When the number of variables is large and the number of strata is large this may make excessive demands on the capacity of the computer. The estimates of variance discussed so far are based essentially on k degrees of freedom and it may be that estimation based on fewer degrees of freedom may be adequate. For instance, Deming (1960) suggests that ten degrees of freedom should be adequate though this seems to me to be on the low side.

If the number of strata seems too large it is an easy matter to aggregate strata and thus to reduce the amount of data-processing needed. For instance, suppose the contributions from the first two strata to t are y_{11}, y_{12} and y_{21}, y_{22}. Instead of taking as their contribution to the variance the two degrees of freedom $(y_{11} - y_{12})^2 + (y_{21} - y_{22})^2$, one merely takes the single degree of freedom $(y_{11} + y_{21} - y_{12} - y_{22})^2$. By aggregating strata in pairs in this way one reduces the number of degrees of freedom from k to $\frac{1}{2}$ k. When the designs of the last two sections are used it is clear that it doesn't matter whether a particular stratum is contributing a first-stage or a second-stage component, it can still be aggregated in the same fashion. If the number of strata is odd the last

stratum is merely left unaggregated. If a drop in the number of degrees of freedom approximately to k/3, k/4 etc. is required the strata can be aggregated in threes, fours, etc. in a similar way.

8. Applications to election data.

In order to test some of the methods in practice, studies were made of their performance on a sample layout based on British 1964 General Election data. The fact that the Election results are published in full, constituency by constituency, makes them particularly suitable for the study of the validity of sampling methods since quantities of interest can be calculated exactly from the whole population instead of having to be based on a particular sample. A further point is that many sampling schemes used in practice in Britain are based on election data so the layout used here can be regarded as fairly comparable with sample designs in current use.

A scheme similar to that employed for political polling by National Opinion Polls Ltd., who generously helped with information and data, was set up. The sample consisted of 100 constituencies, two from each of 50 strata selected with probability proportional to the 1964 electorate. From each constituency 30 electors were to be chosen at random giving 3000 electors altogether. This scheme differs slightly in certain respects from that used in practice by N.O.P. In particular they do not choose electors at random within constituencies but use a clustering method.

The results for Method 1, with and without grouping are compared with the results for with-replacement sampling in Table 1. It is clear that the grouping device has a negligible effect on the variance. The first-stage component of the with-replacement variance is a little under 10% greater than the Method 1 value. The second-stage component is, of course, the same.

The results for ℓ and c are exact whereas those for ℓ' and c are approximate since they are ratio estimators. The comparison shows the increase of variance that results from the use of ratios instead of linear estimators.

Values for ℓ-c and ℓ'-c' are given since they are of special interest in discussing election statistics. The strong negative correlation between ℓ and c and between ℓ' and c' causes these to be substantially higher than the sums of variances of ℓ and c and ℓ' and c'. This is a point that is often neglected in discussions of sampling errors of voting comparisons.

The last two lines of Table 1 show the gain of efficiency that can be achieved by the use of known data from the previous election. The idea employed here is to use the survey to estimate the change from the last election to the next rather than to estimate the result of the next election in a single operation. The estimate of change is then combined with the known result of the previous election to give forecasts of greater efficiency. This seems a simple idea but as far as I know it is not used by polling organizations in their work on election forecasting. It appears that the overall variances of ℓ and ℓ-c are reduced by about a third by using the 1959 data. Incidentally, one of the side advantages of the routine calculation of sampling errors along the lines suggested in this paper would be that it would be a fairly straightforward matter to estimate the gain from alternative designs or estimation procedures. One has the feeling that many inefficient procedures remain in use simply because there is no objective measure available of how inefficient they are.

In applying Method 1 with grouping the factors $\pi'_{hi}\,\pi'_{hj}\,\pi'^{-1}_{hij} - 1$ are of interest. A frequency distribution of the 696 values of this factor for the 193 groups of the scheme - there were 154 groups of three units and 39 groups of four units - is shown in Table 2.

Table 2. Distribution of values of the factor $\pi'_{hi}\,\pi'_{hj}\,\pi'^{-1}_{hij} - 1$.

0-0.2	-0.4	-0.6	-0.8	-1.0	-1.2	-2.0	-5.0	>5.0	Total
108	244	190	97	25	13	11	5	3	696

The number greater than one is 32 which is 4.6% of the total. To assist in the assessment of the general level of values it is worth pointing out that with three equi-sized units in a group the value of the factor is $\frac{1}{3}$ and with four equi-sized units the value is $\frac{1}{2}$.

The expected proportion of cases in which units are chosen from the same group turned out to be 0.270 for the 618 constituencies arranged in 50 strata. This may be compared with a theoretical proportion of 0.250 for 600 equisized constituencies in 50 strata each of four groups of three units. For sampling with replacement the expected proportion of cases in which the same constituency would be chosen twice turned out to be 0.085.

Table 3 gives the results obtained when Method 2 was used to estimate ℓ, c and ℓ-c . For comparison the first-stage component of variance for Method 1 (without grouping) is also given. The bias is evidently completely negligible for this particular set of data. The mean-square error comparisons are rather surprising. They show Method 2

to be more accurate than Method 1 for two of the three quantities estimated On investigation it turned out that there was a slight negative correlation in the proportion voting Labour between constituencies in the same stratum having similar sizes. This seems to be rather a freak result which could not be expected to occur generally. Nevertheless the results are certainly very encouraging so far as the accuracy of Method 2 is concerned.

Table 2. Bias and mean-square error of Method 2.

(Each entry should be multiplied by 10^{-5}).

Quantity estimated	Bias	First-stage component only	
		Method 2 mean-square error	Method 1 variance (without grouping)
Proportion $\frac{\text{Labour vote}}{\text{Electorate}}$ (ℓ)	0.012	4.83	4.86
Proportion $\frac{\text{Conservative vote}}{\text{Electorate}}$ (c)	-0.961	3.70	3.69
Difference ℓ - c	0.973	12.51	12.56

Acknowledgements.

I am indebted to Leslie Kish for some helpful discussions and to Peter G. Hyett., managing director of National Opinion Polls Ltd., for providing the data used in section 8. The calculations were done by Clive D. Payne under the supervision of Susannah Brown.

References.

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. Australian J.Statist. 5, 5-13.

Cochran, W.G. (1963). Sampling Techniques. 2ndEd. New York: John Wiley and Sons.

Deming, W.E. (1960). Sample Designs in Business Research. New York: John Wiley and Sons.

Des Raj (1964). The use of systematic sampling with probability proportionate to size in a large-scale survey. J.Amer.Statist.Assoc. 59, 251-255.

Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. J.Roy.Statist.Soc. B, 15, 262-269.

Durbin, J. (1965). A method of sample selection with unequal probabilities without replacement. (abstract). Ann.Math.Statist. 36, p. 1327.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities without replacement. Ann.Math.Statist. 33, 350-374.

Kish, L. (1965). Survey Sampling. New York: John Wiley and Sons .

Mahalanobis, P.C. (1946). Recent experiements in statistical sampling in the Indian Statistical Institute. J.Roy.Statist.Soc. 109, 325-378.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. J.Roy.Statist.Soc. B, 25, 482-491.

Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. J.Roy.Statist.Soc. B, 15, 235-261.

Stevens, W.L. (1958). Sampling without replacement with probability proportional to size. J.Roy.Statist.Soc.B, 20, 393-397.

Table 1. Variance for Method 1 compared with sampling with replacement

Each entry should be multiplied by 10^{-5} (except for last column)

Quantity estimated	First-state component of variance			Second-stage component of variance (2)	Total variance (1)+(2)	True value of quantity estimated
	Without replacement no grouping	Without replacement with grouping (1)	With replacement			
Proportion $\frac{\text{Labour vote}}{\text{Electorate}}$ (ℓ)	4.86	4.87	5.33	7.12	11.99	0.346
Proportion $\frac{\text{Conservative vote}}{\text{Electorate}}$ (c)	3.69	3.70	4.03	7.15	10.85	0.332
Difference ℓ - c	12.56	12.59	13.75	21.50	34.09	0.014
Proportion $\frac{\text{Labour vote}}{\text{Total vote}}$ (ℓ')	8.76	8.76	9.59	9.93	18.69	0.448
Proportion $\frac{\text{Conservative vote}}{\text{Total vote}}$ (c')	5.28	5.29	5.77	10.24	15.53	0.430
Difference ℓ' - c'	21.14	21.18	23.15	36.07	57.25	0.018
$\ell - \ell_o$	0.94	0.94	1.04	7.12	8.06	
ℓ - c - (ℓ_o - c_o)	1.41	1.41	1.54	21.50	22.91	

$\ell_o(c_o)$ is the proportion of the 1964 electorate that would have voted Labour (Conservative) if the constituency proportions voting Labour (Conservative) had been the same as they were in 1959 .