A CONTROLLED SELECTION WHICH PERMITS UNBIASED ESTIMATES OF SAMPLING VARIANCES

Roe Goodman, Bureau of the Census

Controlled selection has been utilized in the sampling for a number of statistical surveys during the dozen or so years since this method of sampling was introduced. The original application was in the selection of a nation-wide sample of primary sampling units by the Survey Research Center of the University of Michigan. (1) That sample of PSU's, or an up-to-date version of it, was used extensively by the Survey Research Center over a period of perhaps ten years. Other uses of controlled selection have been made by the Census Bureau in the Current Population Survey (5), by the Bureau of Labor Statistics in the city sample for the Consumer Price Index (4), and by the University of Michigan's Bureau of Hospital Administration in the sampling of hospitals and hospital patients (2). A current application of the method is that being made by the Bureau of the Census in the selection of PSU's for agricultural surveys to be conducted as a part of the program of the 1964 Census of Agriculture.

Present and past uses of the method of controlled selection have been undertaken despite the fact that there has generally been no information in the particular instance regarding the gains (or losses) which may have resulted from the use of this sampling procedure as compared to some alternative sampling method. Evaluation of the gains (if any) which are achieved with the use of controlled selection has generally been difficult. Nevertheless the method seems to have been used because it was believed that the survey results would be at least as reliable as those which could have been obtained under an alternative sampling scheme.

As far as estimates of sampling variances are concerned, with controlled selection recourse is ordinarily made to the same types of approximations as are utilized for other sampling designs for which the sample data alone do not yield unbiased estimates of variances. One of the methods which is widely used is the method of collapsed strata. The method of collapsed strata was first devised for the case of stratified random sampling when the stratification was carried to the point that only one unit had been selected within each stratum. Similarly, with controlled selection sampling variances may be computed considering the sample as though consisting of sets of two or more units selected at random within collapsed strata, the estimated variances then being computed based on the variability among sample units within a set, with the use of suitable weighting factors. This method seems generally to serve well enough for purposes of approximating standard errors of survey estimates both for the ultimate case of stratified sampling (one unit per stratum) and for the case of controlled selection as well.

At some point however it becomes necessary to go farther than merely to approximate sampling variances and sampling standard errors and to determine rather definitely the relationship between the sampling variances found under controlled selection and those which obtain if some alternative sampling procedure is used. A logical standard for comparison in such instances is that of the sampling variance under stratified random sampling. The need for these comparisons of the variances is especially real due to the fact that variance estimates under the method of collapsed strata are on the average over-estimates of the true variances. Moreover, the theory shows that the greater the gain which has been achieved by the last stage of stratification or by the refinements of controlled selection the greater will be the degree of over-estimation of the variances with the method of collapsed strata. From the estimates of sampling variances and sampling standard errors obtained in practice then we are left with no clue whatsoever concerning the gains which may be achieved by the use of the controlled selection.

It has frequently been recognized that complete data for a population, such as Census data, permit the drawing of repeated samples and the preparation of estimates for each sample and that estimates of sampling variances can then be computed directly simply by computing the variance among the different sample estimates. This approach has frequently been used experimentally and, in fact, use of this method accounts for what little is known about the gains from the use of controlled selection to date. Considerable further use may yet be made of precisely this approach to the problem. However, it is undeniable that during the past decade very little work has been done along this line - in part no doubt for reasons to be explained below.

The present study is an attempt to progress toward the preparation of more useful estimates of sampling variances with controlled selection than have been available in the past. The results to be presented here and those being derived in our present studies (but not so far available) do depend upon information relating to a number of PSU's in addition to those in the original sample. However, the approach is more one of estimating variances from sample data and does not require the use of repeated samples. The present approach has considerable promise in that often times the drawing of repeated samples is not practicable. In the first place the sampling process may be extremely intricate involving several steps and the use of controls at each step. For this reason the drawing of additional samples becomes very laborious. In addition, more and more there is the desire to do the kind of thing Dr. Kish has just been talking about, namely, to build a new sample upon an old sample, retaining as many unite of the previous

sample as possible. As discussed by Dr. Kish the process involves changing the composition of strata and even re-defining some of the primary sampling units in the population. Moreover, controlled selection may have been used in the original sampling and it is now to be used again in the revision of the sample in order to bring it up to date. The procedure about to be described therefore seems a logical way to go about studying the gains achieved from the use of controlled selection.

For the benefit of those who may not be entirely clear on the exact distinction between some of these sampling procedures let us begin with a simple illustration of the use of controlled selection. See Table 1. In this example it is assumed that it is desired to select a sample of Standard Metropolitan Statistical Areas for the United States, or some major region, for use in multi-purpose sample surveys of households. In the illustration there are given three strata consisting of PSU's in Virginia, Maryland, Delaware, New Jersey and the eastern part of Pennsylvania but excluding the major cities of Washington, Baltimore, Philadelphia and New York City (the part in New Jersey). It is assumed that these large cities would be selected with certainty and hence that there would be no sampling of them at the first stage. It is assumed that there are other strata containing the SMSA's in other parts of the Region covered by the surveys, and that these other strata would be sampled also.

In the illustration the sampling probabilities were computed so as to be proportionate to the total population of each SMSA. In order that the sum of the probabilities for each stratum should be exactly 1.000 the chances of selection for occasional PSU's are divided a part being placed in one stratum and the remainder being placed in another stratum. The SMSA Newport News-Hampton is split in this way; it has a total probability of 0.150 of which 0.011 is placed in the first stratum and 0.139 in the second. The main consideration in the original stratification was the size of the central cities within the SMSA. In the columns to the right can be seen the possible samples which may be selected under one controlled selection scheme. Each possible sample consists of three places, reading across. The probability given beside the city in the last column indicates the probability of selection of the particular sample. Thus a sample consisting of Norfolk, Reading and Wilkes-Barre has a probability of 0.160 and finally one of Newport News, Scranton and Lancaster has a probability of selection of 0.011. The samples containing Norfolk have probabilities adding to 0.387, those containing Richmond have probabilities adding to 0.273 - and so on - and the sum of the probabilities for all possible samples is of course 1.000. From examination of the PSU's it can be seen that SMSA's located in Virginia have probabilities adding to 0.990 (0.387 plus 0.273 plus 0.150 plus 0.106 plus 0.074). Therefore every sample contains one and only one PSU in Virginia with the exception of the one Allentown, Wilmington and Lancaster which has a probability of 0.010. Looking at the coastal cities and seaports it can be seen that the sum of the probabilities for Norfolk, Newport News, Wilmington and Atlantic City is 0.890. Again no two of these cities appear in the same sample except for the sample consisting of Norfolk, Scranton and Atlantic City with a probability of 0.11. Samples having total probabilities of 0.121 then contain none of these port cities. On the whole a good control of the selection of port cities has been achieved. The fact that one possible sample contains both Norfolk and Atlantic City illustrates another common characteristic of controlled selection, namely, that with this method the goals sought are approached but usually not fully achieved. (Often if one goal is fully achieved another goal has to be sacrified somewhat).

Now for comparison with stratified random sampling it may be noted that, if the selections were to be made independently within the three strata, Norfolk and Wilmington would have a probability of (0.387) (0.245) or 0.095 of appearing in the same sample whereas with the controlled selection the probability of this joint occurrence is zero. In the case of Norfolk and Atlantic City the probability of their appearing in the same sample with stratified random sampling is (0.387) (0.108) or 0.042 compared with the 0.011 mentioned above. Numerous other comparisons can easily be made in the example at hand which would show clearly the effects of using controlled selection as an alternative to the well-known stratified random sampling.

Let us turn now to the method of estimating sampling variances which has been devised at the Bureau of the Census for purposes of the present analysis. See Table 2. In a generalized solution for sampling with varying probabilities, given by Horvitz and Thompson (3) ten years ago it was shown that to obtain unbiased estimates of sampling variances from sample data it is necessary that each pair of units in the population should have had a chance of appearing in the same sample. The solution in the present instance then is to supplement a sample chosen by the method of controlled selection by adding one or more units in such a way that every pair in the population has a chance of being in the amended sample. The idea of supplementing a sample, say a systematic sample, in such a way as to fulfill this condition is not new although it is doubtful if results based on this approach

have been published. The exact method used now is first to select a simple random sample of ℓ strata out of the L strata and then within each stratum so selected to select one additional PSU. Prior to the selection of the additional PSU the one originally selected is "replaced", thus permitting the selection of the same one a second time. Under these conditions the formulas as given in the second table apply and it is now theoretically at least possible to obtain unbiased estimates of the sampling variances using data for the sample as amended.

A few remarks concerning the formulas and their purposes may be useful at this point. First, please note that in the estimated sample total, $\hat{X}$, only the data for PSU's in the original sample are used. $\hat{X}$ is defined in this way because the purpose is to estimate the variance for precisely the sample as originally selected. Data for the additional PSU's are to be used only for purposes of estimating the variance but not for estimating the population total, X. The reasons for this decision stem from the fact that the estimated variance is itself subject to a great amount of variability and the kind of analysis being made must therefore be considered as a laboratory project rather than a procedure which would be used in the conducting of actual sample surveys. Since the formulas are to be applied in cases in which Census data are available for every PSU in the population the possibility exists of supplementing the original sample with a large number of additional units in order to obtain the desired degree of precision of the estimated variances.

From the variance formula itself the importance of the difference, $x'_{h(1)} - x'_{h(2)}$, may be noted. Basically, what is done is to obtain this difference between the estimated stratum total in each stratum for which a second PSU is selected and then to use this difference, both to obtain sums of squares (the first term) and sums of cross-products (the second term). In the second term the sign of the difference is important of course since it is multiplied by the estimated stratum total, based on the original sample for every <u>other</u> stratum, and then summed. A feature of the formula is that the first term reflects the variance of an estimated total for stratified random sampling and the second term the gain (or loss) from the use of controlled selection. In order to have a gain then the sum of the crossproducts, the second term of the formula, must be negative.

It may be noted that it is easily possible to set ℓ equal to L, that is, to select an additional unit within each stratum. In any use of the formula when ℓ is not extremely small it is to be expected that in some strata the second PSU selected will be the same as the original selection. In such cases both the squared term and the sum of crossproducts naturally become zero for the particular stratum. The proof that the estimated variance is unbiased is a simple one as shown in C. From the second step to the third step the formula becomes simplified due to the fact that the expected values of so many of the product terms are equal to zero. The selection of the additional PSU's independently from stratum to stratum accomplishes this result even though the PSU's in the original sample are not selected independently within the different strata.

The results obtained to date with the use of this formula have been found to be of little value due to the extreme variability of the estimated variances. It has been found that estimated variances computed from samples of no more than 36 PSU's and supplemented by an additional set of 36 PSU's, still do not yield meaningful results. Until a sample of adequate size is used many estimated variances turn out to be negative and it is clear that no satisfactory measures of gains or losses can be derived unless much larger samples are used.

At the Census Bureau experience has now been gained with the use of this formula as applied to data for past Censuses and a computer program has been tested and utilized in the experimental work done to date. We will now proceed to utilize the sample of some 400 PSU's, supplemented by as many as 400 PSU's, and perhaps even supplemented all over again by an additional 400, in an attempt to obtain reliable estimates of sampling variances with the particular controlled selection being used in the new agricultural sample. Meanwhile, there is room also for the possible development of other estimates of sampling variances, estimates which need not necessarily be unbiased provided the estimates are consistent and the bias is not unduly large. It appears then that satisfactory solutions to this problem will be possible, especially with the aid of the computers.

Table 1.--CONTROLLED SELECTION -- ILLUSTRATION
Population Consisting of SMSA's, m = 3
(Assumed to be part of larger population and sample)

Standard Metropolitan Statistical Area	P_{hj}	Approx. pop. of central cities (000's)	Possible samples under one controlled selection scheme and corresponding probabilities (read across)		
			Stratum I	Stratum II	Stratum III
Stratum I			.387 Norfolk-Portsmouth	.184 Reading	.024 Wilkes-Barre
Norfolk-Portsmouth, Va.	.387	420			.160 York
Richmond, Va.	.273	220		.127 Scranton	.116 Harrisburg
Allentown-Bethlehem-Easton, Pa.	.329	185			.011 Atlantic City
Newport News-Hampton,Va.	.011	200		.076 Trenton	.076 Wilkes-Barre
	1.000				
Stratum II			.273 Richmond	.152 Wilmington	.132 Wilkes-Barre
Newport News-Hampton, Va.	.139	200			.020 Lancaster
Wilmington, Del.	.245	95		.102 Trenton	.006 Lancaster
Trenton, N. J.	.178	115		.019 Scranton	.115 Harrisburg
Scranton, Pa.	.157	110			
Reading, Pa.	.184	100			
Roanoke, Va.	.097	95			
	1.000		.329 Allentown-Bethlehem-Easton	.139 Newport News-Hampton	.139 Lancaster
Stratum III					
Roanoke, Va.	.009	95		.097 Roanoke	.097 Atlantic City
Lynchburg, Va.	.074	55			
Atlantic City, N. J.	.108	60		.093 Wilmington	.074 Lynchburg
Harrisburg, Pa.	.231	80			.010 Lancaster
Lancaster, Pa.	.186	60			.009 Roanoke
Wilkes-Barre, Pa.	.232	65			
York, Pa.	.160	55			
	1.000		.011 Newport News-Hampton	.011 Scranton	.011 Lancaster

Table 2.--CONTROLLED SELECTION EXTENDED

Estimators and Expected Value of Estimated Variance

A. Estimated total - based on original sample only.

$$\hat{X} = \sum_{h=1}^{L} \sum_{j=1}^{1} \frac{x_{hj(1)}}{P_{hj(1)}}$$

Subscript (1) indicates units originally selected.
L = number of strata.
P_{hj} = probability of selection for j-th unit in h-th stratum.

B. Estimated variance.

$$s^2_{\hat{X}} = \frac{L}{2\ell} \sum_{h=1}^{\ell} (x'_{h(1)} - x'_{h(2)})^2 + \frac{L}{\ell} \sum_{h=1}^{\ell} \sum_{\substack{g=1 \\ g\neq h}}^{L-1} (x'_{h(1)} - x'_{h(2)})(x'_{g(1)})$$

where

$$x'_{h(1)} = \frac{x_{hj(1)}}{P_{hj(1)}}, \text{ etc.}$$

Subscript (2) indicates unit selected subsequently, with replacement.
ℓ = number of randomly chosen strata in which second units are selected independently.

C. Expected value of $s^2_{\hat{X}}$.

$$E(s^2_{\hat{X}}) = E\left[\frac{L}{2\ell} \sum_{h=1}^{\ell} (x'_{h(1)} - x'_{h(2)})^2 + \frac{L}{\ell} \sum_{h=1}^{\ell} \sum_{\substack{g=1 \\ g\neq h}}^{L-1} (x'_{h(1)} - x'_{h(2)})(x'_{g(1)})\right]$$

$$= \frac{L}{2\ell} E \sum_{h=1}^{\ell} \left[(x'_{h(1)} - X_h) - (x'_{h(2)} - X_h)\right]^2$$

$$+ \frac{L}{\ell} E \sum_{h=1}^{\ell} \sum_{\substack{g=1 \\ g\neq h}}^{L-1} \left[(x'_{h(1)} - X_h) - (x'_{h(2)} - X_h)\right]\left[(x'_{g(1)} - X_g) + X_g\right]$$

$$= \frac{1}{2} \sum_{h=1}^{L} E\left[(x'_{h(1)} - X_h)^2 + (x'_{h(2)} - X_h)^2\right]$$

$$+ \sum_{h=1}^{L} \sum_{\substack{g=1 \\ g\neq h}}^{L-1} E\left[(x'_{h(1)} - X_h)(x'_{g(1)} - X_g)\right]$$

$$= \sum_{h=1}^{L} E(x'_{h(1)} - X_h)^2 + \sum_{h=1}^{L} \sum_{\substack{g=1 \\ g\neq h}}^{L-1} E\left[(x_{h(1)} - X_h)(x'_{g(1)} - X_g)\right]$$

$$= E\left[\sum_{h=1}^{L} (x'_h - X_h)\right]^2 = E\left[\sum_{h=1}^{L} x'_{h(1)} - E \sum_{h=1}^{L} x'_{h(1)}\right]^2$$

$$= \sigma^2_{\hat{X}}, \text{ by definition.}$$

REFERENCES

1. Goodman, Roe and Kish, Leslie, "Controlled Selection - A Technique in Probability Sampling", Journal of the American Statistical Association, 45: 350-372 (1950).

2. Hess, I., Riedel, D. C., and Fitzpatrick, T. B., "Probability Sampling of Hospitals and Patients". Ann Arbor: The University of Michigan, Bureau of Hospital Administration, Research Series No. 1, 1961.

3. Horvitz, D. G. and Thompson, D. J., "A Generalization of Sampling without Replacement from a Finite Universe", Journal of the American Statistical Association, 47: 663-685 (1952).

4. Wilkerson, Marvin, "The Revised City Sample for the Consumer Price Index", U.S. Department of Labor, Bureau of Labor Statistics, Monthly Labor Review, Oct., 1960.

5. U.S. Department of Commerce, Bureau of the Census, "The Current Population Survey A Report on Methodology" Technical Paper No. 7.