SURVEY RESEARCH CENTER
UNIVERSITY OF MICHIGAN

THE USE OF CONTROLS BEYOND SIMPLE STRATIFICATION IN THE PROBABILITY SELECTION OF A SAMPLE

by Roe Goodman and Leslie Kish

A paper presented at the 108th Annual Meeting of
the American Statistical Association, Cleveland, Ohio

December 29, 1948

THE USE OF CONTROLS BEYOND SIMPLE STRATIFICATION IN THE PROBABILITY SELECTION OF A SAMPLE

by Roe Goodman and Leslie Kish

SUMMARY

In the selection of a sample of n units from a population of N sampling units, various procedures which are in accord with the principles of probability sampling are available. As compared to unrestricted random sampling, the other procedures may be said to introduce controls in the sense that the probabilities of selection of a part of the possible combinations of n out of N are reduced (perhaps to zero), while those of other combinations are increased. The aim of these controls is, of course, the reduction in the variability of sample estimates; this is achieved, it is hoped, by increasing the probability of selection for combinations which will yield the more accurate estimates.

Heretofore use of controlled selection has been confined chiefly to the elimination of possibility of selection of many combinations through the use of stratification. However, introduction of additional controls will result in reduced probabilities of selection of other combinations. Past attempts to introduce further controls by means of "deep stratification" have frequently led to the use of biased estimates. Systematic sampling, though a highly controlled procedure, has not been well utilized in that the possible re-ordering of the sampling units prior to selection seems to have escaped attention. The use of additional controls and the consequent change in probabilities of selection of different combinations of sampling units does not necessarily lead to sample estimates of increased accuracy. But by reference to the theory of systematic sampling, it may be seen that substantial gains in precision may be obtained under certain conditions.

Various procedures of controlled selection are discussed and illustrated by the selection of primary units for a sampling of the North Central States. The sampling variability of a number of items obtained by this method is compared with that of a stratified random selection.

THE USE OF CONTROLS BEYOND SIMPLE STRATIFICATION

IN THE PROBABILITY SELECTION OF A SAMPLE

At the heart of the problem of sampling is the following: Given a population consisting of N rigorously defined sampling units, how is one to select a sample of n of these units?

The question of how to select a sample is one that may arise several times in designing a sample. In the sampling of human populations, with which this paper is primarily concerned, there is usually first the selection of primary sampling units, then the selection of sub-units, and so forth. The successive stages of sample selection in surveys of this type may properly be regarded as separate processes. That is, in each instance one is faced anew with the problem of how to select the sample.

Method of Selection One Aspect of Sample Design

It is of course to be recognized at the outset that a decision on how to select the sample is but one of a number of decisions which should preferably be made jointly. Assuming that the design involves cluster sampling, decisions have to be made regarding the size of clusters and their precise definition. There is also the question of whether the clusters are to be selected with equal or varying probabilities,[1] and if with varying probabilities, of determining what the probabilities assigned the different sampling units are to be. Finally, there are the necessary decisions regarding the choice of unit for sub-sampling, for sub-sub-sampling if necessary, etc. Throughout the analyses of these questions, the minimizing of costs in relation to the determined standards of accuracy must ever be the primary concern.

Despite the complexity of decisions finally determining the design of a sample, the question of alternative methods of selection is a fundamental one, and one that may profitably be studied alone. While in this paper the emphasis is upon the selection of primary sampling units, it will perhaps become apparent that the techniques discussed have wide, general application.

Stratification Viewed as a Method of Control

A selection procedure which is widely used today is one that is called "stratified random."[2] It is to be noted that stratification introduces restrictions or controls in the process of selection. With stratified random sampling the selection of units for a sample os size n is in a sense partially controlled in that the probability of selection of the C(N,n) possible combinations of n units will vary, depending upon the strata with which the n units are associated. For example, if only two units are to be selected from each stratum, all combinations of n units in which exactly two units are associated with each stratum have a positive probability of

[1] This paper is confined to the type of sampling in which every element in the population or universe sampled has a known probability of selection, that is, probability sampling. The notion of making a selection among units with varying assigned probabilities was introduced by Morris Hansen about five years ago. This technique has been found increasingly useful.

[2] Although the word random is generally taken to connote equal probability, its use in this paper implies merely specific known probabilities, not necessarily equal.

selection, whereas all other combinations have a zero probability of selection. Thus the number of possible sample combinations is reduced. In unrestricted random sampling, on the other hand, the probability of selection of any combination of n units is a function solely of the probabilities of selection assigned each of the n units; moreover, if all units are assigned equal probabilities, the probability of selection is the same for all the possible combinations of n out of N units. See lines 1 and 2 in table I A.

Use of Controls - Definition

Before going further it seems necessary to state what is meant by the expression "use of controls in the probability selection of a sample." This expression is defined to mean any process of selection in which, while maintaining the assigned probabilities of each unit, the probabilities of selection of some or all of the possible combinations of n out of N units differ from those which would maintain under unrestricted random sampling. In short, any selection by probability methods except unrestricted random sampling is included in this definition.

As may be readily observed from the above definition, the use of controls in the selection of a sample does not necessarily involve a procedure of stratification. For example, if the N units in the universe is an even multiple k of the n units desired for a sample (assuming all units are to have equal probabilities of selection) the N units can be divided arbitrarily into k groups of n units each and one of the k groups then selected at random for the sample. (See line 4 in table I A). This is one possible method of controlled selection and yet it involves no stratification. Notwithstanding possibilities of this kind, stratification seems in practice to be a convenient step in the process of selecting most samples. And since it is itself a method of controlled selection, the emphasis in this paper, as indicated by the title, is on the use of controls beyond simple stratification.

An Extension of Purposive Selection

Conceptually, the use of controls in selecting a sample may be viewed as an extension of the technique known as purposive selection. One of the modes of extension may be the use of more judgment than in the conventional purposive sampling. If, however, the estimates to be derived from the sample are to be unbiased, an additional step not ordinarily considered to be a part of purposive selection is required. In order that the sampling may be probability sampling, the sampler must carry through at least the equivalent of the procedure mentioned in the last example. That is, he must select not just one but many purposive samples, until every unit in the universe is included in one or more samples. The number of samples in which each unit appears must be exactly proportionate to its assigned probability of selection (See table I B).

After the complete set of purposive samples has been established, the random selection of one of them constitutes a probability sample. As will be seen, such a procedure is not in conflict with the use of stratification but, on the contrary, can be more readily accomplished after strata have been established.

The preceding describes what appears to be the ultimate in controlled selection. In the process of purposive selection one could, conceivably, use considerable judgment and also make numerous checks in regard to various known characteristics of the sampling units, finally establishing samples each of which was as nearly as possible in accord with the population as a whole with respect to each characteristic.

I. Controlled Selection Illustrated

A. Equal probabilities

N = 2000 n = 40
Each unit has .02 probability of selection.

Method of selection	Number of possible sample combination (approximate)	Probability of selection of each combination (approximate)
1. Unrestricted random	90 x 10^{82}	.011 x 10^{-82}
2. Stratified random, 20 strata of 100 units each	78 x 10^{72}	.013 x 10^{-72}
3. 20 x 20 Latin square, 2 units taken at random from each selected cell	24 x 10^{35}	.04 x 10^{-35}
4. Population grouped into 50 possible samples	50	.02

As the method of selection becomes more and more controlled the number of possible sample combinations is drastically reduced until, in method 4, only 50 combinations are possible. Each of these 50 combinations has a chance of selection in method 1, although an extremely small chance. As the last column shows, the probability of selection of each of the 50 combinations in method 4 is .02.

If for the items to be estimated, method 4 will yield estimates which have smaller variances than those derived from use of the other methods, then the controlled selection would be worthwhile, provided costs were not increased disproportionately.

I. Controlled SelectionIllustrated (continued)

B. Varying probabilities

Selection of 1 unit from each stratum, n = 3

Stratum	1	6 units
"	2	4 units
"	3	5 units

Units and their assigned probabilities

Stratum 1

Unit	Probability
A	.10
B	.15
C	.10
D	.40
E	.05
F	.20
	1.00

Stratum 2

Unit	Probability
K	.40
L	.20
M	.15
N	.25
	1.00

Stratum 3

Unit	Probability
R	.10
S	.25
T	.20
U	.30
V	.15
	1.00

Method of selection

Stratified random		Controlled*	
Combination	Probability of selection	Combination	Probability of selection
AKR	.004	AKR	.10
AKS	.010	BKS	.15
AKT	.008	CKS	.10
...	....	DKT	.05
...	....	DLT	.15
...	....	DLU	.05
...	....	DMU	.15
CKR	.004	ENU	.05
CKS	.001	FNU	.05
...	....	FNV	.15
...	....		1.00
...	....		
FNV	.008		
	1.000		

*Meaningful ordering of units, samples selected systematically

Note that with stratified random sampling the probability of selection of any combination consisting of one unit from each stratum is equal to the product of the probabilities of the three units. For example, AKR has a probability of (.10)(.40)(.10) = .004 of being the chosen combination. If all the combinations and their probabilities were entered, the sum of the probabilities would, of course, be one.

On the other hand it is possible, if so desired, to introduce a measure of control by giving certain combinations any probability not greater than that of the unit in the combination which has the smallest assigned probability. This is subject to the restriction that the sum of the probabilities for combinations containing a particular unit must equal the probability assigned to it. In the illustrative example, the combination AKR receives a probability of selection of .10 which is 25 times as great as it has in stratified random sampling. With the controlled selection illustrated, only 10 combinations have any chance of selection. At the same time the original probabilities assigned each unit are not violated.

II. Introduction of an Elementary Control to Avoid an Undesirable Combination

Assume theme are two strata of 100 units each, two units to be selected from each stratum as on page IA.

Stratum	Sampling unit numbers, units randomly ordered before numbering	
I	1,2,3...........49,50	51,52............99,100
II	1,2,3...........49,50	51,52............99,100

Assume that, because unit number 14 in stratum I and unit number 67 in stratum II are geographically so close together, it is deemed undesirable that both should be in the same sample. Their being selected together can be avoided by determining by a random draw whether the selections within each of the strata are to be confined to the first 50 units or to the last 50 units. Having made this initial determination, the units for the sample are then selected at random within the half strata.

This example shows the flexibility of methods of controlled selection in that a technique of this kind can be simply applied without violating principles of probability sampling.

In practice the sampling procedure just described may not be feasible. But controls may be introduced in a less inclusive manner. Suppose, for example, that two sampling units are to be selected from each stratum; then we may merely arrange the sampling units in two strata in the same meaningful order and then determine at random whether the selections are to come from the top half of the first stratum and the second half of the second stratum, or vice versa. Following this the selection of specific sampling units would be made at random within the selected half strata. A slight variation of this is shown in table II. Many other variations in the manner and extent of controls in selection are possible. Since the use of controls frequently may cost relatively little, the question then becomes one of whether or not this procedure will result in reductions in sampling variability. When elaborate procedures of controlled selection are introduced, reductions (if any) in the variance of the sample estimates to be derived have to be balanced against the costs of doing the work in order to determine whether this additional work is worthwhile.

Previous Investigations of Selection Techniques

Earlier research workers seem to have dealt with somewhat limited applications of controlled selection. There is, of course, a wide range of published material on stratification, including a fundamental contribution by Neyman[3] in 1934. Neyman's also analyzed the usual type of purposive selection which he found to be generally inferior to stratified random sampling. The purposive selection he considered, however, was to a large degree objective, whereas it may be noted that the methods discussed here permit the use of considerable judgment.

Previously Bowley[4] and Jensen[5] had reported their analyses of stratified random sampling and purposive selection but their findings had been far from conclusive. Jensen[6] in 1928, described the purposive selection of a sample from records of the 1923 Danish Agricultural Census and showed that it represented the population well in respect to distributions of several farm variables. Strand and Jessen[7] compared the use of purposive and stratified random selection of townships in Iowa counties and concluded that "purposive selection does not provide samples of greater accuracy than stratified random selection" for situations of the type investigated. However, none of these investigators attempted to combine purposive selection with probability sampling.

Frankel and Stock[8], in 1942, suggested the use of a Latin square design which, within the framework of the modes of stratification used, extended the controls of stratification to a second dimension.

[3] Neyman, J., "On the Two Different Aspects of the Representative Method," Journal of the Royal Statistical Society, Vol. 97, 1934, pp. 558-625.

[4] Bowley, A. L., "Measurement of the Precision Attained in Sampling," Bulletin de l'Institut International de Statistique, Tome XXII, 1926.

[5] Jensen, A., "The Representative Method in Practice," Bulletin de l'Institut International de Statistique, Tome XXII, 1926, pp. 381-439.

[6] Jensen, A., "Purposive Selection," Journal of the Royal Statistical Society, Vol. 91, 1928, pp. 541-547.

[7] Strand, Norman V. and Jessen, Raymond J., "Some Investigations of the Suitability of the Township as a Unit for Sampling Iowa Agriculture," Iowa State College Research Bulletin 315, 1943.

[8] Frankel, Lester R. and Stock, J. Stevens, "On the Sample Survey of Unemployment," Journal of the American Statistical Association, Vol. 37, 1942, pp. 77-80.

The following year Tepping, Hurwitz, and Deming[9] reported extensive analyses on techniques of this kind which they designated as "deep stratification." There is an important limitation of a Latin square design, namely, that the probability of selection of each of the 1 x 1 cells into which the population units are grouped is the same. Tepping, Hurwitz, and Deming accordingly considered estimates, derived as though the probabilities for the various cells were equal, whether or not the combined probabilities for the units in the different cells were equal in all cases. That is, they considered the use of biased, as well as unbiased, estimates and in some cases found the bias to be an important source of error. Yates[10], in 1946, reported a selection procedure called balancing in which additional random selections are substituted for units originally drawn until finally "the mean value of the balanced factor in the sample is equal to the mean of the factor in the whole population." With the procedures just described samplers attain a balance with respect to certain controls but accept the possibility of biased estimates. In contrast, the emphasis in this paper is upon securing unbiased estimates through random sampling procedures, at the same time securing a partial balance in respect to the control factors.

Despite certain similarities of controlled selection to deep stratification and balancing there appears to be a still greater resemblance between this procedure and a method which has been increasingly discussed in the literature in the last few years, namely systematic sampling. Madow and Madow[11] and Cochran[12] have shown the theoretical conditions under which systematic sampling can be expected to be useful and have discussed the application of the method to certain real situations. In the research on systematic sampling to date, the emphasis has been upon the selection of a sample systematically from a more or less "naturally ordered" sequence of sampling units. Thus far, no attention seems to have been given the question of how the sampling units might be arbitrarily ordered in such a way that a systematically selected sample would yield estimates of greatest accuracy.

Controlled Selection Viewed as Systematic Sampling

The similarity between the highly controlled selection discussed a few minutes ago and systematic selection is apparent when it is considered that if the N units in the population are numbered in the proper sequence the selection of a sample systematically will automatically yield one of the purposive samples. In other words, under this condition the two methods become the same.

Systematic sampling, in fact, appears to be a good common denominator for various methods of probability selection. Unrestricted random selection may be viewed as systematic selection in which the sampling units in the universe have been ordered

[9] Tepping, Benjamin J., Hurwitz, William N., and Deming, W. Edwards, "On the Efficiency of Deep Stratification in Block Sampling," Journal of the American Statistical Association, Vol. 38, 1943, pp. 93-100.

[10] Yates, F., "A Review on Recent Statistical Developments in Sampling and Sampling Surveys," Journal of the Royal Statistical Society, Vol. 109, 1946, pp. 12-43.

[11] Madow, William G., and Madow, Lillian, "On the Theory of Systematic Sampling, I," Annals of Mathematical Statistics, Vol. XV, 1944, pp. 1-24.

[12] Cochran, W. G., "Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations," Annals of Mathematical Statistics, Vol. XVII, 1946, pp. 164-177.

at random.[13] Assuming a uniform sampling rate within all strata, stratified random selection is systematic selection in which the strata are placed in some sequence and the sampling units within each strata are randomly ordered.[13] The selection with deep stratification, assuming equal probabilities for all cells, becomes a systematic one in which the cells within each stratum are ordered in a randomly chosen sequence to accord with the restriction of a Latin square and the sampling units within each cell are also randomly ordered.[13] Thus controlled selection may be viewed as systematic selection in which the ordering of the sampling units is to some degree not random.

Viewed in this way the vast amount of flexibility inherent in the methods of controlled selection is evident. Moreover, the theory of systematic selection can be expected to provide valuable clues regarding the conditions under which various procedures of controlled selection may be useful.

The Underlying Theory

As the Madows[14] have stated, systematic sampling may be viewed as cluster sampling, in which each possible sample is one of the clusters. If the expected intra-class correlation for a cluster of n units is negative[15] (less than $-\frac{1}{N-1}$ to be precise) the estimated mean has a smaller variance than that of an unrestricted random sample of n units. Moreover the variance decreases with decreasing values (increasing negatively) of the intra-class correlation. (See table III).

In stratified random sampling the intra-class correlation is never positive, and usually it is less than $(-\frac{1}{N-1})$. But the Madows and Cochran have pointed out that under certain conditions a systematic sample will have a smaller intra-class correlation than a stratified random sample (which is a systematic sample with the units ordered at random within strata); and that under these conditions the sample estimate of the mean will have a smaller variance in the case of the former than of the latter.

Now, unless the intra-class correlations corresponding to all of the possible arrangements of units within all of the strata are identical, the values corresponding to some of the arrangements must be less than the average of all of them, which is that of stratified random sampling. It follows that in practice there always

[13] In the selection of units with varying probabilities the analogy to systematic sampling holds only when exactly one unit is to be selected from each group. This is, however, the usual manner in which varying probabilities are utilized in practice. In such instances, one may proceed as follows: Express the probabilities assigned the various units in fractions having a common denominator, d. The sum of the numerators of these fractions is, of course, d. Then arrange the d chances of selection in a random order. It is well to recognize that with this procedure the various chances of selection relating to a single unit may be scattered throughout the sequence.

[14] Op. cit.

[15] As indicated in footnote 13, units with varying probabilities may be viewed as having varying numbers of chances of selection, the sum of which was given as d. Henceforth in this paper the total number of chances of selection will be considered as N. The problem of selection is then restated as that of selecting n units from a population in which the sum of the chances of selection for all units is N, in which case the number of units as such in the population becomes irrelevant. Thus, when we do this, the conclusions to be reached apply alike to selection with equal and varying probabilities subject to the restriction that when varying probabilities are used one and only one unit is to be selected from each stratum or sub-group.

III Variance of an Estimated Mean with Systematic Sampling

$$\sigma_{\overline{X}'}^{2} = \frac{\sigma^2}{n}\left\{1+\rho\ (n-1)\right\} \text{ where } \sigma^2 = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X})^2$$

X_{ij} = jth unit in the ith cluster (each possible systematic sample is one cluster)

N = number of units in the population

n = number of units in a cluster

k = N/n = number of clusters in the population

$$\rho = \frac{1}{kn(n-1)\sigma^2}\sum_{i=1}^{k}\sum_{j \neq h=1}^{n}(X_{ij}-\overline{X})(X_{ih}-\overline{X})$$

Special case: Units are randomly ordered within each of the n sets of k units prior to each selection of a systematic sample. (The jth set consists of X_{1j} X_{2j}, X_{3j} . . . . X_{kj}). This is equivalent to stratified, random sampling.

Then $\rho = -\frac{1}{n(n-1)\sigma^2}\sum_{j=1}^{n}(\overline{X}_j-\overline{X})^2$ where $\overline{X}_{j} = \frac{1}{k}\sum_{i=1}^{k}X_{ij}$

$\left[\frac{1}{n}\sum_{j=1}^{n}(\overline{X}_{j}-\overline{X})^2\right.$ is the usual between strata variance $\left.\right]$

Now ρ has a maximum value of o, when $\sum_{j=1}^{n}(\overline{X}_j-\overline{X})^2 = o$

and a minimum value of $-\frac{1}{n-1}$ when $\frac{1}{n}\sum_{j=1}^{n}(\overline{X}_j-\overline{X})^2 = \sigma^2$

Substituting the maximum and minimum values of ρ in the formula first given, $\sigma_{\overline{X}'}^{2} = \frac{\sigma^2}{n}$ and o, respectively. These values coincide with the well known limits for stratified, random sampling.

By use of the formula it can be easily demonstrated that the variance for systematic sampling is sometimes less and sometimes more than that with stratified random sampling.

Assume the following small population of values ordered as shown below:

Stratum I	Stratum II
8 11 14 7 10	10 6 5 11 8

For a systematic sample of size 2, ρ sys. = -.48*

For a stratified random sample, ρ str. = -.15*

On the other hand by re-ordering the units in the above population it is possible, for a systematic sample of size 2, to secure a value of ρ as large as +.24.

*Details of computations are:

$$\rho_{syst} = \frac{1}{(5)(2)(2-1)(6.6)}\left[(-1)(+1) + (+2)(-3) + (+5)(-4) + (-2)(+2) + (+1)(-1)\right]$$

$$= -\frac{32}{66} = -.48$$

$$\rho_{str.} = -\frac{1}{2(2-1)(6.6)}\left[(10-9)^2 + (8-9)^2\right] = -\frac{2}{13.2} = -.15$$

exist orderings of the units which have a smaller intra-class correlation (and therefore smaller variance) than that of stratified random sampling. Equally inescapable is the fact that ordering may increase the intra-class correlation above that of stratified random sampling or even cause it to become positive.

Many questions are still to be answered. There is the problem of how to obtain optimum ordering in respect to a variable for which information is sought. Then there is the need to resolve the probable conflict between optimum orderings for two or more variables to be investigated in the same survey. For example, the optimum ordering in respect to one variable may result in less accuracy than would stratified random sampling in respect to other variables. The question becomes one of finding an ordering of the units by means of which the intra-class correlations will be reduced on the average.

It is also pertinent to consider the possible reductions of the variances for the different sizes of sample. The Madows found that with a given ordering of the units, the effect of change in sample size was somewhat complex and that further investigation of this problem was necessary.

In summary, the theory shows that the use of additional controls may increase the accuracy of estimates for some items in a survey but also that the result may be a reduced accuracy for these or other items. It remains then for empirical studies and additional theoretical developments to determine what the results are likely to be in practice.

Results of Limited Tests

We have completed a very small number of tests by computing variances for samples selected by the controlled procedures. The results of these computations show that very substantial reductions in variances are sometimes secured by these methods. The reductions ranged from 5 to 79 percent, except for one item on a very limited size of sample--the variance was increased by 80%. This was the only time we found that the variance was increased, as compared with stratified random sampling.

Controlled Selection Illustrated

Before giving further details regarding the variances, it is desirable to describe the method of controlled selection which was used for this particular problem. As will be seen, the procedure is somewhat complex, although its concurrence with the rules of probability sampling is apparent. The description may also be helpful in showing the logical basis for the use of controlled selection. In giving the details of the method that was used there is no pretense of showing the best way of controlling the sample selection but merely one way that it may be done.

The objective was the selection of 21 primary sampling units to represent the North Central States. The Detroit, Cleveland, Chicago, and St. Louis metropolitan areas were taken as separate strata, each of which was then selected with certainty. There remained the selection of 17 primary sampling units to represent all areas in the region outside these metropolitan areas. The primary sampling unit consisted of a single county or a county joined with part(s) or all of an adjacent county (or counties). In forming the primary sampling units, there were 3 objectives as follows:

(1) To increase diversity within the unit with regard to concentration of population.

(2) To limit the distance from the central city, in which the interviewer would live, to the periphery to a range of 25 or 30 miles.

(3) Not to increase the size of the unit beyond a single county if doing so would decrease the proportion of the population residing in the central city to less than 40% of the total for the unit.

The sampling units were grouped into 17 strata of about 1.8 million each; primarily on the basis of the size of largest city, and secondarily for the more rural counties; by major type of farming areas. The strata were numbered from 1 to 17, the ordering being from those with largest cities to those with rural populations only. Within strata the probability of selection assigned each unit was proportional to its 1940 population.

It was thought possible, by the use of controls, to insure a greater geographic spread and balance than one could expect from simple stratified random sampling. Another aim was to achieve a better balance with regard to percent urbanization in the middle 7 strata, in which there remained considerable variation within the strata with respect to this variable. The geographic balance was intended to yield approximately proportionate representation of the various states in the sample and to make sure that the larger cities, as well as the less densely populated places, were well distributed geographically.

Selection in Successive Stages - Step One

In order to simplify the mechanics of the work it was found desirable to divide the procedure into two successive steps. The various schemes used to accomplish control in these two stages will be discussed with the aid of the tables.

The first step was designed to assure proper representation among four groups of states as shown in the heading of table IV. The table shows the sums of the probabilities of selection assigned the different sampling units within each of the four groups (A, B, C, and D) for each of the 17 strata, and for various groups of strata 1 to 4, 5 to 8, etc. It will be remembered that the ordering of the strata is meaningful, representing decreasing size of the central city within the sampling units. The data in the table may be interpreted in terms of expected number of selections, as well as probabilities of selection. For example, the total 1.1718 under B for strata 5 to 8 means that of the 4 units to be selected in these strata, on the average 1.1718 is to be selected in Ohio or Indiana. By the use of controlled selections it is possible to make sure that in any one sample either one or two of the units selected will be drawn from this group. By this method, when a draw is made the chances are .1718 that two units will be selected from these states and .8282 that a single unit will be selected. As will be seen, this is the major type of control achieved by the first step.

Table V is set up in such a way as to conform with Table IV. The number of cells in each box has been limited to 16 in order to simplify the work. The numbers in the cells represent the number of sampling units to be drawn from each group of states. The P's which are the probabilities of selection for each combination add to 1.0000. Thus the first combination (represented by the first box) has a probability of .1391 of selection. If this combination were selected, the drawings for the first four strata would be taken as follows: 2 from the B group, 1 from the C group,

IV. Assigned Probabilities of Sampling Units Classed in Four State Groups in each of 17 Strata.

Population: North Central States excluding Chicago, Detroit, Cleveland, St. Louis Metropolitan areas

Sample size: One unit to be drawn from each stratum, 17 units in all.

Strata of approximately equal total populations, assigned probabilities adding to one.

Strata ordered by decreasing size of cities. For example, stratum 1 consists of 3 densely populated units, stratum 6 consists of 13 units having medium sized towns, and stratum 17 consists of 195 entirely rural units.

Stratum No.	Mich. Wisc. A	Ind. Ohio B	Ill. Iowa Minn. C	Mo., N.B., S.D., Neb., Kan. D	Total
1			.4946	.5054	1.0000
2	.4146	.5854			1.0000
3		1.0000			1.0000
4	.2827	.3296	.2602	.1275	1.0000
Total 1-4	.6973	1.9150	.7548	.6329	4.0000
5	.1059	.1093	.5165	.2683	1.0000
6	.4937	.3213	.1850		1.0000
7	.1034	.7412	.1554		1.0000
8	.4361		.3478	.2161	1.0000
Total 5-8	1.1391	1.1718	1.2047	.4844	4.0000
Total 1-8	1.8364	3.0868	1.9595	1.1173	8.0000
9	.3040	.4550	.1601	.0809	1.0000
10		.2222	.4737	.3041	1.0000
11	.2915	.0759	.2981	.3345	1.0000
12	.0146	.2476	.4657	.2721	1.0000
13	.2666	.2149	.1922	.3263	1.0000
Total 9-13	.8767	1.2156	1.5898	1.3179	5.0000
14	.0261	.2719	.4297	.2723	1.0000
15	.2914	.1891	.2963	.2232	1.0000
16	.0363	.0663	.3409	.5565	1.0000
17	.2164	.0299	.1570	.5967	1.0000
Total 14-17	.5702	.5572	1.2239	1.6487	4.0000
Total 9-17	1.4469	1.7728	2.8137	2.9666	9.0000
Total 1-17	3.2833	4.8596	4.7732	4.0839	17.0000

The probabilities entered above were derived as follows: The 1940 population for each unit was expressed as a decimal fraction of the total 1940 population for the stratum. The fractions (probabilities) for all units in a single cell were then added to obtain the total for a cell in the table.

V. Thirteen Alternative Combinations which Jointly Satisfy the Requirements of Table on Page IV.

P = Probability of selection for the combinations; A,B,C,D denote state groups. Entries show the number of units to be selected in each cell.

Strata	No. of Strata	P= .1391	P= .0033	P= .0835	P= .0334	P= .0434	P= .0850
		A B C D	A B C D	A B C D	A B C D	A B C D	A B C D
1-4	4	0 2 1 1	0 2 1 1	0 2 1 1	0 2 1 1	0 2 1 1	1 1 1 1
5-8	4	2 1 1 0	1 2 1 0	1 2 1 0	1 1 1 1	1 1 1 1	1 2 1 0
9-13	5	1 1 2 1	1 1 2 1	1 1 1 2	1 1 2 1	1 2 1 1	0 2 2 1
14-17	4	0 1 1 2	1 0 1 2	1 0 2 1	1 1 1 1	1 0 1 2	1 0 1 2
Total	17	3 5 5 4	3 5 5 4	3 5 5 4	3 5 5 4	3 5 4 5	3 5 5 4

Strata	P=.2047	P= .0022	P= .0383	P= .0838	P= .1404	P= .0940	P= .0489
1-4	1 2 0 1	1 2 0 1	1 2 0 1	1 2 1 0	1 2 1 0	1 2 1 0	1 2 1 0
5-8	1 1 2 0	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
9-13	1 1 2 1	1 1 2 1	0 2 2 1	1 1 2 1	1 1 1 2	1 1 1 2	1 2 1 1
14-17	0 1 1 2	0 1 1 2	1 0 1 2	0 1 1 2	1 0 2 1	1 1 1 1	1 0 1 2
Total	3 5 5 4	3 5 4 5	3 5 4 5	3 5 5 4	4 4 5 4	4 5 4 4	4 5 4 4

If the first combination is selected (P= .1391) the sample for the first four strata will contain no units from group A, two units from group B, and one unit each from groups C and D; and so on for strata 5-8, 9-13, and 14-17.

The product of (1) the entry in a cell and (2) the probability for the combination, when summed over all combinations, equals the entry for the corresponding cell in the table on page IV. For example, for group A strata 1-4 we have 0 × .1391 + 0 × .0033 + ... + 1 × .0940 + 1 × .0489 = .6973. It may be noted that the entries in a given cell of all the combinations never differ from that in the corresponding cell of the previous table by more than a fraction. Likewise the sums of entries in pairs of adjacent cells in general agree equally well with the corresponding sum in the previous table.

Thirteen was the least number of combinations found to be necessary to satisfy the requirements of the seven lines of totals shown in table, page IV.

and 1 from the D group, and so on for the other groups of strata. There are various devices by which a table of this kind can be derived. The solution given in Table V is not unique. A convenient way of arriving at such a table is demonstrated in Table Va.

We begin by writing down a desirable combination such as the one in the first box in Table V. To do this, we place a zero on the first line under A, since from Table IV it is found that the probability for this cell is only .6973. In other words, the number in this cell must sometimes be zero. (We could have just as well written a first combination in which this entry was 1, it doesn't matter). When the first entry is zero the second should be 2, since the sum of the entries in Table IV (.6973 + 1.9150) is greater than 2, hence the sum of the first two entries should never be less than 2, (since a control was attempted in respect to A + B and C + D). Having made these entries, the other two entries on the first line should each be 1, since the total must be 4 (the number of selections from the first four strata) and neither entry should be greater than 1 since the probabilities for the cells (.7548 and .6329) are each less than 1. On the second line we place a 2 under A, since the entry for the cell in Table IV (1.1391) is greater than 1 and the total for the first 8 strata under A (1.8364) means that the sum of the first 2 entries under A should usually be 2. Having entered the 2 under A, the entries of 1,1,0 under B, C, D, respectively, automatically follows since the sum of the four entries must be 4 and neither of the entries under B or C should be less than 1, since the probabilities in Table IV (1.1718 and 1.2047) are both greater than 1. The entries on the third line are those in the corresponding cells in Table IV rounded to whole numbers. They add to the required total of 5. The total for the last 9 strata in Table IV (1.4469) indicates that the sum of the last two entries under A should more often be 1 than 2. A zero is accordingly entered on the last line under A. The entry under B should then be 1, since the sum of entries in Table IV (.5702 +.5572) is slightly greater than 1. The entry of 1 under C and 2 under D results in totals of 3 for the last two lines in each of these columns, which totals are in harmony with the corresponding totals of 2.8137 and 2.9666 in Table IV. It may be noted also that the totals for the entire columns in this combination (3,5,5,4) are in harmony with totals in Table IV. The probability of .1391 is the largest desirable probability for this combination, since the entry on the second line under A can be a 2 only this proportion of the time, if it is never to be less than 1, and the assigned probabilities are not to be violated.

Table Va is useful at this point in that it shows the permissive probabilities for each desired number of selections in each cell. The .1391 assigned the combination just discussed is then subtracted from each permissive probability corresponding to the cell numbers of selection appearing in this particular combination. As additional combinations are set down the probabilities are successively subtracted until the remainders are all zero. Thus we are guided in assigning probabilities to the different combinations in that the restrictions imposed by Table IV are made ironclad by means of Table Va. For example, it is obvious from the table that the maximum probability for the first combination is .1391, since the subtractions in the first column leave a zero on line A, strata 5-8. The second and third combinations were originally written as one, that is, the second combination was assigned the probability of .0868; however, it was later found convenient to split this combination and slightly alter the arrangement in the lower right corner yielding the third combination. The probability of .0868 was the maximum for a combination having a total of 4 for the first eight strata under B.

It is perhaps fairly easy to see how certain "improvements" may be made in the combinations and their probabilities as entered in Table V. For example, the combinations could probably have been written in such a way as to harmonize with totals

Va WORK SHEET SHOWING RESIDUAL PROBABILITIES FOR EACH DESIGNATED NUMBER OF SELECTIONS

Cell designation		Selections in each cell		Probabilities of combinations to be successively deducted												
Strata	Group	Number	Probability	.1391	.0033	.0835	.0334	.0434	.0850	.2047	.0022	.0383	.0838	.1404	.0940	.0489
1-4	A	0	.3027	.1636	.1603	.0768	.0434	.0000								
		1	.6973						.6123	.4076	.4054	.3671	.2833	.1429	.0489	.0000
	B	1	.0850						.0000							
		2	.9150	.7759	.7726	.6891	.6557	.6123		.4076	.4054	.3671	.2833	.1429	.0489	.0000
	C	0	.2452							.0405	.0383	.0000				
		1	.7548	.6157	.6124	.5289	.4955	.4521	.3671				.2833	.1429	.0489	.0000
	D	0	.3671										.2833	.1429	.0489	.0000
		1	.6329	.4938	.4905	.4070	.3736	.3302	.2452	.0405	.0383	.0000				
5-8	A	1	.8609		.8576	.7741	.7407	.6973	.6123	.4076	.4054	.3671	.2833	.1429	.0489	.0000
		2	.1391	.0000												
	B	1	.8282	.6891			.6557	.6123		.4076	.4054	.3671	.2833	.1429	.0489	.0000
		2	.1718		.1685	.0850			.0000							
	C	1	.7953	.6562	.6529	.5694	.5360	.4926	.4076		.4054	.3671	.2833	.1429	.0489	.0000
		2	.2047							.0000						
	D	0	.5156	.3765	.3732	.2897			.2047	.0000						
		1	.4844				.4510	.4076			.4054	.3671	.2833	.1429	.0489	.0000
1-8	A	1	.1636		.1603	.0768	.0434	.0000								
		2	.8364	.6973					.6123	.4076	.4054	.3671	.2833	.1429	.0489	.0000
	B	3	.9132	.7741			.7407	.6973	.6123	.4076	.4054	.3671	.2833	.1429	.0489	.0000
		4	.0868		.0835	.0000										
	C	1	.0405								.0383	.0000				
		2	.9595	.8204	.8171	.7336	.7002	.6568	.5718	.3671			.2833	.1429	.0489	.0000
	D	1	.8827	.7436	.7403	.6568			.5718	.3671			.2833	.1429	.0489	.0000
		2	.1173				.0839	.0405			.0383	.0000				
9-13	A	0	.1233						.0383			.0000				
		1	.8767	.7376	.7343	.6508	.6174	.5740		.3693	.3671		.2833	.1429	.0489	.0000
	B	1	.7844	.6453	.6420	.5585	.5251			.3204	.3182		.2344	.0940	.0000	
		2	.2156					.1722	.0872			.0489				.0000
	C	1	.4102			.3267		.2833						.1429	.0489	.0000
		2	.5898	.4507	.4474		.4140		.3290	.1243	.1221	.0838	.0000			
	D	1	.6821	.5430	.5397		.5063	.4629	.3779	.1732	.1710	.1327	.0489			.0000
		2	.3179			.2344								.0940	.0000	
14-17	A	0	.4298	.2907						.0860	.0838		.0000			
		1	.5702		.5669	.4834	.4500	.4066	.3216			.2833		.1429	.0489	.0000
	B	0	.4428		.4395	.3560		.3126	.2276			.1893		.0489		.0000
		1	.5572	.4181			.3847			.1800	.1778		.0940		.0000	
	C	1	.7761	.6370	.6337		.6003	.5569	.4719	.2672	.2650	.2267	.1429		.0489	.0000
		2	.2239			.1404								.0000		
	D	1	.3513			.2678	.2344							.0940	.0000	
		2	.6487	.5096	.5063			.4629	.3779	.1732	.1710	.1327	.0489			.0000
9-17	A	1	.5531	.4140					.3290	.1243	.1221	.0838	.0000			
		2	.4469		.4436	.3601	.3267	.2833						.1429	.0489	.0000
	B	1	.2272		.2239	.1404								.0000		
		2	.7728	.6337			.6003	.5569	.4719	.2672	.2650	.2267	.1429		.0489	.0000
	C	2	.1863					.1429							.0489	.0000
		3	.8137	.6746	.6713	.5878	.5544		.4694	.2647	.2625	.2242	.1404	.0000		
	D	2	.0334				.0000									
		3	.9666	.8275	.8242	.7407		.6973	.6123	.4076	.4054	.3671	.2833	.1429	.0489	.0000
1-17	A	3	.7167	.5776	.5743	.4908	.4574	.4140	.3290	.1243	.1221	.0836	.0000			
		4	.2833											.1429	.0489	.0000
	B	4	.1404											.0000		
		5	.8596	.7205	.7172	.6337	.6003	.5569	.4719	.2672	.2650	.2267	.1429		.0489	.0000
	C	4	.2268					.1834			.1812	.1429			.0489	.0000
		5	.7732	.6341	.6308	.5473	.5139		.4269	.2242			.1404	.0000		
	D	4	.9161	.7770	.7737	.6902	.6568		.5718	.3671			.2833	.1429	.0489	.0000
		5	.0839					.0435			.0383	.0000				

for strata 5 to 13 in Table IV, as well as with the other totals. 'Also it is readily seen that there was freedom to use a much greater number of combinations than were used, if so desired. The procedure that was used, however, was based on the assumption that until it is established that techniques of this kind result in gains in sampling precision, the complications to be introduced should be somewhat limited in the interests of economy.

Table V is intermediate to Table VI, which provides a basis for selection of specific groups of sampling units. Using the data in Tables IV and V, a table such as Table VI can quite readily be written down. It will be noted that for each box in Table V there is a corresponding column (or perhaps a set of columns) in Table VI. In setting forth the data in Table VI is is possible to do more than to conform with the restrictions of Tables IV and V, namely, to see that the selections in any group of states are scattered more or less uniformly throughout the strata. For example, in the first column, in which there are to be two selections from group A in strata 5 to 8, the arrangement is such that these are not taken from numerically adjacent strata. The same holds for the two selections from group C in strata 9 to 13, and the two selections from group D in strata 14 to 17. The two selections from group B in strata 1 to 4, on the other hand, are adjacent, due to the fact that all of stratum 3 is in group B; hence, whenever the selection for stratum 2 is taken from group B, the two selections must come from the same group.

In writing down the selection patterns, as in Table VI, the manner in which Table IV is used may be illustrated in the following way: According to Table V, 1 selection in the first 4 strata is to be taken from units in group C in all but three of the combinations. Table IV shows that the letter C must be entered opposite stratum 1 in selection patterns having a combined probability of .4946 and opposite stratum 4 in selection patterns having a combined probability of .2602. Entries for each letter can therefore be made one at a time for the first four strata until all of the patterns are completed as far as the first four strata are concerned. A similar procedure is then used for the other sets of strata.

In view of the small number of units in each of the first six strata (there are only three units each in the first two) a still more exacting system of controls seemed feasible for these strata. Here the process was extended to the setting down of specific combinations of sampling units with a probability of selection determined for each such combination in accord with Tables I, II, and III. In this way, it was possible to emphasize desirable combinations of sampling units, that is, those yielding a good geographical scattering of the larger cities.

At this point a random selection was made. Number .8184 was drawn, and if Table VI were complete the selected pattern could be found, namely, the first one in which the cumulated probability equals or excedes .8184.

One may rest at this point, believing that he has obtained the substantial part of whatever gains controlled selection may yield. He would then make his selection of one unit for each stratum at random among the sampling units of each selected group. That is what was done for purposes of the sampling error calculations.

Second Stage of Controlled Selection

On the other hand one may introduce further restrictions within the restrictions of the groups selected from table VI. In this instance, it was decided to secure an approximate balance in respect to distribution among individual states and also in respect to per cent urbanization. At this stage this can only be done, however, by

VI. Selection Patterns Showing the State Groups from which the Various Sampling Units will be Selected.

P = probability of selection of the pattern
P cum = cumulated probabilities of patterns

P	.0200	.0061	.0038	.0195	.0124	.0657	.0116	.0002	.0031	...	.0489
P cum.	.0200	.0261	.0299	.0494	.0618	.1275	.1391	.0002	.0033	...	.0489
"							.1391		.1421	...	1.0000
Stratum No.											
1	C	C	C	C	C	C	D	D	D	...	C
2	B	B	B	B	B	B	B	B	B		A
3	B	B	B	B	B	B	B	B	B		B
4	D	D	D	D	D	D	C	C	C		B
5	A	A	A	A	A	C	A	B	B		D
6	C	C	C	C	C	A	C	C	C		B
7	B	B	B	B	B	B	B	B	B		C
8	A	A	A	A	A	A	A	A	A		A
9	B	C	C	C	C	C	C	C	D		B
10	C	B	B	B	B	B	B	D	B		D
11	D	C	C	C	C	C	C	A	A		A
12	C	D	D	D	D	D	D	C	C		C
13	A	A	A	A	A	A	A	B	B		B
14	D	D	C	C	D	D	D	D	C		D
15	C	C	D	D	B	B	B	C	A		C
16	D	D	D	B	C	C	C	D	D		D
17	B	B	B	D	D	D	D	A	C	...	A

The table means that the pattern in the first column has a probability of .0200 of being chosen, the second one a probability of .0061, and so on. As indicated by the dots, many of the patterns are omitted here for lack of space. If the first pattern were selected the sampling unit for the first stratum would be selected from those units in state group C, the sampling unit for the second stratum in state group B, the sampling unit for the third stratum in state group C, and so on.

The patterns have been written down in such a way as to conform with tables IV and V. The sum of the probabilities of selection of the patterns in which a certain letter appears on a given line (stratum) is equal to the probability shown for that letter and stratum on page IV. For example, for C in stratum 1: .0200 + .0061 + .0038 + ... + .0489 = .4946. The patterns are grouped in sets, each set corresponding to a combination on page V, as indicated by the cumulated probabilities on the second line of the above table. Within each set within each group of strata (1-4, 5-8, etc.) each letter will occur the number of times required for that combination on page V. For example, in each pattern of the first set (P= .1391) in strata 1 to 4, there are zero A's, two B's, one C and one D. Moreover in writing down the patterns an attempt was made to avoid writing the same letter in two adjacent strata in the same pattern. Thus in each pattern the different letters (state groups) are well distributed.

Each of these patterns of course would have a chance of selection if ordinary stratified sampling were used. However the sum of the probabilities of all of them under stratified sampling would be only .0000058 in contrast to the 1.0000 here.

applying a procedure of controlled selection to the groups selected in step one. All the units in other cells of Table IV are henceforth out of the picture altogether. Thus, the balancing of the ultimate sample in respect to these additional factors cannot be better on the average than that of the groups already selected.

A table, somewhat comparable with Table VI was prepared for the second stage of sample selection. It set forth the possible selection patterns for selection of sub-groups and again is not a unique system. Prior to the preparation of the table, sub-groups of units within the selected groups were formed on the basis of state, and within state, on the basis of whether the units were plus, average, or minus, in respect to per cent urbanization. For these purposes the probabilities of selection for each sub-group were scaled upward in such a way that the total probability for each group was 1.0000. Frequently, the units in a given sub-group belong to only one of the states in the group; hence, the selection from the group could be drawn only from this one state. It would have been possible, of course, to have obtained a greater control of the sample distribution by states by having done additional work prior to the first selection. Moreover, at this, the second stage, it would be possible to introduce a measure of gepgraphic control within the states if this were considered worth while.

After a random draw was taken, based on this last table, there were eleven sets of sub-groups, some of which contained but two or three sampling units. (The individual sampling units for the first six strata had been automatically determined by the procedure of step one.) Additional random draws were taken within sub-groups until, finally, the selection of seventeen (17) sampling units was completed.

Further Details on Comparisons of Variances

As indicated previously, the tests of variances were based on samples selected at the conclusion of step one, the procedure of which was described in some detail. For this purpose we draw 100 samples of 17 units each, using Table VI. Within the selected groups individual units were chosen with probabilities proportionate to their 1940 populations, as required by the original sample design. Estimated means were then prepared for each sample for a number of items using published 1940 Census data, and finally a variance was computed among the 100 estimates. For the simple stratified case, with which this variance was to be compared, variances were computed for the same items by means of the standard formula which utilizes data for the entire population. To date variances have been computed for three items based on the entire 17 strata and for four items based on the first six strata only. The variances for the first six strata for the controlled as well as the stratified sampling were computed for the entire population. For this purpose estimates were computed for each possible controlled selection and the properly weighted variances were then derived. The comparisons of the variances are given below:

Variance of Estimates Derived from Controlled Sampling
as Percents of the Variances for ordinary Stratified Sampling

1940 Census Data	First six strata only n = 6 %	All 17 strata n = 17 %
Average monthly rent of urban and rural non-farm dwellings	41	81
Ratio of total dwelling units to total population	180	95
Percentage of dwelling units vacant		95
Percentage of population foreign born white	21	
Percentage of population over 65 years of age	25	

In the results as given, a percentage of less than 100 indicates a gain for the use of controlled selection, while one of more than 100 indicates a loss. While the percentage for the ratio of dwellings to population was 180, based on the first six strata only, the loss was converted into a small gain when the computations were based on the entire 17 strata (including the 6). It may be noted that if the percentages were expressed in inverse form, that is, the variance of the stratified random as a percentage of the controlled selection, percentages above 100 would indicate a loss by not having used the additional controls. For example, the percentage for average rent based on all 17 strata would be 123 and that for the percentage of foreign born white, based on the first 6 strata only, would be 484.

Concluding Remarks

The emphasis in this paper has been upon the possibility of using controlled procedures of selection without violating principles of probability sampling. A great deal remains to be done in improving the techniques and additional theoretical developments will be needed.

The procedure described illustrates a selection with uniform rates within strata of approximately equal size. The methods are sufficiently flexible, though, to permit their extension to situations in which different rates are used for different segments of the population.

A problem that will require attention is the estimation of variances based on sample data for cases of controlled selection. This is exactly the problem of estimating variances for systematic sampling and while it presents difficulties it should be possible to develop approximations which may be so formulated as to be on the "safe side."

If the use of controlled selection is found to result in reduced variances of estimates for large scale surveys, efficiency of sample surveys based on a relatively small number of sample areas will be increased. In this connection, the use of an even greater number of control variables may be desirable if one set of sample areas is to be utilized for a wide variety of surveys.