ON THE THEORY OF SAMPLING FROM FINITE POPULATIONS

By Morris H. Hansen and William N. Hurwitz

Bureau of the Census

I—HISTORICAL BASIS FOR MODERN SAMPLING THEORY

The theory for independent random sampling of elements from a population where the unit of sampling and the unit of analysis coincide was developed by Bernoulli more than 200 years ago. The theory that would measure the gains to be had from introducing stratification into sampling was indicated by Poisson a century later. Subsequently, Lexis systematized previous work and provided the theoretical basis for sampling clusters of elements.[1] The adaptation of the work of Bernoulli and Poisson to sampling from finite populations was summarized by Bowley in 1926 [1] approximately a century after the work of Poisson.

An impetus to sampling advancement, following some fundamental statistical contributions of Pearson, Fisher, and others, resulted from the work of Neyman when he published his paper in 1934 on the two different aspects of the representative method [8]. In that paper he introduced new criteria of the optimum use of resources in sampling, including the concept of optimum allocation of sampling units to different strata subject to the restriction that the sample have a fixed total number of sampling units.

If, no matter how a sample be drawn, the cost were dependent entirely on the number of elements included in the sample, there would be little need for theory beyond the classical theories of Bernoulli and Poisson covering the independent random sampling of elements within strata, supplemented by the extension of the theory to finite populations, and the extension to optimum allocation of sampling units. Very often, however, in statistical investigations it is extremely costly, if noτ impossible, to carry out a plan of independent random sampling of elements in a population. Such sampling, in practice, requires that a listing identifying all the elements of the population be available, and frequently this listing does not exist or is too expensive to get. Even if such a listing is available, the enumeration costs may be excessive if the sample is too widespread. Frequently also, there are other restrictions on the sample design, such as the requirement that enumerators work under the close supervision of a limited number of supervisors, and as a consequence the field operations must be confined to a limited number of administrative centers. Techniques such as cluster sampling [2, 3, 4, 5, 6, 7, 8, 10], subsampling, and double sampling [9], have been

[1] The sampling of clusters of elements refers to the sampling of units that contain more than one element. Examples of cluster sampling include the use of the city block or the county as the sampling unit when the purpose of the survey is to determine the properties of the population made up of individual persons or individual households. In these instances, the city block or county is referred to as the cluster of elements, and the individual person or household is referred to as the element.

developed with the aim of making most effective use of available resources, while keeping within existing administrative restrictions, and thus producing the maximum amount of information possible within these resources and restrictions. Neyman [8], Yates and Zacopanay [10], Cochran [2], Mahalanobis [7], and others have made important contributions in this regard.

We can illustrate a number of the developments indicated above in a simple but fairly general subsampling design. This design involves the sampling of clusters of elements from a stratified population and the subsampling of elements from each of the selected clusters, where the number of elements in each of the primary sampling units within a stratum is the same.

Suppose we have a population made up of L strata, with the i-th stratum containing M_i primary sampling units of N_i elements each. The individual element will be the subsampling unit. Let X_{ijk} be the value of some characteristic of the k-th element of the j-th primary sampling unit in the i-th stratum, and assume that the character to be estimated is

$$(1) \qquad \bar{X} = \sum_i^L \sum_j^{M_i} \sum_k^{N_i} X_{ijk} \Big/ \sum_i^L M_i N_i .$$

For example, if $\bar{X}$ is the average income per household in a given city, X_{ijk} might be the income of the k-th household in the j-th city block in the i-th ward; where the household is the subsampling unit, the city block is the primary sampling unit, and the stratification has been by wards. Suppose, further, that we sample m_i primary units from the i-th stratum, and subsample n_i elements from each of the primary units sampled from that stratum.

The "best linear unbiased estimate" [8] of $\bar{X}$ from the sample will be

$$(2) \qquad \bar{X}' = \sum_i^L \frac{M_i N_i}{m_i n_i} \sum_j^{m_i} \sum_k^{n_i} X_{ijk} \Big/ \sum_i^L M_i N_i ,$$

and the variance of $\bar{X}'$ is

$$(3) \quad \sigma^2_{\bar{X}'} = \sum_i^L M_i^2 N_i^2 \left\{ \frac{M_i - m_i}{M_i - 1} \frac{\sum_j^{M_i} (\bar{X}_{ij} - \bar{X}_i)^2}{M_i m_i} + \frac{N_i n_i}{N_i - 1} \frac{\sum_j^{M_i} \sum_k^{N_i} (X_{ijk} - \bar{X}_{ij})^2}{M_i N_i m_i n_i} \right\} \Big/ \left(\sum_i^L M_i N_i \right)^2$$

where $\bar{X}_{ij} = \sum_k^{N_i} X_{ijk}/N_i$ and $\bar{X}_i = \sum_j^{M_i} \sum_k^{N_i} X_{ijk}/M_i N_i$.

These formulas have no practical utility in designing samples unless there are, in addition, some considerations of differential costs. Cost relationships sometimes may be stated explicitly as a function of the m_i and the n_i, or, what is frequently the case, they may be approximated sufficiently through intuition and speculation to guide one to a reasonable decision among the various alternatives implied by the design.

If we know the cost function we proceed to determine the values of the m_i and the n_i that make $\sigma^2_{\bar{x}'}$ a minimum for a fixed total expenditure, and also subject to any other restrictions that may be imposed. This theory provides a basis for determining the optimum allocation of the sampling ratios to the various strata, and to primary and secondary sampling units within each stratum.

Such developments, however, must be regarded as only the first step in sample design. We cannot go forward if we only know that the optimum sample design is some particular mathematical function of the population parameters and the cost factors; we need also to know something about the relative magnitudes of certain parameters in the particular populations under consideration, as well as something about the costs associated with the various sampling and estimating operations.

Thus, considerable work in recent years has been done on the study of the relative magnitudes of variances and covariances between and within various types of sampling units and on the study of costs and types of cost functions that operate. Work is being done in this field by the Department of Agriculture in connection with sampling for agricultural items, and is being done also in the Bureau of the Census, and in other places.

II—THE DIRECTION OF MORE RECENT DEVELOPMENTS

The sampling procedure indicated above involves as a first step the definition of the *system of sampling*, such as whether the sampling method will involve cluster sampling, double sampling, or subsampling, and along with this the definition of the stratification and the sampling units. The second step is that of determining the *method of estimation*, together with *the allocation of the sampling units.*

The first step, that of defining the sampling system is taken with a view to administrative feasibility and sampling efficiency, but no simple procedure exists which leads one uniquely to the selection of a system except perhaps by the impractical method of listing and examining all possible alternatives and accepting one on some criterion of best. However, given the definition of a population character to be estimated, and a sampling system, a simple procedure is available that will provide a unique solution to the second step providing we accept some criterion as to what "best" means, such as the best linear unbiased estimate, subject to any cost or administrative restrictions that may be imposed. Such criteria lead us to both our estimating procedure and our allocation of sampling within the sampling system defined.

While no theory with practical applicability has been developed which indicates a "best" system of sampling, and at the same time indicates the "best" estimating procedure and sampling allocation, some progress in the choice of improved sampling systems and estimating procedures has been made. The developments in the following two directions appear to us to be particularly pertinent.

1. Modifications in some of the fairly generally accepted criteria of good

sample estimates have led to more reliable sample results for some types of sampling systems (some of these are mentioned in Sec. III);

2. Some principles are emerging, that have led to improved determination of the sampling units, the strata, and other aspects of the sampling system (some efforts at formulating such principles are reported in Secs. IV, V, and VI).

We shall summarize, principally, some of the recent work in the Census—and in so doing shall mention some work of others that is closely related. Most of the work that we shall summarize relates to problems where the sampling units are clusters of elements and vary in size.

III—MODIFICATIONS IN THE CRITERIA FOR GOOD ESTIMATES

The estimate given in the general subsampling problem formulated in Sec. I satisfies the criterion of the "best linear unbiased estimate." Also, as far as our experience has indicated, this estimate is frequently the most efficient one for populations of the form described, that is, where the number of elements in each sampling unit within a stratum is the same. However, if the numbers of elements differ between sampling units, a biased but consistent estimate can frequently be found that has a substantially smaller mean square error[2] than the best linear unbiased estimate.

For example, consider the case where clusters of elements are the sampling units and we want to estimate $\bar{X} = \sum_i^M X_i / \sum_i^M N_i$, the average value per element of some specified characteristic. Here M is the number of sampling units in the population, X_i is the aggregate value of the specified character for all elements in the i-th cluster, and N_i is the number of elements in that cluster. The joint distribution of X_i and N_i is unknown, but $\sum_i^M N_i = N$ is known. Under these circumstances the "best linear unbiased estimate" of $\bar{X}$ from a sample of m clusters turns out to be $\frac{M}{m} \sum_i^m X_i/N$. However, a smaller mean square error is often obtained by the use of a ratio estimate from the sample such as $\sum_i^m X_i / \sum_i^m N_i$. This estimate is excluded by the "best linear unbiased" criterion because it is nonlinear and biased, although the bias is usually negligible and the estimate is consistent. Since the best linear unbiased estimate of $\bar{X}$ requires the knowledge of N, the sample ratio has a further advantage in that it can be used even when N is not known.

A recent paper by Cochran [3] gives a number of consistent though biased esti-

[2] In this paper the terms "mean square error" and "variance" are used interchangeably to refer to $E(X - \hat{X})^2$ when EX is equal to $\hat{X}$, the population character to be estimated. When EX is not equal to $\hat{X}$, however, $E(X - \hat{X})^2$ will be referred to only as the "mean square error." Since, under these latter circumstances, $E(X - \hat{X})^2 = E(X - EX)^2 + (EX - \hat{X})^2$, the mean square error is equal to the variance of X plus the contribution due to the bias.

mates of $\bar{X}$ that make use of the least square estimate of the linear regression of X_i on N_i. These estimates generally have a smaller mean square error than either the best unbiased linear estimate or the simple ratio estimate given above. However, they require knowledge of N, as does the best linear unbiased estimate, and in addition may require detailed tabulations and considerable clerical work as a part of the estimating process.

Both types of biased estimates mentioned above are consistent, and usually have a smaller mean square error than the best linear unbiased estimate for sampling systems in which the sampling units vary in size. Thus, improved sample estimates will be obtained by modifying the "best linear unbiased estimate" criterion to include estimates that are nonlinear, consistent, but have a smaller mean square error than the best linear unbiased estimate.

IV—IMPROVEMENTS IN THE SPECIFICATIONS OF SAMPLING SYSTEMS

A great deal can be done to improve sampling designs through improved specification of the sampling system even though one has only a limited knowledge of the manner in which the population is likely to be made up, and no specific information concerning the particular population parameters involved (see Sec. VI).

1. The sizes of sampling units. A number of recent investigations have indicated the desirability, with costs considered, of keeping the size of cluster very small when clusters of elements are used as the sampling unit in field surveys [2, 5, 6, 7, 8]. It is important to point out, however, that this principle is not necessarily applicable to subsampling systems, and that the use of large clusters as the primary sampling units in a system involving subsampling may yield distinct gains over the use of smaller clusters without subsampling. Moreover, one of the often recurring problems in large-scale studies is the designing of sample surveys within stringent administrative restrictions on the number of different locations in which operations can be carried on. Under such restrictions a procedure commonly used is to choose a limited number of existing political units, such as counties, as the primary sampling units, and then to subsample units such as blocks, small rural areas, or households. Under the circumstances, if the numbers of primary subsampling units to be included in the sample are assumed to be held constant, the use of larger primary sampling units than the existing political units would have the effect of decreasing the sampling variance.

The advantage of using large primary units in subsampling is evident in the simple case when the original units, each having the same number of elements, are consolidated to form half as many enlarged primary units, each twice as large as the original units. The variance between the enlarged primary units will be $\sigma_{2b}^2 = \frac{1}{2}\sigma_{1b}^2(1 + \rho)$, where σ_{1b}^2 is the variance between the original primary units, and ρ is the correlation between the units that are paired. The correlation coefficient will be close to zero (exactly equal to $-1/\{M - 1\}$, where M is the number of original primary units) if the pairing is done at random, and it follows that the variance between counties is then cut at least in half. Ordinarily, ρ will be greater than zero if the paired units are required to be contiguous. However, through choosing for consolidation those contiguous units that are as different as possible, ρ is made as small as possible, and in some instances this minimum value may even be negative. In any event, the smaller the value that ρ takes on, the greater the reduction of the sampling variance between primary units from the use of enlarged units. While the sampling variance within primary units is increased by such consolidations, the increase is slight, and the total sampling variance is almost invariably decreased (see Appendix, Section 1).

The restriction on extending the consolidation of primary units is introduced by the increased cost of subsampling within larger and larger areas. This increased cost is to be weighed against the decreased variance. If the cost restriction were not sufficiently severe, consolidation would proceed to the point of eliminating the use of primary sampling units altogether, and the subsampling units would be selected independently throughout the entire stratum.

2. Subsampling where the primary units are of unequal size. Use of probability proportionate to size in subsampling. A subsampling system frequently followed, whether or not the primary sampling units vary in size, involves the selection of one or more primary units from each stratum with the probability of selection the same for each primary unit in the stratum, and the subsampling of a fixed proportion of the subsampling units from the selected primary unit. When the primary units vary in size this subsampling system has some administrative disadvantages that arise because the number of subsampling units to be included in the sample will vary with the number of elements in the selected primary unit. (The term "size" of sampling unit as used in this paper refers to the number of elements in the sampling unit.)

The disadvantages in the above system have led in some instances to the specification of a second subsampling system in which, although the primary units were selected with equal probability, the subsampling has been of a constant *number* rather than of a constant proportion.

A third subsampling system that can be recommended over both the above systems is to make the probability of selection of a primary unit proportionate to its size and then to subsample a constant number of subsampling units.

We shall assume that for all three systems only one primary unit is selected from each stratum. Stratification to this degree leads to a smaller sampling variance than does less extensive stratification. For simplicity in making comparisons, we shall assume, furthermore, that the subsampling unit is the element of analysis and that the sample estimate used is of the form $\bar{X}' = \Sigma N_h \bar{X}'_h / \Sigma N_h$ where $\bar{X}'_h$ is the sample average, for the h-th stratum, of the character being estimated, and N_h is the size of that stratum. This estimate, which is frequently used, is biased for the first two systems but unbiased for the recommended system. However, an unbiased estimate, say the "best" linear unbiased estimate for the first two systems generally has a much larger mean square error than the biased estimates used in these comparisons and hence has not been considered in the comparisons which follow (see Sec. VII, footnote 7).

The first two subsampling systems mentioned are about equally efficient when the number of subsampling units drawn from each primary unit is reasonably large, but each will usually have a larger mean square error than will the recommended system. The difference between the mean square errors of either of the first two and that of the recommended design is given approximately by

$$\frac{1}{N^2} \sum_h Q_h \bar{N}_h \sigma_h^2 \left[\sum_j \rho_{h,} \bar{N}_h - \sum_j \rho_{hj} N_{hj} \right] \tag{4}$$

where, within the h-th stratum, N_{hj} is the number of elements in the j-th primary sampling unit, $\bar{N}_h$ is the average size of primary sampling unit, Q_h is the number of primary sampling units, ρ_{hj} is the intra-class correlation between elements within the j-th unit and σ_h^2 is the variance between individual elements within the stratum; L is the number of strata. (See Section 2 of the Appendix for the development of this difference.)

This difference, which is a multiple of the average covariance between the N_{hj} and ρ_{hj}, will be positive if N_{hj} and ρ_{hj} are negatively correlated, and this is exactly the situation that exists in most practical problems we have encountered in sampling for social and economic statistics (see Sec. VI).

The reduction in the mean square error arises because the recommended design provides a more nearly optimum allocation of sampling as between large and small sampling units than do the other two. It might be possible, of course, as another alternative, to stratify the primary units by size and then allocate sampling to the various strata on the basis of optimum sampling considerations. However, this would mean that some other and perhaps more important modes of stratification would be sacrificed, and moreover, the optimum allocation of sampling between the larger and smaller units could only be guessed at in most practical problems. Furthermore, it usually is not possible to stratify on size to the point that there is no variation in the sizes of units within a stratum.

The sample estimate from the recommended system is unbiased whereas the estimates from the other two are usually biased, and sometimes fairly seriously so. (For a proof of this statement see Appendix, Section 1, and see also Sec. VII for a numerical illustration.)

The use of probability proportionate to size serves to decrease only the sampling variation between primary units and has very little effect on the sampling variance within. Therefore, the recommended design shows its greatest advantage over the two alternatives when the contribution of the mean square error between primary units to the total mean square error is large.

Ordinarily, the actual sizes of the primary sampling units will not be known, but numbers may be known that are highly correlated with the sizes. For example, ordinarily we will not know the populations of blocks or of cities or counties at the time a sample is taken, but we may know their populations at the preceding census. Under these circumstances the primary units may be sampled with probabilities proportionate to the previously known (or their estimated) sizes, but if this is done the subsampling is to be modified in order to take account of the changes in the sizes between the two dates. If the actual sizes are known, the constant number taken from the selected primary unit in the h-th stratum is $n_h = t_h N_h$ where t_h is the sampling ratio assigned to the stratum, and N_h is the total number of elements in the stratum. The subsampling ratio within the selected primary unit, therefore, is $t_h N_h / N_{hj}$, where N_{hj} is the number of elements in the selected unit. On the other hand, if there is available only a measure of size P_{hj} highly correlated with the actual sizes of the units N_{hj} and, if the probability of selection of the primary unit has been proportionate to the P_{hj} the subsampling ratio in the selected primary unit will be equal to $t_h P_h / P_{hj}$, where P_h is the measure of size of the entire stratum, and P_{hj} is the measure of size of the selected primary unit. The variance of a sample estimate where measures of size are used is given subsequently in this paper (see Eq. (9)).

3. The use of area substratification within primary strata in a subsampling system. Another modification, which will be called area substratification within primary strata, may be particularly useful where a relatively small sample is required from a population covering a large area, and where operations must be confined to a limited number of centers.

Some preliminary remarks are necessary before area substratification can be explained. Area substratification requires (a) that the entire population to be sampled be divided into areas that will serve as primary sampling units; (b) that these units be further subdivided into a number of sub-areas; and (c) that certain summary statistical information be available for each of the sub-areas in advance of drawing the sample. The information that must be known for the sub-areas includes a reasonably good measure of their sizes (perhaps the total population, total dwelling units, or total farms) and other information which is indicative of the characteristics of the area, such as whether predominantly farm or nonfarm, predominantly white or colored, etc. The sub-areas, when grouped into homogeneous classes, will serve only to determine the substrata described subsequently, and will not ordinarily serve as the subsampling units, which may be defined independent of the sub-areas.

The definition of the primary sampling units and the classification of them into strata proceed as indicated earlier, with the primary units made as internally heterogeneous as possible within strata that are as homogeneous as possible. It will be assumed that only one primary unit is sampled from each stratum, and that the probability of selecting the j-th primary unit within the h-th stratum is proportionate to P_{hj}, where P_{hj} is the measure of size of the primary unit and is equal to the sum of the measures of size of the sub-areas that it contains. It will be assumed, also, that t_h, the over-all sampling ratio to be used within the h-th stratum, has been determined for all strata on the basis of considerations of optimum allocation.

The introduction of area substratification within primary strata may then be accomplished as follows:

(a) The sub-areas within each primary stratum are classified into substrata on the basis of their characteristics. (For example, they may be classified into predominantly farm and predominantly nonfarm sub-areas, and these further classified on the basis of the average size of farm or average rental value of the dwelling units. In such a case, the sub-areas within the primary stratum that are predominantly farm and that have average rental values lying within a specified interval constitute a substratum.)

(b) The sub-areas within the primary unit selected from each primary stratum are classified into the same substrata.

(c) Subsampling units are defined within each of the substrata within the selected primary units. The number of subsampling units defined within that part of the i-th substratum that is contained within the j-th primary unit is denoted by M_{hij}. (Various types of subsampling units may be defined, such as the individual person, farm, dwelling unit, or structure, a very small area, etc. The subsampling units need be defined only within the selected primary sampling units.)

(d) The number of subsampling units to be included in the sample from the i-th substratum within the selected (j-th) primary sampling unit is

$$m_{hij} = M_{hij}t_hP_{hi}/P_{hij}, \tag{5}$$

where P_{hij} is the measure of size of that part of the i-th substratum that lies within the j-th primary unit, and $P_{hi} = \sum_j P_{hij}$ is the sum of the measures of size of the sub-areas contained in the i-th substratum of the h-th primary stratum. This method of allocating the subsampling provides that the subsample drawn from the selected primary unit is representative, so far as possible, of the entire stratum, rather than of the particular primary unit that happens to be included in the sample from that stratum. To illustrate, suppose the numbers of persons in sub-areas from the 1940 census are used as the measures of their sizes, and that the sub-areas are classified into substrata on the basis of their characteristics in 1940 as indicated by the 1940 Decennial Census of Population. The allocation of the subsampling indicated above then provides that if the proportion of the total population residing in sub-areas that are predominantly farm is 30 percent, the sample will be drawn in such a manner that 30 percent of the 1940 population expected in the sample would be from the predominantly farm sub-areas, even though, in the selected primary sampling unit, perhaps only 15 percent of the 1940 population might reside in such areas.

(e) The population character to be estimated is

$$X = \sum_h^L \sum_i^{S_h} \sum_j^{Q_h} \sum_k^{M_{hij}} X_{hijk}, \tag{6}$$

where X_{hijk} is the aggregate value of a specified characteristic for all of the elements contained within the k-th subsampling unit in the i-th substratum of the j-th primary unit; S_h is the number of substrata and Q_h is the number of primary units in the h-th primary stratum; and L is the number of primary strata. (X might be the total number of workers in the United States, or the total number of farm laborers, etc.) An estimate of X from the sample is

$$X' = \sum_{h}^{L} 1/t_h \sum_{i}^{S_h} \sum_{k}^{m_{hij}} X_{hijk}. \tag{7}$$

No summation over j is involved, because only one primary unit is drawn from the h-th stratum. This is a very simple estimate, involving a sum weighted only at the primary strata level. If the t_h are all set equal to t, i.e., if a constant proportion is sampled from each stratum, the estimate becomes merely the total number of elements in the sample having the specified characteristic multiplied by $1/t$, the reciprocal of the sampling ratio.

The allocation of the subsampling indicated above may be deviated from and the controls of area substratification can still be maintained if proper modifications are made in the sample estimate. In this event, differential weighting must be introduced at the substrata level rather than only for the primary strata.

The definition of heterogeneous primary sampling units, the proper classification of them into strata, and the use of probabilities proportionate to the measures of size in the selection of the primary units are particularly desirable if area substratification is used. If these are not introduced the likelihood of making substantial gains through the use of area substratification is decreased. The definition of the primary strata should be made in conjunction with the definition of the substrata, and should insure that each primary unit has adequate representation of each substratum that is to be defined within that primary stratum. With this restriction observed, the number of significant substrata that can be defined will be limited by the heterogeneity of the primary units. Thus, in order to provide for substratification into predominantly farm and predominantly nonfarm areas, the primary sampling units should be defined so that both farm and nonfarm areas are represented in each unit. This procedure not only makes area substratification more effective, but improves the efficiency of the sample in making separate estimates for such classes of the population. However, if this procedure cannot be adhered to exactly in practice, primary units in which certain of the substrata are not represented will occasionally come into the sample. One alternative when this occurs is to combine certain substrata; another is to exclude such primary units from the sample.

Since the number of primary strata is restricted by the number of primary units to be sampled, it is wasteful to set up strata at the primary level with respect to sources of variation that can be controlled adequately through area substratification. For example, if farm areas and nonfarm areas are to be distinguished in the substrata, the primary strata should not be exhausted by classifying the primary units into a large number of strata by percent farm (percent of total population in primary unit living on farms), since the effect of the substratification is to control the variation in the percentage farm. Limiting the number of percentage farm classes at the primary level makes possible the use of other modes of stratification that will control on farm type, or on the industrial character of the nonfarm population, or on some other similar criteria.

Area substratification is to be distinguished from the fairly commonly used method of specifying the number of elements to come into the sample from each of several different classes of elements—whether such quotas are fixed to make the sample correspond with the specified characteristics of the entire primary stratum or of the selected primary sampling unit. The method of fixing quotas and instructing interviewers or enumerators to obtain a given number of elements (persons, dwelling units, farms, voters, etc.) having various specified characteristics has a fundamental weakness that is avoided in area substratification within primary strata. Such quotas ordinarily must be set on the basis of previous information or rough estimates, and thus cannot accurately reveal changing characteristics of the population. Area substratification, on the other hand, uses previous information to insure the proper representation of various types of areas in the sample. The numbers of elements obtained with various specified characteristics are determined from the population as it is, and not as it was at some previous date. In times of rapid change the fixing of quotas on the basis of previous information may introduce increasingly serious biases.

The gain from using previously available information in stratifying areas arises from the fact that there is a high correlation in the characteristic of an area from time to time over a period of several years. An area that is predominantly farm at one date ordinarily will be predominantly farm a few years later. Similarly, while very substantial shifts in population may occur, the numbers of persons in a set of areas at one time ordinarily will be very highly correlated with the numbers a few years later. However, area substratification does not depend on the fact that no shifts occur. If shifts have occurred it will measure them. If the shifts have been sufficient to completely alter the character of most small areas, it will still provide estimates revealing the changing character of the population, but under these circumstances the efficiency of the method is decreased.

V—EXPECTED VALUES AND VARIANCES FOR THE SUBSAMPLING SYSTEM INCORPORATING THE PRINCIPLES OUTLINED ABOVE

The system of sampling incorporating the principles of enlarged primary units, the selection of primary units with probabilities proportionate to the measures of size and area substratification will be examined more fully below. It will be referred to, for convenience, as the specified subsampling system.

1. The expected value of an estimated total for the specified subsampling system. All summations in the formulas below are over the population unless otherwise indicated. The expected value of X' as defined in Eq. (7) is

$$EX' = \sum_h \sum_j \sum_i \sum_k (1/t_h)(P_{hj}/P_h)(m_{hij}/M_{hij})X_{hijk}.$$

From (5) $t_h = m_{hij}P_{hij}/M_{hij}P_{hi}$, and therefore

$$EX' = \sum_h \sum_j \sum_i \sum_k (P_{hi}/P_{hij})(P_{hj}/P_h)X_{hijk}$$

$$= \sum_h P_h \sum_j \sum_i (P_{hi}/P_h)(P_{hj}/P_h)(X_{hij}/P_{hij}) = \sum P_h R_{h(A)}$$

where

$$P_h = \sum_i P_{hi} = \sum_j P_{hj}; \qquad R_{h(A)} = \sum_j (P_{hj}/P_h)R_{hj(A)};$$

$$R_{hj(A)} = \sum_i (P_{hi}/P_h)R_{hij}; \quad \text{and} \quad R_{hij} = \sum_k X_{hijk}/P_{hij} = X_{hij}/P_{hij}.$$

The $R_{hj(A)}$ will be referred to as the adjusted ratio for the j-th primary unit. It is the weighted average within the j-th unit of the substrata ratios, R_{hij}, where the same set of weights P_{hi} is applied to the R_{hij} in each primary unit within a stratum. The $R_{h(A)}$ is the average, within the h-th stratum of the adjusted ratios. Hence

$$EX' = X + \sum_h P_h(R_{h(A)} - R_h), \tag{8}$$

where

$$R_h = X_h/P_h, \quad \text{with} \quad X_h = \sum_i \sum_j X_{hij},$$

is the ratio of the aggregate value of the specified characteristic for the elements in the h-th stratum to the measure of size of that stratum, and where the population character being estimated (6), is equal to $X = \Sigma X_h = \Sigma P_h R_h$.

From (8), it is seen that X' is a biased estimate of X, although ordinarily, in practice, only slightly so. The bias, equal to $\Sigma P_h(R_{h(A)} - R_h)$, is the sum of the biases for the various primary strata. Under many practical circumstances some of these will be slightly negative and some slightly positive, with the result that the total bias will be relatively small. The bias would be nonexistent if area substratification were not used, or if the form of the sample estimate were properly modified, but here again, as in the case of substituting biased for unbiased estimates discussed in Sec. III, the introduction of a slight bias may result in a substantial reduction in the variance.

A sufficient, although not necessary, condition for the sample estimate (7) with area substratification to be unbiased is for the ratios P_{hij}/P_{hj} to be uncorrelated with the R_{hij} within each substratum. Under these circumstances

$$\sum_j \frac{P_{hj}}{P_h}\frac{P_{hij}}{P_{hj}} R_{hij} = \sum_j \frac{P_{hj}}{P_h}\frac{P_{hij}}{P_{hj}} \sum_j \frac{P_{hj}}{P_h} R_{hij} = \frac{P_{hi}}{P_h} \sum_j \frac{P_{hj}}{P_h} R_{hj}$$

and therefore

$$R_h = \sum_i \sum_j \frac{P_{hj}}{P_h} \frac{P_{hij}}{P_{hj}} R_{hij} = \sum_i \sum_j \frac{P_{hi}}{P_h} \frac{P_{hj}}{P_h} R_{hij} = R_{h(A)} .$$

To illustrate, if the measures of size are the 1940 populations, then the sample estimate will be unbiased if the proportions of the 1940 populations of the primary sampling units that are in the various substrata are uncorrelated with the corresponding R_{hij}. As indicated earlier these conditions are approximated in many practical problems, especially if the primary stratification has been carried out effectively. Moreover, if the conditions are not met approximately, the bias introduced may still be very small. (See Sec. VII for a numerical illustration.)

2. The mean square error of an estimated total for the specified subsampling system. For the development of the mean square error of X' for the specified subsampling system, see the Appendix, Section 2. There it is shown that the mean square error of X' is

$$\begin{aligned}\sigma_{X'}^2 = {} & \sum_h \sum_i \sum_j P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \frac{\sigma_{hij}^2}{m_{hij} \bar{P}_{hij}^2} \\ & + \sum_h P_h^2 \sum_j \frac{P_{hj}}{P_h} (R_{hj(A)} - R_{h(A)})^2 + [\sum P_h (R_{h(A)} - R_h)]^2\end{aligned} \tag{9}$$

where

$$\sigma_{hij}^2 = \sum_k (X_{hijk} - \bar{X}_{hij})^2 / M_{hij}$$

is the variance between subsampling units within a substratum of the aggregate value of a specified characteristic for the subsampling unit and

$$\bar{P}_{hij} = P_{hij}/M_{hij}$$

is the average measure of size of the subsampling units in the h-i-j-th area.

The first term of (9) is the contribution of the variance between subsampling units and may be kept small by proper definition of the subsampling units, and, of course, by increasing the subsampling ratio. The second term of (9) is the contribution of the variance between primary sampling units within strata; and the third term is the contribution of the bias, which, as indicated before, ordinarily will be of negligible size, so that the mean square error and the variance will be approximately equal.

It is the variance between primary sampling units that contributes most heavily to the total variance in many subsampling situations, and it is on this contribution that the modifications proposed in this paper have their principal effect. The effect of area substratification is seen by comparing the variance between primary units given above with that obtained if area substratification were not used but other aspects of the design remained unchanged. In this event the variance between primary units involves the variance of the ratio, $R_{hj} = \sum_i X_{hij}/P_{hj} = X_{hj}/P_{hj}$, instead of the variance of the adjusted ratio, $R_{hj(A)}$.

The relationship between the variance of R_{hj} and that of $R_{hj(A)}$ within the h-th primary stratum is given by

$$\sigma^2_{R_{hj}} = \sigma^2_{R_{hj(A)}} + \sigma^2_{R_{hj}-R_{hj(A)}} + 2\rho\sigma_{R_{hj(A)}}\sigma_{R_{hj}-R_{hj(A)}}, \tag{10}$$

where $\sigma^2_{R_{hj}-R_{hj(A)}}$ is the variance of the difference between the adjusted and the unadjusted ratios, and ρ is the correlation between the adjusted ratio and the amount of the adjustment. Thus, if the correlation is near zero or positive, there will be a gain from the introduction of area substratification, although there may be a loss if the correlation is highly negative. Essentially, the condition for ρ being equal to or near zero is the same as that for the sample estimate being unbiased; namely, that the P_{hij}/P_{hj} be uncorrelated or only slightly correlated with the R_{hij} within each substratum.[3]

The variance of $R_{hj(A)}$ rather than that of R_{hj} occurs in the variance of X' because the subsampling numbers were allocated proportionate to the P_{hi}, no matter what primary sampling unit happened to be selected for inclusion in the sample. The ratio R_{hj} like $R_{hj(A)}$ may be regarded as the weighted average of the R_{hij} but with the weights equal to P_{hij} instead of P_{hi}, and thus varying from primary unit to primary unit. It would appear, therefore, from the relationship of the variances given above, that if the substrata are effective, and if the P_{hij} are highly correlated with the actual sizes of the substrata, the weighted average using fixed weights in all primary units should have a considerably smaller variance than that using variable weights. This turns out to be the case in many practical situations, some illustrations of which will be given later (see Sec. VII).

3. The mean square error of ratio estimates for the specified subsampling system. The need for estimating a ratio from a sample arises in two cases; first, when the ratio is the population character for which an estimate is desired, and second, when the application of a ratio from the sample to a known total uses additional available information for obtaining an improved estimate of the desired total.

Ratio estimates are desired as an end-result when, for example, the change in a characteristic from one time to another is being considered. Thus, if Y' is the estimated total income of farm workers at one date, and X' the corresponding estimated total income at a second date, then $r' = X'/Y'$ is an estimate of the relative change in the total income of farm workers over the period of time covered. Similarly, the estimate of a percentage such as the percentage of the

[3] Actually, a sufficient, although not necessary, condition for ρ to be equal to zero is that P_{hij}/P_{hj} be uncorrelated with both the ratio R_{hij} and the cross-product $R_{hij}\ R_{hgj}$ for all pairs of substrata.

workers unemployed will involve the ratio of two random variables from the sample. Ratio estimates from a sample may be particularly useful in instances where the reliability of the ratio estimate is greater than the reliability of the estimate of either the numerator or the denominator, as is frequently the case.

Ratio estimates may be used as a means of obtaining an estimated aggregate value of a specified characteristic, if Y, the aggregate value of a second characteristic highly correlated with X is known exactly from independent sources, and X' and Y', estimates of X and Y respectively, are available from the sample. Thus

$$X'' = [X'/Y']Y = r'Y \tag{11}$$

is an estimate of the aggregate value of the specified characteristic. If the correlation, in successive samples, between X' and Y' is sufficiently high, the ratio estimate will be a more efficient estimate of X than will X', the simple estimated total given earlier (7); but X' will prove the more reliable estimate when the correlation is low.[4] Thus, X'', when the correlation between X' and Y' is sufficiently high, makes use of more of the relevant available information for estimating X than does X'.

The application of ratio estimates to the specified subsampling system is considered below.

(*a*) *The estimated ratio and its mean square error.* The estimate of the population ratio $r = X/Y$ is:

$$r' = \frac{X'}{Y'} = \frac{\sum_h^L \frac{1}{t_h} \sum_i^{s_h} \sum_j^{1} \sum_k^{m_{hij}} X_{hijk}}{\sum_h^L \frac{1}{t_h} \sum_i^{s_h} \sum_j^{1} \sum_k^{m_{hij}} Y_{hijk}}, \tag{12}$$

where X' is given in (7) above, and Y' is a similar estimate of the total value of a second characteristic. The mean square error of r' is approximately

$$\begin{aligned}\sigma_{r'}^2 = \frac{1}{Y^2} \Biggl\{ & \sum_h \sum_i \sum_j P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \frac{\sum_k Y_{hijk}^2 (r_{hijk} - r_{hij})^2}{m_{hij} M_{hij} \bar{P}_{hij}^2} \\ & + \sum_h \sum_i \sum_j P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \sigma_{hij:Y}^2 \frac{(r_{hij} - r)^2}{m_{hij} \bar{P}_{hij}^2} \\ & + \sum_h P_h^2 \sum_j \frac{P_{hj}}{P_h} R_{hj(A):Y}^2 (\bar{r}_{hj(A)} - \bar{r}_{h(A)})^2 \\ & + \sum_h P_h^2 (\bar{r}_{h(A)} - r)^2 \sum_j \frac{P_{hj}}{P_h} (R_{hj(A):Y} - R_{h(A):Y})^2 \Biggr\}\end{aligned} \tag{13}$$

[4] The variance of the ratio of random variables of the form $r' = X'/Y'$ is approximately $\sigma_{r'}^2 = r^2(V_{X'}^2 + V_{Y'}^2 - 2\rho_{X'Y'} V_{X'} V_{Y'})$ where V indicates the coefficient of variation of the variable designated by the subscript, and $\rho_{X'Y'}$ is the correlation. Hence, if $\rho_{X'Y'}$ is sufficiently large $V_{r'}^2$ will be less than $V_{X'}^2$. The size of $\rho_{X'Y'}$ required depends on the relative magnitudes of the coefficients of variation of X' and Y'.

where

X_{hijk} = the aggregate value of a specified characteristic for the elements in the k-th subsampling unit within the h-i-j-th area, for which a total is to be estimated;

Y_{hijk} = the aggregate value of a second specified characteristic for the elements in the same subsampling unit, and for which the total in the population is known;

$Y_{hij} = \sum_k Y_{hijk}$, and $Y_h = \sum_i \sum_j Y_{hij}$.

$\sigma^2_{hij:Y} = \dfrac{\sum_k (Y_{hijk} - \bar{Y}_{hij})^2}{M_{hij}}$ is the variance of the sampling units in the h-i-j-th area with respect to the second characteristic, and $\bar{Y}_{hij} = Y_{hij}/M_{hij}$.

$R_{hj(A):Y} = \sum \dfrac{P_{hi}}{P_h} \dfrac{Y_{hij}}{P_{hij}}$ is the adjusted average of the Y_{hij}, and

$r_{hijk} = \dfrac{X_{hijk}}{Y_{hijk}}$, $r_{hij} = \dfrac{X_{hij}}{Y_{hij}}$, etc., are the ratios of the X to the Y for the areas indicated by the subscripts, and

$\bar{r}_{hj(A)} = \dfrac{R_{hj(A)}}{R_{hj(A):Y}}$, and $\bar{r}_{h(A)} = \dfrac{R_{h(A)}}{R_{h(A):Y}}$ are the ratios of the adjusted ratios for X and Y indicated by the subscripts;

and the remaining symbols are as defined in the sections above where the expected value and variance of X' are given.

The first and third terms of (13) are, ordinarily, the principal contributing terms. The second and fourth terms contain contributions due to the variation between the means of the substrata and the primary strata respectively even though the sample was stratified with respect to these classes. In some instances, the contributions of these terms will be important. The between strata contributions arise because the primary and subsampling units vary in size with respect to the character Y.

This formula for the mean square error of a ratio is approximately equal to the one more commonly used given in footnote 4. The two formulas, both of which are approximations, would be identical if certain terms which are ordinarily negligible were retained in (13). This latter formula has the advantage of indicating the effect of different aspects of the design of the sample on the variance of the ratio. The derivation of this approximate variance formula is given in the Appendix, Section 3, together with an indication of the accuracy of the approximation.

(*b*) *The estimated totals and their mean square errors.* As mentioned earlier, two estimates of X, the aggregate value of a given characteristic for all elements are X' (7), and X'' (11). The mean square error of X' is given by (9) and that of X'' is simply equal to $Y^2\sigma^2_{r'}$, where $\sigma^2_{r'}$ is given approximately by (13).

The decision as to whether to use X' or X'' as an estimate of X depends, of course, in the first instance, on whether Y is known, and in the second instance, on the relative magnitudes of the respective mean square errors given in (9) and (13). These may be approximated from prior knowledge concerning the relationships in the population under investigation, or they may be estimated from preliminary sample investigations. However, in instances where there is a positive correlation between the X_{hijk} and the Y_{hijk} within substrata, it is fairly safe to assume that if the information necessary for the ratio estimate is available, there will be little to lose and possibly considerable to gain from its use.

The use of (11) instead of (7) is often desirable when Y in (11) is the aggregate value of the actual sizes of the primary units, and Y' is its estimate. This is particularly so if the measures of size used are not fairly precise measures of the actual sizes, and if, at the same time, the actual size is highly correlated with the character being estimated, in which case the use of ratio estimates will yield gains in both the between primary unit contribution and the within primary unit variance. (See Sec. VII for numerical illustrations.) However, if the measures of size are identical with the actual sizes (i.e., $P_{hijk} = Y_{hijk}$) the last two terms of (13) are identical with the between primary unit contribution to the variance of X' (9), and only the within primary unit variance is affected by the ratio estimate.

While it is fairly safe in practice, if Y is known, to make use of X'' instead of X' as the estimate of X, some care must be exercised to make sure that the X_{hijk} has at least a moderately high average correlation with the Y_{hijk}, where the correlations considered are those within substrata within primary sampling units. If this correlation is low, and if the size of the subsampling unit varies considerably, the ratio estimate may be considerably less efficient than the simple total estimate. On the other hand, if the measures of size of the various substrata and of the primary sampling units are fairly close measures of the actual size, and if the subsampling units have been carefully defined so that they do not vary too greatly in size, the two estimates are likely to have about the same efficiency.

VI—SOME PHYSICAL PROPERTIES OF FREQUENTLY OCCURRING POPULATIONS THAT ARE BASIC TO THE SAMPLING PRINCIPLES RECOMMENDED IN THIS PAPER

Many actual populations are characterized by the following physical properties:

(i) The elements within a cluster are positively correlated with regard to a specified characteristic.

(ii) Clusters containing large numbers of elements have greater internal heterogeneity than clusters containing small numbers of elements.

(iii) Increasing the size of the cluster brings in correlated elements (e.g., in population or agriculture surveys larger clusters are formed by including households or farms in adjacent areas).

The first of these properties is recognized implicitly in the literature where the losses of efficiency through the use of large clusters as sampling units are frequently cited. In our experience the second and third properties hold just as commonly in actual populations, and ordinarily fqr the same populations for which the first property holds.

The presence of these physical properties in combination within strata leads to the following mathematical relationships that have been used throughout this paper:

(a) The sizes of the primary sampling units, N_{hj}, are negatively correlated with the ρ_{hj}, the intra-class correlations between elements within the units;

(b) The N_{hj} and $N_{hj}\rho_{hj}$ are positively correlated;

(c) The N_{hj} and σ_{hj}^2 are positively correlated;

(d) The N_{hj} and σ_{hj}^2/N_{hj} are negatively correlated.

The use of these relationships has determined most of the choices among alternative procedures throughout this paper. The relationships, of course, do not necessarily hold, and exceptions to them can be found [5]. The frequent occurrence of populations characterized by such properties justifies further research on the more effective use of these and other properties that may be found to hold.

VII—SOME APPLICATIONS OF THE PRINCIPLES DESCRIBED IN THIS PAPER TO AN ACTUAL SAMPLING PROBLEM

The analyses summarized below were carried out for the purpose of deciding between alternative sampling procedures in the revision of a monthly national sample for labor force and other characteristics. Budgetary and administrative restrictions made it necessary to confine the field operations to a limited number of administrative centers scattered over the country, from which a sample of less than one-tenth of one percent of the population of the United States was to be drawn.

The original sample (the one to be revised) was of a usual subsampling design in which counties were used as the primary sampling units, and households or small clusters of households were used as the subsampling units. In the revised sample contiguous counties were combined wherever administratively feasible, to form more heterogeneous primary units than the individual counties. Approximately 2000 primary sampling units were formed from the 3000 counties in the United States. The combinations of counties, the primary stratification, the area substratification, and the measures of size, were determined on the basis of 1940 Decennial Census data together with more recent data where available.[5]

The applications of the various principles suggested in this paper have been

[5] See [11] for a full description of the proposed revised sample, including an outline of the criteria of stratification used. That paper may be useful as a simple description of an application of the specified subsampling system.

evaluated by estimating 1930 Census labor force characteristics from a sample that was stratified on the basis of 1940 and more recent data. This constituted a particularly severe test of some of the methods, because of the substantial shifts that had taken place during the 10-year interval between 1930 and 1940.

The analyses to be summarized in this section are concerned primarily with the gains obtainable under favorable circumstances by the introduction of three sampling principles; namely,

(1) enlarged primary units (see Sec. IV-1);

(2) the sampling of primary units with probability proportionate to measures of their size (see Sec. IV-2);

(3) area substratification (see Sec. IV-3).

Some comparisons are also given to illustrate the effect of using alternative sample estimating formulas. Computations have been made for six of the principal items that are currently being included in a monthly report of the labor force; namely, total numbers of male and female workers, total numbers of male and female agricultural workers, and total numbers of male and female nonagricultural workers. The comparisons between alternative systems have been made holding constant both the primary stratification criteria and the expected numbers of persons to be drawn into the sample.

The percentage gains given below are the reductions in the *between* primary unit contributions (which include the bias contributions) to the mean square error.[6] Except where otherwise specified, the sample estimate used is given by (7).

1. Gains obtained by introducing enlarged primary units. The gains obtained by using enlarged primary units are calculated by comparing the mean square errors arising from the sampling design in which individual counties are primary units with the mean square errors arising from the design in which combinations of counties are the primary units. In both designs, the primary units are drawn with equal probabilities and no area substratification is used. For this comparison, preliminary computations have been completed for only a limited number of strata and for two of the labor force items given above; namely, total male workers and total female workers. The reduction in the sampling errors obtained by introducing enlarged primary units is estimated to be 48 per cent for total male workers and 26 per cent for total female workers.

2. Further gains obtained by introducing probability proportionate to measures of size. The further gains obtained by using the principle of sampling with probability proportionate to measures of size are calculated by comparing the mean square errors arising from the design in which the units are drawn with

[6] The contribution of the variance *within* the primary units to the total mean square error was relatively small in all instances, and practically unaffected by the introduction of the various principles.

equal probability with the mean square errors arising from the design in which the units are drawn with probability proportionate to measures of size. In both the designs, the primary units are combinations of counties, and in neither of them is area substratification used. The estimated per cent gains are as follows:

Total Workers		*Agricultural Workers*		*Nonagricultural Workers*	
Male	*Female*	*Male*	*Female*	*Male*	*Female*
50	8	77	6	19	21

The gains reflect both decreases in the sampling variance and the elimination of the bias which arises when the primary units are drawn with equal probabilities.[7]

3. Further gains obtained by introducing area substratification. The further gains obtained by using the principle of area substratification are calculated by comparing the mean square errors for the design in which area substratification is not used, with those for the design in which area substratification is introduced. In both these designs the primary units are combinations of counties, and are drawn with probability of selection proportionate to measures of their sizes. The estimated per cent gains are as follows:

Total Workers		*Agricultural Workers*		*Nonagricultural Workers*	
Male	*Female*	*Male*	*Female*	*Male*	*Female*
6	31	46	51	32	22

4. Gains obtained by the integration of the above principles into a single subsampling system (the specified subsampling system). The gains obtained by using all three principles are calculated by comparing the mean square errors for the specified subsampling system (in which all three principles are used) with the mean square errors for the system in which none of these principles is used. In the specified subsampling system, combined counties are the primary units, the primary units are drawn with probability proportionate to measures of their size, and area substratification is used. In the other system, the primary units are individual counties, the sampling is done with equal probabilities and area substratification is not used. Preliminary computations for this comparison are available for only 2 of the 6 labor force items; namely, total male and total female workers. The estimated gains were 76 per cent for male workers and 53 per cent for female workers.

[7] As indicated before, estimate (7) is used in both designs compared above. This estimate is unbiased for the design in which the primary units are drawn with probability proportionate to measures of size, but is biased for the design in which they are drawn with equal probabilities. However, for the latter design, the biased estimate is usually much more efficient than the best linear unbiased estimate. For the six labor force items, the best linear unbiased estimate gives rise to variances that are several times as large as the mean square errors for the biased estimate.

Calculations are available for all 6 items to measure the gains obtained by the use of the last two of the principles in combination; namely, probability proportionate to measures of size and area substratification. For measuring these gains, the systems are as described above, except that in both designs the primary units are combinations of counties. The estimated per cent gains are as follows:

Total Workers		*Agricultural Workers*		*Nonagricultural Workers*	
Male	*Female*	*Male*	*Female*	*Male*	*Female*
54	37	88	54	45	39

While both the specified subsampling system and the alternative to which it was just compared are biased designs, the bias in the specified system is appreciably smaller than the bias in the latter. For example, while the bias of the specified system in the estimation of total male workers was less than one-half per cent of the true total male workers, the bias for the alternative design in the estimation of the same population character was more than one and one-half per cent.

5. The choice of estimate to use with the specified subsampling system. The simple estimate (7) given for the specified subsampling system may be improved on by the use of regression techniques (see Sec. III). However, such techniques may require a great deal of clerical work, so that they frequently cannot be used in practice. As indicated in the last part of Sec. V, however, if certain independent information such as a knowledge of the total population is available, a simple ratio estimate of the form of (12) may sometimes introduce gains over (7). The use of the ratio estimate may be particularly desirable when the correlation between the measures of size and the actual sizes of the primary sampling units is only moderately high, and when, at the same time, the actual sizes are highly correlated with the values for the character being estimated. A small-scale experiment in the sampling for labor force items indicated that for estimating total male workers for 1930, both the variance between primary units and the variance within primary units for the ratio estimate (12) were approximately one-half that for the simple estimate (7). The use of the ratio estimate had very little effect in the estimation of the remaining five labor force characteristics. The reduction in variance of the total male employment figure was brought about because migration since 1930 reduced the correlation between the 1930 and 1940 sizes, and furthermore, the number of male workers is highly correlated with the total population. Similar reductions for the variances of the other five items were not obtained because the correlations with actual sizes for the other items were not as high.

6. Some final remarks. The gains just obtained arose from application of the sampling principles enumerated above. The situations that these principles were applied to are favorable, but are frequently met in practice. The principles differ in their effect depending on the particular attributes of the population being studied. The use of enlarged primary units may be desirable whenever the enlarged units are internally more heterogeneous than are the smaller units. The selection of primary units with probability proportionate to size is desirable for the general classes of populations described in Sec. VI whenever the primary units vary considerably in size. The use of area substratification is limited to sampling situations where large primary units are used. The joint effect of all three principles shows to greatest advantage when subsampling is used, the primary units are large, but variable in size, and the number of primary units included in the sample is limited by cost or administrative conditions. The types of estimates described in Sec. III may be effective in a large number of physical situations other than those mentioned in this paper.

ACKNOWLEDGMENT

The authors are particularly indebted to Harold Nisselson for many helpful suggestions, and also for planning and directing the analyses summarized in Sec. VII. They are heavily indebted also, to W. G. Cochran, W. Edwards Deming, L. R. Frankel, William G. Madow and Frederick F. Stephan for reading the manuscript and making many helpful suggestions. The work has been carried out in the Bureau of the Census, in connection with the development of various sampling surveys, under the general direction of J. C. Capt, Director, and Philip M. Hauser, Assistant Director.

APPENDIX

1. The effect of the consolidation of the primary units on the sampling variance (see Sec. IV-1). Let $\bar{X}_1' = \sum_j^q \sum_k^n X_{jk}/qn$, be the average for the sample where the primary units are the original units and where X_{jk} is the value of the k-th element in the j-th primary unit; q is the number of primary units in the sample, and n is the number of elements sampled from each of the q primary units. The variance of $\bar{X}_1'$ is

$$\sigma^2_{\bar{x}_1} = \frac{N-n}{(N-1)nq}\,\sigma^2_{1w} + \frac{Q-q}{(Q-1)q}\,\sigma^2_{1b} \tag{14}$$

where Q is the number of original primary units in the population; N is the number of elements in each original primary unit; $\sigma^2_{1w} = \Sigma\Sigma(X_{jk} - \bar{X}_j)^2/QN$ is the variance within the original primary units, with $\bar{X}_j = \sum_k X_{jk}/N$; and $\sigma^2_{1b} = \Sigma(\bar{X}_j - \bar{X})^2/Q$ is the variance between the original primary units, with $\bar{X} = \Sigma\bar{X}_j/Q$.

$$\sigma^2 = \Sigma\Sigma(X_{jk} - \bar{X})^2/QN = \sigma^2_{1w} + \sigma^2_{1b}. \tag{15}$$ Then

$$\sigma^2_{1b} = \sigma^2[1 + \rho_1(N-1)/N \tag{16}$$

where $\rho_1 = \left[\sigma_{1b}^2 - \frac{\sigma_{1w}^2}{N-1}\right]\frac{1}{\sigma^2}$ is the intra-class correlation[8] between elements in the original units.

From (15) and (16)

$$(17) \qquad \sigma_{1w}^2 = \frac{N-1}{N}\sigma^2(1-\rho_1).$$

Hence

$$(18) \qquad \sigma^2_{\bar{X}'} = \frac{N-n}{N}\frac{\sigma^2}{nq}(1-\rho_1) + \frac{Q-q}{(Q-1)q}\frac{\sigma^2}{N}[1+\rho_1(N-1)].$$

Similarly, the variance of $\bar{X}_2'$ is

$$(19) \qquad \sigma^2_{\bar{X}_2'} = \frac{CN-n}{CN}\frac{\sigma^2}{nq}(1-\rho_2) + \frac{Q-qC}{(Q-C)q}\frac{\sigma^2}{CN}[1+\rho_2(CN-1)]$$

where $\bar{X}_2'$ is the mean for the enlarged primary units, ρ_2 is the intra-class correlation between elements in the enlarged primary units and C is the number of original units combined to form each enlarged unit. Then

$$(20) \qquad \sigma^2_{\bar{X}_1'} - \sigma^2_{\bar{X}_2} = \frac{\sigma^2}{qN}\left\{\frac{(q-1)(C-1)}{(Q-1)(Q-C)} + \rho_1 a_1 - \rho_2 a_2\right\}$$

where $a_1 = \frac{(Q-q)(N-1)}{Q-1} - \frac{N-n}{n}$ and $a_2 = \frac{(Q-Cq)(CN-1)}{(Q-C)C} - \frac{CN-n}{Cn}$.

Since

$$a_1 - a_2 = \frac{(C-1)(q-1)(QN-1)}{(Q-1)(Q-C)} \geqq 0 \quad \text{and} \quad \frac{(q-1)(C-1)}{(Q-1)(Q-C)} \geqq 0,$$

then a gain is brought about by enlarging primary units whenever $\rho_1 > \rho_2$, where ρ_1 and ρ_2 are both positive.

2. Comparison of variances of certain alternative subsampling systems where the primary units are of unequal sizes. The development of (4), the formula for the difference between the variances of sample estimates compared in Sec. IV-2 is given below. We shall confine ourselves to the simple case where only one primary sampling unit is drawn into the sample from each stratum. Let

$$(21) \qquad \bar{X}' = \Sigma N_h \bar{X}_h'/N$$

be the sample estimate used for each of the three designs to be compared, where $\bar{X}_h' = \bar{X}_{hj}' = \sum_k^{n_{hj}} X_{hjk}/n_{hj}$, and X_{hjk} is the value of the k-th element in the j-th

[8] For definitions and properties of intra-class correlations, see Secs. 38–40 of *Statistical Methods for Research Workers*, R. A. Fisher, and [5].

primary unit in the h-th stratum; L is the number of strata; n_{hj} is the number of elements drawn into the sample from the j-th primary unit in the h-th stratum with N_{hj} the corresponding total number, $N_h = \sum_j^{Q_h} N_{hj}$ with $Q_h =$ the number of primary units in the h-th stratum, and $N = \sum_h^L N_h$. If the subsampling within a stratum is of a constant proportion, C, as in the first of the subsampling systems mentioned, n_{hj} in the above estimate is equal to C N_{hj}. If the subsampling within a stratum is of a constant number, as in the second subsampling system mentioned, as well as in the recommended system, n_{hj} is equal to $\bar{n}_h = C \sum_j N_{hj}/Q_h = C\bar{N}_h$.

We shall denote the sample estimate for the first design by $\bar{X}_1'$, that for the second design by $\bar{X}_2'$, and that for the recommended design by $\bar{X}_3'$.

The expected values of the sample estimates for the first two designs, $\bar{X}_1'$, and $\bar{X}_2'$, are the same, and are equal to

$$E\bar{X}_1' = E\bar{X}_2' = \bar{\bar{X}} = \frac{1}{N} \sum_h \frac{N_h}{Q_h} \sum_j \frac{n_{hj}}{N_{hj}} \sum_k \frac{X_{hkj}}{n_{hj}} = \frac{1}{N} \sum_h \frac{N_h}{Q_h} \sum_j \bar{X}_{hj}$$

where $\bar{X}_{hj} = \sum_k X_{hjk}/N_{hj}$. Thus, since $\bar{\bar{X}}$ is not, in general, equal to $\sum_{h,i,j} X_{hij}/\sum_{h,j} N_{hj} = \bar{X}$, both $\bar{X}_1'$ and $\bar{X}_2'$ are biased estimates of $\bar{X}$.

For the recommended design, in which the primary unit is drawn with probability of selection proportionate to size and a constant number taken from the sampled units within a stratum, the expected value of the sample estimate is

$$(22) \qquad E\bar{X}_3' = \frac{1}{N} \sum_h \sum_j N_h \frac{N_{hj}}{N_h} \frac{\bar{n}_h}{N_{hj}} \frac{X_{hj}}{\bar{n}_h} = \frac{1}{N} \sum_h \sum_j X_{hj} = \bar{X}$$

and therefore the estimate for the recommended design is unbiased.

The mean square error of $\bar{X}_1'$ is

$$(23) \quad \sigma^2_{\bar{X}_1'} = \frac{1}{N^2} \sum_h \frac{N_h^2}{Q_h} \left[\sum_j \frac{N_{hj} - n_{hj}}{(N_{hj} - 1)n_{hj}} \sigma_{hj}^2 + \sum_j (\bar{X}_{hj} - \bar{X}_h)^2 \right] + (\bar{\bar{X}} - \bar{X})^2 - \frac{1}{N^2} \sum_h N_h^2 (\bar{\bar{X}}_h - \bar{X}_h)^2$$

where $\sigma_{hj}^2 = \sum_k (X_{hjk} - \bar{X}_{hj})^2/N_{hj}$ is the variance between elements within the j-th primary sampling unit of the h-th stratum, $\bar{\bar{X}}_h = \sum_j \bar{X}_{hj}/Q_h$, $\bar{X}_{hj} = X_{hj}/N_{hj}$, and $\bar{X}_h = \sum_j \sum_k X_{hjk}/\sum_j N_{hj} = \sum_j N_{hj}\bar{X}_{hj}/\sum_j N_{hj}$. The first term in the square bracket of (23) is the contribution of the variance within primary units. The second term in the square bracket is an approximation to the mean square error between primary units and the remaining terms give the error in this approximation. The mean square error of $\bar{X}_2'$ is given by the same formula but with n_{hj} replaced by $\bar{n}_h$.

The difference between $\sigma^2_{X'_1}$ and $\sigma^2_{\bar{X}'_2}$ is

$$(24) \qquad \sigma^2_{\bar{X}'_1} - \sigma^2_{\bar{X}'_2} = \frac{1}{CN^2} \sum_h \frac{N_h^2}{Q_h} \sum_j \sigma^2_{hj} \frac{N_{hj}}{N_{hj} - 1} \left(\frac{1}{N_{hj}} - \frac{1}{\bar{N}_h} \right),$$

which will be positive if σ^2_{hj}/N_{hj} is negatively correlated with N_{hj}, as is almost invariably the case in practice (see Sec. VI). Thus, since $\sigma^2_{\bar{X}'_1}$ ordinarily is larger than $\sigma^2_{\bar{X}'_2}$, it will suffice to compare $\sigma^2_{\bar{X}'_2}$ with $\sigma^2_{\bar{X}'_3}$ to show that the recommended subsampling system is more efficient than either of the first two mentioned.

The variance for the recommended design is

$$(25) \qquad \sigma^2_{\bar{X}'_3} = \frac{1}{N^2} \sum_h N_h^2 \left[\sum_j \frac{N_{hj}}{N_h} \frac{N_{hj} - \bar{n}_h}{N_{hj} - 1} \frac{\sigma^2_{hj}}{\bar{n}_h} + \sum_j \frac{N_{hj}}{N_h} (\bar{X}_{hj} - \bar{X}_h)^2 \right].$$

For comparing the mean square error of $\bar{X}'_2$ with the variance of $\bar{X}'_3$ we shall define

$$\rho_{hj} = \frac{1}{\sigma_h^2} \left[(\bar{X}_{hj} - \bar{X}_h)^2 - \frac{\sigma^2_{hj}}{N_{hj} - 1} \right]$$

as the intra-class correlation coefficient between elements within the j-th primary unit, where σ_h^2 is the variance between all elements within the h-th stratum. In this comparison, the terms outside the square brackets in (23), have been ignored because their contribution to the mean square error is either positive or negligible. Then,

$$(26) \quad \sigma^2_{\bar{X}'_2} - \sigma^2_{\bar{X}'_3} = \frac{1}{N^2} \sum_h \frac{N_h^2}{Q_h} \left\{ \sum_j \frac{N_{hj}}{N_{hj} - 1} \frac{\sigma^2_{hj}}{\bar{n}_h} \left(1 - \frac{N_{hj}}{\bar{N}_h} \right) + \sigma_h^2 \sum_j \rho_{hj} \left(1 - \frac{N_{hj}}{\bar{N}_h} \right) \right\}.$$

The second term of this difference was given in Sec. IV-2 as the approximate difference, and the first term was neglected. To examine the relative magnitudes of the two terms we shall write

$$(27) \qquad \frac{N_{hj}}{N_{hj} - 1} \sigma^2_{hj} = \sigma_h^2 (1 - \delta_{hj}).$$

Then

$$(28) \quad \sigma^2_{\bar{X}'_2} - \sigma^2_{\bar{X}'_3} = \frac{1}{N^2} \sum_h \frac{N_h^2}{Q_h} \sigma_h^2 \left\{ \frac{1}{\bar{n}_h} \sum_j \delta_{hj} \left(\frac{N_{hj}}{\bar{N}_h} - 1 \right) - \sum_j \rho_{hj} \left(\frac{N_{hj}}{\bar{N}_h} - 1 \right) \right\}.$$

For the general class of populations given in Sec. VI the covariance between δ_{hj} and N_{hj}, and also that between ρ_{hj} and N_{hj}, will be negative. Moreover, in many practical problems of this class the two covariances will be of approximately the same magnitude. In such instances the first term of (27) will be equal to $\frac{1}{\bar{n}_h}$ times the second, and thus smaller than the second term for all $\bar{n}_h > 1$, and much smaller for moderately large values of $\bar{n}_h$. For example, in populations made up of clusters of different sizes for which the conditional probability of an element having a particular property for a fixed size of cluster is the same for all sizes of clusters, the two covariances will be very nearly equal. A number of practical problems approximate this situation. Moreover, even in the situations where the covariance of δ_{hj} and N_{hj} is several times that of ρ_{hj} and N_{hj}, say 5 times as large, then the second term will be larger than the first for all $\bar{n}_h > 5$.

Some numerical illustrations of the gains obtained through the use of the recommended system are given in Sec. VII, and for some of the items for which results are summarized in that section the gains were substantial.

3. The derivation of the variance formulas (13) and (9). The mean square error of a ratio of random variables is generally approximated from Taylor's expansion. If X' and Y' are random variables, $Y' > 0$, and r is the population character of which $X'/Y' = r'$ is an estimate, then

$$(29)\qquad E\left\{\frac{X'}{Y'} - r\right\}^2 = E\,\frac{Y'^2}{(EY')^2}\left(\frac{X'}{Y'} - r\right)^2 + E\left(1 - \frac{Y'^2}{(EY')^2}\right)\left(\frac{X'}{Y'} - r\right)^2.$$

The first term in the right-hand side of (29) is a first approximation to the mean square error from Taylor's expansion, and the second term is the error in this approximation.

Eq. (13), and as a special case (9), is derived as follows:

$$(30)\qquad E(r' - r)^2 = E\left\{\frac{\sum_h^L \frac{1}{t_h}\sum_i^{s_h}\sum_j^{1}\sum_k^{m_{hij}} X_{hijk}}{\sum_h^L \frac{1}{t_h}\sum_i^{s_h}\sum_j^{1}\sum_k^{m_{hij}} Y_{hijk}} - r\right\}^2.$$

Let $\psi_{hijk} = Y_{hijk}(r_{hijk} - r)$, and $Y' = \sum_h^L \frac{1}{t_h}\sum_i^{s_h}\sum_j^{1}\sum_k^{m_{hij}} Y_{hijk}$. Then, setting

$$(31)\qquad \theta = \sum_h^L \frac{1}{t_h}\sum_i^{s_h}\sum_j^{1}\sum_k^{m_{hij}} \psi_{hijk}\Big/E\left(\sum_h^L \frac{1}{t_h}\sum_i^{s_h}\sum_j^{1}\sum_k^{m_{hij}} Y_{hijk}\right) = \frac{Y'}{EY'}\left(\frac{X'}{Y'} - r\right)$$

$$E\theta^2 = EY'^2(r' - r)^2/(EY')^2$$

is the first approximation to the mean square error.

Since EY' is evaluated in the same way as EX'(8), it is merely necessary to evaluate $EY'^2(r' - r)^2$, the numerator of $E\theta^2$. Now

$$\begin{aligned} EY'^2(r' - r)^2 &= E\left[\sum_h \frac{1}{t_h}\sum_i^{s_h}\sum_j^{1}\sum_k^{m_{hij}} \psi_{hijk}\right]^2 \\ &= E\sum_h \frac{1}{t_h^2}\psi_h'^2 + E\sum_{\substack{h,q\\ h\neq q}} \frac{\psi'_h}{t_h}\frac{\psi'_q}{t_q} \end{aligned}$$

where $\psi'_h = \sum_i^{s_h}\sum_j^{1}\sum_k^{m_{hij}} \psi_{hijk} = \sum_i^{s_h} \psi'_{hi}$.

Since $E \sum \frac{1}{t_h^2} \psi_h'^2 = E \sum_h \sum_i \psi'^2_{hi}/t_h^2 + E \sum_h \sum_{\substack{i,r \\ i \neq r}} \psi'_{hi}\psi'_{hr}/t_h^2$

$$EY'^2(r' - r)^2 = E \sum_h \frac{1}{t_h^2} \sum_i \psi_{hi}'^2 + E \sum_h \frac{1}{t_h^2} \sum_{\substack{i,r \\ i \neq r}} \psi'_{hi}\psi'_{hr} + E \sum_{\substack{h,q \\ h \neq q}} \frac{\psi'_h}{t_h} \frac{\psi'_q}{t_q}. \tag{32}$$

The first term in the right-hand side of (32) is

$$\begin{aligned} E \sum_{h,i} \psi'^2_{hi} \frac{1}{t_h^2} = \sum_{h,i,j} \frac{1}{t_h^2} \frac{P_{hj}}{P_h} \frac{m_{hij}}{M_{hij}} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \sum_k \psi^2_{hijk} \\ + \sum_{h,i,j} \frac{1}{t_h^2} \frac{P_{hj}}{P_h} \frac{m_{hij}}{M_{hij}} \frac{m_{hij} - 1}{M_{hij} - 1} (\sum_k \psi_{hijk})^2. \end{aligned} \tag{33}$$

The second term of (32) is

$$E \sum_h \frac{1}{t_h^2} \sum_{\substack{i,r \\ i \neq r}} \psi'_{hi}\psi'_{hr} = \sum_{h,j} \frac{1}{t_h^2} \left(\sum_i \frac{m_{hij}}{M_{hij}} \psi_{hij} \right)^2 - \sum_{h,j} \frac{1}{t_h^2} \frac{P_{hj}}{P_h} \sum_i \frac{m^2_{hij}}{M^2_{hij}} \psi^2_{hij} \tag{34}$$

where

$$\psi_{hij} = \sum_k \psi_{hijk} ;$$

and the third term of (32) is

$$E \sum_{\substack{h,q \\ h \neq q}} \frac{\psi'_h}{t_h} \frac{\psi'_q}{t_q} = \left[\sum_{h,i,j} \frac{1}{t_h} \frac{P_{hj}}{P_h} \frac{m_{hij}}{M_{hij}} \psi_{hij} \right]^2 - \sum \frac{1}{t_h^2} \left[\sum_{i,j} \frac{P_{hj}}{P_h} \frac{m_{hij}}{M_{hij}} \psi_{hij} \right]^2. \tag{35}$$

Therefore $EY'^2 \, (r' - r)^2 = (33) + (34) + (35)$, and when $Y_{hijk}(r_{hijk} - r)$ is substituted for ψ_{hij}, we have

$$\begin{aligned} EY'^2(r' - r)^2 = \sum \frac{1}{t_h^2} \Bigg[\sum_{i,j} \frac{P_{hj}}{P_h} \frac{m_{hij}}{M_{hij}} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \sum Y^2_{hijk}(r_{hijk} - r)^2 \\ + \sum_{i,j} \frac{P_{hj}}{P_h} \frac{m_{hij}}{M_{hij}} \frac{m_{hij} - 1}{M_{hij} - 1} [\sum_k Y_{hijk}(r_{hijk} - r)]^2 \\ + \sum_j \frac{P_{hj}}{P_h} [\sum_i \frac{m_{hij}}{M_{hij}} Y_{hij}(r_{hij} - r)]^2 \\ - \sum_{i,j} \frac{P_{hj}}{P_h} \frac{m^2_{hij}}{M^2_{hij}} - Y^2_{hij}(r_{hij} - r)^2 \\ - \left\{ \sum_{i,j} \frac{P_{hj}}{P_h} \frac{m_{hij}}{M_{hij}} (r_{hij} - r) Y_{hij} \right\}^2 \Bigg] \\ + \left[\sum_{h,i,j} \frac{1}{t_h} \frac{P_{hj}}{P_h} \frac{m_{hij}}{M_{hij}} Y_{hij}(r_{hij} - r) \right]^2. \end{aligned} \tag{36}$$

By substituting $(r_{hijk} - r_{hij} + r_{hij} - r)^2$ for $(r_{hijk} - r)^2$ in the first term of (36) and $P_{hi}M_{hij}/P_{hij}m_{hij}$ for $1/t_h$ in the 1st, 2nd, and 4th terms, the sum of these three terms becomes

$$\begin{aligned}&\sum_{h,i,j,k} \frac{P_{hj}}{P_h}\frac{M_{hij}}{m_{hij}}\frac{P_{hi}^2}{P_{hij}^2} F_{hij} Y_{hijk}^2 (r_{hijk} - r_{hij})^2 \\ &+ 2\sum_{h,i,j,k} \frac{P_{hj}}{P_h}\frac{M_{hij}}{m_{hij}}\frac{P_{hi}^2}{P_{hij}^2} F_{hij} Y_{hijk}^2 (r_{hijk} - r_{hij})(r_{hij} - r) \\ &+ \sum_{h,i,j} \frac{P_{hj}}{P_h}\frac{M_{hij}}{m_{hij}}\frac{P_{hi}^2}{P_{hij}^2} F_{hij}(r_{hij} - r)^2 \left[\sum_k Y_{hijk}^2 - \frac{Y_{hij}^2}{M_{hij}}\right]\end{aligned} \tag{37}$$

where $F_{hij} = (M_{hij} - m_{hij})/(M_{hij} - 1)$ and $r_{hij} = \sum_k X_{hijk}/\sum_k Y_{hijk}$.

When we substitute the appropriate value for $1/t_h$ in the 3rd, 5th, and 6th terms of (36), the sum of these terms becomes

$$\begin{aligned}&\sum_{h,j} \frac{P_{hj}}{P_h}\left[\sum_i \frac{P_{hi}}{P_{hij}} Y_{hij}(r_{hij} - r)\right]^2 - \sum_h \left[\sum_{i,j} \frac{P_{hj}}{P_h}\frac{P_{hi}}{P_{hij}} Y_{hij}(r_{hij} - r)\right]^2 \\ &\qquad + \left[\sum_{h,i,j} \frac{P_{hj}}{P_h}\frac{P_{hi}}{P_{hij}} Y_{hij}(r_{hij} - r)\right]^2.\end{aligned} \tag{38}$$

Now

$$\begin{aligned}\sum_i \frac{P_{hi}}{P_{hij}} Y_{hij}(r_{hij} - r) = \sum_i P_{hi}\left(\frac{X_{hij}}{P_{hij}} - \frac{Y_{hij}}{P_{hij}} r\right) &= P_h(R_{hj(A)} - rR_{hj(A):Y}) \\ &= P_h R_{hj(A):Y}(\bar{r}_{hj(A)} - r)\end{aligned} \tag{39}$$

where $\bar{r}_{hj(A)} = R_{hj(A)}/R_{hj(A):Y}$, and

$$\begin{aligned}\sum_{i,j} \frac{P_{hj}}{P_h}\frac{P_{hi}}{P_{hij}} Y_{hij}(r_{hij} - r) = \sum_j P_{hj}(R_{hj(A)} - rR_{hj(A):Y}) &= P_h(R_{h(A)} - rR_{h(A):Y}) \\ &= P_h R_{h(A):Y}(\bar{r}_{h(A)} - r)\end{aligned} \tag{40}$$

where $\bar{r}_{h(A)} = R_{h(A)}/R_{h(A):Y}$.

Substituting (39) and (40) in (38), we have

$$\begin{aligned}&\sum_{h,j} (P_{hj}/P_h)P_h^2 R_{hj(A):Y}^2(\bar{r}_{hj(A)} - r)^2 - \sum_h P_h^2 R_{h(A):Y}^2(\bar{r}_{h(A)} - r)^2 \\ &\qquad + [\sum_h P_h R_{h(A):Y}(\bar{r}_{h(A)} - r)]^2.\end{aligned} \tag{41}$$

By substituting $(\bar{r}_{hj(A)} - \bar{r}_{h(A)} + \bar{r}_{h(A)} - r)^2$ for $(\bar{r}_{hj(A)} - r)^2$ in the first term in (41) and expanding, (41) becomes

$$\begin{aligned}&\sum_{h,j} P_h^2 \frac{P_{hj}}{P_h} R_{hj(A):Y}^2(\bar{r}_{hj(A)} - \bar{r}_{h(A)})^2 + 2\sum_{h,j} P_h^2 \frac{P_{hj}}{P_h} R_{hj(A):Y}^2(\bar{r}_{hj(A)} - \bar{r}_{h(A)}(\bar{r}_{h(A)} - r) \\ &+ \sum_h P_h^2(\bar{r}_{h(A)} - r)^2\left[\sum_j \frac{P_{hj}}{P_h} R_{hj(A):Y}^2 - R_{h(A):Y}^2\right] + [\sum_h P_h R_{h(A):Y}(\bar{r}_{h(A)} - r)]^2.\end{aligned} \tag{42}$$

Hence, since $(EY')^2 \, E\theta^2 = (37) + (42)$,

$$
\begin{aligned}
(EY')^2E\theta^2 = {} & \sum_{h,i,j} P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \frac{\sum_k Y_{hijk}^2 (r_{hijk} - r_{hij})^2}{m_{hij} M_{hij} \bar{P}_{hij}^2} \\
& + 2 \sum_{h,i,j} P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \frac{\sum_k Y_{hijk}^2 (r_{hijk} - r_{hij})(r_{hij} - r)}{m_{hij} M_{hij} \bar{P}_{hij}^2} \\
& + \sum_{h,i,j} P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \sigma_{hij:Y}^2 \frac{(r_{hij} - r)^2}{m_{hij} \bar{P}_{hij}^2} \\
& + \sum_{h,j} P_h^2 (P_{hj}/P_h) R_{hj(A):Y}^2 (\bar{r}_{hj(A)} - \bar{r}_{h(A)})^2 \\
& + 2 \sum_{h,j} P_h^2 (P_{hj}/P_h) R_{hj(A):Y}^2 (\bar{r}_{hj(A)} - \bar{r}_{h(A)})(\bar{r}_{h(A)} - r) \\
& + \sum_{h,j} P_h^2 (\bar{r}_{h(A)} - r)^2 (P_{hj}/P_h)(R_{hj(A):Y} - R_{h(A):Y})^2 \\
& + [\sum_h P_h R_{h(A):Y} (\bar{r}_{h(A)} - r)]^2
\end{aligned}
\tag{43}
$$

where $\sigma_{hij:Y}^2 = \sum_k^{M_{hij}} (Y_{hijk} - \bar{Y}_{hij})^2 / M_{hij}$ and $\bar{Y}_{hij} = \sum_k^{M_{hij}} Y_{hijk}/M_{hij} = Y_{hij}/M_{hij}$.

The approximation to $E(r' - r)^2$ is given by (43) divided by $(EY')^2$. By ignoring the 2nd, 5th, and 7th terms which are negligible for a large class of populations, we obtain (13).

The variance of X' is derived from (43) by simply substituting $\bar{P}_{hij}/P$ for Y_{hijk} in (43). This follows from the considerations given below:

Since $r' = X'/Y'$, and X' is the numerator of r', $\sigma_{X'}^2$ is given by $\sigma_{r'}^2$ when the denominator, Y', is identically equal to unity in repeated samplings.

$$\text{Since } \frac{1}{t_h} = \frac{M_{hij} P_{hi}}{m_{hij} P_{hij}} = \frac{P_{hi}}{m_{hij} \bar{P}_{hij}} \text{ from (5),}$$

the denominator of r' which is equal to

$\sum_h^L \sum_i^{s_h} \sum_j^1 \sum_k^{m_{hij}} \frac{P_{hi}}{m_{hij} \bar{P}_{hij}} Y_{hijk}$, will be identically equal to unity in repeated sampling when Y_{hijk} is set equal to $\bar{P}_{hij}/P$ where $P = \Sigma P_h$.

The formula for the mean square error of X' (9), of course is exact since the error term

$$E\{Y'^2/(EY')^2\}\{r' - r\}^2 = 0.$$

It may be pointed out that $\sigma_{X'}^2$ may be obtained directly and more simply without the use of (29) since X' is not estimated from the ratio of random variables.

From (29), the error term for the approximation to $E(r' - r)^2$, $(43)/(EY')^2$, is given by $E\left(1 - \frac{Y'^2}{(EY')^2}\right)\{r' - r\}^2$. This cannot be expressed as a simple function of the individual observations, but useful maxima and minima for it may be obtained. A method for obtaining the upper and lower bounds of the variance of r' is simply attained from the following inequalities which hold independent of the joint distribution of X' and Y'.

$$\frac{EY'^2}{Y^2_{\max}}(r'-r)^2 \leq E(r'-r)^2 \leq \frac{EY'^2}{Y^2_{\min}}(r'-r)^2 \tag{44}$$

where $Y_{\max}$ is the maximum value of the Y' obtained simply by choosing or estimating the largest Y_h' for each stratum. $Y_{\min}$ (the minimum value of Y') is obtained in a similar manner.

Eq. 44 when evaluated turns out to be

$$\frac{(EY')^2E\theta^2}{Y^2_{\max}} \leq E(r'-r)^2 \leq \frac{(EY')^2E\theta^2}{Y^2_{\min}} \tag{45}$$

where $(EY')^2E\theta^2$ is given by (43).

Eq. (45) will serve adequately as an indicator of the accuracy of $E\theta^2$ for sampling systems in which the variability of the Y's within strata is restricted. However, in other designs, where stratification is not used and the variability in the Y's is not restricted the limits given by (45) may be too broad to be useful.

REFERENCES

[1] A. L. Bowley, "Measurement of the precision attained in sampling," *Bulletin de L'Institut International de Statistique* Tome XXII, (1926), 1 ere Livraison.

[2] W. G. Cochran, "The use of analysis of variance in enumeration by sampling," *Jour. Amer. Stat. Assoc.*, Vol. 34 (1939), pp. 492–510.

[3] W. G. Cochran, "Sampling theory when the sampling units are of unequal sizes," *Jour. Amer. Stat. Assoc.*, Vol. 37 (1942), pp. 199–212.

[4] Calvert L. Dedrick and Morris H. Hansen, *Final Report on Total and Partial Unemployment, Vol. IV. The Enumerative Check Census*, U. S. Govt. Printing Office (1938).

[5] Morris H. Hansen and William N. Hurwitz, "Relative efficiencies of various sampling units in population inquiries," *Jour. Amer. Stat. Assoc.*, Vol. 37 (1942). pp. 89–94.

[6] R. J. Jessen, *Statistical investigation of a sample survey for obtaining farm facts*, Research Bulletin 304 (1942), Iowa State College, Ames, Iowa.

[7] P. C. Mahalanobis, "A sample survey of the acreage under jute in Bengal," *Sankhyā*, Vol. 4 (1940), pp. 511–530.

[8] J. Neyman, "On the two different aspects of the representative method; a method of stratified sampling and the method of purposive selection," *Jqur. Royal Stat. Soc.*, New Series, Vol. 97 (1934), pp. 558–606.

[9] J. Neyman, "Contribution to the theory of sampling human populations," *Jour. Amer. Stat. Assoc.*, Vol. 35 (1938), pp. 101–116.

[10] F. Yates and I. Zacopanay, "The estimation of the efficiency of sampling, with special reference to sampling for yield in cereal experiments." *Jour. Agr. Sci.*, Vol. 25 (1935), pp. 543–577.

[11] Bureau of the Census, "A revised sample for current surveys," February 1943.