regression estimator is also more efficient than the ratio estimator. If ρ is near -1, the product estimator should be considered.

The use of auxiliary information in the estimator must be in the form of quantitive variables. In addition, it must be available for the total of all units in the population prior to the data collection phase unless double sampling is being employed.

6.8.2 Choice of Stratification Criterion

Information useful for formation of strata is generally of two kinds; that which is based on

(1) the arrangement of the elements in the universe such as a listing structure, or

(2) some knowledge about individual elements, such as on a variate X_i related to Y_i.

In many types of listings, the principle of proximity in grouping units to attain a lower within strata variance is useful based on geographical areas such as by county, city, or minor civil division which correspond to political subdivisions. However, subdivisions shown on maps which correspond to major soil types, medical areas, socio-economic class, or value of housing are examples of types of information which may also be useful in forming strata.

For the second type of information, a universe of homes may have data available on assessed value of individual homes and buildings as well as for entire political units. For universes of business establishments, dollar volume of business in the previous year may be available as well as type of business, number of employees, and various kinds of other information. This later type of information may be either quantitative or categorical in nature.

In many practical situations, the statistician is confronted with several potential stratification "factors." Frequently, geographic location and size of business, based on volume of sales and number of employees, are available for forming strata. Sometimes the number of potential strata becomes so large, it is necessary to drastically reduce either the number of stratification factors or the number of levels, or both. In this case some rough and simple rules for deciding on preference may be useful.

(1) In general, qualitative and non-measurable characteristics should be preferred over quantitative characteristics for use in stratification. Qualitative information is difficult to use anywhere except in stratification whereas quantitative data may be more fully utilized in the estimator or in selection probabilities.

(2) If the quantitative information is not related to Y_i in a simple manner (say linear) then it may be better to utilize it in stratification rather than in the estimator or selection probabilities.

(3) If more than one characteristic is being surveyed and each is roughly of equal importance, then it is better to forego use of quantitative information thought to be correlated with one or only a few of the characteristics under measurement in either the estimator or selection phase and use it in stratification.

6.8.3 Use in Assigning Selection Probabilities

Equal probability schemes are quite popular and applicable to a wide range of problems because of their basic simplicity. However, the use of unequal probabilities in selection can result in a considerable increase in efficiency. It will be found that the variance is a minimum when $P_i = Y_i/Y$. That is, when the probabilities of selection are proportional to values being observed. This is an interesting fact, but difficult to apply in practice since the Y_i's are unknown, otherwise we would not need the survey. For a survey with many characteristics, this condition cannot be satisfied for all characteristics since P_i will be determined based on a single set of X_i representing some measure of size for the sampling unit; that is $P_i = \frac{X_i}{X}$ where X_i is correlated with Y_i. However, two types of size measures have proved to be useful over rather general conditions. The first is the use of information on the Y characteristic for a previous point in time, such as censuses, as a measure of size of the current Y's. The second depends on the existence of sub-elements, such as number of farms, housing units, etc., within the units to be selected.

If such information does not exist on the number of subunits, it is frequently possible to substitute "eyeball estimates" or cruise counts which are current and correlated with the Y's. Of course, the same information might be employed in an alternate way by forming clusters of units of approximately equal size. The use of the information in this manner is perhaps more properly referred to as frame construction or modification.

6.9 Periodic Surveys (Sampling Over Several Occasions)

Many surveys are made periodically of the same population to measure change in the same characteristic over time or to estimate the average characteristic over the combined periods. In some cases, this information might be obtained in a single survey by requesting respondents to provide information for two or more periods. While a single survey would be less expensive in terms of dollars spent, many respondents are unable to provide accurate information for several periods of time either due to problems of memory recall or records are not retained so they can be referred to where necessary. However, periodic surveys provide opportunities to make use of experience gained from earlier surveys to change the sample allocation and make other improvements in the survey over time. Repetitive surveys basically employ auxiliary data and double sampling concepts. Two types of problems are of special interest in periodic surveys:

(1) Choosing the appropriate estimator(s) to use since repeated information on the same characteristic(s) is usually available for some or all of the same sampling units, and

(2) Whether to replace all or a part of the initial sample selected to represent the population for subsequent surveys.

6.9.1 Replacement of Sampling Units

(1) Fixed Sampling Units (Panel Method)

If the main emphasis in the surveys is to estimate change over time (i.e., trends), it is best to use a fixed sample since there will generally be a high positive correlation between observations on the same sampling unit on successive occasions. If there is no correlation over time, then at least partial replacement of sampling units is preferred.

In using a fixed sample, there are disadvantages which develop after several periods due to non-sampling error problems which arise because of: (a) respondent fatigue due to repeated requests for information resulting in some sampling units not cooperating and the sample becoming unrepresentative, (b) sampling units may be changed by repeated requests for survey information. That is, the respondents may decide they know what information is wanted, and provide data which is different than that being requested; or, the sampling units may change their character because they are being "observed" or become "conscious of their practices" if they are required to participate for too many surveys.

However, there are certain cost advantages which result on the second and subsequent visits due to knowing the location of the sampling units and when to find the respondents at home.

(2) Complete Replacement

This implies an independently selected sample of units on each survey occasion. The correlations for characteristics over time are expected to be low between the observation on the same units on successive occasions because the data relate to different time periods.

In using independent samples, we are generally interested in combining of the characteristic(s) over two or more successive periods. That is, the first survey might conceivably obtain information on the first planting of a crop while the second survey would obtain data relating to a second planting of the crop where under favorable climatic conditions there are two (or more) distinct crop plantings and harvests during a 12-month period. The two surveys would be designed to measure the total production for the entire year.

The disadvantages over time of a fixed sample in terms of non-sampling errors which are related to the respondent are eliminated by the selection of an independent sample each time. However, the costs are also greater when using complete replacements of sampling units due to (a) selection of new units, and (b) locating and enumerating of new units for the first time.

(3) Partial Replacement

Part of the sample is retained, and remainder is replaced for each survey. This type of periodic survey has the advantages of the fixed sample for measuring change and those of the completely replaced sample in estimating the mean relating to the current or most recent survey. If costs of replacement are ignored, the extent of replacement is dependent on the correlation between successive surveys for the same characteristic since the variance is not expected to change. If $\rho = .5$ or larger for a characteristic, than less than 50 percent should be retained where the best estimate is desired for the current survey. Since most surveys have many content items, an iterative or trial-error solution must be sought to optimize the fraction retained for all content items in the survey. However, the fraction retained typically varies between one-fourth to one-half of the previous survey.

6.9.2 Some Useful Estimators for Means (or Totals)

The estimator considered will depend on whether the main purpose is to (a) estimate the change over the time period between surveys, or (b) estimate a combined total or mean for several time periods covered in the surveys, or (c) make the best possible estimate for the last or current survey. These estimation problems will be discussed in terms of two periodic surveys where the two successive surveys being considered might be 6, 12 or 24 months apart and relate to reported data for a similar period of time.

(1) Best Linear Unbiased Current Estimator

A random subsample $m = n\lambda$ units is retained for use on the second occasion and with another independent random sample $\ell = n-m = n\mu$ which is not match with the units in the first survey. λ and μ are the fractions retained and replaced, respectively. Consequently, we have two independent estimates of the current mean (i.e., second survey). The first estimate, $\bar{y}_d$, is based on the difference estimator and $\bar{y}_\ell$ is the simple mean of the new units. In general, the variate of interest will be assumed to have the same variance on both occasions for simplicity though this is not necessary. The variances of the two means are:

$$V(\bar{y}_d) = \frac{S^2}{n\lambda}\ [1 + (1-\lambda)(1-2\rho)] \quad \text{and}$$

$$V(\bar{y}_\ell) = \frac{S^2}{n\mu}\ , \text{ where } S^2 \text{ is the "pooled" variance from the two surveys.}$$

By weighting the two estimates inversely to their variances, we obtain $\bar{y}_n$ and its variance is:

$$V(\bar{y}_n) = \frac{S^2}{n}\ [1 + (1-2\rho)\mu][1 + (1-2\rho)\mu^2]^{-1}$$

which is minimized by taking derivative with respect to μ and solving the resulting equation set equal to zero; that is:

$$\mu = \frac{1}{1 + \sqrt{2}\sqrt{1-\rho}} \quad \text{for which } V_{Min}(\bar{y}_n) = \frac{S^2}{n}\left(\frac{1}{2} + \sqrt{\frac{1-\rho}{2}}\right) .$$

For making current estimates, it is best to replace the sample partially and use the difference estimator if $\rho > \frac{1}{2}$.

However, there exists a minimum-variance unbiased estimator for large populations which can be derived based on general estimation theory in terms of the means for the match and unmatched portions of the sample. This estimator for a characteristic appearing in both surveys can be shown to be

$$\bar{y} = \frac{1}{1-\rho^2\mu^2}\ [\lambda\mu\rho(\bar{X}_1 - \bar{X}_2) + \lambda\bar{y}_2 + \mu(1-\rho^2\mu)\bar{y}_1]$$

and $V(\bar{y}) = \frac{1-\rho^2\mu}{1-\rho^2\mu^2} \cdot \frac{\sigma^2}{n}$, (σ^2 is assumed constant between surveys)

where:

$\bar{X}_1$ = mean of units appearing only in first survey (unmatched units)

$\bar{X}_2$ = mean of units appearing in first survey which can be matched on second survey (matched units)

$\bar{y}_1$ = mean of units appearing only in second survey (unmatched units)

$\bar{y}_2$ = mean of units appearing in second survey which can be matched with first survey (matched units)

(2) Estimation of Change

If the interest centers on estimating the rate of change in the mean value (or estimated total), we consider the estimator based on the mean on each occasion.

$$\hat{R} = \frac{\bar{y} - \bar{x}}{\bar{x}} \quad \text{and the approximate variance is}$$

$$V(\hat{R}) = \{V(\bar{y}) + (1+R)^2 V(\bar{x}) - 2(1+R)\frac{m}{n}\,\mathrm{Cov}(\bar{y},\bar{x})\} \div \bar{x}^2 \ .$$

If we are interested in an unbiased estimate of the absolute change, we estimate (or revise) the characteristic for the first occasion based on the means (or estimated totals), $\bar{X}_\lambda$, based on the difference estimator for the matched portion and $\bar{X}_\mu$ for the unmatched portion using the minimum-variance estimator discussed above.

$$\bar{x} = \frac{1}{1-\rho^2\mu^2}\left[\rho\lambda\mu(\bar{y}_1-\bar{y}_2) + \lambda\bar{X}_2 + \mu(1-\rho^2\mu)\,\bar{X}_1\right]$$

Or, the difference D between surveys is

$$D = \bar{x} - \bar{y} = \frac{1}{1-\mu\rho}\left[\mu(1-\rho)(\bar{y}_1-\bar{x}_1) + \lambda(\bar{y}_2-\bar{x}_2)\right]$$

and

$$V(D) = \frac{2(1-\rho)}{n(1-\mu\rho)}\,\sigma^2 \quad (\sigma^2 \text{ is assumed constant between surveys})$$

(3) Estimation of the Combined Mean (Or Estimated Total) for Two Periods

The minimum-variance estimator for the sum of the two occasions is

$$S = \bar{x} + \bar{y} = \frac{1}{1+\rho\mu}\left[\mu(1+\rho)(\bar{y}_1+\bar{x}_1) + \lambda(\bar{y}_2+\bar{x}_2)\right]$$

and

$$V(S) = \frac{2(1+\rho)\sigma^2}{n(1+\mu\rho)} \quad (\sigma^2 \text{ is assumed constant between surveys})$$

Chapter VII. Use of Several Frames in Sampling

7.0 Introduction

In this chapter, we introduce a general methodology for "multiple frame surveys." The need for several frames arises because: (1) the individual frames do not completely cover all the units in the population but collectively the frames do include all the population units of interest, or (2) even though all the units in the population of interest are covered by a single frame, the use of several frames leads to smaller expected sampling errors per dollar spent. In either case, the use of several frames results in some units being included in more than one frame. For these subdivisions or domains of the population, two or more estimators of the same parameter are available. The material covered in this chapter deals with the general theory of utilizing any number of frames with and without prior knowledge as to the extent of their mutual overlap. The technique of domain estimation described in Section 5.7 is employed. The "overlap domain(s)" provide estimates of the same parameter which arise from each frame; consequently, it is necessary to test the reasonableness of the assumption that the sample estimates of the parameter have the same value before "pooling" the estimates. In the event the assumption of equality of the parameter is rejected, the sample data does not suggest which frame should be used to obtain the estimate of the parameter. This decision must be based on other statistical considerations.

Aside from the theoretical considerations of sampling, multiple frame surveys are more difficult to execute operationally and require more controls to avoid non-sampling errors becoming an important source in sample surveys. This is a direct result of each frame consisting of different types of listing units. In addition, the sampling units in each frame may differ even though both frames contain the same elementary units. Alternatively, the elementary units themselves may differ from one frame to another. Thus, operationally the survey may include two frames with different types of listing units, two different types of sampling units, two different types of elementary units, two different procedures for associating the population of interest with the sampling units, and the necessity of identifying all units or multiples of units which are in two or more frames in the sample.

7.1 Two Frame Surveys

The technique to be employed is that of domain estimation which was discussed in Section 5.7. One of the first published results in the agricultural field was a 1956 poultry study conducted in Maryland. One frame was the area frame consisting of segments of land with which operators of layer flocks were associated and the second frame consisted of a list of operators with 3000 layers or more whose eggs had been graded. This was a two frame survey in which the area sample contained all operators of flocks residing in Maryland (i.e., 100 percent coverage) and the list consisted of all prior known operators residing in Maryland with 3000 layers or more. In other fields of application, the availability of a complete frame may occur less frequently.

7.1.1 Two Frame Methodology

Consider two frames A and B and assume that a sample has been drawn from each frame. The samples may be entirely different in the two frames but the following assumptions are made:

(1) Every unit in the population of interest belongs to at least one of the frames.

(2) It is possible to record for each sampled unit in each frame whether or not it belongs to the other frame.

This means we can divide the units of the sample into three $(2^2 - 1)$ domains.

Domain (a) The unit belongs to Frame A only

Domain (b) The unit belongs to Frame B only

Domain (ab) The unit belongs to both frames

The units in the population are also conceptually divided into the above domains.

7.1.2 Notation for Two-Frame Surveys

There are four different situations concerning our state of knowledge of the total number of units in the frame and in the domains and of our ability to allocate prescribed sample sizes to the domains. We consider only cases 1, 2, and 3 in the discussion. In Case 4, the sample sizes are random variable since the number of units in the frames are unknown. Unless otherwise stated, the type of elementary unit is the same in both frames.

Table 1 Notation

	Frame		Domain		
	A	B	a	b	ab
Population number	N_A	N_B	N_a	N_b	N_{ab}
Sample size	n_A	n_B	n_a	n_b	n_{ab} & n_{ba}
Population total	Y_A	Y_B	Y_a	Y_b	Y_{ab}
Population mean	$\bar{Y}_A$	$\bar{Y}_B$	$\bar{Y}_a$	$\bar{Y}_b$	$\bar{Y}_{ab}$
Sample total	y_A	y_B	y_a	y_b	y_{ab} & y_{ba}
Sample mean	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_a$	$\bar{y}_b$	$\bar{y}_{ab}$ & $\bar{y}_{ba}$
Cost of sampling unit	C_A	C_B			

Random samples are drawn from each frame and n_{ab} and n_{ba} are the subsamples of n_A and n_B respectively which fall into the overlap domain ab where the first letter a or b indicates the frame from which the sample was drawn. The means $\bar{y}_{ab}$ and $\bar{y}_{ba}$ can be computed only if $n_{ab}>0$ and $n_{ba}>0$.

Table 2 Four Cases of Prior Knowledge

Case	Knowledge of population numbers in domains and frames	Possibility of fixed sample allocations to domains and frames	Nature of Domains
1	$N_A, N_B, N_a, N_b, N_{ab}$ known	It is feasible to allocate sample sizes to domains	Domains ≡ Strata
2	$N_A, N_B, N_a, N_b, N_{ab}$ known	It is not feasible to allocate sample sizes to domains	Domains ≡ post-strata
3	Only N_A and N_B known	Sample sizes can only be allocated to frames	Domains ≡ domains proper
4	Neither domain sizes nor frame sizes known	Sampling rates only can be allocated to frames	Domains ≡ domains in populations of unknown sizes

7.1.3 Estimation of Population Totals and Means

In Case 1 the estimation problem is reduced to the standard methodology for stratified sampling covered in Chapter V. For Cases 2 and 3 two approaches leading to identical formula are possible: (a) the theory of domain estimation, or (b) the method of weight variables. For (b) we introduce the following attributes to units in the two frames:

$$\text{Frame A} \quad y_i' = \begin{cases} y_i \text{ if } i^{th} \text{ unit is in domain a} \\ c_i y_i \text{ if } i^{th} \text{ unit is in domain ab} \end{cases}$$

$$\text{Frame B} \quad y_i' = \begin{cases} y_i \text{ if } i^{th} \text{ unit is in domain b} \\ d_i y_i \text{ if } i^{th} \text{ unit is in domain ab} \end{cases}$$

where c_i and d_i are numbers which satisfy for each unit in domain ab $E(c_i+d_i) = 1$. Therefore, the two frames are to be converted into two mutually exclusive strata of sizes N_a and N_{ab} for Frame A and N_b and N_{ab} for Frame B. That is, we have duplicated the N_{ab} units in both frames. The population total will be equivalent to the single frame total of Y. However, the sample estimator of the total and the variance are easily derived only if c_i and d_i are constants. That is, $c_i = p$ and $d_i = q$ where $p + q = 1$ and are determined independently of the parameter being estimated for unbiasedness. Clearly, the population total is equivalent to the original population total since the $N = N_a + N_{ab} + N_b$ units are now $N_a + 2N_{ab} + N_b$ and the totals are:

$$Y = Y_a + Y_{ab} + Y_b$$

$Y' = Y_a + pY'_{ab} + qY''_{ab} + Y_b$ where there are two independent estimators of Y_{ab} which are combined. This notation can be translated directly into that of Section 5.7 by letting $y_i' \equiv {}_j y_i$ and the count variable being ${}_j\mu_i$ where j correspond to the two strata in each frame.

The standard methodology applicable to the survey designs in Frame A and Frame B are therefore applicable to obtain estimates of the two stratum totals for the variate y_i', their variances and variance estimates. Adding the totals for both frames, we obtain the total for the population of interest. To obtain estimates of the population mean $\bar{Y} = Y/N$ apply these formulas to the count variable μ_i' (or ${}_j\mu_i$) to estimate its total N in the way Y' was estimated.

The estimate of the population total given by Hartley for a characteristic when N_a, N_b and N_{ab} are known is:

$$\hat{Y} = N_a\bar{y}_a + N_{ab}p\bar{y}_{ab} + N_{ab}q\bar{y}_{ba} + N_b\bar{y}_b \ .$$

This estimator is in the form of a post-stratified sampling estimator. If the sample is sufficiently large and the f.p.c. factor is not important, the variance is given by

$$V(\hat{Y}) = \frac{N_A^2}{n_A}\{\sigma_a^2 N_a \sigma_{ab}^2 N_{ab} p^2\} + \frac{N_B^2}{n_B}\{\sigma_b^2 N_b + \sigma_{ab}^2 N_{ab} q^2\}$$

where σ_a^2, σ_b^2 and σ_{ab}^2 are the within post stratum variances.

When N_a, N_b, and N_{ab} are unknown, an estimator given by Lund based on the actual subdivisions n_{ab} and n_{ba} is:

$$\hat{Y} = \frac{N_A}{n_A} n_a \bar{y}_a + [\frac{N_A}{n_A} n_{ab} p + \frac{N_B}{n_B} n_{ba} q] \bar{y}_{ab} + \frac{N_B}{n_B} n_b \bar{y}_b$$

where

$\bar{y}_{ab} = \dfrac{n_{ab}\bar{y}_{ab} + n_{ba}\bar{y}_{ba}}{n_{ab} + n_{ba}}$. The approximate variance where $\alpha = N_{ab}/N_A$ and $\beta = N_{ab}/N_B$ is:

$$V(\hat{Y}) = \frac{N_A^2}{n_A}(1-\alpha)\sigma_a^2 + \frac{N_A N_B}{\alpha n_A + \beta n_B}\sigma_{ab}^2 + \frac{N_B^2}{n_B}(1-\beta)\sigma_b^2 +$$

$$\frac{N_A^2(1-\alpha)\alpha}{n_A}[\bar{Y}_a - p\bar{Y}_{ab}]^2 + \frac{N_B^2(1-\beta)\beta}{n_B}[\bar{Y}_b - q\bar{Y}_{ab}]^2$$

An alternative approach proposed by Fuller and Burmeister uses a multiple regression type estimator for samples selected from two overlapping frames. It is assumed that the sampling is such that unbiased estimators of the item totals and the total number of units in each domain are available as well as the same observational unit being used in each frame. The estimator suggested for the population total of the content item is as follows:

$$\hat{Y} = \hat{Y}_a + \hat{Y}_B + \hat{\beta}_1(\hat{N}_{ab} - \hat{N}_{ba}) + \hat{\beta}_2(\hat{Y}_{ab} - \hat{Y}_{ba}) \text{ where } \hat{Y}_B = \hat{Y}_b + \hat{Y}_{ba}.$$

When Frame B is complete and Frame A incomplete, we do not have domain a, hence the estimator is

$$\hat{Y} = \hat{Y}_B + \hat{\beta}_1(\hat{N}_{ab} - \hat{N}_{ba}) + \hat{\beta}_2(\hat{Y}_{ab} - \hat{Y}_{ba})$$

where

$\hat{Y}_B$ = an unbiased estimator of the total constructed from the sample in Frame B,

$\hat{Y}_{ab}$ = an unbiased estimator of the total of domain "ab" constructed from the sample of Frame A,

$\hat{Y}_{ba}$ = an unbiased estimator of the total of domain "ab" constructed from the sample of Frame B,

$\hat{N}_{ab}$ = an unbiased estimator of the number of observational units in domain "ab" constructed from the sample of Frame A,

$\hat{N}_{ba}$ = an unbiased estimator of the number of observational units in domain "ab" constructed from the sample of Frame B, and

$\hat{N}_{b}$ = an unbiased estimator of the number of observational units in domain "b" constructed from Frame B.

The optimal values of $\hat{\beta}_1$ and $\hat{\beta}_2$ are given by

$$\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \begin{bmatrix}\hat{V}(\hat{N}_{ab}-\hat{N}_{ba}) & \hat{Cov}(\hat{N}_{ab}-\hat{N}_{ba},\hat{Y}_{ab}-\hat{Y}_{ba})\\ \hat{Cov}(\hat{N}_{ab}-\hat{N}_{ba},\hat{Y}_{ab}-\hat{Y}_{ba}) & \hat{V}(\hat{Y}_{ab}-\hat{Y}_{ba})\end{bmatrix}^{-1} \cdot \begin{bmatrix}-\hat{Cov}(\hat{Y}_a,\hat{N}_{ab})+\hat{Cov}(\hat{Y}_b,\hat{N}_{ba})\\ -\hat{Cov}(\hat{Y}_a,\hat{Y}_{ab})+\hat{Cov}(\hat{Y}_b,\hat{Y}_{ba})\end{bmatrix}$$

A consistent estimator of the variance is

$$\hat{V}(\hat{Y}) = \hat{V}(\hat{Y}_a) + \hat{V}(\hat{Y}_b) + \hat{\beta}_1[\hat{Cov}(\hat{Y}_a,\hat{N}_{ab}) - \hat{Cov}(\hat{Y}_b,\hat{N}_{ba})]$$

$$+ \hat{\beta}_2[\hat{Cov}(\hat{Y}_a,\hat{Y}_{ab}) - \hat{Cov}(\hat{Y}_b,\hat{Y}_{ba})] .$$

It is also suggested that if other y characteristics are observed in the survey, it may be possible to further decrease the variance of the estimator by including other unbiased estimators of zero in the regression type equation.

7.1.4 Determination of Fixed Weights (p and q)

The value of p is to be determined independently of the parameter being estimated, $\bar{Y}$ or Y. If the sample sizes n_A and n_B are determined, the value of p might be determined as: $\frac{n_A}{n_A+n_B}$. However, it is possible to contemplate finding the values of n_A, n_B and p that will give a minimum value for the variance whenever the cost is fixed or vice versa. Assuming a simple cost function $C = C_A n_A + C_B n_B$ where C is the total cost of sampling, C_A is the cost of an observation from Frame A and C_B is the cost of an observation from Frame B. After some labor, the optimum value of p was found by Hartley to be one of the solutions of:

$$p^2\rho[\phi_B(1-\beta)+\beta q^2] = q^2[\phi_A(1-\alpha)+\alpha p^2]$$

where

$$\rho = \frac{C_A}{C_B}\ ,\ \phi_B = \frac{\sigma_b^2}{\sigma_{ab}^2}\ ,\ \phi_A = \frac{\sigma_a^2}{\sigma_{ab}^2}\ ,\ \alpha = \frac{N_{ab}}{N_A} \text{ and } \beta = \frac{N_{ab}}{N_B}\ .$$

Once the value of p has been determined, the values of n_A and n_B can be found from

$$\frac{n_A}{N_A} = \Theta\ \{(\sigma_a^2(1-\alpha) + \alpha p^2\sigma_{ab}^2)/C_A\}^{\frac{1}{2}}$$

$$\frac{n_B}{N_B} = \Theta\ \{(\sigma_b^2(1-\beta) + \beta q^2\sigma_{ab}^2)/C_B\}^{\frac{1}{2}}$$

where Θ would be determined by the budget available. The foregoing derivation requires knowledge of the costs, variances, and population domain sizes N_a, N_b, and N_{ab}. An alternate derivation for p due to Lund, when N_{ab} is known, is given by the simpler solution for p by the expression $p_0 = \frac{\alpha n_A}{\alpha n_A + \beta n_B}$. While n_A and n_B can be expressed by the iterative system

$$r_1 = \sqrt{\frac{C_B}{C_A}}\left(\frac{\beta}{\alpha}\right) \text{ and } r_{i+1}^2 = \frac{C_B}{C_A}\left(\frac{\beta}{\alpha}\right)^2\ \frac{(r_i+\frac{\beta}{\alpha})^2(1-\alpha)\sigma_a^2 + r_i^2\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2(1-\beta)\sigma_b^2 + (\frac{\beta}{\alpha})^2\beta\sigma_{ab}^2} \text{ where } r = \frac{n_A}{n_B}\ .$$

Thus, the optimum value for p is the ratio of the expected value of the "overlap domain" size in Frame A with respect to the sum of the expected values of the "overlap domain" in both frames.

When N_a, N_b and N_{ab} are unknown, it is necessary to insert unbiased estimates of these three parameters. The minimization of the variance expression in the middle of Page 5 as a function of p, n_A and n_B subject to the cost equation specifies

$$p_0 = \frac{\frac{N_A(1-\alpha)}{n_A}\bar{Y}_a + \frac{N_B(1-\beta)}{n_B}(\bar{Y}_{ab} - \bar{Y}_b)}{\left[\frac{N_A(1-\alpha)}{n_A} + \frac{N_B(1-\beta)}{n_B}\right]\bar{Y}_{ab}}\ .$$

The sample allocation among the two frames can be expressed by an iterative system

$$r_1 = \sqrt{\frac{C_B}{C_A}} \cdot \left(\frac{\beta}{\alpha}\right)$$

$$r_{i+1}^2 = \frac{C_B}{C_A}\left(\frac{\beta}{\alpha}\right)^2 \left\{ \frac{(1-\alpha)\sigma_a^2 + \dfrac{r_i^2\alpha\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2} + \dfrac{r_i^2\alpha(1-\alpha)(\bar{Y}_a+\bar{Y}_b-\bar{Y}_{ab})^2}{[r_i+\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2}}{(1-\beta)\sigma_b^2 + \dfrac{(\frac{\beta}{\alpha})^2\beta\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2} + \dfrac{[\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2\beta(1-\beta)(\bar{Y}_a+\bar{Y}_b-\bar{Y}_{ab})^2}{[r_i+\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2}} \right\}$$

where $r = \frac{n_A}{n_B}$. Generally only a few iterations are required to obtain r starting from a reasonable "guess" for r_1. The estimator and its variance are not sensitivity to deviations from r_0 (optimum) of 10 percent or less. An estimator of the optimum p(i.e. p_0) from the sample data is:

$$\hat{p} = \frac{\dfrac{N_A n_a}{n_A^2}\bar{y}_a + \dfrac{N_B n_b}{n_B^2}(\bar{y}_{ab}-\bar{y}_b)}{\left(\dfrac{N_A n_a}{n_A^2}+\dfrac{N_B n_b}{n_B^2}\right)\bar{y}_{ab}}$$

But $\hat{p}$ is now a function of several sample statistics which disturbs the unbiasedness of the estimator. However, the degree of bias is considered to be negligible. An alternative estimator of p is available, but requires the parameter σ_a^2, σ_{ab}^2 and σ_b^2. This is the bi-quadratic solution given by Hartley.

7.1.5 Assumption of Equality Means for "Overlap" Domains

In practice, we face the problem of pooling of independent estimates of the parameter Y_{ab} or $\bar{Y}_{ab}$ from different frames. Each estimate is given with its sample size and estimated standard error. Can the estimates be considered as homogeneous? That is, are they estimating the same quantity? Let $n = n_1+\ldots+n_K$ equal the samples corresponding to each frame and denote by π_i the ratio n_i/n. The asymptotic distribution of $\sqrt{n_i}\,(T_i-\Theta_i)$ is $N[0,S_i^2(\Theta_i)]$.

Consider, $H = \sum^K \dfrac{n_i(T_i-\hat{\theta})^2}{S_i^2(T_i)} = n\sum^K \dfrac{\pi_i(T_i-\hat{\theta})^2}{S_i^2(T_i)}$

where T_i is the estimate of the parameter θ from the i^{th} frame, and $\hat{\theta}$ is given by

$$\hat{\theta} = \frac{\sum_{i}^{K} \pi_i T_i}{S_i^2(T_i)} \div \sum_{i}^{K} \frac{\pi_i}{S_i^2(T_i)}$$

H is distributed as x^2 with (K - 1) degrees of freedom as $n \to \infty$.

7.1.6 The Special Case of Frame A With 100 Percent Coverage

If Frame A is complete (covers all the units in the population) then $N_A = N$, $N_{ab} = N_B$, $N_a = N_A - N_B$, $N_b = 0$

so we are in case 2. Since $N_a = N_A - N_B > 0$, Frame B must have fewer units than Frame A.

7.1.7 Different Units in Frames With Overlapping Characteristics

In this case, the elementary units which make up the frame are different. Consider a survey in a city to estimate the total cost expanded on the laundering of clothes; both private households and commercial laundries will have launder items which we refer to as "clothes." A portion of "clothes" belonging to a household may be sent to a laundry and the rest washed in the home. A commercial laundry handles clothes from households and from some "commercial nstitutions" which send all their laundry out. That is, the characteristic pertaining to the elementary unit is partitioned rather than assigning the unit to either domain a, ab, or b. The three domains are: (1) household clothes laundered in the home, (2) household clothes laundered in commercial laundries, and (3) commercial institution clothes laundered in commercial laundries. The characteristic of interest might be dollars spent or pounds of clothes, or both.

For each frame the characteristic of interest is defined as follows:

Frame A $${}_n y_i = \begin{cases} y_i & \text{if the clothes in the } i^{th} \text{ home are laundered in the home } (j^{th} \text{ domain} = a) \\ py_i & \text{if the clothes in the } i^{th} \text{ home laundered in a commercial laundry (i.e., } j^{th} \text{ domain} = ab) \end{cases}$$

Frame B $${}_j y_K = \begin{cases} y_K & \text{if clothes in the } K^{th} \text{ commerical laundry are from commercial institutions } (j^{th} \text{ domain} = b) \\ qy_K & \text{if clothes in the } K^{th} \text{ commercial laundry are from a home } (j^{th} \text{ domain} = ab) \end{cases}$$

The unbiased estimate of the population total is given by

$$\hat{Y} = \frac{N_a}{n_A} \{\Sigma_a^{n_A} y_i + p\Sigma_{ab}^{n_A} y_i\} + \frac{N_B}{n_B} \{\Sigma_b^{n_B} y_K + q\Sigma_{ba}^{n_B} y_K\}$$

$$V(\hat{Y}) = \frac{N_A^2}{n_A} (1 - \frac{n_A}{N_A}) \; S_{j y_i}^2 + \frac{N_B^2}{n_B} (1 - \frac{n_B}{N_B}) \; S_{j y_K}^2$$

and the sample estimator of the variance is a copy of $V(\hat{Y})$.

Another example might be the total costs of veterinary drugs purchased. Drugs are used by farm operators, and institutional farms as well as by licensed veterinarians. Additional frames might need to be considered if costs for nonfarm purchases of veterinary drugs for home pets, riding stables etc., were to be included.

7.2 Surveys With More Than Two Frames

The concepts for two frames can be extended to K-frames. In this section, the methodology is described for K = 3. The number of domains created by K-frames is $2^K - 1$ or $2^3 - 1 = 7$ for three frames. We consider simple random sampling from the three frames. It is necessary to directly estimate only the number of units in the four "overlap domains;" that is; N_{ab}, N_{ac}, N_{bc} and N_{abc}. In many of the applications to date, the main interest has centered on estimating the population size. Examples are the number of animals in a population, the number of housing starts in a month or year, etc. In this latter case, the frames might conceivably be: (1) New applications for gas, (2) new applications for electricity, and (3) building permits issued.

7.2.1 Three Frame Estimators

Using the obvious extension of the notation and procedures of the two frame case, the following estimates of domain sizes are:

$$\hat{N}_{ab} = p_{ab} \frac{N_A}{n_A} n_{ab} + q_{ab} \frac{N_B}{n_B} n_{ba} ,$$

$$\hat{N}_{ac} = {}_{ac} \frac{N_A}{n_A} n_{ac} + q_{ac} \frac{N_C}{n_C} n_{ca} ,$$

$$\hat{N}_{bc} = p_{bc} \frac{N_B}{n_B} n_{bc} + q_{bc} \frac{N_c}{n_c} n_{cb} ,$$

$$\hat{N}_{abc} = P_A \frac{N_A}{n_A} n_{abc} + P_B \frac{N_B}{n_B} n_{bac} + P_C \frac{N_c}{n_c} n_{cab} ,$$

$$\hat{N}_a = N_A - (\hat{N}_{ab} + \hat{N}_{ac} + \hat{N}_{abc}) ,$$

$$\hat{N}_b = N_B - (\hat{N}_{ab} + \hat{N}_{bc} + \hat{N}_{abc}) ,$$

$$\hat{N}_c = N_C - (N_{ac} + \hat{N}_{bc} + \hat{N}_{abc}) , \text{ and}$$

$$\hat{N} = \hat{N}_a + \hat{N}_b + \hat{N}_c + \hat{N}_{ab} + \hat{N}_{ac} + \hat{N}_{bc} + \hat{N}_{abc}$$

where the variances are:

$$V(\hat{N}_{ab}) = p_{ab}^2 \frac{N_A^2}{n_A} \alpha_1(1-\alpha_1) + q_{ab}^2 \frac{N_B^2}{n_B} \alpha_2 (1 - \alpha_2)$$

$$\alpha_1 = \frac{N_{ab}}{N_A} , \quad \alpha_2 = \frac{N_{ab}}{N_B} ;$$

$$V(\hat{N}_{ac}) = p_{ac}^2 \frac{N_A^2}{n_A} \gamma_1(1 - \gamma_1) + q_{ac}^2 \frac{N_c^2}{M_c} \gamma_2(1 - \gamma_2)$$

$$\gamma_1 = \frac{N_{ac}}{N_A} , \gamma_2 = \frac{N_{ac}}{N_C} ;$$

$$V(\hat{N}_{bc}) = P_{bc}^2 \frac{N_B^2}{n_B} \beta_1(1 - \beta_1) + q_{bc}^2 \frac{N_c^2}{n_c} \beta_2(1 - \beta_2)$$

$$\beta_1 = \frac{N_{bc}}{N_B} , \quad \beta_2 = \frac{N_{bc}}{N_c} ;$$

and

$$V(\hat{N}_{abc}) = P_A^2 \frac{N_A^2}{n_A} \delta_1(1 - \delta_1) + P_B^2 \frac{N_B^2}{n_B} \delta_2(1 - \delta_2) + P_C^2 \frac{N_C^2}{n_C} \delta_3(1 - \delta_3)$$

$$\delta_1 = \frac{N_{abc}}{N_A} , \quad \delta_2 = \frac{N_{abc}}{N_B} , \quad \delta_3 = \frac{N_{abc}}{N_C} .$$

The values of the p's that minimize these variances are:

$$P_{ab} = \frac{V(\hat{N}_{ba})}{V(\hat{N}_{ab}) + V(\hat{N}_{ba})} \quad , \quad q_{ab} = 1 - P_{ab}$$

$$P_{ac} = \frac{V(\hat{N}_{ca})}{V(\hat{N}_{ac}) + V(\hat{N}_{ca})} \quad , \quad q_{ac} = 1 - P_{ac}$$

$$P_{bc} = \frac{V(\hat{N}_{cb})}{V(\hat{N}_{bc}) + V(\hat{N}_{cb})} \quad , \quad q_{bc} = 1 - P_{bc}$$

$$P_A = \frac{\frac{1}{V(\hat{N}_{abc})}}{\frac{1}{V(\hat{N}_{abc})} + \frac{1}{V(\hat{N}_{bac})} + \frac{1}{V(\hat{N}_{cab})}}$$

$$P_B = \frac{\frac{1}{V(\hat{N}_{bac})}}{\frac{1}{V(\hat{N}_{abc})} + \frac{1}{V(\hat{N}_{bac})} + \frac{1}{V(\hat{N}_{cab})}}$$

$$P_C = \frac{\frac{1}{V(\hat{N}_{cab})}}{\frac{1}{V(\hat{N}_{abc})} + \frac{1}{V(\hat{N}_{bac})} + \frac{1}{V(\hat{N}_{cab})}}$$

and the variance of N_a (similarly N_b and N_c) by:

$$V(N_a) = \frac{N_A^2}{n_A} \{ P_{ab}^2 \alpha_1(1 - \alpha_1) + P_{ac}^2 \gamma_1(1 - \gamma_1) + P_A^2 \delta_1(1 - \delta_1) -$$

$$2P_{ab} P_{ac} \alpha_1\gamma_1 - 2P_{ab} P_A \alpha_1\delta_1 - 2P_{ac} P_A \gamma_1\delta_1 \} +$$

$$+ \frac{N_B^2}{n_B} \{ q_{ab}^2 \alpha_2(1 - \alpha_2) + P_B \delta_2(1 - \delta_2) - 2 P_B q_{ab} \alpha_2 \delta_2 \}$$

$$+ \frac{N_C^2}{n_C} \{ q_{ac}^2 \gamma_2(1 - \gamma_2) + P_C^2 \delta_3(1 - \delta_3) - 2 P_C q_{ac} \gamma_2 \delta_3 \} .$$

For a characteristic other than the population size, such as value of housing starts, the mean of the characteristic for each domain would need to be determined.

For the total of the domain ab, we have

$$\hat{Y}_{ab} = \hat{N}_{ab} \hat{\bar{Y}}_{ab} \quad \text{where } \hat{\bar{Y}}_{ab} = \frac{n_{ab} \bar{y}_{ab} + n_{ba} \bar{y}_{ba}}{n_{ab} + n_{ba}}$$

and in a similar manner the totals for the other six domains can be obtained.

Hence, for Y we obtain

$$\hat{Y} = \hat{Y}_a + \hat{Y}_b + \hat{Y}_c + \hat{Y}_{ab} + \hat{Y}_{bc} + \hat{Y}_{ac} + \hat{Y}_{abc} .$$

The variance of $\hat{Y}_{ab}$ can be obtained as the variance of a product of two independent quantities $\hat{N}_{ab}$ and $\bar{y}_{ab}$. Hence, the variance of $\hat{Y}$ can be obtained as a sum of the seven variances and their covariances of the linear estimator.

Chapter VIII. Sample Size and Allocation for Surveys

8.0 Introduction

The first question which a statistician is frequently called upon to answer is about the size of the sample. Before this question can be answered, the purpose of the survey, variances, costs, and the desired precision of the estimates of the population parameters must be specified.

The purpose (or purposes) of the survey can have a profound effect on how the sample size question is answered. Most persons who ask the question about sample size cannot be expected to realize the answer will be different depending on the main purpose of the survey. If the main purpose of the survey is to estimate a population parameter with a specified precision, we have the classical problem which all sampling books answer. However, the answer is different when the main purpose of the survey is to compare returns per acre or per establishment for irrigated lands versus non-irrigated lands, or for the yield of a fruit crop grown on the mountain slopes versus fruit grown on the valley floor. The answer to this latter type problem is found in books on experimental design and in some of the newer books on sampling under the topic of analytic surveys or "domain estimation."

The availability of data on costs and variances is necessary if the sample size is to be determined accurately based on sampling theory. Where such data is not available, a preliminary sample is generally recommended for improving the design of the survey when it is important to achieve the desired precision.

The specification of the desired precision is arbitrary since there is generally no means of determining a loss function based on the magnitude of the survey error. However, frequently the choice of estimator used in estimating the population parameters is overlooked in determining the sample size. There are many situations in which the estimator is very important, and the opportunity for consideration of this factor should always be investigated when a preliminary sample is required to obtain estimates of variances and costs.

In the discussion which follows the main emphasis is on the classical sample size problem where the population parameters are to be estimated with a specified precision.

8.1 Single Stage Sample Surveys

The number of population parameters to be estimated determines the ease with which the sample size can be determined. Initially, the precision is usually specified in terms of the margin of error permissible in the estimate of a single survey parameter and the coefficient of confidence with which one wants to make sure that the estimate is within the permissible margin of error. The confidence interval statement for the mean of a quantitative variable is given by the following form:

$$\bar{y}_n - t_{(\alpha,\infty)}\sqrt{\frac{N-n}{Nn}}\ \sigma \leq \bar{Y} \leq \bar{y}_n + t_{(\alpha,\infty)}\sqrt{\frac{N-n}{Nn}}\ \sigma$$

where $t_{(\alpha,\infty)}$ is the value of the normal variate corresponding to the value $1 - \frac{\alpha}{2}$ of the tabled normal probability integral N(0,1), to hold on the average of the mean with a probability $1 - \alpha$. From this statement we can find the sample size "n"

$$n = \frac{\frac{t^2_{(\alpha,\infty)}}{E^2} \cdot \frac{\sigma^2}{\bar{Y}^2}}{1 + \frac{1}{N}\frac{t^2_{(\alpha,\infty)}}{E^2}\ \frac{\sigma^2}{\bar{Y}^2}}$$

where $\sigma/\bar{Y}$ is the population coefficient of variation and E is the margin of error specified as a fraction of the mean. Even when $\sigma/\bar{Y}$ is known, n is underestimated since $t_{(\alpha,\infty)}$ is less than $t_{(\alpha,n-1)}$ to be used in calculating the sample confidence interval. This can be corrected by increasing the calculated "n" by the ratio $t^2_{(\alpha,n-1)}/t^2_{(\alpha,\infty)}$. The correction is not likely to be important unless "n" is small.

When σ is unknown and the margin of error is specified as $E\cdot\bar{Y}$, a preliminary sample of size n_1 for improving the design of the survey is selected and the total sample size n is calculated from the pilot survey by

$$n = \frac{t^2_{(\alpha,n_1-1)}S^2_1}{E^2\bar{Y}^2}$$

where s_1^2 is the variance calculated from the n_1 units and N is assumed to be large. The additional units required to give the desired accuracy is $n-n_1$.

The size of sample required for estimating a population proportion with a specified precision is

$$n = \frac{t^2_{(\alpha,\infty)}q}{E^2P}$$

where P is the population proportion while q = 1-P and E·P is the error permissible when the degree of assurance is 1 - α; N is assumed large and E not too small. The knowledge of P is not as critical here since the sample size may be determined for a range of P values and the largest value of "n" used.

When the number of parameters being estimated is two or more, the sample size needs to be determined for each according to the methods just described. The survey characteristic which requires the largest "n" determines the sample size needed to meet the specified margins of error for all variables.

It will be noted that costs did not directly enter into any of the equations. Where the total survey costs are $C = c_0 + c_1n_1$ and the maximum dollars available C_M is less than C, either the sample size will need to be reduced or the margin of error will need to be increased. If the sample size is to be reduced so the dollars spent will be C_M, then the calculated n will be reduced by the ratio:

$$r = \frac{C_M-c_0}{C-c_0}$$

where C_0 is the overhead cost for the survey and C_1 is the cost incurred in acquiring the information for a selected unit.

If it is planned to compare means of certain subdivisions for the population, a larger sample size will be required. We specify the magnitude of the difference in two means we wish to detect as D,

$$V(\bar{y}_i - \bar{y}_j) \leq D^2$$

To satisfy this requirement, the pair with the largest sample size is used:

$$n \doteq \max_{i,j} \frac{t^2_{(\alpha,\infty)}}{D^2} \left(\frac{\sigma_i^2}{\pi_i} + \frac{\sigma_j^2}{\pi_j}\right)$$ If σ_i and σ_j are not very different, we replace them by a pooled estimate σ^2 and

$$n \doteq \max_{i,j} \frac{t^2_{(\alpha,\infty)}}{D^2} \left(\frac{1}{\pi_i} + \frac{1}{\pi_j}\right)\sigma^2$$

where π_i and π_j are the fraction on the population units in the i^{th} and j^{th} domains.

When the K domains are of equal size

$$n \doteq \frac{2Kt^2_{(\alpha,\infty)}\sigma^2}{D^2}$$

8.2 Stratified Sample Surveys

In stratified sampling the population of N units is divided into nonoverlapping subpopulations of $N_1, N_2, \ldots N_H$ units where $N_1+N_2+\ldots+N_H = N$. These subpopulations are called strata and all must be represented in any sample which is to be representative of the population. consequently, the sample size for each of the strata n_h and the total sample size $\Sigma n_h = n$ are to be determined. We wish to do this in such a way as to either minimize the variance to be used in the confidence interval for a specified cost or to minimize the cost for a specified margin of error. This problem is answered first for a single variable and then for two or more characteristics.

8.2.1 Univariate or Single Parameter Allocation

The cost function most frequently used is

$$\text{Cost} = C_0 + \sum_{h}^{H} C_h n_h \ .$$

where C_0 is the overhead cost and C_h the cost incurred in acquiring the information for a selected unit in the h^{th} strata. First, we seek to minimize the variance of the mean

$$V(\bar{y}) = \sum_{h}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{n_h} \left(\frac{N_h - n_h}{N_h}\right)$$

subject to the restriction

$$C_1 n_1 + C_2 n_2 + \ldots + C_H n_H = C - C_0 = \text{Fixed Cost}$$

Using the calculus method of Lagrange multiplier or the Cauchy-Schwarz inequality, we can obtain a solution for single stage designs within strata. For more complex designs within strata the C-S method cannot be used.

8.2.2 Cauchy-Schwarz Inequalities

These are frequently used in determining optimum allocations and making efficiency comparisons.

(1) $(\sum_{i}^{n} x_i y_i)^2 \leq (\sum_{i}^{n} x_i^2)(\sum_{i}^{n} y_i^2)$ where x_i and y_i are any two sets of real numbers. The equality holds if and only if $x_i = Ky_i$.

(2) A generalization of C-S

Let μ and V be n-vectors of real numbers, then

$(\mu' V)^2 \leq (\mu' M \mu)(V' M^{-1} V)$ where the matrix M is positive definite and has an inverse. The equality holds if and only if $M\mu$ is proportional to V.

(3) Probabilistic Version of C-S

Let μ and V be two random variables, then

$$[E(\mu V)]^2 \leq E(\mu^2) \cdot E(V^2) \ .$$

The equality holds if and only if $\mu = KV$ with probability one.

8.2.3 Application of C-S to Optimum Allocation

The variance formula for the population total can be written as

$$V(\hat{Y}) = \sum_{h} N_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) = \sum_{h} \frac{N_h^2 S_h^2}{n_h} - \sum_{h} N_h S_h^2$$

where the second term on the far right does not involve n_h. Hence, the variance is composed of a constant and a term involving n_h which we wish to find an optimum solution for based on some criterion.

(A) Minimum Variance for Fixed Cost

Using (1) of 8.2.2 and

letting $x_i = \frac{N_h S_h}{\sqrt{n_h}}$, $y_i = \sqrt{c_h n_h}$ we have

$$(\sum_h \frac{N_h S_h}{\sqrt{n_h}} \cdot \sqrt{c_h n_h})^2 \leq (\sum_h \frac{N_h^2 S_h^2}{n_h})(\sum_h c_h n_h)$$

The minimum will be achieved when the equality holds or when $c_h n_h$ is proportional to $\frac{N_h^2 S_h^2}{n_h}$, or, $C_h n_h = \lambda^2 \frac{N_h^2 S_h^2}{n_h}$. Substituting this into the preceding formula for $C_h n_h$ we can verify the equality. Hence, we may write the equality as

$$(\sum_h N_h S_h \sqrt{C_h})^2 = (\sum_h \frac{N_h^2 S_h^2}{n_h})(\sum_h C_h n_h),$$

or

$$\sum_h \frac{N_h^2 S_h^2}{n_h} = \frac{(\Sigma N_h S_h \sqrt{c_h})^2}{\sum_h C_h n_h}$$

To find the proportionality constant λ, we use the cost constraint (dollars available), or

$\Sigma C_h n_h = C - C_0$ and substitute for n_h in the equation above involving λ we have $n_h = \frac{\lambda N_h S_h}{\sqrt{C_h}}$

$$\sum_h C_h \frac{\lambda N_h S_h}{\sqrt{C_h}} = C - C_0$$

Gives:

$$\lambda = \frac{C - C_0}{\sum_h N_h S_h \sqrt{C_h}}$$

Hence, we obtain (ignoring f.p.c.)

$$n_h = \frac{C-C_0}{\sum_h N_h S_h \sqrt{C_h}} \cdot (N_h S_h \div \sqrt{C_h})$$

and $$n = \sum_h n_h = \frac{C-C_0}{\sum_h N_h S_h \sqrt{C_h}} \cdot (\sum_h N_h S_h \div \sqrt{C_h})$$

(B) Minimizing Cost for Fixed Variance

Proceeding as before, but λ is now found by using the constraint which fixed the variances as V_0

$$\lambda = \frac{1}{V_0} \sum_h \frac{N_h S_h}{\sqrt{C_h}}$$

and

$$n_h = \frac{N_h S_h}{\sqrt{C_h}} \sum_h \frac{N_h S_h \div \sqrt{C_h}}{V_0}$$

$$n = \sum_h n_h = \frac{(\sum_h N_h S_h \div \sqrt{C_h})^2}{V_0} \qquad \text{(ignoring f.p.c.)}$$

8.2.4 Application of Calculus to Optimum Allocation

The variance formula for the population mean can be used to obtain the solution. We use the same cost function as before except we let

$$C_1 = C - C_0 = \sum_h C_h n_h .$$

We consider a function based on variance and cost which is applicable to any type of survey design

$$\phi = V(\bar{y}) + \mu C$$

where μ is some constant to be determined from the constraints used in obtaining the optimum solution for n and n_h .

For a stratified random sample, the variance of $\bar{y}$ and cost

$$\phi = \sum_h (\frac{1}{n_i} - \frac{1}{N_i}) (\frac{N_h}{N})^2 S_h^2 + \mu (\sum_h C_h n_h)$$

$$\phi = \sum_h \left(\frac{N_h}{N}\frac{S_h}{\sqrt{n_h}} - \sqrt{\mu C_h n_h}\right)^2 + \text{terms not involving } n_h$$

For fixed cost C_1, the minimum value of ϕ is when the derivative is set equal to zero, or solving

$$\frac{\partial \phi}{\partial n_h} = 0 \text{ for } n_h, \text{ gives } n_h = \frac{N_h S_h}{N\sqrt{\mu C_h}} .$$

To find the exact value of the n_h, we calculate $\frac{1}{\sqrt{\mu}}$ under fixed cost conditions

$$C_1 = \sum_h \frac{N_h S_h}{N\sqrt{\mu C_h}} C_h$$

and

$$\sqrt{\mu} = \left(\sum_h \frac{N_h S_h \sqrt{C_h}}{N}\right) \div C_1$$

Hence

$$n_h = \frac{N_h S_h}{\sqrt{C_h}} \cdot \frac{C_1}{\sum_h \frac{N_h S_h \sqrt{C_h}}{N}}$$

and

$$n = \sum_h n_h = \frac{C_1}{\sum_h \frac{N_h}{N} S_h \sqrt{C_h}} \cdot \left(\sum_h \frac{N_h}{N} S_h \div \sqrt{C_h}\right)$$

For fixed variance, the proportionality constant μ, considering the terms in the variance not involving n_h, is:

$$\sqrt{\mu} = \frac{V_0 + \frac{1}{N}\sum_h \frac{N_h}{N} S_h^2}{\sum_h \frac{N_h}{N} S_h \sqrt{C_h}}$$

Hence

$$n_h = \frac{N_h S_h}{N\sqrt{C_h}} \cdot \frac{\sum_h \frac{N_h}{N} S_h \sqrt{C_h}}{V_0 + \frac{1}{N}\sum_h \frac{N_h}{N} S_h^2}$$

and

$$n = \sum n_h$$

If the costs of obtaining information is constant across strata, i.e. $C_h = \bar{C}$, then we have the Neyman allocation which under a fixed variance constraint gives a total sample size

$$n = \frac{\sum_h \frac{N_h}{N} S_h^2}{V_0 + \frac{1}{N} \sum_h \frac{N_h}{N} S_h^2}$$

In the event the calculated value(s) of some n_h exceeds N_h, we selected all units in the strata and allocate the remaining sample units, $n - N_h$, to the H - 1 strata using the allocation formula. However, the formula for the expected variance must also be modified.

8.3 Multivariate Allocation

While the problem of optimum allocation has a unique analytical solution which is easily obtained for a single parameter, the above approach for surveys with two or more variables, i.e., the need to estimate two or more parameters, is not easily solved analytically. However, several "compromise solutions" have been suggested based on applying the optimum allocation to individual survey parameters for which the individual survey parameters for which the individual n's (and n_h's) have been computed based on the results discussed for a single survey parameter, i.e., mean or total for a specific survey characteristic.

8.3.1 Some Approximate Solutions

(A) Use the optimum allocation for the individual survey characteristic requiring the largest sample size. This method will almost surely not satisfy the individual variance restrictions for all the means unless there are only a few survey items. However, this method does indicate a minimum value or lower bound for the sample size n.

(B) For each strata, choose the maximum n_h obtained from the optimum allocation for each of the survey characteristics (or the maximum Neyman allocation). This method will satisfy all the individual fixed variances restrictions for each mean. The sum of the maximum n_h's provides the maximum value or upper bound for the sample size n. It is somewhat larger than is required.

(C) A third method is to calculate the percent n_h is of n for each of the individual optimum allocations and then average the percentage allocation for each stratum. However, a problem still remains in how to choose n. One procedure is to average the minimum and maximum n obtained in (1) and (2). This method will not necessarily satisfy all the variance restrictions on the means, but will satisfy most of the restrictions. A second procedure is to determine an average cost per sampling unit, i.e., $C_h = \bar{C}$, and use a fixed cost $C - C_0 \div \bar{C}$ to determine n. This procedure will not satisfy all the variance restrictions.

8.3.2 Iterative Solution for Optimum Allocation

While an analytical solution is not available, it is possible by "trial and error" to find a solution for n which will satisfy the variance restrictions at minimum costs. A mathematical programming technique for convex functions will yield a solution since the cost and variance function satisfy the mathematical conditions. We formulate all restrictions on the individual totals and any restrictions we may wish to impose on the n_h's. These restrictions would be as follows:

$$V(Y_j) = \sum_h N_h^2 S_{hj}^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \leq V_j$$

for each of the j characteristics in the survey, and for each strata

$2 \leq n_h \leq N_h$. The last requirement insures that all strata are to be represented and the mean and variance can be estimated. In addition, it insures that the allocation to a stratum does not exceed N_h. We also which to minimize the cost function (i.e., the objective function)

$$C - C_0 = \sum_h C_h n_h .$$

8.3.3 Formulation of Convex Programming Problem

The general convex programming problem may be described as: find the vector X that will

maximize g(X), subject to the

constraints $f_i(X) \leq 0 \quad i = 1, 2, \ldots m.$

where g(X) is concave and the $f_i(X)$ are convex, real-valued functions of the n-vector X for all real X and the functions are differentable. There is no loss of generality in describing the problem as a maximization problem, since maximizing g(X) = -h(X) is equivalent to minimizing h(X). In the current problem we wish to find the vector X, where $X' = (x_1, x_2, \ldots x_H)$ is the vector of sample sizes for the strata (i.e., $n_h = x_h$) that will, minimize the cost

$$h(X) = C_0 + C'X$$

or equivalently

maximize g(X) = -h(X).

In addition, we must satisfy certain constraints

$$\sum_h \frac{a_{hj}}{x_h} - V_j^* \leq V_j \quad j = 1, 2, \ldots J, \text{ plus } X \geq 2$$

and

$$x_h \leq N_h \quad i = 1, 2, \ldots H.$$

Where the strata cost per sampling unit are represented by the vector $d' = (C_1 \ C_2, \ldots C_H)$, and $a_{ij} = (\frac{N_h}{N})^2 S_{hj}^2$ are known constants determined for each characteristic and strata.

The above formulation results in a bounded convex feasible region; the concave function g(X) is also bounded over the feasible region, if fact $g(X) \leq 0$. Now the problem, in the form to which an algorithm of Hartley and Hocking will apply is

$$\text{maximize} \quad x_{H+1}$$

$$\text{subject to } f_h(X) = -x_h + 2 \leq 0 \quad h = 1,2,\ldots H$$

$$f_{H+h}(X) = x_h - N_h \leq 0 \quad h = 1,2,\ldots H$$

$$f_{2H+h}(X) = x_{n+1} - g(X) = x_{n+1} + C_0 + \sum_h C_h x_h \leq 0$$

$$f_{2H+j+1}(X) = \sum_h \frac{a_{hj}}{x_h} - v_j \leq 0 \quad j = 1,2,\ldots J,$$

and

$$v_j = V_j^* + V_j \ .$$

8.4 Multistage Sample Surveys

In the preceding sections, either single-stage sampling of the entire population was employed or was assumed within each of the strata for which an optimum allocation was sought. If the sample mean is estimated using a two-stage design, the variance depends on the distribution of the sample between the two stages. In the solutions for the preceding sections, if a two-stage design had been employed, the number of second-stage units was assumed to be known and fixed so the variance depended only on the number of first-stage units to be selected. We now address ourselves to the problem of how to allocate our sample units between the first and second stage units. To determine this allocation, we require detailed information on variance components and costs.

The units of sampling at the first-stage are assumed to be clusters of equal number of second-stage units(i.e., equal size clusters). The procedure is easily generalized to three or more stages and termed multistage sampling. For two stages, the population is composed of N first stage units each of which have M second stage units. We let n denote the number of first-stage units in the sample and m the number of second-stage units to be drawn from each selected first-stage unit. Further, we suppose that the units at each stage are selected with equal probability. The survey cost and precision will depend on the choice of n and m. If we use a simple cost function:

Cost = $c_2 nm$ where c_2 is cost per secondary unit.

If total cost is fixed, say C_0, then the variance upon replacing m by $m = \frac{C_0}{c_2 n}$ is

$$V(\bar{y}_{nm}) = (\sigma_b^2 - \frac{\sigma_w^2}{M}) \frac{1}{n} - \frac{\sigma_b^2}{N} + \frac{c_2 \sigma_w^2}{C_0}$$

where

σ_b^2 is the variance between first-stage units, and

σ_w^2 is the mean square within first stage units.

This expression is a monotonic function of n that reaches a minimum when n assumes the maximum value and $m = 1$ for $(\sigma_b^2 - \frac{\sigma_w^2}{M}) > 0$; and if $(\sigma_b^2 - \frac{\sigma_w^2}{M}) < 0$ the variance is a minimum when n is a minimum given by $n = C_0/c_2 M$ (i.e., no subsampling).

If we fix the variance desired as V_0, rather than the cost, we have

$$V_0 = \left(\frac{1}{n} - \frac{1}{N}\right) \sigma_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \cdot \frac{\sigma_w^2}{n}$$

which give

$$n = \frac{\sigma_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \sigma_w^2}{V_0 + \frac{\sigma_b^2}{N}}$$

If we substitute this value of n into our cost function, we obtain

$$C = c_2 m \left(\frac{\sigma_b^2 - \frac{\sigma_w^2}{M}}{V_0 + \frac{\sigma_b^2}{N}} \right) + \frac{c_2 \sigma_w^2}{V_0 + \frac{\sigma_b^2}{N}}$$

C attains a minimum when $m = 1$ for $\sigma_b^2 - \frac{\sigma_w^2}{M} > 0$, or when $m = M$ for $\sigma_b^2 - \frac{\sigma_w^2}{M} < 0$.

Next, we examine a more general case based on the cost function $C = C_1 n + c_2 nm$ where c_1 and c_2 represent the respective costs of including first and second stage units.

For $(\sigma_b^2 - \frac{\sigma_w^2}{M}) > 0$, the optimum allocation give m as the positive

integer closest to $\sqrt{\frac{c_1}{c_2} \cdot \frac{\sigma_w^2}{\sigma_b^2 - \frac{\sigma_w^2}{M}}}$ or $\sqrt{\frac{c_1}{c_2} \left(\frac{1}{\rho} - 1\right)}$

where ρ is the intra-class correlation within first-stage units.

For $\sigma_b^2 - \frac{\sigma_w^2}{M} \leq 0$, the value of m for total fixed cost $C_0 > c_1 + c_2 M$, $m = M$ and n is the greatest integer not exceeding $C_0/(c_1 + c_2 M)$; if $C_0 < c_1 + c_2 M$, m is the greatest integer not exceeding $\frac{C_0 - c_1}{c_2}$ and n is 1.

When the primary units vary in size, we have the following costs (based on an average cost per secondary unit):

$$C = C_1 n + C_2 \frac{n}{N} \sum_{i=1}^{N} m_i \text{ , and variance}$$

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}}\right)^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 .$$

We obtain a minimum variance for fixed costs, the number of secondary units m_i is the closest positive integer to

$$\hat{m}_i = \sqrt{\frac{C_1}{C_2 \Delta}} \cdot \frac{M_i}{\bar{M}} S_i$$

where

$$\Delta = S_b^2 - \frac{1}{N\bar{M}} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} S_i^2 \text{ is assumed}$$

positive. Since m_i depends on S_i, some prior knowledge of S_i is required. S_i is frequently related to M_i, possible $S_i^2 = \sqrt{M_i}$. Or, to reduce the dependency of S_i on M_i, try to place first stage units with approximately the same size into the same strata. Then $m_i = KM_i$ where K may be approximated by

$$\hat{K} = \sqrt{\frac{C_1}{C_2} \cdot \frac{1}{\bar{M}^2} \cdot \frac{1-\bar{\rho}}{\bar{\rho}}}$$ where $\bar{\rho}$ is an average intra-class correlation over all units in the stratum.

In the preceding allocation problems, the calculus method of Lagrange multiplier was not always demonstrated. However, this method of minimizing a function $\emptyset$ by adding the cost function multiplied by a proportionality factor μ to the variance of the parameter being estimated provides a general approach for problems of optimum allocation for a single parameter.

The foregoing discussion was based on the assumption the necessary information on costs and variances was available or could be obtained in a pilot survey. Lacking this information, the experience in similar surveys provides the best substitute. In other situations, the expertise

of sampling people in the field can usually provide guidance for the subject matter specialist in arriving at an approximate answer for sample size and allocation. Some knowledge of the general nature of the distribution of the characteristic(s) being estimated is helpful since the mean, variance and range are frequently related to provide a reasonable basis for variance estimation. Likewise, the nature of the cost function may be obtained by having some knowledge of the operating organization and physical dispersion of the universe and frame being employed.

BIBLIOGRAPHY

1. Deming, W.E., Some Theory of Sampling (1950), John Wiley & Sons, N.Y.
2. Deming, W.E., Research in Business Statistics, (1952), John Wiley & Sons, N.Y.
3. Hansen, M.H., Hurwitz, W., & Madow, W.G. (1953), Sampling Theory, Vol. II, John Wiley & Sons.
4. Hendricks, W., Theory of Sampling, Scare Crow Press (1955).
5. Huddleston, H.F., Point Sampling for Potatoes in Colorado's San Luis Valley (1955) ERS - Journal of Agricultural Research.
6. Hartley, H.O., Theory of Advanced Design in Surveys, Lecture Notes (1959) Iowa State University.
7. Des Raj, Sampling Theory, (1963) McGraw-Hill, N.Y.
8. Cochran, W.G., Sampling Techniques, (1963) John Wiley & Sons, N.Y., All Editions.
9. Kish, Leslie, Survey Sampling (1965), John Wiley & Sons, N.Y.
10. Rao, J.N.K., Advanced Sampling Theory, Lecture Notes, Texas A&M University, (1966).
11. Claypool, Hocking, and Huddleston, Optimum Allocation to Strata Using Convex Programming (1966), J.R. Statistical Society.
12. Sukhatme, P.V., Sukhatme, B.V., Sampling Theory of Surveys with Applications (1970), Iowa State University Press, Ames, Iowa, All Editions.
13. Huddleston, H.F., A Training Course for Sampling Concepts in Agricultural Statistics (1976), SRS-USDA, # 21.
14. Huddleston, H.F., Sampling Techniques for Measuring & Forecasting Crop Yields (1978), ERS # 09.
15. Jansen, Raymond, Statistical Survey Sampling, (1978), John Wiley & Sons, N.Y.
16. Hocking, Ron, Analysis of Linear Models (1985), Brooks/Cole.
17. Guide to Small Business Computing, Digital Corp.
18. Huddleston, H.F., (1990), Estatisica Vol. 3, Problems with Area Sampling in Trinidad and Tobago.

* U.S. G.P.O:1990-281-097:40002/NASS