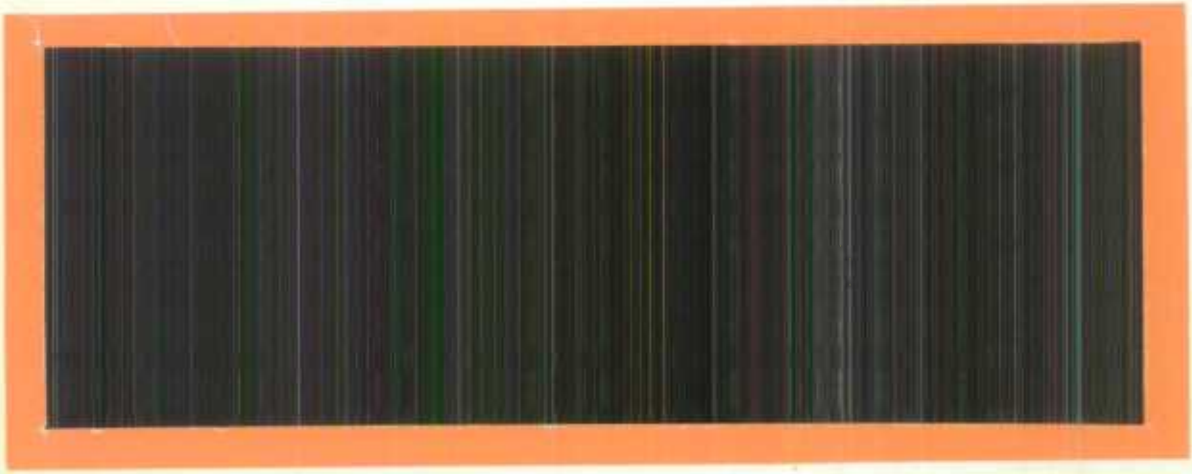
Statistics Canada / Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

Canada

SAMPLE SIZE DETERMINATION AND ALLOCATION FOR BUSINESS SURVEYS

M.A. Hidiroglou

Working Paper No. BSMD 85-005E

Statistics Canada Statistique Canada

Ottawa

SAMPLE SIZE DETERMINATION AND ALLOCATION FOR BUSINESS SURVEYS

M.A. HIDIROGLOU
BUSINESS SURVEY
METHODS DIVISION
JANUARY 1983

TABLE OF CONTENTS

1. Introduction

Stratified sampling involves the division or stratification of a population into relatively homogeneous groups called strata, and the selection of samples is performed independently in each of these strata.

Basically, stratified sampling attempts to restrict the possible samples to those which are 'less extreme' by ensuring that all parts of the population are represented in the sample. It follows that the more homogeneous the groups, the greater the precision of the sample estimates.

The selection of samples in Business Surveys frequently uses simple random sampling techniques applied to strata. The strata are usually based on geography (i.e. provinces and major metropolitan centres), standard industrial classification (i.e. restaurants, agents and brokers, garages, department stores), and some measure of size (i.e. number of employees, gross business income, net sales).

The sample size determination and its subsequent allocation to the strata is then carried out using different criteria. These criteria are, in general, dictated by the existing stratification of the frame and requirements made by the users of the data. These requirements are often in terms of reliability criteria for purposes of tabulation or in terms of affordability of total sample size. These two basic requirements quite often lead to different allocation schemes.

The information required for stratification and sample size determination is obtained from censuses that are periodically carried out or may have to be obtained from existing sample surveys. Stratified sampling is a technique which uses such information in order to increase efficiency.

Stratified sampling is administratively convenient. It can enable a survey organization to control the distribution of fieldwork among its regional offices. Also, for large complex surveys, stratified sampling can facilitate sample design work by enabling such work to be carried out within operationally manageable units.

2. Determining the Sample Size (*)

One of the first considerations in the planning of a sample survey is the size of the sample. Since every survey is different, there can be no hard and fast rules for determining size.

Generally, the factors which decide the scale of the survey operations have to do with cost, time, operational constraints and the desired precision of the results. Once these points have been appraised and individually assessed, the investigators are in a better position to decide the size of the sample.

---

* Extracted from "Survey Sampling, a non-mathematical guide", Federal Statistical Activities Secretariat and Census and Household Survey Methods Division, report written by Alvin Satin and Wilma Shastry.

a. Desired Precision of Sample Estimates

One of the major considerations in deciding sample size has to do with the level of error that one deems tolerable and acceptable.

Measures of sampling error such as standard deviation or coefficient of variation are frequently used to indicate the precision of sample estimates. Since it is desirable to have high levels of precision, it is also desirable to have large sample sizes, since the larger the sample, the more precise estimates will be.

The sample size can be determined by specifying the precision required for each major finding to be produced from the survey, and the level of disaggregation to which the precision must apply.

Often estimates are required not only on a global basis, but for sub-populations as well. Such sub-populations might be defined in terms of employment groups or geographic areas. The sample size falling into each sub-population should be large enough to enable estimates to be produced at specified levels of precision. Sometimes, it will simply cost too much to take the size of sample required to achieve a certain level of precision. In this case, decisions must be made on whether to relax precision levels, reduce data requirements, increase the budget, or find other areas of the survey where cost cutting can be carried out.

b. Factors Which Affect Precision

In any decision related to the precision expected of the sample survey, a number of factors must be taken into account. Such elements as population size, variability of characteristics in the population and the sample plan itself will all affect the precision of the estimates. Consequently, all these factors are identified in the statistical formulae which ultimately relate sample size to the desired level of precision. In the following sub-sections, these factors are considered individually.

i) Size of the Population

Contrary to popular belief, the size of a sample does not increase in proportion to the size of the population. In fact the population size plays only a moderate role as far as medium-sized populations are concerned and an almost non-existent role as far as large populations are concerned.

Consider, for example, a simple random sample of 500 from a population of 200,000. Those 500 units will provide, for most practical purposes, the same precision as a simple random sample of 500 from a population of 10,000.

For very small populations, the relationship is more direct, and often more substantial proportions of the population must be

surveyed in order to achieve the desired precision. In some cases, it is more prudent to consider taking a census rather than a sample.

ii) Variability of Characteristics in the Population

Since the magnitude of differences between members of a population with respect to characteristics of interest is not generally known in advance, it must often be approximated, on the basis of previous surveys or pilot test results.

In general, the greater the difference between population units, the larger the sample size required to achieve specific levels of reliability.

iii) Sample Plan

Many surveys involve moderately complex or very complex sampling and estimation procedures. A more complex design such as stratified multi-stage sampling with ratio estimation can often lead to higher variance in resulting estimates than might a simple random sample design. If, then, the same degree of precision is desired, it is necessary to inflate the sample size to take account of the fact that a simple random sampling is not being used. This is often done by the use of a factor known as a 'design effect' in the calculation of sample size. Design effect refers to the ratio

of the variance of the estimate for a particular design to the variance of the estimate for a simple random sample of the same size. The value of the design effect depends upon the sample plan, as well as the characteristics being measured. It can be estimated from similar past surveys, pilot surveys or using conservative judgement.

iv) Non-Response

Non-response can occur for many reasons. Sometimes, members of a population being surveyed may not be available. Sometimes, they may refuse to answer questionnaires or take part in interviews. It is rare, indeed, when a 100% response rate is achieved. If non-response is not taken into account, the effective number of units in the sample will be smaller than expected. Consequently, the precision of the estimates produced will also be lowered.

To overcome this, the sample size is sometimes inflated at the design stage to account for an anticipated rate of non-response. While this procedure is effective in reducing the variance, it does not reduce the bias resulting from the non-response. In fact, the magnitude of the bias is a function of the size of the non-response and the difference in characteristics between respondents and non-respondents. Since, however, there is a point beyond which non-response cannot be further reduced without an unreasonable expenditure of time and

money, compensation should be considered at the time of estimation (e.g. adjustment of sampling weight of respondents).

Unfortunately, it is not often possible to know in advance what the non-response rate will be. This is especially true of surveys that are breaking new ground. In some instances, the response rate can be estimated with the help of a pilot survey or from past experience with similar surveys.

c. Cost and Time

It is a rare world in which considerations of time and cost are not paramount and most survey takers are not exempt from such restrictions. It almost goes without saying that the time and cost involved have a very definite effect on the size of a sample.

In many studies, funds are allocated and time deadlines set even before the specifics of the study have been decided. It may turn out that the sample size required to implement a survey is larger than existing funds can accommodate. In this case, if more money cannot be found, obviously the sample size must be reduced, thus lowering the precision of the estimates. The same is true for time considerations. If the time allowed is simply not sufficient, the size and scale of the sample may have to be limited to accommodate the deadlines.

d. Operational Constraints

Since surveys require properly trained field staff, coding and editing staff, as well as processing facilities, any limitations on these resources will mean that the size of the sample must be reduced.

In practice, the sample size is evaluated in terms of data requirements, precision, cost, time and operational feasibility. Such an exercise often results in a re-examination and possible modification of the original objectives, data requirements, levels of precision and elements of the survey plan. The survey designer in the process of interrelating such factors will generally attempt to develop a number of feasible design options for consideration and choose the one that best meets all these often conflicting requirements and constraints.

3. Some Notation

In stratified sampling, a finite population of N units is divided into L non-overlapping sub-populations or strata, of size N_1, N_2, ..., N_L, respectively, so that

$$N = N_1 + N_2 + \ldots + N_L.$$

A sample of size n_h ($h = 1, 2, \ldots, L$) is independently drawn from each stratum, so that the total sample size is: $n = n_1 + n_2 + \ldots + n_L$.

For the purpose of this report, simple random sampling without replacement is assumed within each stratum.

Let x_{hi} denote the i-th observation arising from stratum h, so that the true stratum total X_h, the true population total X, and the true stratum variance S_h^2, can be denoted as:

$$X_h = \sum_{i=1}^{N_h} x_{hi}$$

$$X = \sum_{h=1}^{L} X_h$$

$$S_h^2 = \sum_{h=1}^{N_h} (x_{hi} - \bar{X}_h)^2 \; / \; (N_h - 1)$$

with $\bar{X}_h = X_h \; / \; N_h$.

For stratified sampling, the unbiased estimate for the true population total is expressed as

$$\hat{X} = \sum_{h=1}^{L} N_h \; \bar{x}_h$$

where $\bar{x}_h = \sum_{i=1}^{n_h} x_{hi} \; / \; n_h$. The variance of the estimated total is one of the components required for sample size determination and allocation. This variance is of the form

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

$$= \sum_{h=1}^{L} A_h / n_h - D$$

where $A_h = N_h^2 S_h^2$ and $D = \sum_{h=1}^{L} N_h S_h^2$.

In the event that only a sample of the population has been used as a pilot survey, $V(\hat{X})$ will not be available, but an estimate $v(\hat{X})$ based upon a sample of size n_h' can be used as a substitute, where

$$v(\hat{X}) = \sum_{h=1}^{L} A_h' / n_h' - D'$$

with $$A_h' = N_h^2 s_h^2 , \quad D' = \sum_{h=1}^{L} N_h s_h^2$$

$$s_h^2 = \sum_{i=1}^{n_h'} (x_{hi} - \bar{x}_h')^2 / (n_h' - 1)$$

$$\bar{x}_h' = \sum_{i=1}^{n_h'} x_{hi} / n_h' .$$

One of the advantages in conducting a sample survey is that the data can be obtained more quickly and economically than surveying the whole population. A reduction in work load implies a lowering in the cost of conducting a survey and at the same time a lowering of the precision for the required estimates. The allocation problem may be formulated in one of two ways. One may specify a tolerance on the precision of the estimate $\hat{X}$ in terms of a predetermined coefficient of variation, c: that is, find n_h ($h = 1, 2, \ldots, L$) where n_h obeys a given allocation rule in order that

$$\frac{V(\hat{X})}{\hat{X}^2} \leq c .$$

One may also specify that the variance of $\hat{X}$ should be minimal given that the sample size n is fixed.

Cost Considerations also enter into the sample size determination and allocation. In this report, it is assumed that costs for collecting the data from each unit in the population are the same.

In the sections that follow, the terms take-all and take-some stratum will often be referred to. A take-all stratum is a stratum of the population all of the elements of which are included in the sample with probability one: it will be briefly referred to as "TA" and will be enoted with a SUPERSCRIPT "a". The take-all stratum is very necessary in business surveys in order to isolate the large units of highly skewed distributions. The recognition of such large units will save the designer the trouble of later having to adjust for "outliers". The take-some stratum consists of the remaining elements of the population not classified to take-all. These elements will be included in the same with a probability between zero and one. The take-some stratum will be referred to as "TS" and will be denoted with a SUPERSCRIPT "s".

4. Allocation Schemes Criteria

As was mentioned in the introduction, the user may have two ways to allocate a sample to a set of strata. One way is to provide the survey statistician with a fixed sample size n, and the other is to require a given level of precision from the overall sample.

a. Fixed Sample Size n

In this case, the sample size n is to be allocated to the L strata in a specified manner. The portion allocated to the h-th stratum will be denoted as a_h , where $0 \leq a_h \leq 1$ and $\sum_{h=1}^{L} a_h = 1$. Hence, for this case

$$n_h = n\, a_h , \quad h = 1, 2, \ldots, L \tag{4.a.1}$$

b. Fixed Coefficient of Variation for the Sample

In this case, the sample size n is not known and must be computed using the chosen allocation scheme to the strata. Once more, denoting a_h as the portion to be allocated to the h-th stratum, we have that $n_h = n\, a_h$ where n is now unknown.

To solve for n and subsequently n_h, note that

$$c^2 = \frac{V(\hat{X})}{\hat{X}^2}$$

where c is the required coefficient of variation and

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

$$= \sum_{h=1}^{L} A_h / n_h - D \tag{4.b.1}$$

with $A_h = N_h^2 S_h^2$ and $D = \sum_{h=1}^{L} N_h S_h^2$.

Substituting $n_h = n\, a_h$ and $V(\hat{X}) = c^2\, \hat{X}^2$ into equation (4.b.1) one obtains that

$$c^2\, \hat{X}^2 = \sum_{h=1}^{L} A_h / (n\, a_h) - D$$

or solving for n

$$n = \frac{\sum_{h-1}^{L} A_h / a_h}{c^2\, \hat{X}^2 + D} \quad ; \qquad (4.b.2)$$

Hence $$n_h = a_h \frac{\sum_{h=1}^{L} A_h / a_h}{c^2\, \hat{X}^2 + D} \quad . \qquad (4.b.3)$$

5. Allocation Schemes

Equations (4.a.1) and (4.b.3) are the basic tools for allocation of sample size for the two contexts mentioned in 4.a and 4.b. The choice of allocation schemes (choice of a_h for each stratum) may then be classified into two schemes: allocation along a proportionate basis or disproportionate basis.

a. N - proportional Allocation (Proportionate Basis)

This scheme is usually used when information on stratum variances is not available or when one wishes to make the design self-weighting. For this type of allocation,

$$a_h = N_h / N \quad \text{for } h = 1, 2, \ldots, L, \qquad (5.a.1)$$

and $W_h = N_h / n_h$

$= N / n$

$=$ Constant

where W_h is the weight associated with the i-th observation in the h-th stratum. Self-weighting designs on business surveys may be a desirable characteristic to safeguard the estimates against the effect of the movement of units between strata, referred to as stratum jumping.

Proportional allocation will be superior to simple random sampling of the whole population if the strata averages $\bar{X}_h$ differ considerably from each other. If the strata made are such that their means are about the same, stratification (along with proportional allocation) will afford only a slight reduction in the variance.

Disproportionate allocation will be next discussed. The following four schemes are of interest: X - proportional allocation, $\sqrt{X}$ - proportional allocation, $\sqrt{N}$ - proportional allocation and Neyman allocation.

b. X - proportional Allocation

If the measures of size x_{hi} are available for all units in the population, the sample sizes n_h may be found as a proportion of X_h

(aggregate measure of size of stratum h) rather than of N_h.

This allocation will be called X - proportional. In this case

$$a_h = \hat{X}_h \; / \; \hat{X} \quad \text{for} \quad h = 1, 2, \ldots, L, \tag{5.b.1}$$

where $\hat{X}_h$ may be an estimate or the value of X_h and $\hat{X} = \sum_{h=1}^{L} \hat{X}_h$. The variance for $\hat{X}$ in this case is

$$V_X(\hat{X}) = \sum_{h=1}^{L} A_h \frac{1}{X_h \; / \; X} - D$$

for the case that $\hat{X}_h = X_h$ for all strata whereas

for N - proportional allocation it is

$$V_N(\hat{X}) = \sum_{h=1}^{L} A_h \frac{1}{N_h \; / \; N} - D \; .$$

The X - proportional allocation is a very popular allocation scheme to use in business surveys because the distributions associated with such surveys are very skew. Examples are 'employment' in manufacturing industries, and 'retail sales' or 'income' of companies to mention a few examples. The stratum containing the very large units will be found to be many times more variable than other strata. In the case of N - proportional allocation the contribution of this stratum to the total variance will be very considerable. In the case of X - proportional allocation, the factor X_h/X in the denominator will exert a dampening effect on the variance. "X - proportional allocation pushed to the limit" implies that the largest units of a skewed population are sampled with certainty and that the remaining units are sampled. The determination of cutoff points (beyond which to include all units with certainty) have been provided by Glasser (1962) in the case of

fixed sample size n and by Hidiroglou (1979) in the case of predetermined level of reliability c (see Section 7).

c. $\sqrt{N}$ and $\sqrt{X}$ - proportional Allocation

In some cases, the survey data user may be interested in having good reliability attached to the stratum estimates $\hat{X}_h$ as well as to the overall estimate $\hat{X}$. For instance, if strata are provinces, provincial as well as national estimates are considered to be important. Allocation to strata using $\sqrt{N}$ - or $\sqrt{X}$-proportional allocation tends to achieve this goal. The allocation parameter a_h is $\sqrt{N_h} \; / \; \sum_{h=1}^{L} \sqrt{N_h}$ for $\sqrt{N}$ - proportional allocation and $\sqrt{X_h} \; / \; \sum_{h=1}^{L} \sqrt{X_h}$ for $\sqrt{X}$ - proportional allocation. Allocations to strata using these schemes, although not as efficient in terms of minimal overall sample size allocated, are nevertheless serving an objective to provide better estimates at the stratum level. These types of allocation were first proposed by Carroll (1970) and have been used by Bankier (1981) and Tryon (1983).

d. Neyman - allocation

Neyman - allocation or optimal allocation provides an allocation of the total sample size to strata which minimizes the overall variance. For this type of allocation

$$a_h = N_h S_h \; / \; \sum_{h=1}^{L} N_h S_h \; .$$

This type of allocation allocates more sample units to the strata which are the larger ones and/or have the highest variances. In practice, a difficulty with Neyman allocation is that the variances S_h^2 (h = 1, 2, ..., L) may not be known. One way to overcome this limitation is to estimate S_h^2 from a preliminary sample of size n_h' (h = 1, 2, ..., L). Another way is to assume that $S_h / \bar{X}_h$ is constant across strata so that

$$a_h = X_h \Big/ \sum_{h=1}^{L} X_h \quad (X \text{ - proportional allocation}).$$

Another difficulty with using Neyman - allocation is that S_h^2 or any estimate of it may not be stable hence giving rise to unstable samples.

6. Some Special Considerations

a. Non-Response

Non-response, that is the lack of response from some of the units surveyed, has the effect of reducing the effective sample size required to achieve a given level of reliability across all strata. Non-response must therefore be taken into account when designing a survey. If a sample of size n was provided by the user to the survey statistician to collect the data and the response rates were known to be r_h (h = 1, 2, ..., L) within each stratum from previous experience, then the effective sample size would be $n_{EFF} = \sum_{h=1}^{L} n_h (1 - r_h)$ after data collection. If a given level of

reliability were to be achieved and n_h was the required sample size within each stratum required to satisfy the objective, a sample size of $n_h' = n_h / (1 - r_h)$ would have to be selected from each stratum.

b. Overallocation

The formulae for optimum allocation (Neyman), X - proportional or $\sqrt{X}$ allocation may produce sample sizes n_h in some strata that are larger than the corresponding population sizes N_h. If nothing is done, the overall sample size resulting from such overallocation will be smaller than the original sample size. Denote the set of strata where overallocation has taken place as "OVER" and let the remaining set of strata be denoted as "NORM" where "NORM" stands for normal allocation. If a fixed sample size n is used the allocation will be as follows (given that there was overallocation):

$$n_h = N_h \quad \text{for } h \in \text{OVER}$$

$$n_h = (n - n_{OVER}) \; a_h' \quad \text{for } h \in \text{NORM}$$

where a_h' is computed according to the given allocation scheme over the set of strata belonging to NORM and $n_{OVER} = \sum_{h \in OVER} n_h$.

If a fixed coefficient of variation c is used, the allocation is as follows:

i) Determine $n_h = a_h \dfrac{\sum_{h=1}^{L} A_h / a_h}{c^2 \hat{X}^2 + D}$.

ii) Determine the set of strata "OVER" where n_h is greater or equal to N_h.

iii) $n_h = N_h$ for $h \in$ OVER

$$n_h = (n - n_{OVER})\, a_h' \frac{\sum_{h \in NORM} A_h / a_h'}{c^2 \hat{X}^2 + D'}$$

where a_h' is computed according to the given allocation scheme over the set of strata belonging to NORM, $n_{OVER} = \sum_{h \in OVER} n_h$, and $D' = \sum_{h \in NORM} N_h S_h^2$.

c. Minimal Sample Size

A minimal sample size within each stratum is a requirement in order to protect against empty strata occurring as a result of non-response. A minimal size of 3 to 5 units is quite often used in large scale surveys. The minimal sample size within the h-th stratum will be denoted as m_h ($h = 1, 2, \ldots, L$): m_h will most likely be the same across all strata. The minimum size criterion may be applied before or after the given sample size has been allocated. If it is applied before, a sample size m' is initially set aside for minimum size requirements across all strata, where

$m' = \sum_{h=1}^{L} m_h'$ and $m_h' = \min \{N_h, m_h\}$. The remaining sample size $n-m'$ is allocated to the population strata of size $N-m'$ using the chosen allocation method. If the minimum size criterion is applied after, the sample size for the h-th stratum $n_h' = \min \{\max[n_h, m_h], N_h\}$. Note that for this case $\sum_{h=1} n_h'$ may be slightly higher than n;whereas for the other case, there is a better chance that the sum of the selected sample sizes within each strata is exactly equal to the required sample size n.

7. Construction of Self Representing Strata

Stratification of a population into natural strata based on geography and industrial activity is a required step for increasing the efficiency of a sample design. Business surveys are typically characterized with populations whose distribution exhibits a marked positive skewness, with a few large units (accounting for a good portion of the total for the variable of interest) and many small units. Stratification of such population into a take-all stratum (the take-all stratum being the stratum containing the largest units being surveyed with certainty) and a take-some stratum (the take-some stratum being the stratum containing the remaining units being drawn into the sample with a given probability) is highly desirable. Failure to recognize that highly skewed populations should be stratified in the above manner may result in overestimation of the population characteristics to be estimated. Note, that some of the natural strata may be sufficient as take-alls.

The cut-off boundary between the take-all and take-some strata may be derived using one of two criteria. One criterion is to minimize the

variance of the estimate of total given a fixed sample size (Glasser 1962) and the other criterion is to minimize the overall sample size given that a level of reliability has to be achieved (Hidiroglou 1979).

The problem may be defined as follows. Consider a population of size N with units 1, 2, ..., N whose X-characteristic of interest are X_1, X_2, ..., X_N. Define ordered statistics $X_{(1)}$, $X_{(2)}$, ..., $X_{(N)}$ where $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(N)}$. Given that ℓ units are to be included in the take-all stratum, the total X may be broken up into a take-all and take-some portion $X^{(a)}$ and $X^{(s)}$ respectively where

$$X^{(a)} = \sum_{i=N-\ell+1}^{N} X_{(i)} \quad \text{and} \quad X^{(s)} = \sum_{i=1}^{N-\ell} X_{(i)}$$

$$\hat{X} = X^{(a)} + \hat{X}^{(s)} \quad \text{where} \quad X^{(s)} = \frac{N-\ell}{n(\ell)-\ell} \sum_{i=1}^{n(\ell)-\ell} X_{(i)}$$

$$\bar{X}_N + \sqrt{\frac{N}{n}}\, S \ .$$

An estimator of the total (given that the sample has been selected with S.R.S. without replacement in the take-some stratum) would be:

$\hat{X} = X^{(a)} + \hat{X}^{(s)}$ where $\hat{X}^{(s)} = \frac{N-\ell}{n(\ell)-\ell} \sum_{i=1}^{n(\ell)-\ell} X_i$

and $n(\ell)$ is the overall sample size to be allocated. Glasser's rule for determining an optimum cut-off point given a fixed sample size n is to place all units whose x value exceeds $\bar{X}_N + \sqrt{\frac{N}{n}}\, S$ into the take-all stratum where

$\bar{X}_N = \sum_{i=1}^{N} X_i/N$ and $S^2 = \sum_{i=1}^{N} (X_i - \bar{X}_N)^2 / (N-1)$.

Hidiroglou's rule for determining an optimum cut-off point, given that the estimates of total $\hat{X}$ is to have a level of reliability c, is given by putting all units whose x value exceeds $\bar{X}_N + \left[\frac{c^2 X^2}{N} + S^2\right]^{\frac{1}{2}}$ into the take-all stratum.

8. Special Allocations

Two special allocation procedures will be presented for some non-standard requests of reliability criteria.

a. Equalization of the Coefficient of Variation between Strata

A sample of size n (KNOWN) is to be distributed amongst L strata so that the coefficient of variation for the estimates of total between each of the strata is as equal as possible (Hidiroglou, 1983). This type of allocation implies that

$$\frac{V(\hat{X}_1)}{\hat{X}_1^2} = \ldots = \frac{V(\hat{X}_L)}{\hat{X}_L^2} = c$$

where c is not known. The above equation implies that

$$c = \frac{A_h / n_h - D_h}{\hat{X}_h^2}$$

or that $n_h = A_h / (c X_h^2 + D_h)$ where $D_h = N_h S_h^2$.

In order to solve for this coefficient of variation,

solve

$$n = \sum_{h=1}^{L} A_h / (c \hat{X}_h^2 + D_h)$$

for c where everything else is known.

The steps involved in solving the above non-linear equation are as follows:

Letting $f(c) = n - \sum_{h=1}^{L} A_h / (c \hat{X}_h^2 + D_h)$

(i) If $f(0) > 0$, no solution exists,

(ii) If $f(0) \leq 0$ but $f(1) < 0$, no solution exists,

(iii) Define DELTA = 0.000025 (for convergence criterion), and let T1 = 0 and T2 = 1 to initialize the problem,

(iv) Let T = (T1 + T2) / 2,

(v) Find $f(T) = n - \sum_{h=1}^{L} A_h / (c \hat{X}_h^2 + D_h)$,

(vi) If $f(T) = 0$, then let c = T and go to (xii),

(vii) If $f(T) > 0$, then let T2 = T,

(viii) If $f(T) < 0$, then let T1 = T,

(ix) If T1 = 0 or T2 = 1, then go to (iv),

(x) If Abs (T2 - T1) is greater than DELTA, then go to (iv),

(xi) Let c = T2,

(xii) c is the solution with respect to the given DELTA,

(xiii) Compute $n_h = A_h / (c \hat{X}_h^2 + D_h)$ for each h = 1, 2, ..., L.

b. Simultaneous Level of Reliability for Two Stratification Variables

Assume that the population has been stratified using two stratification variables (e.g. provinces and Standard Industrial Classification) and that specified levels of reliability are required for each of the strata of one stratification across the strata of the other stratification (e.g. reliability by province across Standard Industrial Classifications and reliability by Standard Industrial Classification across provinces). Some notation for this problem is now introduced. Assume that there are L strata for stratification variable 1 and M strata for stratification variable 2 yielding LM cross-strata. Subscript h will be used for stratification variable 1 and subscript k will be used for stratification variable 2. For the hk-th cross-stratum let:

$\hat{X}_{hk}$ = estimate of total for the variable of interest,

N_{hk} = number of population elements,

n_{hk} = number of units to be sampled,

S^2_{hk} = variance for the variable of interest.

Furthermore, let

$$N_{h.} = \sum_k N_{hk} \; ; \quad (h = 1, 2, \ldots, L),$$

$$N_{.k} = \sum_h N_{hk} \; ; \quad (k = 1, 2, \ldots, M),$$

$$n_{h.} = \sum_k n_{hk} \; ; \quad (h = 1, 2, \ldots, L),$$

$$n_{.k} = \sum_{h} n_{hk} \; ; \quad (k = 1, 2, \ldots, M),$$

$c_{h,.}$ = required coefficient of variation for the h-th stratum stratified by variable 1,

$c_{.,k}$ = required coefficient of variation for the k-th stratum stratified by variable 2.

The steps required for achieving the required marginal coefficients of variation are as follows:

(i) Compute within coefficients of variation for each stratification variable. That is,

$$cw_{.,k} = c_{.,k} \hat{X}_{.k} \Big/ \left\{ \sum_{h=1}^{L} \hat{X}_{hk}^2 \right\}^{\frac{1}{2}}$$

for $k = 1, 2, \ldots, M$;

and $$cw_{h,.} = c_{h,.} \hat{X}_{h,.} \Big/ \left\{ \sum_{k=1}^{M} \hat{X}_{hk}^2 \right\}^{\frac{1}{2}}$$

for $h = 1, 2, \ldots, L$;

(ii) Compute a compromise first-round coefficient of variation for the (h,k)-th cross-stratum as

$$c_{h,k}^{(o)} = (cW_{h,.} + cW_{.,k}) \; / \; 2$$

where (o) stands for initialization,

(iii) Use the following iterative formula:

$$c_{h,k}^{(r)} = \frac{c_{h,k}^{(r-1)} \; c_{.,k} \; c_{h,.} \; \hat{X}_{h.} \; \hat{X}_{.k}}{\left[\sum_{h=1}^{L} \left\{c_{h,k}^{(r-1)} \; \hat{X}_{h,k}\right\}^2\right]^{\frac{1}{2}} \left[\sum_{k=1}^{M} \left\{c_{h,k}^{(r-1)} \; \hat{X}_{h,k}\right\}^2\right]^{\frac{1}{2}}}$$

where r = 1, 2, ..., 10.

The iterative process revises the coefficients of variation at the cross-stratum level so that they approximate in the best way the marginal required coefficients of variation. In practice, 5 iterations stabilize the $c_{h,k}^{(r)}$ values. The sample size required to achieve the required marginal coefficients of variation for each (h,k)-th cell is then

$$n_{hk} = \frac{A_{hk}}{cf_{h,k}^2 \; \hat{X}_{hk}^2 + D_{hk}}$$

where

$$A_{hk} = N_{hk}^2 \; S_{hk}^2$$

$$D_{hk} = N_{hk} \; S_{hk}^2 \; .$$

and

$cf_{h,k}$ = final iteration coefficient of variation.

This type of allocation scheme was used for the Monthly Restaurants, Caterers and Taverns Survey (Hidiroglou et al 1980).

REFERENCES

Bankier, M. (1981). The Square Root Allocation, Business Survey Methods Division Technical Memorandum, Statistics Canada.

Carroll, J. (1970). Allocation of a Sample Between Status Report Originating from the Australian Bureau of Census and Statistics.

Cochran, W.G. (1963). Sampling Techniques. Second Edition. John Wiley and Sons, New York.

Glasser, G.J. (1962). On the Complete Coverage of Large Units in a Statistical Study. Review of the International Statistical Institute, Vol. 30, p. 28-32.

Hidiroglou, M.A. (1979). On the Inclusion of Large Units in Simple Random Sampling. Proceedings of the American Statistical Association, Survey Research Methods Section, p. 305-308.

Hidiroglou, M.A. , Bennett R., Eady J. and Maisonneuve L. (1980). Sample Design of the Monthly Restaurants, Caterers and Taverns Survey. Survey Methodology Journal, Vol. 6, No. 1, p. 57-83.

Hidiroglou, M.A. (1982). Sample Size Determination and Allocation for the Retail Trade Survey. Business Survey Methods Division report. Statistics Canada.

Hidiroglou, M.A. (1983). Equalization of Coefficients of Variation Between Strata given a fixed sample size for all strata. Business Survey Methods Division Technical Memorandum. Statistics Canada.

Rao, J.N.K. (1975). Analytical Studies of Sample Survey Data. Survey Methodology 1, Supplementary Issue.

Raj, Des (1968). Sampling Theory. Published by Tata McGraw-Hill Publishing Company.

Satin A. and Shastry W. (1981). Survey Sampling, a non-mathematical guide. Federal Statistical Activities Secretariat and Census and Household Survey Methods Division report. Statistics Canada.

Sukhatme, P.V. and Sukhatme B.V. (1970). Sampling Theory of Surveys with Applications. Iowa State University Press.

Tryon, C. (1983). Allocation for the TRACC-2 Survey at the KOB/Province/Room Size Level. Business Survey Methods Division Report. Statistics Canada.

Zirger, B. (1976). The Sampling Allocation Problem - A Review of Literature. Business Survey Methods Division report. Statistics Canada.