STRATIFICATION INDEX: METHODOLOGY & ANALYSIS (CANADIAN LFS)

G.B. Gray, Statistics Canada

Introduction

Large scale surveys, such as the Canadian LFS, the Current Population Survey and many others utilize stratification in setting up the sample frame. This means that, instead of drawing a sample only to represent the whole population, separate independent samples are drawn to represent sub-populations called strata, which may be counties, states, provinces, or cities in the case of area samples (of which the Canadian LFS is a typical example), or lists of establishments in a certain industry group, size groups according to the number of employees or assets in the case of list samples.

To paraphrase Kish [6], there are 3 basic reasons for stratifications and these include:

i) to decrease the variances of the sample estimates,
ii) to employ different methods and procedures within them, and
iii) to employ the sub-populations representing the strata as separate domains of study.

The first reason leads to what may be called optimal stratification whereby the maximum proportion of the MSE between strata is removed so as to minimize the sampling variances within strata. The second and third lead to what one might call administrative stratification whereby special administrative procedures must be undertaken in certain sub-populations.

In this paper, we will concentrate on the methodology in which the sampling variance of characteristic totals (or means) is derived when strata are delineated compared with the sampling variances when they are not delineated.

The Canadian LFS is a typical area sample which has been stratified at several levels: (a) province, (b) type of area within province (NSRU and SRU), (c) economic regions, (d) groups of urban and rural enumeration areas within economic regions, (e) census metropolitan areas and large cities and (f) subunits within met areas and large cities.

Case (d) is the only one that belongs to the optimal stratification category. Case (f) could have been undertaken by optimal stratification but in practice was not because of the time factor and complexities involved. Furthermore, growth and characteristic changes in met areas tend to negate the advantages of optimal stratification while rural areas and small towns tend to remain stable over time where time in the future is of primary importance in continuous surveys.

Whatever the reason for stratification, it is desirable to determine whether or not we are getting our money's worth in effecting a significant reduction in the sampling variance as a result of stratification. A stratification index, measuring the fractional reduction in the variance because of stratification is developed and some empirical examples based on the Canadian LFS (10 months' data from March-December 1975 just prior to redesign) are presented.

Development of Stratification Index

(a) Simple Estimate

$$\hat{X} = \sum_{h=1}^{L} \hat{X}_h, = \sum_{h=1}^{L} \sum_i{}' \frac{1}{n_h p_{i/h}} \hat{X}_{hi}, \text{ where}$$

$\hat{X}_{hi}$ = estimate of characteristic total in psu i, stratum h

n_h = no. of selected psu's out of N_h in stratum h

$p_{i/h}$ = relative size of psu i in stratum h

(b) Variance of Simple Estimate (with stratification)

$$V(\hat{X}) = \sum_{h=1}^{L} V(\hat{X}_h), \text{ where } V(\hat{X}_h) \text{ may be stated in several ways, two of which are:}$$

i) $$V(\hat{X}_h) = \sum_{i<i'} (\pi_{i/h} \pi_{i'/h} - \pi_{ii'/h})(X_{hi}/\pi_{i/h} - X_{hi'}/\pi_{i'/h})^2 + \sum_{i=1}^{N_h} \sigma_{hi}^2/\pi_{i/h}, \text{ where}$$

Yates-Grundy [8]

$\pi_{i/h}$ and $\pi_{ii'/h}$ respectively denote individual inclusion probability of unit h and joint inclusion probability of i and i'. Note that $\pi_{i/h} = n_h p_{i/h}$ and

ii) $$V(X_h) = N_h^2 \sigma_h^2/n_h [1 + (n_h - 1) r_{FP:h}] + \sum_{i=1}^{N_h} \sigma_{hi}^2/n_h p_{i/h}, \text{ where}$$

σ_h^2 = pop'n variance between psu's

$r_{FP:h}$ = finite pop'n correlation.

Here, $$\sigma_h^2 = \sum_i p_{i/h} \left(\frac{X_{hi}}{N_h p_{i/h}} - \frac{X_h}{N_h}\right)^2$$ Gray [3] & Sukhatme [7]

and

$$r_{FP:h} \sigma_h^2 = \frac{1}{n_h(n_h-1)} \sum_{i \neq i'} \pi_{ii'/h} \left(\frac{X_{hi}}{N_h p_{i/h}} - \frac{X_h}{N_h}\right) \left(\frac{X_{hi'}}{N_h p_{i'/h}} - \frac{X_h}{N_h}\right)$$

if sampling without replacement has occurred, and $r_{FP:h} = 0$ if sampling with replacement has occurred.

(c) Ratio Estimate

$$\hat{X} = \sum_{a=1}^{A} P_a (\hat{X}_a/\hat{P}_a) = \sum_{a=1}^{A} P_a \hat{R}_a, \text{ where}$$

P_a = independent source data of category a (eg. age-sex pop'n), and

$\hat{X}_a/\hat{P}_a = \hat{R}_a$ = est'd ratio of characteristic among category a. Then in variance and covariance formulas, X_{hi} may be replaced by

$\sum_a (X_{hia} - R_a P_{hia})$, where $R_a = E\hat{X}_a/E\hat{P}_a$

$= X_a/P_a$

Suppose now that instead of an area delineated into L strata, we have no stratification in the area but simply select $n = \Sigma_{nh}$ psu's with or without replacement with pps and sub-sample within in the same manner.

(d) Then,

$$V_{\bar{S}}(\hat{X}) = N^2\sigma^2/n\cdot[1 + (n-1)r_{FP}] + \sum_{i=1}^{N} \sigma^2_{hi}/np_hp_{i/h}, \text{ where}$$

the same psu delineation is assumed to have occurred and where $n = \sum_h n_h$ psu's have been selected from $N = \sum_h N_h$ psu's with or without replacement.

r_{FP} is the finite population correlation for the sampling without replacement scheme that would have been undertaken over the h strata (eg., systematic pps with units in random order).

If srs is applied $r_{FP} = -1/(N-1)$

If sampling with replacement, $r_{FP} = 0$.

$$\sigma^2 = \sum_h \sum_i p_hp_{i/h} \left(\frac{X_{hi}}{N\ p_hp_{i/h}} - \frac{X}{N}\right)^2 \quad \text{and}$$

$$= \sum_h \sum_i (X_{hi} - X/N)^2 \text{ if srs is applied.}$$

(e) Difference Between $V(\hat{X})$ and $V_{\bar{S}}(\hat{X})$

In order to derive an expression for $V_{\bar{S}}(\hat{X}) - V(\hat{X})$, the relationship between σ^2 and σ^2_h's must be derived.

By an adaptation of Sukhatme [7], it can readily be shown that

$$N^2\sigma^2 = \sum_{h=1}^{L} \frac{1}{p_h} N^2_h\sigma^2_h + L^2\sigma^2_{BS}, \text{ where}$$

σ^2_{BS} = between stratum MSE,

and $L^2\sigma^2_{BS} = \sum_{h=1}^{L} p_h \left(\frac{X_h}{p_h} - X\right)^2$.

By substituting the above relationship, we find that: $V_{\bar{S}}(\hat{X}) - V(\hat{X})$ can be partitioned into 3 distinct components;as follows:

i) $\frac{L^2\sigma^2_{BS}}{n}[1 + (n-1)r_{FP}]$ ... effect of M.S.E. between strata,

ii) $\sum_{h=1}^{L} \left(\frac{n_h}{np_h} - 1\right)V(\hat{X}_h)$ ... effect due to different size strata and/or different no. of selected psu's per stratum,

and

iii) $\sum_{h=1}^{L} \frac{N^2_h\ \sigma^2_h}{n_h}[(n-1)r_{FP} - (n_h-1)r_{FP:h}]$ ... effect of different f.p.c.'s with stratification (would be zero if sampling undertaken with replacement).

The stratification index is defined by $[V_{\bar{S}}(\hat{X}) - V(\hat{X})]/V_{\bar{S}}$ and the main component in the difference is $L^2\sigma^2_{BS}/n$ and we shall assume that $V_{\bar{S}}(\hat{X}) - V(\hat{X}) \doteqdot L^2\sigma^2_{BS}/n$.

(f) Estimates of Variance and Stratification Index

A couple major square deviation expressions are available for estimation purposes

$$\sum_{h=1}^{L} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} \left(\hat{X}_{hi} - \frac{1}{n_h}\hat{X}_h\right)^2 \quad \text{estimates}$$

$$\sum_{h=1}^{L} [N^2_h\sigma^2_h/n_h \cdot(1-r_{FP:h}) + \sum_{i=1}^{N_h} \sigma^2_{hi}/n_hp_{i/h}]$$

instead of the true variance $V(\hat{X})$; however, since $r_{FP:h}$ is usually negative in most ppswor schemes, the expression usually gives a slight over-estimate of the variance.

Another expression which brings in the MSE between strata, vis.,

$$\sum_{h=1}^{L} p_h \left(\frac{\hat{X}_h}{p_h} - \hat{X}\right)^2 \text{ estimates } L^2\sigma^2_{BS} +$$

$$\sum_{h=1}^{L} \left(\frac{1}{p_h} - 1\right)V(\hat{X}_h)$$

so that a slightly biased estimate of $L^2\sigma^2_{BS}$ may be obtained.

The stratification index may be estimated in two stages, as follows:

$$I' = \frac{1}{n} \frac{\Sigma\ p_h \left(\frac{\hat{X}_h}{p_h} - \hat{X}\right)^2 - \sum_{h=1}^{L} \left(\frac{1}{p_h} - 1\right)\hat{V}(\hat{X}_h)}{\hat{V}(\hat{X})}$$

and the stratification index I is finally estimated by:

$\hat{I} = I'\hat{V}(\hat{X})/[\hat{V}(\hat{X}) + I'\hat{V}(\hat{X})] = I'/(1 + I')$, an index which can never exceed one although the estimate $\hat{I}$ could be negative.

A composite index $\hat{\hat{I}}$ over several sub-population domains may be obtained by summing the numerators and denominators of the individual $\hat{I}$'s.

TABLE I: Stratification Indexes Estimated by Levels of Stratification and Sub-populations (Canadian LFS data) and Guide to Tables II to IV

8 Characteristics studied:

Employed (Emp); Unemployed (Unemp); Employed: Agriculture (Emp Ag); Employed Non-agriculture (Non-ag); Employed: Manufacturing (Manuf); Employed: Construction (Constr); Employed: Transportation and Public Utilities (TPU) and Employed: Trade (Trade).

For the above characteristics, the following individual and composite indexes of stratification were calculated for each of 10 months (March - December, 1975) and averaged.

TABLE	SUB-POPULATION	INDEX AND STRATIFICATION	COMPOSITE INDEX
II	Province p, type of area T,	$\hat{I}_{pT1}$, by Economic Region E and deeper strata h's	$\hat{\hat{I}}_{T1}$ (Canada by type of area; over provinces p)
III	Province p, NSRU areas and econ. region E	$\hat{I}_{p2E}$, by strata h's (not included in Table II)	$\hat{\hat{I}}_{p2}$ (province NSRU, over econ. regions E)
IV	Met Area M	$\hat{I}_{M3}$ by sub-units h's within met area M	$\hat{\hat{I}}_{3}$ (for area covered by 10 met areas as indicated in table)

Type of area (self- and non-self representing unit areas or SRU/NSRU)

Observations and Conclusions

The economic regions were used as primary strata across Canada in the NSRU areas while the Census metropolitan areas and large cities were used as primary strata in SRU areas. If an economic region's population was small enough, the NSRU portion of an ER contained one stratum. Otherwise, it was sub-divided into 2 to 5 smaller strata using 3 major employed by industry groups as the stratifying variables. The met areas and large cities were divided into so-called sub-units which are also strata. These were delineated on maps containing blocks and block faces with dwelling counts. Contiguous sub-units were thus formed by drawing boundaries in areas of approximately equal dwelling counts noting the census tract boundaries and the potential for growth. Since optimal stratification was not applied in SRU areas, one would expect smaller reductions in the sampling variance as a result of stratification in SRU areas than as a result of stratification in NSRU areas.

In Table II, stratification first be economic regions within province NSRU areas and second, deeper strata within reduced the variance from that which would occur without stratification within province - NSRU areas between 10% and 44% for 7 of the 8 characteristics, the greatest reductions exceeding 40% for Employed: Agriculture and Employed: Manufacturing. In province SRU areas primary stratification by met areas or large cities and secondary stratification by sub-units within reduced the variance 4 to 17% for the same 7 characteristics from the variance with no stratification. In the same 7 of 8 cases, the exception being Employed: Trade, the summary stratification index for Canada NSRU areas was higher than for Canada SRU areas. At province type of area levels, the index was higher in the NSRU areas than in the SRU areas in 5 to 9 out of 10 provinces, depending upon the characteristic.

In Table III, where each primary stratum was the NSRU portion of each economic region rather than of each province, as in Table II,the Canada summary index was lower than the corresponding index of Table II for 5 out of 8 characteristics and the reduction varied between 2.5% and 26.5% vs. 9.3% and 43.7%, excluding Employed: Trade. The reductions are expected to be lower since Table II results include the effect of stratifying by economic regions as well as within economic regions while Table III includes the effect of only stratifying within economic regions.

Finally, in Table IV, the effect of delineating sub-units within 10 metropolitan areas comprising about 2/3 of the whole SRU area of Canada was considered. The summary indexes over the 10 met areas were not too different from those of Canada SRU in Table II. The comparisons between Tables II and IV are somewhat blurred because Table II considers the whole Canada SRU area while Table IV only the 10 major met areas. Most of the high indexes occurred among Employed: Agriculture which may be concentrated in certain fringe area sub-units. Finally, in Ottawa and Quebec City a high stratification index indicates that Employed: Manufacturing is concentrated in certain districts and hence sub-units of these two cities. Apart from these cases, the variance reductions did not appear substantial.

Conclusion

Stratification resulted in significant reductions in the sampling variance in Employed: Agriculture and Employed: Manufacturing - perhaps 15% to 20% at the primary stratification stage of delineating economic regions (obtained by subtracting $\bar{I}_{T2}$ from $\bar{I}_{T1}$) and another 20% to 25% through deeper stratification within economic regions. In the case of Employed, the comparable results are about 10% and 18% while for Unemployed they are only about 6% and 6%. These percentages are rough estimates because of the relatively few degrees of freedom available to estimate the MSE between strata and perhaps because of non-normal deviations. In the SRU areas the reductions are not nearly so striking.

Even if stratification results in small reductions in the variance, it does not mean that it should not be employed. As pointed out in the introduction, there are other reasons for delineating the universe into strata, not the least of which is to monitor a sample control in compact areas and perhaps redesign the sample in small areas without affecting the universe as a whole.

References and Bibliography

[1] Fellegi, I.P., "Sampling With Varying Probabilities Without Replacement: Rotating and Non-Rotating Samples", Journal of the American Statistical Association; Vol. 58 (1963), pp. 183-201: a description of method used to select psu's in NSRU areas (prior to 1971 redesign).

[2] Fellegi, I.P., Gray, G.B., and Platek, R., "The New Design of the Canadian Labour Force Survey", Journal of the American Statistical Association, Vol. 62 (1967), pp. 421-453: a description of the sample design based on 1961 Census.

[3] Gray, G.B., "Components of Variance Model in Multi-Stage Stratified Samples", Household Surveys Development Staff Survey Methodology, Vol. 1, No. 1 (June 1975), pp. 27-43.

[4] Gray, G.B., "Stratification Index: Methodology and Analysis", submitted to Household Surveys Development Staff Survey Methodology journal for possible publication.

[5] Hansen, M.H., Hurwitz, W.N., and Madow,W.G., Madow, W.G., "Sample Survey Methods and Theory", Vols. 1 and 2, New York: John Wiley & Sons (1953).

[6] Kish, L., "Survey Sampling", New York: John Wiley & Sons (1965), pp. 75-77.

[7] Sukhatme, P.V., "Sampling Theory of Surveys With Applications", The Iowa State University Press, Ames Iowa (1960), pp. 132-137.

[8] Yates, F., and Grundy, P.M., "Selection Without Replacement From Within Strata With Probability Proportional To Size", Journal Of The Royal Statistical Society, Series B, Vol. 15 (1953), pp. 253-261.

TABLE II: $\hat{I}_{pT1}$ by Province and Type of Area $\hat{\bar{I}}_{T1}$ for Canada and Type of Area, by Characteristic (Mar-Dec, 1975 Averages)

	Characteristic							
Sub-pop	Emp	Unemp	Emp Ag	Non-Ag	Manuf	Constr	TPU	Trade
Nfld SRU	.112	.110	.022	.101	.205	.079	.263	.090
Nfld NSRU	.212	.177	.016	.180	.244	.351	-.068	-.081
PEI SRU	-.013	-.006	.079	.007	.054	.032	-.031	-.010
PEI NSRU	.250	.392	-.152	.000	.327	.071	.061	-.229
NS SRU	.146	.030	.056	.196	.172	.032	.123	.085
NS NSRU	.354	.192	.247	.130	.114	.063	.215	-.067
NB SRU	.020	.063	.053	.019	.160	.009	.136	-.004
NB NSRU	.546	.572	.437	.350	.163	.162	.365	-.119
Que SRU	.053	.044	.066	.053	.112	.023	.048	.017
Que NSRU	.336	.217	.415	.246	.534	.055	-.058	-.060
Ont SRU	.067	.060	.146	.076	.225	.071	.075	.022
Ont NSRU	.319	.005	.541	.017	.431	.056	.296	.005
Man SRU	-.014	.024	.069	.002	.117	-.003	.074	.088
Man NSRU	.154	.023	.019	.096	.308	.096	-.214	.313
Sask SRU	-.011	.014	.032	.048	.022	.017	.034	.054
Sask NSRU	.086	-.022	.018	.282	.077	.208	-.097	.139
Alta SRU	-.024	.016	.058	-.001	.041	.024	.006	.046
Alta NSRU	-.044	.144	-.013	.050	.463	.292	.041	-.028
BC SRU	.047	.046	.018	.063	.167	.004	.078	.066
BC NSRU	.114	.018	.507	-.106	-.016	.026	.225	.029
Can SRU	.055	.050	.102	.063	.173	.045	.065	.028
Can NSRU	.287	.129	.437	.135	.429	.093	.142	.000

TABLE III: $\hat{I}_{p2}$ and $\hat{I}_{pT1}$, respectively, by province NSRU and $\bar{I}_2$ and $\bar{I}_{T1}$ respectively for Canada NSRU (Mar-Dec 1975 Averages)

	Characteristic							
Sub-pop	Emp	Unemp	Emp Ag	Non-Ag	Manuf	Constr	TPU	Trade
Nfld	.175	.215	-.063	.105	.167	.124	-.090	.012
	.212	.177	.016	.180	.244	.351	-.068	-.081
PEI	.250	.392	-.152	.000	.327	.071	.061	-.229
	.250	.392	-.152	.000	.327	.071	.061	-.229
NS	.174	.033	.019	.103	.160	.026	.111	.015
	.354	.192	.247	.130	.114	.063	.215	-.067
NB	.355	.283	.149	.449	.164	.191	.300	.053
	.546	.572	.437	.350	.163	.162	.365	-.119
Que	.152	.103	.110	.122	.318	.088	-.071	.156
	.336	.217	.247	.246	.534	.055	-.058	-.060
Ont	.243	.002	.325	.224	.324	-.001	.110	.045
	.319	.005	.541	.017	.431	.056	.296	.005
Man	.196	-.008	.086	.109	-.077	-.055	-.007	.115
	.154	.023	.019	.096	.308	.096	-.214	.313
Sask	.134	.044	.097	.377	.085	.290	-.108	.271
	.086	-.022	.018	.282	.077	.208	-.097	.139
Alta	-.079	.045	.059	.041	.275	.198	.031	.142
	-.044	.144	-.013	.050	.463	.292	.041	-.028
BC	.148	.063	.400	-.062	-.052	.213	-.014	.006
	.114	.018	.507	-.106	-.016	.026	-.225	.029
Can	.181	.064	.224	.168	.265	.095	.025	.104
	.287	.129	.436	.135	.429	.093	.142	.000

TABLE IV: $\hat{I}_{M3}$ by Met Areas and $\hat{\hat{I}}_3$ Over 10 Met Areas, By Characteristic With $\hat{\hat{I}}_{T1}$ For Comparison (Mar-Dec, 1975 Averages)

	Characteristic							
Met Area	Emp	Unemp	Emp Ag	Non-Ag	Manuf	Constr	TPU	Trade
Halifax	.050	-.028	.140	.064	.058	.090	.047	.047
Quebec City	.179	.169	.234	.301	.269	.049	.100	.000
Montreal	.014	.060	.070	.032	.031	.028	.049	.031
Ottawa	.017	.108	.238	.033	.315	-.023	-.032	.002
Toronto	.048	.065	.209	.057	.108	.156	.071	.016
Hamilton	.003	.028	.085	-.002	.150	.036	-.046	.047
Winnipeg	.022	.029	.166	.038	.129	-.029	.061	.073
Calgary	-.040	.039	.204	-.001	.049	.053	-.028	.067
Edmonton	.009	.011	.035	.031	.041	.017	.003	.029
Vancouver	.028	.034	.020	.047	.063	.006	.020	.077
Can SRU $\hat{\hat{I}}_3$	.036	.064	.142	.059	.107	.069	.045	.031
Can SRU $\hat{\hat{I}}_{T1}$	.055	.050	.102	.063	.173	.045	.065	.028