

DISCUSSION OF PAPERS BY KOTT AND ELTINGE

M.A.Hidioglou¹

1. PAPER BY KOTT

The sample design may be described as follows. In the first-phase a stratified PPSWR cluster sample s_1 (with the likelihood of double hits trivially small) was drawn from the universe some *universe* U . The population was split into are H strata in that phase, and n_h clusters are drawn with replacement from each stratum.

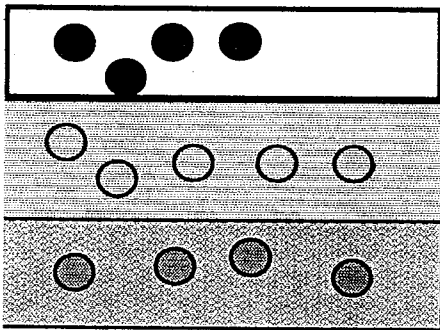


Diagram 1: Sample design for Phase 1

Diagram 1 provides a picture of the sampling process. Arbitrary element sampling occurred in the second phase, and the probability of selecting the same element twice in the second phase was assumed to be so small as to be ignorable. The resulting sample s_1 was further stratified into G strata. It was assumed that model-errors of elements from different first-phase PSU were uncorrelated, implying that the second-phase was an element sample. The empirical work was based on drawing a simple random with a replacement cluster sample in the first-phase followed by a re-stratified SRS element sample in the second.

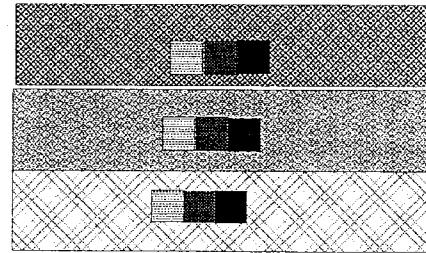


Diagram 2: Sample design for Phase 2

Diagram 2 represents the sampling process. Note that the resulting sample in each second-phase stratum g ($g=1, \dots, G$) is the union of sub-samples drawn from the first-phase stratified sample.

The estimators "t" that were studied included the two-phase regression estimator, and the ratio of two-phase regression estimators. The author derived the randomization-model variance of t , by using a model-based procedure for the second phase, and a randomization-based procedure for the first-phase resulting in

$$v_1(t) = var_1(x_1 \beta) + \sum_h \sum_j var_e(e_{2hj} + [x_1 - x_2] Z^{-1} u_{hj})$$

where $var_1(\bullet)$ is the sample-based (randomization inference of \bullet) with respect to the first phase of sampling, and $var_e(\bullet)$ denotes the model variance of (\bullet). He showed that the jackknife estimator for t

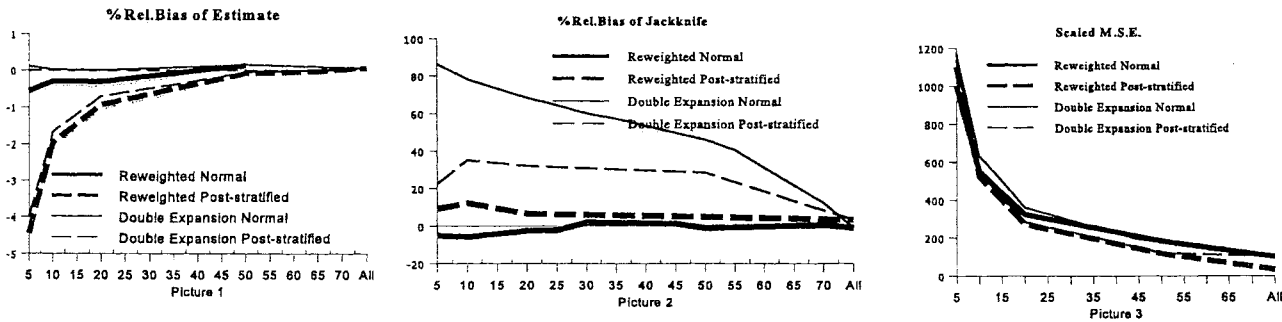
$$v_j(t) = \sum_h \left[\frac{n_h - 1}{n_h} \sum_j (t_{(hj)} - t)^2 \right]$$

was a good estimator for $v_1(t)$: that is $E_e(v_j(t)) = v_1(t) + O(n^{-2})$. The theory showed that the same conclusion held for the ratio of two-phase derived estimators: $\hat{R} = t_1 / t_2$ where y_1, y_2 were regressed on the same x_1 .

¹ Mike Hidioglou, Business Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6

A simulation was carried out using Canadian Labour Force data to investigate the properties of the proposed jackknife procedure. Stratified wtr srs samples (4,000 replicates) of 2 area clusters per stratum (out of 18 strata) were drawn. Individuals within selected clusters were re-stratified into 5 age groups, and a wtr srs sample drawn from each with sizes 50, 20, 10, and 5. A regression estimator for the total t was based on auxiliary data x_i where x_i is a 5-component row vector reflecting the group membership of the i -th individual.

Here, x_1 is a vector of estimates for the age-group sizes based on the first-phase sample. β is a vector of group means estimated based on the second-phase sample. Two estimators were considered: the Double expansion (and the corresponding post-stratified version) and Reweighed expansion (and the corresponding post-stratified version) Estimators. The following variables were investigated: Total employed and Employment Rate = Total employed/Total eligible.



The behaviour of the jackknife procedure for the total employed was summarized via: (A) relative percentage bias of the estimate (Picture 1), (B) relative bias of the jackknife variance (Pictures 2), and (C) scaled mean square error (Picture 3).

The conclusions that may be drawn from the simulation study were as follows:

A. Relative percentage bias of the estimate

The poststratified versions were more negatively biased than their straight counterparts. The bias becomes more pronounced when the second-phase sample sizes are small. The bias of the normal double expansion and reweighed estimators were negligible.

B. Relative percentage bias of the jackknife variance estimate

Better behaviour of the estimator for the Reweighed estimator is better than that of the Double Expansion estimator. There is a better control with poststratification for the Double Expansion estimator. Having no poststratification is best (current one may be too fine) when it comes to variance estimation using the jackknife variance estimator.

C. Mean Square Error

The scaled MSE's were very similar across all estimators considered. The ranking of these estimators is

as follows. The MSE of the double expansion normal estimator is larger than that of the reweighed version. The MSE of the reweighed normal estimators is larger than that of the reweighed poststratified version. The MSE of the reweighed normal estimators is equivalent to that of the double expansion poststratified version.

The jackknife procedure was used to compute the estimated variance. However a Taylor variance procedure could also have been used as opposed to the jackknife. The Taylor method requires fewer computations, and is asymptotically correct. Furthermore, the cumbersome justification required to show that the jackknife procedure is a reasonable approach to use is not required. The derivation of this variance fit in the general frame work of Hidiroglou and Särndal (1995). Table 2 provides available information at the population, and for each level of the multiphase design.

Set of Units	Data available
Population	$\{x_{1k} : k \in U\}$
First-phase sample	$\{(x_{1k}, x_{2k}) : k \in s_1\}$
Second-phase sample	$\{(x_{1k}, x_{2k}) : k \in s_2\}$

Table 2: Available auxiliary information

In the current paper, x_i were post-strata variables and x_2 were counts that were available once the first-phase sample has been drawn. The estimator of total can be written as:

$$t = \sum_{g=1}^G \sum_{s_{2g}} c_{1k} c_{2k} w_{1k} w_{2k} y_k$$

where the calibration factors c_{1k} and c_{2k} incorporate the auxiliary information available for phases one and two respectively.

$$v(t) = \sum_{s_2} \sum_{s_2} w_{2kj} (w_{1k} w_{1j} - w_{1kj}) (c_{1k} e_{1k}) (c_{1j} e_{1j}) + \sum_{s_2} \sum_{s_2} w_{1k} w_{1j} (w_{2k} w_{2j} - w_{2kj}) (c_{2k} e_{2k}) (c_{2j} e_{2j})$$

where w_{1kj} and w_{2kj} are weights that are the inverses of the inclusion probabilities of units k and j . The residuals e_{1k} and e_{2k} are the residuals obtained at the first and second-phase for the implied models that support the estimation.

The above expression can be simplified to single sums. Further details of this variance estimation are given in Binder, Brodeur, Hidiroglou, and Jocelyn (1997, ASA: Los Angeles).

2. PAPER BY ELTINGE

A two-phase design was used to obtain physical measurements. The first phase measured health and nutritional status of "regular" people / US, whereas the second-phase measured error estimates for physical measurement component. Examples of the second type of measurement include low bone mineral density in femur neck, trochanter, and intertrochanter. Two problems were addressed in this paper.

The first one is that the second-phase sample is not a random sample of phase one sample. Given this difficulty, how should the weights be computed? John's solution was to use propensity methods to estimate the probability that a given person would cooperate. The reason for using the propensity procedure is that straight weighting for nonresponse, n_1 / n_2 , could induce bias in the estimates. The ideas of propensity can be traced back to Cochran (1965). Statistics of interest are compared across different populations by splitting each population into subgroups and using linear combinations of the subgroup means. The propensity score $P(r_{hij} = 1 \text{ or } 0)$ is the conditional probability of a particular treatment assignment (responds, does not respond) to a unit given a vector of observed covariates (x). Propensity modelling has several advantageous features. These include (i) a significant potential for reducing bias, as it

Note that c_{1k} is equal to one for the case of no post-stratification. The reweighted expansion estimator is really a separate Hájek-type estimator with the following model:

$$y_k = \beta_g + \varepsilon_k \text{ for } k \in s_{2g}.$$

Note that combined versions of the Hájek ratio estimator are also possible. The estimated variance for t is:

has been shown by Rosenbaum and Rubin (1983); (ii) removing the effects of concomitant variables such as age, sex differences between populations, resulting in "balanced comparisons" and (iii) the potential for modelling the non-response using an appropriate logistic

$$\text{model } \log \left[\frac{p_{hij}}{1 - p_{hij}} \right] = x_{hij} \beta.$$

The second problem is how to reduce large second-phase design component. The question is whether it should be reduced, since the overall CV of the estimated measurement error covariance matrix is quite good. Suggestions include (i) using a Bernoulli second-phase design with constant expected selection probabilities (\bar{p}) implying a negligible loss in efficiency, or (ii) increase the sampling rate at the second phase.

Recommendation (i) can only be carried by stratifying the first phase sample into groups ($g=1, 2, \dots, G$), where (\bar{p}_g) is constant within the groups. However it should be recognized that, even with the application of such a scheme, there will be some non-response within the newly formed groups. The resulting loss should be quite negligible if \bar{p}_g^* is used to account for nonresponse, where \bar{p}_g^* represents the average selection probability after accounting for nonresponse. The stratification of the first-phase sample into groups should be obtained from the propensity analysis resulting from the current study. For instance, these groups could be: Region (North / South); Race (Black / Non-black); Sex (male / females); Age groups (30- / 30+).

$$\text{Assuming that } \hat{V}_2(\hat{\theta} | c=d) = \frac{\hat{V}_2(\hat{\theta} | c=1) \hat{V}_2(\bar{\theta} | c=d)}{\hat{V}_2(\bar{\theta} | c=1)},$$

carrying out of recommendation (ii) would yield the results given in Table 2 for Trochanter, where

$\hat{V}_2(\hat{\theta}|c=d) = \sum_{(hij)} w_{hij} y_{hij}^2 r_{hij}(c) (1 - cp_{hij}) / (c^2 p_{hij}^2)$ and $r_{hij}(c)$ is an independent Bernoulli ($c p_{hij}$) random variable. The greatest gains are made when the second phase sample size is tripled. Usually, this would suggest decreasing the first phase sampling rate and increasing the second phase sampling rate given a fixed budget given that measurement errors are of primary interest.

The sample allocation between the phases will depend on what analyses are important and whether the inclusion of measurement error into these analyses significantly alters the conclusions. Another suggestion, for improving the reliability of phase two estimates, would be to model the measurement error using covariates known at phase 1. This would correspond to using calibration procedures.

Table 2: Impact of increasing sample sizes on second phase variance for the variable Trochanter

Sample Sizes	$\hat{V}(\hat{\theta}) \times 10^9$	$\frac{\hat{V}_2(\hat{\theta})}{\hat{V}(\hat{\theta})}$
1108	0.687	0.856
2266	0.372	0.734
3324	0.267	0.629
4432	0.214	0.537

REFERENCES

- Binder D.A., Brodeur M., Hidiroglou M.A., and Jocelyn W. (1997). Variance Estimation for Two-phase Stratified Sampling. Paper presented at the Annual American Statistical Association meetings in Los Angeles.
- Cochran (1965). The planning of observational studies of human populations (with discussion). *J.R. Statistical Society, A*, **128**, 234-255.
- Hidiroglou, M.A. and Särndal, C.E. (1995). Use of Auxiliary Information for Two-phase Sampling. Proceedings of the Section on Survey Research Methods, *Annual American Statistical Association*, 873-878.
- Rosenbaum P.R. and Rubin D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.