Design Effects for Totals in Multi-stage Samples

Keith Rust[1] and Pamela Broene[1]
[1]Westat, 1600 Research Boulevard, Rockville MD, 20850

Abstract

After the completion of a survey with a given design, or when designing a new survey using available data, often one wishes to develop the design, based on the data collected, so as to increase the efficiency of the design for future use. In multi-stage samples it is common to use the concept of a design effect to summarize the efficiency of a design for a particular survey estimator. For estimators of totals this concept is not straightforward to work with, which results from the fact that for a simple random sample, the basis for comparison when calculating design effects, the population size is known. We present an approach for evaluating potential revisions to a two-stage design, where the sample estimators of primary interest are population and subgroup totals. An example is shown for a design to sample emergency room visits from hospitals. The approach relies on the fact that estimators of population totals and subgroup proportions often have small correlations. We demonstrate that, in our application, this assumption is not always tenable even approximately. Nevertheless the proposed approach does offer the possibility of improvement over the naïve approach of assuming that the design effect for a total is the same as for the corresponding mean value.

Key Words: Sample redesign, establishment survey, occupational injuries and illnesses

1. Introduction

The concept of a design effect in sample surveys is a very useful one. The idea, popularized by Kish (1965), can be used to summarize the relative efficiencies of two or more alternative sample designs. The concept is most useful, however, if it can be, at least approximately, expressed in a functional form that makes clear the respective impacts of the structure of the population and the features of the sample design, because in this case it can be used to guide improvements in design efficiency in future applications.

Consider the simple, well-known, example of the design effect for a mean, $\hat{\bar{Y}}$ from a two-stage design using simple random sampling with replacement at each stage. The design effect for the mean estimator is defined as:

$$Deff\left(\hat{\bar{Y}}\right) = Var\left(\hat{\bar{Y}}\right) \Big/ \left\{S_y^2 / E(n)\right\}$$

Where $Var\left(\hat{\bar{Y}}\right)$ denotes the true sampling variance of $\hat{\bar{Y}}$ under the design used, and the denominator represents the sampling variance of $\hat{\bar{Y}}$ were a simple random sample of the same size to be used. One can define the simple random sampling variance as including a

finite population correction (for simplicity we omit that). In the case of designs with a fixed sample size of elements, $E(n)$ reduces to n.

In the case of a two-stage sample design, that is self-weighting (that is, the population elements have equal probabilities of inclusion in the sample), and with (approximately) equal sample sizes of elements per primary sampling unit (PSU), $\bar{n}$, the design effect for an estimate of a mean can be expressed as:

$$Deff\left(\bar{\bar{Y}}\right) = 1 + (\bar{n} - 1)\rho_y$$

where ρ_y denotes the intraclass correlation of the y. Given a fixed population of PSUs, this formula decomposes the design effect into that component due to the parameters of the design, $\bar{n}$, and that due to the population structure with respect to the y. This permits the designer to consider the effect of changing the design for a future application of the survey. Most obviously it can be used to consider the effects of changing $\bar{n}$, and by considering the costs of sampling PSUs and within PSUs, one can develop an efficient design that produces a relatively low variance estimator for a given cost. The formula can also be used to consider the effect on ρ_y, and therefore the design effect, of changing the definitions of the PSUs, or stratifying them differently, since the formula above extends to stratified two-stage designs if the intraclass correlation is determined within strata.

Although often used in connection with parameters that are not a function of population size, such as means, proportions, and subgroup means, or parameters of models, as Park and Lee (2004), and Särndal, Swensson and Wretman (1992) show, the design effect can be defined for any estimator of interest. For a design p and an estimator, $\hat{\theta}$, the design effect is defined as:

$$Deff_p\left(\hat{\theta}\right) = V_p\left(\hat{\theta}\right) / V_{srs}\left(\hat{\theta}\right)$$

Where $V_p\left(\hat{\theta}\right)$ denotes the sampling variance of $\hat{\theta}$ under the design p; $\hat{\theta}$ denotes the estimator of θ, and $V_{srs}\left(\hat{\theta}\right)$ denotes the variance of $\hat{\theta}$ that would be used with a simple random sample (again, one can argue for the use of either with or with replacement in the definition).

However, difficulties arise when θ denotes the size of the population, N. In this case the simple random sample estimator has zero variance. Yet, the estimation of the population total size can be a legitimate goal for a two-stage sample. One can also consider totals for population characteristics, Y. In this case, the simple random sample variance is, in general, defined. The design effect can be written as:

$$Deff\left(\hat{Y}\right) = Var\left(\hat{Y}\right) \Big/ \left\{N^2 S_y^2 / E(n)\right\}$$

Where S_y^2 denotes the population variance of y.

For such estimators it is of interest to consider whether the design effect can be expressed in way that makes it feasible to consider the effects on the sampling variances for such total estimates of modifying the sample design. The key to such an approach is to recognize that a total Y can be expressed as the product of the mean of y, $\overline{Y}$, and the population size, N. Using a first-order Taylor series approximation one can express the variance of $\hat{Y}$ in terms of the variances of $\hat{\overline{Y}}$ and $\hat{N}$.

2. An Expression for the Design Effect for Totals

Park and Lee (2004) give a formula for the design effect for an estimator of a total using the Horvitz-Thompson estimator for a two-stage design where PSUs are selected with unequal probabilities with replacement, and a simple random sample of fixed size is selected with replacement from each selected PSU. This is equation (4.23) of their paper. It is essentially a re-expression of the first-order Taylor series approximation for the variance of the product of two estimators, as presented, for example, in Hansen, Hurwitz, and Madow (1953). This formula can be re-expressed in the following way:

Let

$$f_y\left(\hat{N}, n\right) = n.\mathrm{Re}\, lVar\left(\hat{N}\right)\overline{Y}^2 \Big/ S_y^2$$

Then

$$Deff\left(\hat{Y}\right) \approx Deff\left(\hat{\overline{Y}}\right) + f_y\left(\hat{N}, n\right) + 2\rho\sqrt{Deff\left(\hat{\overline{Y}}\right) f_y\left(\hat{N}, n\right)}$$

where

$$\rho = Cov\left(\hat{N}, \hat{\overline{Y}}\right) \Big/ \sqrt{Var\left(\hat{N}\right) Var\left(\hat{\overline{Y}}\right)}$$

Note that several terms in the expression are undefined if the total in question is the population size (i.e. $y = 1$ for all population elements). On the other hand, if the sample design is such that the population size N is known, then the last two terms in the expression for the design effect of $\hat{Y}$ reduce to zero, so that the design effect for the total of y is the same as for the mean of y.

However this expression shows that, in general, the design effect for a total for a two-stage sample involves a much more complex relationship between the parameters of the sample design and the structure of the population than is the case for a mean or proportion. In addition to the components of the design effect for a mean, the design effect for a total involves the correlation between the estimate of the mean of the variable of interest and the estimate of the size of the population, the variance of the estimate of the size of the population, and the ratio of the relative variance of the estimate of the size of the population, to the relative variance of the mean of y that would be achieved under

a simple random sample. The expression then combines these quantities in a complex and non-intuitive form.

This means that it is no simple matter to evaluate what modifications to an existing design will improve estimates of population and subgroup totals. In this paper we address that issue, but making a simplifying assumption in a relatively straightforward case. The problem is simplified in the case that we consider because, although the design is a two-stage one, all of the variance in the estimate of the population size comes from the first-stage. That is, although we do not know how many elements are in the population, it is straightforward to obtain the number of elements in each sampled PSU. This means that the variance of the estimate of the population total is only affected by the design and sample size of PSUs, and not the within-PSU design.

3. Application: Study of Underreporting of Occupational Injuries and Illnesses by Workers

The sample design for the proposed Underreporting of Occupational Injuries and Illnesses by Workers study requires sampling Emergency Department (ED) admissions from within a pre-existing sample of hospitals, referred to as NEISS-Work hospitals. These NEISS-Work hospitals were already selected for another, related, survey. The proposed study, to be conducted by the National Institute for Occupational Safety and Health (NIOSH), will involve administering a questionnaire to the sampled patients, to obtain additional information about their admissions (and in particular information about insurance and worker's compensation), as well as basic demographic and other information.

Since certain kinds of workers are of particular interest, the proposed sampling plan is based on taking all eligible self-employed and farm workers into the sample with certainty and subsampling the remaining eligible ED patients in each of the NEISS-Work sample hospitals (excluding Children's hospitals). Each hospital is to be assigned a within-hospital sampling rate for NEISS-Work eligible ED patients based on its stratum (Small, Medium, Large, Very Large). The same sampling rate will initially be assigned to all hospitals in the stratum, based on the rate needed to minimize variation in the final patient weights and obtain the total required sample size. The sampling will be done on a flow basis throughout the year with the interview to follow shortly afterwards.

The sampling frame will consist of all work-related injury/illness ED patient records for persons 18 years of age and above reported at the NEISS-Work sample hospitals over the course of the year. In the 2009 NEISS-Work files, this consisted of 58 hospitals.

The goal when setting sampling rates is to minimize variation in final patient weights and obtain the required total initial sample size, as well as to produce enough cases to make subgroup estimates. Given the importance of the self-employed and farm workers for the underreporting study and their very low prevalence in the sampling frame, all such persons should be taken into the sample with certainty to provide enough cases for producing estimates that meet the NIOSH precision requirement.

4. Design Procedure

For evaluating possible design options we had available complete data for 2009 for the 58 sampled hospitals. The hospitals were sampled with equal probabilities within four strata (Small, Medium, Large, Very Large), which were created based on the number of Emergency Department visits per year. The target population is work-related injuries reporting to a hospital Emergency Department. Using the complete data for one year from these 58 hospitals we wished to evaluate the likely standard errors for different approaches to subsampling patients within hospitals. In particular we wished to estimate what the new standard errors for subgroup totals would be, for various key subgroups. Thus in these cases the y variable reduces to a dichotomous variable, and we can express the population variance for Y as:

$$S_y^2 = \overline{Y}\left(1-\overline{Y}\right)$$

We wish to consider the likely optimal design for a range of total sample sizes, from 1000 to 4000. Keeping in mind both that the number of PSUs in sample is fixed, and that certain cases are included with certainty at the second stage, it can be understood that in this case the design effects are likely to vary considerably across the range of total sample sizes.

5. Simplifying the Design Effect in the case of PSUs of Known Size

In our survey of interest, the total population of admissions to ED departments nationally for work-related injuries is not known. This total is known, however, for each of the 58 hospitals selected as PSUs. This means that the second-stage sample design has no effect on the sampling error of $\hat{N}$. Thus in the expression for the design effect for totals, $\hat{Y}$, given above, we can regard all terms involving the variance of $\hat{N}$ as invariant to the second-stage sample design. Furthermore, given that, we can use the complete data that we have for the 58 PSUs for 2009 to estimates the relative variance of $\hat{N}$.

Considering the expression for the design effect for a total, in this setting we can put

$$f_y\left(\hat{N}, n\right) = nf_y^*,$$

so that

$$Deff\left(\hat{Y}\right) \approx Deff\left(\overset{\wedge}{\overline{Y}}\right) + nf_y^* + 2\rho\sqrt{Deff\left(\overset{\wedge}{\overline{Y}}\right)nf_y^*}$$

We can obtain an estimate of f_y^* for each of the characteristics of interest from the 2009 data. Nevertheless, we are still left with the difficulty of evaluating the correlation between the estimate of the total population size N and the mean of the variable of interest, y. Unfortunately the variance of the estimate of N is invariant to the second stage sample design, in general this will not be the case for ρ. To derive a tractable expression from the equation above we need to determine whether conditions apply that would allow us to treat the third term as negligible with regard to the first two, or at least constant with respect to the second stage design, like the second term.

In practice it might often be reasonable to assume that ρ is close to zero. It is easy to imagine that there will be little relationship between the sampling variance for the estimation of population size, and that of the mean of the characteristic of interest. In our current example, it is variation in sizes of hospitals within strata that generates the sampling variance of the estimate of population size. Thus if the variations across hospitals in the proportions of cases with the particular characteristic of interest, is, within strata, largely unrelated to the size of the hospital, then ρ is likely to be close to zero.

If we can make such an assumption validly, then by obtaining f_y^* from past data, we can adjust estimates of the design effects for the proportion with characteristic of interest, y, by adding to that nf_y^*. That is, we use the approximation:

$$Deff^*\left(\hat{Y}\right) \approx Deff\left(\hat{\bar{Y}}\right) + nf_y^*$$

6. Numerical Example: Study of Underreporting of Occupational Injuries and Illnesses by Workers

A numerical example is shown in Table 1, presenting results for nine subgroups. In each case we estimated the standard error for the subgroup total number of work-related injuries using the estimated standard errors for the overall total and the subgroup proportions. The new design was simulated by sampling patients within hospitals at rates designed to produce 2,000 completed interviews, with minimal variation in the final patient weights. Ten samples were selected and the standard errors, design effects and coefficients of variation were averaged across the ten samples.

This was repeated with sample sizes of 1,000, 3,000 and 4,000, but in essence the results were very similar to those produced for the sample size of 2,000, and so only results for that one sample size are presented in Table 1.

The table shows the estimated proportion for each subgroup, together with its design effect. The next column shows the values of the quantity nf_y^*. The adjacent column to the right shows the sum of design effect of the proportion and nf_y^*; that is, the approximation given in Section 5. This is followed by the true design effect for the total number of injuries, obtained by averaging the estimated sampling variance of the total, across the ten simulations. That is, we did not use the actual variance of the ten estimates of total in each case, as this would have been unstable with only ten simulations.

Next we present a measure of the effectiveness of the proposed approximation: the ratio of the approximate design effect for the total to the actual design effect. The second to right-most column shows the correlation between the respective estimate of proportion and the estimate of the overall total number of injuries. This correlation is assumed to be zero in the approximation given in Section 5. Finally we present a summary measure of the effect of just relying on the design effect for the proportion as a proxy of the design effect for the total, by presenting the ratio of these two quantities.

Table 1: Estimates of Design Effects for Various Population Subgroups; Simulations from Design for NIOSH Study of Underreporting of Occupational Injuries and Illnesses by Workers. Estimated Total Number of Injuries (N) = 1,376,000; variance of N = 2.7 x 1010; n = 2,000

Subgroup	*Proportion* $\left(\hat{\bar{Y}}\right)$	$Deff\left(\hat{\bar{Y}}\right)$	nf_y^*	$Deff^*\left(\hat{Y}\right)$	$Deff\left(\hat{Y}\right)$	$\frac{Deff^*\left(\hat{Y}\right)}{Deff\left(\hat{Y}\right)}$	$\rho\left(\hat{N},\hat{\bar{Y}}\right)$	$\frac{Deff\left(\hat{\bar{Y}}\right)}{Deff\left(\hat{Y}\right)}$
Self-Employed	3%	1.68	0.87	2.55	2.00	1.28	-0.23	0.84
Hispanic	11%	8.36	3.60	11.95	10.79	1.11	-0.11	0.77
Female	36%	1.84	15.75	17.59	15.96	1.10	-0.15	0.12
Govt. Employee	12%	5.08	3.98	9.05	11.15	0.81	+0.23	0.46
Farm Worker	2%	2.36	0.59	2.94	1.72	1.71	-0.52	1.37
Worker's Comp.	64%	23.96	50.74	74.70	110.08	0.68	+0.51	0.22
Self-Insured	18%	15.58	6.19	21.77	10.12	2.15	-0.59	1.54
Employee-Insured	4%	6.48	1.04	7.53	7.72	0.97	+0.04	0.84
Private Insurance	1%	1.26	0.21	1.47	1.20	1.23	-0.26	1.05

The results indicate that the approximation given in Section 5 works quite well for some subgroups, but poorly for others. The reason for this limited effectiveness can be seen in the column showing the correlations between the estimates of proportion and total injuries. These vary considerably across subgroups, both in magnitude and in sign. Thus the idea that the estimates of subgroup proportion and total number of injuries would have little correlation is not borne out. Clearly the estimate of the size of the total number of injuries is a function of the sizes of the hospitals in the sample, since the hospital sample was not selected with probability proportional to size. In fact it would seem that, given that the key estimates of interest from the study are totals, a more efficient first-stage design might be considered in future. However, it seems that certain subgroup proportions are also associated with hospital size. Evidently the proportions female and Hispanic are not highly correlated with hospital size, whereas the proportions of farm workers (with negative correlation), worker's compensation cases (with a positive correlation), and self-insurance (negative correlation) are highly correlated with hospital size. In the case of farm workers at least this could have been anticipated.

Comparing the third to last column of the table with the last indicates that, overall, using the approximation is likely to be better than merely assuming that design effect for the proportion applies to the total. Particularly in the cases of the totals for females, worker's compensation, and government employees the design effect for the proportion is substantially smaller than the design effect for the corresponding total.

7. Summary and Conclusions

This research project was motivated by the realization that, when designing (or redesigning) a study in which a two-stage sample design is used and for which many or most of the estimates of key interest are of subgroup totals, it is likely to be inappropriate make decisions based on the design effects for the corresponding mean values or population proportions. We sought a way to obtain a tractable approximation to the design effect for a total that would enable us to usefully evaluate the sample variance

properties of alternative designs, especially in the case where the first-stage design is fixed and not subject to manipulation or redesign. We did derive an approximation that we felt might be appropriate for the kind of study that we were developing. This approach takes advantage of the fact that the variance of the estimator of the population size is a function only of the first-stage design, if the population size can be determined for each selected primary sampling unit. This situation is likely to be the case in surveys where the first-stage units are establishments or institutions, such as hospitals, schools, or businesses.

However, using a simulation-based approach, we demonstrated that our approximation was not robust, and perhaps the best that could be said for it in our situation is that it was superior to an approach of just assuming that the design effect for a total is the same as for the corresponding mean or proportion. In fact our findings suggest that in general a simulation-based approach is likely to give more realistic results. Of course such an approach is not always possible – we were fortunate enough to have census data available for our sampled primary sampling units that we could use to study different design choices.

It may be that in many other applications the assumption that the estimator of population total has a low correlation with the means and proportions corresponding to the totals of interest in the survey is tenable. Then the proposed approximation is likely to be useful. But our research demonstrates that it would seem that this low correlation needs to be demonstrated, and cannot be taken for granted.

Acknowledgements

We wish to acknowledge Larry Jackson from the National Institute for Occupational Safety and Health for permission to present results using the agency's data, and to him and David Marker for helpful comments on our presentation, and consequently this manuscript.

References

Hansen, M.H., Hurwitz W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*. Vol. I, New York: Wiley.

Kish, L. (1965) *Survey Sampling*. New York: Wiley.

Park, I., and Lee, H. (2004). Design Effects for the Weighted Mean and Total Estimators Under Complex Survey Sampling. *Survey Methodology*, Vol. 30, No. 2, pp. 183-193.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.