Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

ESTIMATING INTERVIEWER EFFECTS FOR BINARY RESPONSES

Alastair Scott[1] and Peter Davis[2]

ABSTRACT

In surveys where interviewers need a high degree of specialist knowledge and training, we are often forced to make do with a small number of highly trained people, each having a high case load. It is well-known that this can lead to interviewer variability having a relatively large impact on the total error, particularly for estimates of simple quantities such as means and proportions. In a previous paper (Davis and Scott (1995)) we looked at the impact for continuous responses using a linear components of variance model. However, most responses in health questionnaires are binary and it is known that this approach results in underestimating the intra-cluster and intra-interviewer correlations for binary responses. In this paper we use a multi-level binary model to explore the impact of interviewer variability on estimated proportions.

KEY WORDS: Interviewer variance; Design effect; Multi-level models.

1. INTRODUCTION

Many surveys, particularly in the health area, require the use of interviewers or observers with a great deal of specialist expertise and training. Since people with the appropriate skills are difficult to find, investigators are often forced to use a relatively small number of interviewers, each having a high case load as a consequence. For example, we are currently involved in a study of adverse events in hospitals in New Zealand. The study uses a two-stage design in which a stratified PPS sample of hospitals is drawn at the first stage and a systematic sample of patient records is drawn from each of the chosen hospitals at the second stage. Each selected record is then reviewed carefully for evidence of any medical errors. This review requires a high level of medical knowledge and an intensive period of training to become familiar with the study protocols. Because it was very difficult to find people with the required expertise who were free to participate during the period of the study, we were forced to make do with a small number of reviewers which led to an average case load of about 300 records. This is fairly typical of such studies.

It is well-known that the impact of a high interviewer load on the variance of estimates of population means can be severe, even when interviewer variability is small (see Groves (1982, chapter 8) or Lessler & Kalsbeek (1992, Section 11.3), for example). The impact on more complex statistics such as regression coefficients or odds ratios tends to be less severe (Kish & Frankel (1974)) but can still be serious in some circumstances. In a previous paper (Davis & Scott (1995), we looked at the impact of interviewer variability on means and, more particularly, on comparisons between domain means using a components of variance model of the type pioneered in the classic paper of Hartley & Rao (1969). We then applied the results to data from a large-scale dental survey where the expert interviewers had high case loads. In our empirical study we largely ignored the fact that many of the variables were binary, apart from the

[1]Alastair Scott, Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1, New Zealand

[2]Peter Davis, Department of Public Health & General Practice, Christchurch School of Medicine, Christchurch, New Zealand

throwaway remark that "It is now well-known that this leads to an under-estimate of the variance components for binary data, so that our estimated design effects should be regarded as lower bounds". Similar comments can be found in Anderson and Aitken (1985). Since binary variables are the norm in most health surveys, it is important to investigate the nature and extent of this under-estimation. This is the subject of this paper. We develop some theory based on a simple normal-logistic model in the next section and look at some empirical results from the adverse events study mentioned above in Section 3. Some tentative conclusions are sketched in the final section. In particular, our remark about the underestimation of design effects turns out to need some qualification.

2. BASIC THEORY

2.1 Simple Random Sampling

For simplicity, we look first at the case of simple random sampling where we can consider interviewer effects without the additional complication of clustering. We start by reviewing standard results for the mean of a continuous variable (see Kish (1962), for example). If we have measurements on a continuous variable, then a natural model for the observation from the *rth* respondent with the *ith* interviewer is the simple linear mixed model

$$Y_{ir} = \mu + a_i + e_{ij} \qquad (1)$$

where a_i denotes the *ith* interviewer effect and e_{ij} denotes the individual respondent effect. Assume that the a_Is and the e_{Ir} s are uncorrelated random variables with zero means and variances σ_a^2 and σ^2 respectively. Then the variance of the sample mean under this model is

$$Var\{\bar{Y}\} = Var\left\{\frac{1}{n}\sum\sum Y_{ir}\right\} = \frac{1}{n^2}\sum_i Var\{n_i a_i + \bar{e}_{i.}\}$$
$$= \frac{1}{n^2}\left\{\left(\sum_i n_i^2\right)\sigma_a^2 + n\sigma^2\right\} = \frac{1}{n}\left(\sigma^2 + \tilde{n}_I\sigma_a^2\right),$$

where n_i is the number of responses handled by the *ith* interviewer and $\tilde{n}_I = \sum n_i^2 \Big/ \sum n_i$. Now consider what would happen if the observations were drawn independently from model (1). The variance of the sample mean would now be

$$Var\{\bar{Y}\} = Var\left\{\frac{1}{n}\sum\sum Y_{ir}\right\} = \frac{1}{n}\left(\sigma^2 + \sigma_a^2\right) = V_0,$$

say. Thus the effect of multiple interviews is to inflate this variance by a factor D_0 given by

$$D_0 = \frac{Var\{\bar{Y}\}}{V_0} = 1 + (\tilde{n} - 1)\rho_I$$

where $\rho_I = \dfrac{\sigma_a^2}{\sigma^2 + \sigma_a^2}$ is the intra-interviewer correlation coefficient. We shall call D_0 the design effect, although it represents the inflation in variance under the random effects model (1) rather than for repeated sampling from a fixed finite population as in the usual definition. We see that the impact of interviewer variability can be high when the average case-load is high, even if the intra-interviewer correlation, ρ_I, is fairly small.

Now consider what happens when the response is binary. Model (1) is no longer appropriate but we can extend it in a natural way by supposing that there is an underlying continuous latent variable, Y* say, obeying model (1). We do not observe Y* directly but only the binary response Y, which takes the value 1 if and only if Y* falls below some (unknown) cut-off value, c say. Thus Y_{ir}, the value for the rth respondent with the ith interviewer, is equal to 1 if and only if

$$Y_{ir}^* - c = \theta + a_i^* + e_{ir}^* \le 0, \qquad (2)$$

where $\theta = \mu^* - c$. Note that model (2) is only defined up to a multiplicative constant since multiplying the left side of (2) by an arbitrary positive constant does not affect its sign and hence does not affect the value of Y_{ir}. This means that we can only estimate the ratio, $\frac{\sigma_a^{*2}}{\sigma^2}$, rather than the individual variance components, σ_a^{*2} and σ^{*2}. Fortunately, most of the quantities of interest (such as ρ_I^*) are functions of this ratio and are thus estimable from the observed binary data, given appropriate deployment of interviewers.

To obtain the value of $P\{Y_{ir}=1\}$ from model (2), we need to specify the distributions of a_i^* and e_{ir}^*. We shall assume that a_i^* is normally distributed and that e_{ir}^* has a logistic distribution. (It would be more in keeping with the spirit of the original Hartley-Rao model to assume that e_{ir}^* is also normally distributed. However, the results are almost identical under either model unless we are well out in the tail of the distributions, and the theory and numerical computations are both somewhat simpler under the logistic assumption.) A good review of software for fitting such a logistic-normal model is given in De Leeuw and Kreft (2001). Since we can rescale arbitrarily without affecting the value of Y_{ir}, we can assume without loss of generality that e_{ir}^* has a *standard* logistic distribution, which has variance $\pi^2/3$. This means that Var{a_i*} in (2) corresponds to $\sigma_a^2 \Big/ \frac{\sigma^2}{\pi^2/3} = \frac{\pi^2}{3}.\frac{\sigma_a^2}{\sigma^2}$ in (1). Thus our estimate of σ_a^{*2} is actually an estimate of $\frac{\pi^2}{3}.\frac{\sigma_a^2}{\sigma^2}$, which leads to $\hat{\rho}_I^* = \frac{\hat{\sigma}_a^{*2}}{\hat{\sigma}_a^{*2} + \pi^2/3}$ as our estimate of ρ_I^*, the intra-interviewer correlation for the underlying latent variable.

We are not primarily interested in interviewer variability for its own sake but rather for the effect that it has on the variance of the estimated population proportion. We can write the estimated proportion, $\hat{p}$, in the form

$$\hat{p} = \frac{1}{n}\sum_i \sum_r Y_{ir} = \frac{1}{n}\sum_i n_i \hat{p}_i,$$

where $\hat{p}_i$ denotes the observed proportion of "Yes" responses among the n_i recorded by the *ith* interviewer. Since the responses for different interviewers are independent, it follows that $Var\{\hat{p}\} = \frac{1}{n^2}\sum_i n_i^2 Var\{\hat{p}_i\}$. Now, from standard results for conditional means and variances, we can write

$$Var\{\hat{p}_i\} = E\left\{Var\{\hat{p}_i \mid a_i^*\}\right\} + Var\left\{E\{\hat{p}_i \mid a_i^*\}\right\}$$

where, from model (2),

$$E\{\hat{p}_i \mid a_i^*\} = P\{Y_{ir} = 1 \mid a_i^*\} = e^{\theta + a_i^*} / (1 + e^{\theta + a_i^*}) = p(a_i^*), \qquad (3)$$

say, and

$$Var\{\hat{p}_i \mid a_i^*\} = \frac{p(a_i^*)(1 - p(a_i^*))}{n_i}.$$

It follows that

$$Var\{\hat{p}\} = \frac{1}{n^2} \sum_i n_i^2 Var\{\hat{p}_i\} = \frac{1}{n}\left[E^* + \tilde{n} V^*\right],$$

where $\tilde{n}_I = \sum_i n_i^2 / n$, as before, $E^* = E^*(\theta, \sigma_a^{*2}) = E\{p(a_i^*)(1 - p(a_i^*))\}$ and $V^* = V^*(\theta, \sigma_a^{*2}) = Var\{p(a_i^*)\}$.

If the observations had been made independently (i.e. with a different interviewer for each respondent), then the variance of the estimated proportion $\hat{p}$ would have been

$$V_0 = \frac{1}{n}\left[E^* + V^*\right].$$

Thus the effect of the interviewers carrying out multiple interviews is to inflate the variance by a factor

$$D_0^* = \frac{Var\{\hat{p}\}}{V_0} = 1 + (\tilde{n}_I - 1)\phi$$

and $\phi = \frac{V^*}{E^* + V^*}$. Exact evaluation of E^* and V^* requires numerical integration, but we can derive approximate values using standard linearization methods when the interviewer effects are small. Expanding $p(a) = e^{\theta + a} / (1 + e^{\theta + a})$ about a=0 gives

$$p(a) \approx p_0 + p_0(1 - p_0)a + \frac{1}{2} p_0(1 - p_0)(1 - 2p_0)a^2$$

where $p_0 = p(0)$. This leads to

$$V^* = Var\{p(a_i^*)\} \approx p_0^2 (1 - p_0)^2 \sigma_a^{*2}$$

and, after some algebraic manipulation,

$$E^* = E\{p(a_i^*)[1 - p(a_i^*)]\} \approx p_0(1 - p_0)\left\{1 + \left[\frac{(1 - 2p_0)^2}{2} - p_0(1 - p_0)\right]\sigma_a^{*2}\right\}.$$

Inserting these expressions into the expression for ϕ and simplifying gives

.

$$\phi \approx \frac{\sigma_a^{*2}}{\frac{(1 - 2p_0)^2}{2}\sigma_a^{*2} + \frac{1}{p_0(1 - p_0)}} = \frac{\rho^*}{\frac{(1 - 2p_0)^2}{2}\rho^* + \frac{3(1 - \rho^*)}{\pi^2 p_0(1 - p_0)}} \approx \frac{\pi^2 p_0(1 - p_0)\rho^*}{3}.$$

for small ρ^*. Note that ϕ is always less than ρ^* and tends to be much less unless p_0 is close to 0.5 For example if $p_0 = .1$, as is typical in our studies, then $\phi \approx .3\rho$. (On the other hand, if $p_0 = 0.5$, then $\phi \approx .82\rho$.) This means that interviewer variability tends to have a much smaller impact on the variance of the categorized binary variable than on the variance of the underlying continuous latent variable.

What happens if we try to fit the inappropriate linear random effects model (1) to binary data generated by model (2)? Then the estimate of the interviewer variance component, $\hat{\sigma}_a^2$ say, is an estimate of $\mathrm{Var}\{\mathrm{E}\{\mathrm{Y} \mid \mathrm{a}\}\} = \mathrm{Var}\{\mathrm{p(a)}\} = \mathrm{V}^*$. Similarly, the sum of the estimated variance components, $\hat{\sigma}^2 + \hat{\sigma}_a^2$, is estimating $\mathrm{Var}\{\mathrm{Y}\} = \mathrm{E}^* + \mathrm{V}^*$. It follows that the estimate of the intra-interviewer correlation, $\hat{\rho}_I = \dfrac{\hat{\sigma}_a^2}{\hat{\sigma}^2 + \hat{\sigma}_a^2}$, is actually an estimate of $\phi = \dfrac{V^*}{E^* + V^*}$ rather than the underlying correlation, ρ^*. As we have seen above, ϕ is typically much smaller than ρ^* so that fitting an inappropriate linear model does indeed produce an underestimate of the true intra-interviewer correlation. However, the resulting estimate of the design effect is still giving a consistent estimate of the true design effect, contrary to the statement in our earlier paper.

2.2 Two-Stage Sampling

Now suppose that our observations are drawn using a two-stage self-weighting design. Davis & Scott (1995) considered an additive components of variance model for continuous responses of the form

$$Y_{ipr} = \mu + a_i + b_p + e_{ipr}, \qquad (4)$$

where b_p denotes the *pth* PSU effect. Assume that the $b_p s$ have mean zero and variance σ_b^2 and that they are uncorrelated with each other and with the other random effects. Using the same arguments as in the previous section, we find that the variance of the sample mean is inflated by a factor

$$D_1 = 1 + (\tilde{n}_I - 1)\rho_I + (\tilde{m}_C - 1)\rho_C \qquad (5)$$

compared to the variance that we would get with the same number of independent observations Here $\tilde{m}_C = \sum_p m_p^2 \Big/ n$, where m_p denotes the number of observations in the *pth* PSU, $\rho_C = \dfrac{\sigma_C^2}{\sigma^2 + \sigma_I^2 + \sigma_C^2}$. and ρ_I is now defined as $\rho_I = \dfrac{\sigma_I^2}{\sigma^2 + \sigma_I^2 + \sigma_C^2}$. Thus the linear model (4) leads to a nice additive decomposition of the design effect, with one component due to interviewer variance alone and the other to the clustering alone.

We can extend this model to binary responses in exactly the same way as in the previous section by supposing that there is an underlying latent variable, Y_{ipr}^* say, satisfying model (6) and that our binary variable, Y_{ipr}, takes the value one if and only if $Y_{ipr}^* \le c$ for some unknown cut-off point c. Just as in the previous section, we assume that a_i and b_p are independent normal random variables and that e_{ipr} has a logistic distribution. Again, the model is only defined up to an arbitrary scale parameter so we can only identify the ratios $\dfrac{\sigma^{*2}_a}{\sigma^{*2}}$ and $\dfrac{\sigma^{*2}_c}{\sigma^{*2}}$ rather than the individual variance components. Since ρ_I^* and ρ_p^* are both functions of these ratios we can estimate both correlation coefficients from the observed binary data.

Our primary interest is in the effect of interviewer and PSU variability on the variance of the estimated population proportion, $\hat{p}$, which is simply the sample proportion here since we are assuming that we have a self-weighting design. Let k_{ip} denote the number of interviews that the *ith* interviewer carries out in the *jth* PSU, and let $\tilde{k}_{IC} = \sum\sum k_{ip}^2 / n$. Then applying the standard expression for the variance in terms of the conditional mean and variance given the vectors of random effects, $\underline{a}$ and $\underline{b}$, gives (after some algebraic manipulation)

$$Var\{\hat{p}\} = E\left\{Var\left\{\hat{p} \mid \underline{a}^*, \underline{b}^*\right\}\right\} + Var\left\{E\left\{\hat{p} \mid \underline{a}^*, \underline{b}^*\right\}\right\}$$

$$= \frac{1}{n}\left\{E^* + \tilde{k}_{IP} V_{12}^* + \tilde{n}_I V_1^* + \tilde{m}_P V_2^*\right\}$$

where $E^* = E\left\{p(a^* + b^*)\left(1 - p(a^* + b^*)\right)\right\}$ and $V_{12}^* = Var\left\{p(a^* + b^*)\right\} - V_1^* - V_2^*$, with $V_1^* = Var\left\{p_1(a^*)\right\}$ and $V_2^* = Var\left\{p_2(b^*)\right\}$, where $p(a^* + b^*) = e^{\theta + a^* + b^*} / (1 + e^{\theta + a^* + b^*})$ as in expression (3) and $p_1(a^*)$ and $p_2(b^*)$ denote the expected values of $p(a^* + b^*)$ with respect to a^* and b^* respectively. Using the same arguments as in the previous sections, we find that this leads to a design effect of the form

$$D_1^* = 1 + (\tilde{k}_{IC} - 1)\phi_{IC} + (\tilde{n}_I - 1)\phi_I + (\tilde{n}_C - 1)\phi_C.$$

Note that, because $p(a^* + b^*)$ is not an additive function of a^* and b^*, the design effect can no longer be decomposed simply into an interviewer effect plus a cluster effect. Instead there is an additional term that depends on the way in which interviewers are deployed across the clusters.

Again, finding exact values would involve numerical integration so we limit ourselves to the case when both interviewer and PSU effects are small. Expanding $p(a + b) = e^{\theta + a + b} / (1 + e^{\theta + a + b})$ about a=0, b=0 as in the previous section, we find that $\phi_{IC} \approx 0$ to this order and the design effect can be expressed as the sum of two components,

$$D_1^* \approx \left(1 + (\tilde{n}_I - 1)\phi_I + (\tilde{n}_C - 1)\phi_C\right)$$

where

$$\phi_I \approx \frac{\sigma_a^{*2}}{\frac{(1 - 2p_0)^2}{2}(\sigma_a^{*2} + \sigma_b^{*2}) + \frac{1}{p_0(1 - p_0)}}, \text{ and } \quad \phi_C \approx \frac{\sigma_b^{*2}}{\frac{(1 - 2p_0)^2}{2}(\sigma_a^{*2} + \sigma_b^{*2}) + \frac{1}{p_0(1 - p_0)}}.$$

Overall, at least when interviewer and cluster effects are small, the conclusions are very similar to those in the previous section. Values of $\phi_I (\phi_C)$ are smaller than the corresponding values of $\rho_I^* (\rho_C^*)$ so that the design effect for the binary variable is less than it would have been for the underlying continuous variable. Moreover, estimates of the intra-interviewer and intra-cluster correlations obtained by fitting the inappropriate linear random effects model (2) are actually estimating the corresponding ϕ_I and ϕ_C values which are smaller than ρ_I and ρ_C. This means that using these estimates does actually produce an estimate of the correct design effect, in spite of the fact that ρ_I and ρ_C are underestimates of the corresponding correlations in the underlying continuous model. We are currently investigating what happens when the interviewer and cluster effects are not small enough for the linear approximation to be adequate.

3. EXAMPLE

The New Zealand Quality of Health Care Study is a study looking at adverse events (defined as "unintended injuries resulting in temporary or permanent disability caused by health care management rather than theunderlying disease process") in NZ hospitals. (See Davis et al (2001) for a full description of the study and details of the sample design.)

At the first stage, a pps sample of hospitals was drawn from all public hospitals in New Zealand with at least 300 beds. At the second stage a systematic sample of 575 medical records was drawn from all admissions in 1998. Each sampled record was then checked for evidence of an adverse event. Thirteen medical experts were used in the study, with at least three experts assigned to each hospital. There was sufficient mixing of reviewers across hospitals for the variance components to be formally identifiable. However, the assignment of reviewers to hospitals was not done at random so we need to treat the results with some caution since they depend heavily on the model. We are grateful to Eliza Chan who carried out the analysis in SAS using PROC MIXED to fit the linear random effects models and PROC NLMIXED to fit the normal-logistic models.

The main variable of interest was the proportion of hospital admissions associated with an adverse event. The estimated intra-reviewer correlation was high, with $\hat{\rho}_I^* = .13$, while the intra-hospital correlation was substantially lower with $\hat{\rho}_C^* = .007$. As predicted, the corresponding ϕ values were considerably smaller with $\phi_I = .04$ and $\phi_C = .002$. The impact of reviewer variability in particular on the variance of the sample proportion is still large, however, because of the very high average case-load. We have $\tilde{n} = 687$ leading to an overall design effect D*=29.0, almost all of which is due to reviewer variability.

If we fit a standard linear random effects model to the data we get estimated correlation coefficients $\hat{\rho}_I = .03$ and $\hat{\rho}_C = .0045$, which are much smaller than $\hat{\rho}_I^*$ and $\hat{\rho}_C^*$ but reasonably close to $\phi_I = .04$ and $\phi_C = .002$ respectively. The estimated design effect is D=21.4, which is slightly lower than the value of D*=29.0 obtained from the binary model. All this is reasonably consistent with the theory in Section 2.2.

Publicity about varying rates among hospitals was something that was of concern to the participating hospitals at the outset of the study although, to their great credit, all hospitals chosen in the sample agreed to participate. We see that both approaches above suggest that hospital differences are very small indeed. If we had used the standard survey variance estimate for two stage sampling, ignoring the reviewers altogether, we would get an estimate of 0.01 for the intra-hospital correlation coefficient, ρ_C, and an estimated design effect of D=6.4. Some of the missing reviewer variability leaks into the inter-hospital term, inflating the estimate of ρ_C by a factor of more than two, although the resulting estimate is still relatively small in this particular case. This leakage increases the estimated design effect to some extent but it is still a very considerable underestimate of the true value.

4. CONCLUSIONS

We have shown that, although interviewer variability has a smaller effect on the variance of sample proportions than on the means of continuous variables, it can still have a big impact on the total error when the average case-load is high. We also show that fitting an ordinary linear components of variance model will give a reasonable estimate of the design effect when either interviewer or cluster effects are small and possibly more generally, although this needs further investigation.

The effect of interviewer variability is usually ignored in the analysis of large health surveys. (In fact, we have not been able to find a single published account of such a study in which interviewer effects are included.) When we ignore interviewer effects, we inflate estimates of the intra-cluster correlation and (except in the extreme case when interviewers are completely nested within clusters) underestimate the true design effect. It is very expensive to train more interviewers in these specialized situations so that increasing the number of interviewers costs money and means that we have to take a smaller sample. The effect of a smaller sample size on the accuracy of sample estimates is obvious to any health researcher. However, since the interviewer effect is rarely included explicitly in their estimates of standard errors, most of them are still unaware of the true consequences of a decision to train fewer interviewers and take more observations. A big education effort seems warranted.

REFERENCES

Anderson, D.A. and Aitken, M. (1985). "Variance component models with binary response; interviewer variability", *Journal of the Royal Statistical Society*, Ser B, 47, pp203-210.

Davis, P.D. Lay-Yee, R., Briant, R., Schug, S. and Scott, A.J. (2001), *Adverse Events in New Zealand Public Hospitals*. Wellington: NZ Ministry of Health.

Davis, P.D. and Scott, A.J. (1995), "The effect of interviewer variance on domain comparisons", *Survey Methodology*, 21, pp 99-106.

De Leeuw, J. and Kreft, I.G.G. (2001), "Software for Mulitlevel Modelling", in A.H. Leyland and H. Goldstein (eds) Multilevel Modelling of Health Statistics. New York: Wiley, pp 206-223.

Groves, R. (1989). *Survey Errors and Survey Costs.* New York: Wiley.

Hartley, H.O. and Rao, J.N.K. (1968), "Estimation of nonsampling variance components in sample surveys", in N.K. Namboodiri, (ed), *Survey Sampling and Measurement*,. New York: Academic Press, pp 35-44.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, pp 92-115.

Kish, L. and Frankel. M.R. (1974), "Inferences from complex samples", *Journal of the Royal Statistical Society*, B, 36, pp 1-37.

Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsanpling Errors in Surveys*. New York: Wiley

Pannekoek, J. (1988), "Interviewer variance in a telephone survey", *Journal of Official Statistics*, 4, pp 375-384.