# VARIANCE ESTIMATION FOR COMBINED SURVEY ESTIMATES USING THE EXTENDED DELETE-A-GROUP JACKKNIFE

James L. Reilly (reilly@stat.auckland.ac.nz)
Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

**Key Words:** variance estimation, jackknife, combined estimates, statistical matching.

## Variance Estimation for Combined Survey Estimates

Variance estimates for surveys with complex sample designs are typically obtained using linearization or resampling methods. Survey programs sometimes produce estimates by combining data from two or more separate surveys. Simple cases include aggregations of different populations or geographical areas, or cumulations or differences over time. Standard variance estimation methods can often be adapted to simple combined surveys such as these, as reviewed by Kish (2002), especially if the samples for the surveys are selected independently. However more complex methods of combining surveys can pose greater problems.

A particular difficulty with linearization methods is that new variance estimation formulas would need to derived and implemented for each method of combining the component surveys and for each general type of sample design. Although resampling methods can in theory handle a wide variety of combined survey estimates, most such methods can only do this if the surveys in question have very similar sample designs, at least for the first stage of sample selection. For example, balanced repeated replication would require each survey to use the same number of strata.

This paper will briefly review the extended delete-a-group jackknife described by Kott (1999, 2001) and explain how it provides an effective method of variance estimation for combined survey estimates even when the component surveys have different sample designs, if the samples for these surveys are selected independently. It also describes two applications of this method. The first application involved producing harvest estimates for recreational fishing by using data from two short surveys to develop weights for data from a third diary survey, while the second application involved the statistical matching of two media surveys. The first type of application was alluded to by Kott (2001), but the second application is believed to be novel.

This research was supported in part by Marsden Fund Grant 01-UOA 157, the New Zealand Ministry of Fisheries and ACNielsen (NZ) Ltd.

## The Extended Delete-a-Group Jackknife and Combined Survey Estimates

The extended delete-a-group jackknife is described fully by Kott (1999, 2001), but will be reviewed briefly here. It is assumed that expansion estimators are used, with the inverse probability weights being calibrated against totals from external sources or from earlier stages of the survey.

First the primary sampling units (PSUs) are divided into $R$ variance groups. The construction of appropriate variance groups is discussed by Kott (1998) and Wolter (1985). Jackknife replicates are defined to be the remainder of the sample after removing the corresponding variance group. However the EDAGJK is different from most other jackknife methods, in that a unit outside a EDAGJK replicate can still have a non-zero replicate weight.

Initial replicate weights are created by adjusting the inverse probability weights as described by Kott (1999). Briefly, if respondent $k$ in PSU $j$ within stratum $h$ has inverse probability weight $w_{hjk}$, the weight for this respondent in replicate $r$ remains unchanged if no PSUs were in stratum $h$ for variance group $r$. However if some PSUs did fall within stratum $h$ for variance group $r$, the initial $r$-replicate weights for respondents in these PSUs are given by $w_{hjk}\left(1-(n_h-1)\sqrt{Z_h}\right)$, while the initial $r$-replicate weights for respondents in the other PSUs are $w_{hjk}\left(1+\sqrt{Z_h}\right)$. Here $n_h$ is the number of PSUs selected in stratum $h$, and $Z_h = R/\left((R-1)n_h(n_h-1)\right)$. Calibration is then carried out on each set of initial replicate weights to produce final replicate weights. (Jackknife variance estimates may have an upward bias if calibration groups are not nested within strata (Kott 1998). This potential bias has been ignored in the following applications.)

Once each of the component surveys have been

divided into the same number of variance groups, replicate weights are obtained for each survey as described above. Combined survey estimates are then recalculated for each jackknife replicate, taking these replicate weights into account. Finally the results for all the replicates are combined as suggested in Kott (1999) to produce variance estimates.

For the purpose of variance estimation for combined survey estimates, a key feature of the EDAGJK is that it does not require that the number of PSUs selected in each stratum be large, in contrast to the delete-a-group jackknife. Also the number of variance groups is not determined by the sample design (in contrast to the stratified delete-one-PSU jackknife), and so the same number of variance groups can be used for quite different designs. In particular, the same number of variance groups may be used for two surveys from which combined survey estimates have been produced, even though they may have quite different sample designs.

In contrast, the usual delete-a-group jackknife requires that the number of PSUs selected from each stratum be large (preferably five or more). If one of the component surveys has a highly stratified sample design with many small stratum sample sizes, this means the usual delete-a-group jackknife would overestimate the variance of the combined survey estimates if used as described above for the EDAGJK. Such designs are commonly used in practice. For example, the face to face survey described in the section below had 19 strata from which fewer than five PSUs were selected.

## Application 1: New Zealand National Marine Recreational Fishing Survey

Harvest estimates for marine recreational fishing around New Zealand for the period from 1 December 1999 to 30 November 2000 were produced by combining data from three separate surveys, based on two independent samples of households. An estimate of fisher prevalence derived from a nation-wide face to face survey was combined with detailed diary data on fishing behavior (including numbers harvested) recorded by a nation-wide sample of recreational fishers recruited by telephone, to estimate the numbers of fish harvested recreationally broken down by species and region.

The weighting process used to combine these surveys involved several steps, including:

- calculation of selection probabilities and inverse probability weights;
- calibration of face to face survey data against population totals from the 2001 Census;
- non-response adjustment of diarists using recruitment survey data;
- calibration of diarist data against population totals from the face-to-face survey; and
- an adjustment for fishers entering the fishery based on diary and face-to-face survey data.

This process is described fully in Reilly (submitted).

Similar surveys had been used previously to estimate harvests, as described by Bradford (1998) and Teirney et al. (1997), but these did not take aspects of the sample design into account. The variances of the earlier harvest estimates were also calculated using computationally convenient but highly dubious distributional assumptions, and it was believed these variance estimates were substantially underestimating the true variances.

Variance estimates for the 1999-2000 harvest estimates were instead calculated using the extended delete-a-group jack-knife method. There does not appear to be any reference in the literature to the appropriate number of variance groups specifically for the extended delete-a-group jackknife, and there are conflicting suggestions about the number of variance groups that should be used for the related delete-a-group jackknife method. Although Kott (1998) suggests that 15 groups is sufficient for the delete-a-group jackknife, Smith (2001) found that 40 groups were advisable for the New Zealand Household Labour Force Survey (and that using 120 groups gave even better results).

The data for the face to face survey and the telephone recruitment survey was divided into 120 variance groups, as described below. Since the diarists are a subset of the recruitment survey respondents, this effectively selected 120 groups of diarists as well. Although the variance estimates discussed here (and reported by Boyd and Reilly (submitted)) were based on 120 variance groups, initial estimates were calculated using only 30 groups and this gave very similar results.

In the first stage of sampling for the face to face survey, areas were selected using stratified systematic sampling. There were 94 strata based on detailed regions and level of urbanization within these regions. The 120 variance groups for this survey were formed by ordering the areas by strata, in their usual order within strata, and then systematically selecting every $120^{\text{th}}$ area. The telephone recruitment survey was assumed to be a stratified random sample of households, with equal probabilities of selection within strata. Only one randomly

selected respondent was interviewed in each household. The stratification for this survey was actually implemented through centrally managed quotas for the number of interviews conducted in each of 40 area codes. These 40 areas were also based on region and level of urbanization. The 120 variance groups for this survey were created by randomly ordering the diarist recruitment survey respondents within each strata and selecting every $120^{\text{th}}$ respondent.

Initial replicate weights were created by first adjusting the inverse probability weights for each survey as described above. All the remaining steps in the weighting process were then repeated for each replicate, based on these initial replicate weights, to produce the final replicate weights. Each set of final replicate weights was used to calculate harvest estimates for that replicate. The variation amongst these replicate harvest estimates enabled the calculation of variance estimates for harvests, as described in Kott (1999).

Using the EDAGJK approach for calculating the harvest variance estimates had the advantage that there was no need for distributional assumptions. The EDAGJK could accommodate both sample designs using the same number of replicates, even though they had different numbers of strata with varying numbers of units selected from each stratum.

There were also some potential drawbacks to this approach, particularly regarding the face to face survey component. It effectively assumed that the sample formed by leaving out a systematic sub-sample consisting of every $120^{\text{th}}$ area (as was done to form the jack-knife replicates) will have similar statistical properties to the full NRS sample selected by systematic sampling. However some assumptions are always needed to calculate sampling variation based on a single systematic sample, as discussed by Wolter (1985). The assumption made here seemed fairly reasonable given the slowly varying geographical trends observed in recreational marine fishing prevalence. This EDAGJK approach also did not include any adjustment for sampling from a finite population, although 28% of PSUs were selected for the face to face survey. As a result, the EDAGJK variance estimates for results solely based on this survey would probably have been over-estimated by approximately 15%.

However these issues affect only the face to face component, which has a relatively small impact on the variability of harvest estimates compared to that arising from the skewed distribution of diarists harvests. These issues are therefore believed to have a negligible impact on the variance estimates for recreational harvests.

Based on deviations from proportionality to the inverse square root of the sample size, it appears that the variance estimates may have underestimated the true sampling variances when the calculation was based on a very low number of diarists. This was particularly evident for regions and species which were harvested by only one diarist, but smaller deviations were also noted for species and regions where there were fewer than five successful diarists.

The resulting variance estimates were substantially larger than those reported for earlier surveys. For example, the coefficient of variation for the most reliable harvest estimate increased from 5% to 11%, even though the diarist sample sizes were similar. (Variances for the 1999-2000 harvest estimates were also calculated using the previous approach, and the most reliable harvest estimate had a coefficient of variation of only 2%, so the difference between the two methods is bigger than it might first appear.) The primary reason for this increase was that the earlier variance estimates assumed that the number of fish harvested followed a Poisson distribution, which was a highly unrealistic assumption, and the EDAGJK did not. A secondary reason was that the EDAGJK took the sample designs into account, which will also have increased the variance estimates.

This highlights the need to use variance estimation methods that are appropriate for the surveys and data at hand, and to check their underlying assumptions.

## Application 2: Statistical Matching

Statistical matching techniques involve finding, for each respondent in one survey, one or more similar respondents in another survey. Data for certain variables (denoted Y) from the second survey is transferred to the matching respondents in the first survey. The relationships between the Y variables and other variables (denoted Z) not gathered in the second survey can then be examined in the resulting fused dataset. Similarity is defined for matching purposes in terms of common variables (denoted X) gathered in both surveys. Statistical matching differs from record linkage (sometimes known as exact matching) where the goal is to link records for the same person (not similar people) in the two databases.

There are many techniques for statistical matching, most of which effectively assume that Y is independent of Z given X. Bias may result if this assumption is violated. Evaluations of results from

statistical matching have often been limited to internal diagnostics about the matches and checks that the marginal distributions of the Y variables have not been distorted. Some studies have also investigated the extent of bias or potential bias, but very few authors seem to have investigated the extent of sampling variation. Potential bias is certainly an important issue, but the level of sampling variation is also important. For example, one would like to know how variable the results based on subsamples would be to decide which analyses should be performed.

Zaslavsky and Thurston (1994) used a simple grouped jackknife to incorporate sampling error in their analysis of errors in microsimulation models based on statistically matched data, and Reilly (1996, 2000) summarised the observed variation in statistical matched results over time using a generalised variance function. Ingram et al. (2000) obtained approximate significance tests by dividing the statistically matched data into 16 "replicates", and examining the range of chi-square values calculated on the replicates. Rässler (2002, section 2.6.3) produced Monte Carlo estimates of variances in a simulation study comparing nearest neighbour and propensity score matching methods, and also used multiple imputation methods. Even though statistical matching is widely used, and there are many applications documented in the literature, there do not seem to be any other studies addressing the sampling variation of statistically matched results. This may well be due to the apparent lack of practical variance estimation methods.

Multiple imputation approaches (as proposed by Rubin 1986 and investigated by Rässler 2002) would appear to have wide potential application, but in most situations it seems very difficult to confirm that the imputations would be proper and that the variance estimates proposed by Rubin (1987) would therefore be consistent. Recent work (Reilly 2003, to appear) confirms that Rubin's variance estimates can substantially overestimate or underestimate the sampling variation for survey data when the analysis model is not identical to that used for imputation, or when the imputation model is misspecified. This would typically be the case in statistical matching situations, because many analyses are conducted on a single matched dataset, and the conditional independence assumption underlying the matching process would usually be violated to some degree.

The EDAGJK provides an alternative method of variance estimation for results from statistical matching techniques such as constrained statistical matching that incorporate both sets of survey weights. (Some other statistical matching techniques ignore the survey weights during the matching process, which would seem to preclude the use of the EDAGJK.) The same number of variance groups must be created for each component survey, and statistical matching carried out for each corresponding pair of replicates based on the replicate weights. The statistically matched results from all the replicates are then combined in the usual way to calculate variance estimates.

This method has been applied to a folded database (where one survey is split in half and one half is statistically matched with the other). The resulting variance estimates are generally somewhat larger than for similar results in the component databases. It would be interesting to compare these variance estimates with those based on multiple imputation, using both the traditional method due to Rubin (1987) and the model-robust approach due to Robins and Wang (2000). It is also intended to validate the EDAGJK method through a simulation study, and to apply it to results from an ongoing statistical matching programme in media research.

## References

[1] R. O. Boyd and J. L. Reilly. 1999–2000 national marine recreational fishing survey: harvest estimates. Draft New Zealand Fisheries Assessment Report 2002/XX, Ministry of Fisheries, Wellington, New Zealand, 2002. 26p. Submitted for review.

[2] E. Bradford. National marine recreational fishing survey 1996: scaling the diary survey results to give the total recreational harvest. Technical Report 17, NIWA, Wellington, New Zealand, 1998. 33p.

[3] D. D. Ingram, J. O'Hare, F. Scheuren, and J. Turek. Statistical matching: A new validation case study. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2000.

[4] L. Kish. Combining multipopulation statistics. *Journal of Statistical Planning and Inference*, 102:109–118, 2002.

[5] P. S. Kott. Using the delete–a–group jackknife variance estimator in NASS surveys. RD Research Report RD–98–01, USDA, Washington, D.C., 1998.

[6] P. S. Kott. The extended delete–a–group jackknife. In *Bulletin of the International Statistical Institute, 52nd Session, Contributed Papers*, 1999.

[7] P. S. Kott. The delete–a–group jackknife. *Journal of Official Statistics*, 17:521–526, 2001.

[8] S. Rässler. *Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches*. Springer–Verlag, New York, 2002.

[9] J. L. Reilly. Stability of Panorama results: initial findings. Technical Report 1996-02, AGB McNair (NZ) Ltd, Auckland, New Zealand, 1996.

[10] J. L. Reilly. The development and evaluation of statistical matching applications. Statistics, University of Auckland, Auckland, New Zealand, 2000.

[11] J. L. Reilly. 1999–2000 national marine recreational fishing survey: weighting methodology for harvest estimates. Draft New Zealand Fisheries Assessment Report 2002/XX, Ministry of Fisheries, Wellington, New Zealand, 2002. 24p. Submitted for review.

[12] J. L. Reilly. An estimating equations technique for valid inference from imputed survey data. In *Bulletin of the International Statistical Institute, 54th Session, Contributed Papers*, 2003 (to appear).

[13] J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87:113–124, 1986.

[14] D. B. Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4:87–95, 1986.

[15] D. B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987.

[16] H. Smith. Investigation of the delete–a–group variance estimator for the HLFS. Technical report, Statistics New Zealand, Wellington, New Zealand, 2001.

[17] L. D. Teirney, A. R. Kilner, R. B. Millar, E. Bradford, and J.D. Bell. Estimation of recreational harvests from 1991–92 to 1993–94. New Zealand Fisheries Assessment Report Document 97/15, Ministry of Fisheries, Wellington, New Zealand, 1997.

[18] K. M. Wolter. *Introduction to Variance Estimation*. Springer–Verlag, New York, 1985.

[19] A. Zaslavsky and S. Thurston. Error analysis of food stamp microsimulation models. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1994.