JACKKNIFING FOR VARIANCE COMPONENTS IN COMPLEX SAMPLE SURVEY DESIGNS

Ralph E. Folsom, David L. Bayless, and Babu V. Shah
Research Triangle Institute, Research Triangle Park, N. C. 27709

1. INTRODUCTION

The variance components methodology presented in sections two and three of this paper extends results of Seeger (1970) which were developed for sampling designs with equal probability selection from effectively infinite populations at each stage of sampling. We have shown that Seeger's simple analysis of unweighted means also works for linear statistics from a class of highly stratified three stage designs allowing PPS selection. The trick will be to use properly expanded-up last stage responses as the basic variables of analysis.

In order to accommodate the non-linear statistics which are commonly used with such complex designs, we have developed in section four a multi-stage extension of the Quenouille (1956)-Tukey (1958) Jackknife.

2. NOTATION AND MODEL

The class of sampling designs that we have considered are stratified three stage designs with PPS selection at the first two stages and equal probability sampling at the last stage. To simplify our presentation, we assume that first stage units are sampled "with" replacement and are subsampled independently each time they are selected. Second and third stage units are selected "without replacement."

To establish the link with Seeger's variance components methodology, we will work with the expanded up last stage responses in equation (1).

$$y_{ijk} = M_{ij} Y_{ijk} / p_i p_{j/i} \qquad (1)$$

where the small p's are relative size measures for the first and second stage units and M_{ij} is the number of third stage units in the (ij)-th secondary unit. The cap-Y represents some characteristic of population unit (ijk). Notice that if only one unit was selected at each stage of sampling then y_{ijk} would be the Horvitz-Thompson (1952) estimator for the population total

$$Y_{+++} = \sum_{i=1}^{N} \sum_{j=1}^{S_i} \sum_{k=1}^{M_{ij}} Y_{ijk} \qquad (2)$$

In general, the Horvitz-Thompson estimator for Y_{+++} can be written as the average of our small-y variables; that is,

$$\hat{Y}_{+++} = y... = \sum_{i=1}^{n} \sum_{j=1}^{s_i} \sum_{k=1}^{m_{ij}} y_{ijk} / n s_i m_{ij} \qquad (3)$$

We have defined five variance components associated with the various stages of sampling in our design. Four of these components can be defined simply in terms of the "effects" presented in the "model identity" of equation four.

$$y_{ijk} = \mu + \rho_i + \eta_{j/i} + \epsilon_{k/ij} \qquad (4)$$

where

$$\mu = Y_{+++}$$

$$\rho_i = (Y_{i++}/p_i - Y_{+++})$$

$$\eta_{j/i} = (Y_{ij+}/p_i p_{j/i} - Y_{i++}/p_i)$$

$$\epsilon_{k/ij} = (y_{ijk} - Y_{ij+}/p_i p_{j/i}).$$

For a balanced sample selected "with replacement" at each stage, one can show that

$$\mathrm{Var}(\hat{Y}_{+++} = y...) = \sigma_P^2/n + \sigma_{S/P}^2/ns + \sigma_{K/S}^2/nsm \qquad (5)$$

where the components are defined

$$\sigma_P^2 = \sum_{i=1}^{N} p_i \rho_i^2$$

$$\sigma_{S/P}^2 = \sum_{i=1}^{N} \sum_{j=1}^{S_i} p_i p_{j/i} \eta_{j/i}^2 = \sum_{i=1}^{N} p_i \sigma_{S/P}^2(i)$$

$$\sigma_{K/S}^2 = \sum_{i=1}^{N} \sum_{j=1}^{S_i} \sum_{k=1}^{M_{ij}} p_i p_{j/i} \epsilon_{k/ij}^2 / M_{ij}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{S_i} p_i p_{j/i} \sigma_{K/S}^2(ij).$$

For our "without" replacement sampling at stages two and three, we need two more components. The second stage component involves normalized joint inclusion probabilities $\theta_{jj'/i} = \{\pi_{jj'/i}/s_i(s_i-1)\}$ and squared differences $\epsilon_{jj'/i}^2 = (Y_{ij+}/p_i p_{j/i} - Y_{ij'+}/p_i p_{j'/i})^2$ as follows

$$V_{S/P}^2 = \sum_{i=1}^{N} p_i \sum_{j=1}^{S_i} \sum_{j'<j} \theta_{jj'/i} \epsilon_{jj'/i}^2$$

$$= \sum_{i=1}^{N} p_i V_{S/P}^2(i).$$

The third stage "without replacement" component is

$$V^2_{K/S} = \sum_{i=1}^{N} \sum_{j=1}^{S_i} p_i p_{j/i} \sum_{k=1}^{M_{ij}} \epsilon^2_{k/ij}/(M_{ij}-1)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{S_i} p_i p_{j/i} \, V^2_{K/S}(ij). \qquad (7)$$

The variance of the Horvitz-Thompson estimator for a balanced version of our design can now be written in the simple form of equation (8) where the cap-sigmas are linear composites of our five separate components

$$\mathrm{Var}(\hat{Y}_{+++} = y_{...}) = \Sigma^2_P/n + \Sigma^2_{S/P}/ns + \Sigma^2_{K/S}/nsm \qquad (8)$$

with

$$\Sigma^2_P = \{\sigma^2_P - (V^2_{S/P} - \sigma^2_{S/P})\}$$

$$\Sigma^2_{S/P} = \{V^2_{S/P} - (V^2_{K/S} - \sigma^2_{K/S})\}$$

$$\Sigma^2_{K/S} = V^2_{K/S} \, .$$

For explicit derivations of these results, see Folsom, Bayless, and Shah (1971).

3. UNBIASED ESTIMATION

With the sampling structure and components definitions outlined in section 2, we can show that the following simple unbiased estimators are available for our cap-sigmas

$$\hat{\Sigma}^2_P = (MS_P - M'_{S/P})$$

$$\hat{\Sigma}^2_{S/P} = (MS_{S/P} - MS'_{K/S}) \qquad (10)$$

$$\hat{\Sigma}^2_{K/S} = MS_{K/S}$$

where the MS's denote the following "analysis of unweighted means" type mean squares

$$MS_P = \sum_{i=1}^{n} (y_{i..} - y_{...})^2/(n-1)$$

$$MS_{S/P} = \sum_{i=1}^{n} \sum_{j=1}^{s_i} (y_{ij.} - y_{i..})^2/n(s_i-1)$$

$$MS'_{S/P} = \sum_{i=1}^{n} \sum_{j=1}^{s_i} (y_{ij.} - y_{i..})^2/ns_i(s_i-1) \qquad (11)$$

$$MS_{K/S} = \sum_{i=1}^{n} \sum_{j=1}^{s_i} \sum_{k=1}^{m_{ij}} (y_{ijk}-y_{ij.})^2/ns_i(m_{ij}-1)$$

$$MS'_{K/S} = \sum_{i=1}^{n} \sum_{j=1}^{s_i} \sum_{k=1}^{m_{ij}} (y_{ijk}-y_{ij.})^2/ns_i m_{ij}(m_{ij}-1)$$

The derivation of expected mean squares which leads to the estimators in equation (10) is detailed in Folsom, Bayless, and Shah (1971). Unbiased estimates for the five separate components are also presented in the report cited above.

4. MULTIPLE-STAGE JACKKNIFING

Our contribution to the Jackknife procedure involves partitioning the variance of a non-linear statistic such as $\hat{\theta}$ in equation (12)

$$\hat{\theta} = f[y_{+...}(1), \cdots, y_{+...}(g)] \qquad (12)$$

into components like our cap-sigmas. The (plus) on the little y's (sample totals) in (11) indicate summation over h = 1(1)H strata. Estimates for θ are first formed from pseudo-replicates obtained by successively deleting the data from sampling units at a particular level of the design. These estimates as they occur in equation (13) are subscripted by a minus sign followed by labels for the deleted sampling unit.

$$J\theta_{hijk} = n_h s_{hi} m_{hij} \hat{\theta} - (n_h-1)\hat{\theta}_{-hi} - n_h(s_{hi}-1)\hat{\theta}_{-hij} - n_h s_{hi}(m_{hij}-1)\hat{\theta}_{-hijk} \qquad (13)$$

Equations (14) and (15) demonstrate the form of the replicate estimator when a first stage unit is deleted

$$\hat{\theta}_{-hi} = f\{Y_{-hi}(1), \ldots, Y_{-hi}(y)\} \qquad (14)$$

with

$$Y_{-hi}(r) = y_{+...}(r) - [y_{hi..}(r) - y_{h...}(r)]/(n_h-1) \qquad (15)$$

If results from classical theory hold up in this finite population context, then we would expect the average of our pseudo-values shown in equation (16) to have less bias than $\hat{\theta}$ in (12).

$$\hat{\theta}_{JK} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{s_{hi}} \sum_{k=1}^{m_{hij}} J\theta_{hijk}/Hn_h s_{hi} m_{hij} \qquad (16)$$

$$= J\theta_{....}$$

For a linear statistic, the Jackknife estimate in (15) reduces to (12). To estimate variance components for the jackknifed statistic, we substitute unweighted means of the pseudo-values into the mean squares in equations (11). This is spelled out for the first-stage component in

equation (16)

$$J\Sigma_P^2 = (JMS_P - JMS'_{S/P}) \qquad (16)$$

where

$$JMS_P = \sum_{h=1}^{H} \sum_{i=1}^{n_h} (J\theta_{hi..} - J\theta_{h...})^2/(n_h-1)$$

$$JMS'_{S/P} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{s_{hi}} (J\theta_{hij.} - J\theta_{hi..})^2/n_h s_{hi}(s_{hi}-1)$$

5. EMPIRICAL RESULTS

The P-Values in Table I represent ratio estimates computed from a stratified three-stage sample of High School Seniors conducted by the Research Triangle Institute for the National Center for Educational Statistics. Although the small sample sizes involved in this pretest make it impossible to draw any general empirical conclusions, it is interesting to note that

1. The Jackknife and Standard P-Values are numerically equivalent, indicating little or no bias in the combined ratio estimate.

2. The Jackknife Components for the last two stages are numerically equivalent to corresponding "Taylor Series" estimates with only a slight difference at the PSU stage.

The "Taylor Series" linearization alluded to in point 2 above is a direct extension of Tepping's (1968) results to our variance components setting.

6. DISCUSSION

Although our variance components methodology was developed for a particular sample, it applies to a fairly wide class of stratified three-stage designs. The "with replacement" at the first-stage simplifies the mean squares, but it is not crucial to the application of our Multi-stage Jackknife. This Jackknife shares with Taylor series linearization the property of producing a pseudo-value which is associated with a particular last stage unit. By borrowing the form of variance and variance components estimators already available for linear statistics, the Jackknife and Taylor series linearizations provide direct extensions of these results to non-linear statistics. Our limited empirical results show that these two methods produce very similar results for ratios. In summary, we feel that the Jackknife replication technique with our extension will prove to be a very useful method of variance and variance components estimation for complex sample statistics.

REFERENCES

1. Folsom, Ralph E., Bayless, David L., and Shah, Babu V. Jackknifing for Variance Components in Complex Sample Survey Designs. Presented at the American Statistical Association Meeting at Fort Collins, Colorado, August 23, 1971.

2. Quenouille, M. H. (1956). Notes on bias in estimation. Biometrika 43, 353-360.

3. Seeger, P. (1970). A method of estimating variance components in unbalanced designs. Technometrics 12, 207-218.

4. Tepping, B. (1968). The estimation of variance in complex surveys. Proceedings of the Social Statistics Section of the American Statistical Association. 11-18.

5. Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 47, 663-685.

6. Tukey, J. W. (1958). Bias and confidence in not-quite large samples: Abstract. Ann. Math. Statist. 29, 614.

TABLE I

TAYLOR SERIES (TS) AND JACKKNIFE (JK)
P-VALUES AND VARIANCE COMPONENTS

Item Code	Description of Item	P-Value[1] x 10^2		Variance Components x 10^4							
		P		PSU		Pair		Student		Total[2]	
				$V_1=\hat{\Sigma}^2_P(+)$		$V_2=\hat{\Sigma}^2_{S/P}(+)$		$V_3=\hat{\Sigma}^2_{K/S}$		$V=Var(\hat{P})$	
		$\hat{P}$	$JP_{....}$	TS	JK	TS	JK	TS	JK	TS	JK
A	Highest Education of Parents is Less Than High School	15	15	35.61	35.46	8.72	8.73	91.18	98.18	14.84	14.92
B	Definite or Likely Goer to College	43	43	3.66	3.93	21.49	21.49	246.4	246.4	8.81	9.01
C	Plan to Attend College	59	59	13.20	13.47	17.67	17.67	242.9	242.9	11.39	11.50
D	Don't Belong to a Minority Group	88	88	4.43	4.44	9.55	9.55	88.77	88.77	4.55	4.56
+	Average over Four Items	51	51	14.22	14.32	14.36	14.36	169.1	169.1	9.90	10.00

[1] $P = \frac{\text{Sum of Student Weights with Attribute}}{\text{Sum of All Student Weights}}$

[2] $10^4 V = MS_p(+)/3$. See Section 5 for the formulas of $MS_p(+)$.