A COMPARISON OF REPLICATED AND PSEUDO-REPLICATED COVARIANCE MATRIX ESTIMATORS FOR THE ANALYSIS OF CONTINGENCY TABLES

VIRGINIA RICHARDS AND DANIEL H. FREEMAN, JR., Yale University

Introduction

The method of balanced half-sample replication has been utilized by many major survey programs for the estimation of the variances of estimates from large scale multi-stage surveys. This estimation procedure is an outgrowth from the method of direct replication. Both of these methods for the estimation of a covariance matrix will be discussed. Data from the Connecticut High Blood Pressure (CHBP) Survey will be used to illustrate the effect of the differing techniques on the estimation of a log-linear model.

Design of CHBP Survey

The 169 towns (primary sampling units-PSUs) in Connecticut cover the entire area of the state in a non-overlapping manner. Furthermore each of the towns belongs to one and only one Health Service Area (HSA). Stratification was employed at two levels; first by HSA and second by population size within HSA. An attempt was made to form strata of nearly equal population size. Ultimately then there were 32 strata, 8 which were comprised of only one PSU and 24 which were comprised of 2 or more PSUs. Within each of the 24 non-certainty strata one PSU was chosen with probability proportional to size. Within each PSU, four independent systematic samples of segments were chosen. These determined the 4 replications to be used in variance estimation by the method of direct replication.

Direct Replication

The central idea to the method of direct replication, first proposed by Mahalanobis (1946) as interpenetrating subsamples, is that replicates (subsamples) of the over-all sample can be designed so as to mimic the over-all survey design in every respect except sample size. The ease in variance estimation with this method is due to the assumption that the variability of a statistic based on the entire sample can be estimated from the variability of the statistic among replicates.

Consider the estimation of the covariance matrix for Table 1. Each replicate is considered as a sample unto itself and a 'Table 1' is generated from each.

Table 1.
Proportion of White Adults at least 18 years of age by Sex, Age Group, Hypertension Status: Connecticut 1978-79

		Hypertension Status		
Sex	Age	Normal	Border	High
Males	<50	.2306	.0421	.0231
	≥50	.0867	.0512	.0374
Females	<50	.2098	.0189	.0084
	≥50	.1117	.0606	.0386

If the over-all total were a fixed number n and each replicate total were n/k then the covariance of the vector of cell entries of proportions for Table 1 would be

$$\sum_{i=1}^{k} (\underset{\sim}{p}_i-\underset{\sim}{p})'(\underset{\sim}{p}_i-\underset{\sim}{p})/k(k-1)$$

where k=the number of replicates

$\underset{\sim}{p}$=the vector of estimates from the entire sample

$\underset{\sim}{p}_i$=the vector of estimates from the i-th replicate.

Under these conditions $\sum_{i=1}^{k} \underset{\sim}{p}_i/k=\underset{\sim}{p}$ and the formulation given is analogous to the variance of the mean of a simple random sample of size k.

The denominator for the proportions in Table 1 is the over-all total and since this number is not fixed, the proportions are really ratio estimates. Deming (1960) gives a formula for an estimate of the variance of a ratio as

$$\hat{\sigma}^2 = \frac{k}{(k-1)y^2} \sum_{i=1}^{k} (x_i-ry_i)^2$$

where r=x/y is the estimate from the entire sample and x, y are replicate estimates. This can be modified to give an estimate of the covariance matrix of a vector of ratios with the same denominator.

$$\hat{\text{Cov}}\ (r_j r_{j'}) = \frac{k}{(k-1)y^2} \sum_{i=1}^{k} (\underset{\sim}{x}_i-\underset{\sim}{y}_i r)(\underset{\sim}{x}_i-\underset{\sim}{y}_i r).$$

In the CHBP Survey, non-certainty strata were paired in a manner to be described in the next section. This created two separate zones, one for certainty strata within which there were 4 replicates and one for non-certainty strata within which there were 8 replicates. Again referring to Deming (1960) an estimate of the variance of a ratio when calculated from two zones is

$$\hat{\sigma}^2_r = \frac{k}{(k-1)y^2}\left\{\sum_{i=1}^{k}\left[(x_{i1}-\bar{x}_1)-r(y_{i1}-\bar{y}_1)\right]^2 + \sum_{i=1}^{k}\left[(x_{i2}-\bar{x}_2)-r(y_{i2}-\bar{y}_2)\right]^2\right\}$$

where x_{ij} = x value in zone j replicate i

$\bar{x}_j$= average x value in zone j.

Modification of this formula to estimate the covariance of a vector of ratios as well as a differing number of replications in the two zones yields the following equation.

$$\hat{\sigma}^2_r = \frac{k_1}{(k_1-1)y^2} \sum_{i=1}^{k_1}\left[(\underset{\sim}{x}_{i1}-\bar{\underset{\sim}{y}}_1 r)-(\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{y}}_1 r)\right]\left[\cdot\right] + \frac{k_2}{(k_2-1)y^2} \sum_{i=1}^{k_2}\left[(\underset{\sim}{x}_{i2}-\underset{\sim}{y}_i r)-(\bar{\underset{\sim}{x}}_2-\bar{\underset{\sim}{y}}_2 r)\right]\left[\cdot\right]$$

where k = the number of replicates in zone j.

Balanced Half-Sample Replication

The balanced half-sample method of estimating variances as developed by McCarthy (1966), requires a design of 2 PSUs per stratum. In order to meet this restriction the method of collapsed strata was used in non-certainty strata. Pairing

was done on the basis of a subjective determination of the homogeneity of the strata, since the bias of the balanced half-sample estimator of the variance of a ratio is directly related to the heterogeneity of the two strata forming the collapsed stratum (Stanek and Lemeshow, 1977). In the certainty strata, the PSUs were halved so as to form two pseudo-PSUs. Since 4 independent systematic samples had been selected within each PSU, these were randomly paired to create the 2 pseudo-PSUs. With 20 strata with 2 PSUs per strata, 20 balanced half-sample replications were necessary. These were determined from an orthogonal matrix produced by the method of Plackett and Burman (1946). This estimate of the variance of a vector of ratios is obtained as follows:

$$\hat{Cov}(r_j r_{j'})_{BHS} = \frac{1}{20}\sum_{i=1}^{20} (\underset{\sim}{r}_i - \underset{\sim}{r})'(\underset{\sim}{r}_i - \underset{\sim}{r}).$$

Koch and Lemeshow (1972) noted that balanced half-sample replication produces valid and consistent estimates of variance for statistics calculated from sample survey data.

Comparison of Estimates of Standard Errors

The estimates of standard errors in Table 2 were obtained by taking the square root of the diagonal elements of the respective covariance matrix.

Table 2.
Standard Error of Proportions in Table 1 by Direct Replication (DR) and by Balanced Half-Sample Replication (BHS)

		Hypertension Status		
Sex	Age	Normal	Border	High
Males	<50	.0075	.0037	.0022
		.0088	.0030	.0024
	≥50	.0040	.0035	.0021
		.0051	.0032	.0046
Females	<50	.0054	.0020	.0014
		.0068	.0018	.0020
	≥50	.0046	.0037	.0030
		.0039	.0056	.0038

The standard errors range from a minimum of .0014 to a maximum of .0075 for the direct replication (DR) method. The range for the balanced half-sample (BHS) method is .0018 to .0088. The BHS estimate exceeds the DR estimate 8 out of 12 times with the largest difference being more than twice the DR estimate. In the 4 instances where the DR exceeds the BHS estimate the differences were relatively small with the DR being no more than 1.24 times the BHS estimate. In general it would seem that differences of this magnitude could affect inferences made about the estimated proportions in the table. This is investigated in the next section.

Log-Linear Model

Methodologies have been developed (Grizzle, Starmer, and Koch - 1969: Koch and Lemeshow - 1972; Koch, Freeman, Freeman - 1975) to analyse contingency tables generated by data from complex surveys by using a linear model approach. This approach involves the use of weighted least squares for the computation of Wald statistics.

A transformation of the proportions in Table 1 was performed so that the model would be estimating f, the log of the ratio of the individual cell proportions to the proportion in the cell for females, ≥50, with high blood pressure.

$\underset{\sim}{f} = \underset{\sim}{A} \log \underset{\sim}{p}$ where $\underset{\sim}{A} = \left[I_{11} \; -1_{11x1} \right]$

$\underset{\sim}{p}$ = vector of proportions.

The model can be expressed as follows,

$\log \underset{\sim}{p} = \underset{\sim}{X}\underset{\sim}{B}$

$\underset{\sim}{f} = \underset{\sim}{A} \log \underset{\sim}{p} = \underset{\sim}{AXB}$

where $\underset{\sim}{B}$ = vector of parameters

$\underset{\sim}{X}$ = design matrix characterizing the model.

The GENCAT computer program was used to formulate models through design matrices, and test, through contrast matrices, hypotheses pertaining to the estimated parameters. The vector of cell proportions and their estimated covariance matrix can be entered directly into this program. Thus a parallel development of the modeling was carried out using the two different covariance matrices.

First a model including all interaction terms was fit. Using A=age, S=sex, BP_1=normal bloodpressure, and BP_2=borderline hypertensive, the model can be expressed in terms of the following vector of parameters (S A BP, BP_2 SBP_1 SBP_2 ABP_2 AS $ASBP_1$ $ASBP_2$). Since this is a saturated model the fit of the model was perfect.

A contrast matrix to test the hypothesis of no three-way interaction was rejected. Thus it was necessary to include the three-way interaction terms and no reduction of the model was possible. This was true regardless of which method of covariance estimation was used.

Contrast	df	X^2	method
$ASBP_1=0, ASBP_2=0$	2	32.6	DR
$ASBP_1=0, ASBP_2=0$	2	29.6	BHS

Looking at all interactions with sex and all interactions with age, it was found that a larger X^2 value was obtained with the age interactions. This suggested a model conditioning on age, so a saturated model including S and BP was formed within the <50 year age group (L) and within the ≥50 year age group (G). This model can be expressed in terms of the following parameters. [μ S(L) BP_1(L) BP_2(L) SBP_1(L) SBP_2(L) S(G) BP_1 (G) BP_2(G) SBP_1(G) SBP_2(G)].

The S and BP interactions were still highly significant in the <50 year age group.

Contrast	df	X^2	method
$SBP_1(L)=0, SBP_2(L)=0$	2	97.9	DR
$SBP_1(L)=0, SBP_2(L)=0$	2	65.6	BHS

In the ≥50 year age group, the amount of model reduction varied depending upon which method of covariance estimation was used. Describing first the results obtained when the covariance matrix was estimated by direct replication, the joint effect of sex and BP interactions was non-significant but when one looked at the sex and normal bloodpressure interaction alone it was significant at the .05 level. Thus it was decided that these interaction terms could not be left out of the model.

Contrast	df	X^2	method
$SBP_1(G)=0, SBP_2(G)=0$	2	4.83	.09
$SBP_1(G)=0$	1	4.71	.03*

Contrast	df	X^2	P
$SBP_2(G)=0$	1	.85	.36

When the BHS covariance matrix was used, the joint effect of sex and BP interactions within the $\geq$ 50 year age group was non-significant. In addition each individual S and BP interaction term was not significantly different from zero. These interaction terms could therefore be left out of the model.

Contrast	df	X^2	P
$SBP_1(G)=0, SBP_2(G)=0$	2	3.43	.18
$SBP_1(G)=0$	1	3.43	.06
$SBP_2(G)=0$	1	.87	.35

The final model using the direct replicated estimate of the covariance matrix required the following vector of parameters
$[\mu\ S(L)\ BP_1(L)\ BP_2(L)\ SBP_1(L)\ SBP_2(L)\ S(G)\ BP_1(G)\ BP_2(G)\ SBP_1(G)\ SBP_2(G)]$.

The final model using the BHS estimate of the covariance matrix needed two less parameters. Specifically, $SBP_1(G)$ and $SBP_2(G)$ could be omitted.

Bibliography

1. Deming, W.E. Sample Design in Business Research. New York: John Wiley and Sons, 1960.
2. Grizzle, J.E., Starmer, C.F., Koch, G.G. (1969). Analysis of categorical data by linear models. Biometrics. 25, 489-504.
3. Koch, G.G., Freeman, D.H., Freeman, J.L. (1975) Strategies in the multivariate analysis of data from complex surveys. International Statistical Review. 43, 59-78.
4. Koch, G.G., Lemeshow, S. (1972). An application of multivariate analysis to complex sample survey data. Journal of the American Statistical Association. 67, 780-782.
5. Mahalanobis, P.C. (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society. 109, 325-370.
6. McCarthy, P.J. (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. National Center for Health Statistics. Series 2, No. 14.
7. Plackett, R.L., Burman, J.P. (1946). The design of optimum multifactorial experiments. Biometrika, 33: 305-325.
8. Stanek III, E.J., Lemeshow, S. (1977). The behavior of balanced half-sample variance estimates for linear and combined ratio estimates when strata are paired to form pseudo-strata. ASA Proceedings of the Social Statistics Section 1977. 837-842.