Variance Estimation and Inference for Stratified Sample Designs with Nontrivial Sample Fractions and Nonresponse

John L. Eltinge
Office of Survey Methods Research, Bureau of Labor Statistics, PSB 1950
2 Massachusetts Avenue NE, Washington, DC 20212 Eltinge_J@bls.gov

Key Words: **Balanced repeated replication with Fay factors; Influence function; Quasirandomization model; Sparse-effect model; Stratum collapse; Variance estimator stability.**

Abstract:
In establishment surveys, sample designs often include stratification by the size of the establishment, with large establishments included in strata that have nontrivial sample fractions. For these designs, development and evaluation of appropriate variance estimators generally involve trade-offs among several factors, including the following.
(1) The intended use of the variance estimator, e.g., for formal inference or for construction of weights in a weighted-least-squares procedure.
(2) The extent, if any, to which one must account for nonresponse effects.
(3) The relative magnitudes of the effects of standard variance approximation methods, e.g., stratum collapse.
(4) Simplicity and ease of implementation within a given agency production environment.

This paper explores issues (1)-(4), with principal emphasis on diagnostics to identify acute problems with variance estimator bias or instability. The paper also adapts sparse-effect models from the experimental design literature to evaluate some properties of the proposed diagnostics.

1. Introduction

1.1 Multivariate Inference from Complex Survey Data

In the analysis of sample survey data, it is often important to carry out inference for a k-dimensional parameter vector

$$\theta_U = (\theta_{U1}, \ldots \theta_{Uk})' \qquad (1.1)$$

or an associated superpopulation vector,

$$\theta_\xi = (\theta_{\xi 1}, \ldots \theta_{\xi k})' \qquad (1.2)$$

For some general background, see, e.g., Fuller (1975), Rao and Scott (1981, 1984, 1987), Binder (1983), Skinner, Holt and Smith (1989), Korn and Graubard (1990), Pfeffermann (1996) and references cited therein.

In the current discussion, three points will be of principal interest. First, a complex sample design C is used to collect data from which a point estimator $\hat{\theta}_C = (\hat{\theta}_{C1}, \ldots \hat{\theta}_{Ck})'$ is obtained. Let $E_p(\bullet)$ and $V_p(\bullet)$ denote, respectively, the expectation and the variance-covariance matrix of a random vector evaluated with respect to the sample design. We assume that

$$E_p(\theta_C) \cong \theta_U \qquad (1.3)$$

in the sense that the difference $E_p(\theta_C) - \theta_U$ is small relative to other sources of variability.

Second, suppose that the variance-covariance matrix $V_p(\theta_C)$ is nonsingular, and that

$$\{V_p(\theta_C)\}^{-1/2}(\hat{\theta}_C-\theta_U)\to N(0,I_k) \qquad (1.4)$$

in law, where $A^{-1/2}$ represents the inverse of the symmetric matrix square root of a symmetric positive definite matrix A, and where the convergence in expression (1.4) refers to the limit of the multivariate distribution induced by the complex sample design C, conditional on a given realized finite population U.

Then we may use the distributional approximation (1.4) to carry out design-based inference for θ_U. For example, given a prespecified vector $\theta_0=(\theta_{01},\ldots\theta_{0k})'$, one may test the null hypothesis

$$H_0:\theta_U=\theta_0$$

using the test statistic

$$T_{C0}=(\hat{\theta}_C-\theta_0)'\{V_p(\hat{\theta}_C)\}^{-1}(\hat{\theta}_C-\theta_0) \qquad (1.5)$$

Under H_0 and condition (1.4), T_{C0} is distributed as a chi-square random variable on k degrees of freedom.

Third, in practical applications, we generally do not know $V_p(\hat{\theta}_C)$, but we often can compute a variance-covariance matrix estimator $\hat{V}_p$ (through linearization or replication methods) such that

$$d_p\hat{V}_p \to \text{Wishart } (d_p,V_p) \qquad (1.6)$$

for some appropriate scalar degrees-of-freedom term d_p, independent of $\hat{\theta}_C$. Then the associated quadratic-form test statistic,

$$\hat{T}_{C0}=(\hat{\theta}_C-\theta_0)'(\hat{V}_p)^{-1}(\hat{\theta}_C-\theta_0) \qquad (1.7)$$

is distributed as a multiple of an F random variable on k and d_p-k+1 degrees of freedom under H_0.

Fourth, if d_p-k+1 is relatively small, then $\hat{T}_{C0}$ may perform relatively poorly. In addition, in some cases it may be difficult or impossible to compute $\hat{V}_p$ due to confidentiality constraints or limitations on available software. Under such circumstances, one may consider an alternative test statistic,

$$T^*_{C0}=(\hat{\theta}_C-\theta_0)'(V^*_p)^{-1}(\hat{\theta}_C-\theta_0) \qquad (1.8)$$

where V^*_p is believed to be relatively stable (in a sense specified in Sections 2 and 3), and is computed readily from available data.

1.2 Multivariate Generalized Variance Functions

Properties of the test statistic T^*_{C0}, or of associated test-inversion confidence sets, will depend heavily on the properties of the random matrix V^*_p. The remainder of this paper studies these properties through extensions of ideas developed previously within the literature on univariate generalized variance functions. Section 2 develops a general framework for components of error associated with variance-covariance matrix estimators like $\hat{V}_p$ or V^*_p. Specifically, Subsection 2.1 considers general distinctions among sampling error, equation error and smaller-order parametric estimation error. Subsection 2.2 explores these distinctions further within the context of generalized variance function models. Subsection 2.3 discusses Rao-Scott type adjusted test statistics as special cases of the test statistic (1.8).

Section 3 applies some of the general ideas of Section 2 to a specific problem in the analysis of employment growth rates estimated from the U.S. Current Employment Survey. Of special interest is the fact that in the variance-covariance matrix estimator, there is a nontrivial component of error associated with error in the estimation of the underlying univariate generalized variance function. Section 4 reviews the main ideas of this paper and suggests some possible extensions.

2. Components of Error in Variance-Covariance Matrix Estimators

2.1. Sampling Error, Equation Error and Parametric Estimation Error

Section 1 focused attention on design properties of the point estimator $\hat{\theta}_C$ and the associated variance-covariance matrix estimator $\hat{V}_p$. For example, under conditions (1.4) and (1.6), we have, conditional on a given finite population U,

$$E_p(\hat{\theta}_C)=\theta_U \qquad (2.1)$$

$$V_p(\hat{\theta}_C)=V_p \qquad (2.2)$$

$$\hat{V}_p=V_p+\varepsilon_p \qquad (2.3)$$

where $E_p(\varepsilon_p)=0$. Extension of previous literature on generalized variance functions (e.g., Wolter, 1985, Chapter 5); Johnson and King, 1987; and Valliant, 1987) suggests that one view the finite population U as having been generated by a superpopulation model ξ such that

$$E_\xi \hat{\theta}_U)=\theta_\xi,$$

and

$$E_\xi(V_p | \theta_\xi, X, \gamma)=f(\theta_\xi, X, \gamma) \qquad (2.4)$$

where X is a matrix of available auxiliary information (e.g., sample sizes, the coefficient of variation of relevant weights, or other relevant design information), γ is an r×1 vector of unknown parameters, and $f(\cdot;\cdot;\cdot)$ is a k×k-dimensional matrix function of known form. Then we may define the *equation error*

$$c=V_p-f(\theta_\xi, X, \gamma), \qquad (2.5)$$

the difference between the realized random matrix V_p and its superpopulation expectation, $f(\theta_\xi, X, \gamma)$.

In addition, the parameters θ_ξ and γ generally are unknown, and X often is also unknown. Given specific estimators $\hat{\theta}_C$, $\hat{X}$ and $\hat{\gamma}$ (see, e.g., Valliant (1987) for discussion of ordinary least squares and generalized least squares methods for estimation of γ), define the parametric estimation error,

$$b=f(\hat{\theta}_C, \hat{X}, \hat{\gamma})-f(\theta_\xi, X, \gamma), \qquad (2.6)$$

Under mild conditions, the error b will depend on both ε_p and c, but generally will be of smaller order of magnitude than either ε_p or c. Now suppose the in computation of the test statistic (1.8), we have used

$$V_p^*=f(\hat{\theta}_C, \hat{X}, \hat{\gamma}), \qquad (2.7)$$

Then we may consider the properties of the test statistic (1.8) for three separate cases.

Case 1: Negligible equation error and estimation error. Suppose that $V_p^{-1}c$ and $V_p^{-1}b$ are both of small order of magnitude. Then the misspecification effect associated with the use of V_p^* in place of V_p is negligible, and the test statistic (1.8) will follow approximately (evaluated with respect to the sample design) a noncentral chi-square distribution on k degrees of freedom, and with noncentrality parameter equal to

$$(1/2)(\theta_U-\theta_0)'\{f(\theta_\xi,X,\gamma)\}^{-1}(\theta_U-\theta_0)$$

Case 2: Negligible equation error and nontrivial estimation error. Suppose that $V_p^{-1}c$ is of small order of magnitude, but that $V_p^{-1}b$ is not small. Thus, the estimation error b depends primarily on sampling error, ε_p , rather than on equation error; and V_p is approximately equal to $f(\theta_\xi,X,\gamma)$. Suppose further that for some d_b>0,

$$d_b f(\hat{\theta}_C,\hat{X},\hat{\gamma}) \rightarrow \text{Wishart}\left[d_b, f(\theta_\xi,X,\gamma)\right],$$

and is approximately independent of $\hat{\theta}_C$. (This final condition would hold, for example, if the function $f(\theta,X,\gamma)$ does not depend on θ. Then standard arguments (e.g., Korn and Graubard, 1990) indicate that, evaluated with respect to the design, the test statistic (1.8) is distributed approximately as a scalar multiple of a noncentral F random variable. The relative operating characteristics of the test statistics (1.8) and (1.7) will then depend primarily on the relative values of d_p and d_b.

Case 3: Negligible estimation error and nontrivial equation error. Suppose that $V_p^{-1}b$ is of small order of magnitude, but that $V_p^{-1}c$ is not small. Then two results are of practical interest. First, conditional on a given realization of the finite population U from the superpopulation ξ , use of V_p^* as an approximation to V_p would lead in general to a nontrivial associated misspecification effect matrix,

$$(V_p^*)^{-1/2}V_p(V_p^*)^{-1/2} \approx I_k + (V_p^*)^{-1/2}c(V_p^*)^{-1/2}$$

Thus, if one wished to use V_p^* (e.g., for stability reasons) and to remain within a design-based framework, the performance of (1.8) would depend on the eigenstructure of

$$(V_p^*)^{-1/2}c(V_p^*)^{-1/2}$$

and may in general be problematic.

Under Case 3, a possible alternative would be to consider the use of the test statistic (1.8) in a restricted form of $p\xi$ inference, i.e., inference with respect to both the design and model. Specifically, recall that evaluation a variance-covariance matrix with respect to both the design and superpopulation sources of variability leads to the expression,

$$V_{p\xi}(\hat{\theta}_C) = E_\xi\{V_p(\hat{\theta}_C)\} + V_\xi\{E_p(\hat{\theta}_C)\} \quad (2.8)$$

Furthermore, assume that

$$\{V_{p\xi}(\hat{\theta}_C)\}^{-1}V_\xi\{E_p(\hat{\theta}_C)\}$$

is negligible. (This would occur, for example, under a standard superpopulation model with independent and identically distributed variates, a negligible sample fraction, and additional regularity conditions.) Then $V_{p\xi}(\hat{\theta}_C)$ is approximately equal to $E_\xi\{V_p(\hat{\theta}_C)\}$, which by expression (2.4) is equal to $f(\theta_\xi, X, \gamma)$. Thus, under the assumption of negligible estimation error in Case 3, $V_p^* = f(\hat{\theta}_C, \hat{X}, \hat{\gamma})$ is approximately equal to $V_{p\xi}(\hat{\theta}_C)$.

If we also assume that the distribution of

$$\left\{V_{p\xi}(\hat{\theta}_C)\right\}^{-1/2}(\hat{\theta}_C - \theta_\xi),$$

evaluated with respect to the $p\xi$ distribution, is approximately $N_k(0, I_k)$, then the test statistic (1.8) has $p\xi$ distribution approximately equal to a noncentral chi-square distribution on k degrees of freedom and with noncentrality parameter

$$(1/2)(\theta_\xi - \theta_0)'\left\{f(\theta_\xi, X, \gamma)\right\}^{-1}(\theta_\xi - \theta_0)$$

Note that the abovementioned $p\xi$ approach to Case 3 used several additional assumptions. If these assumptions are not satisfied, then a $p\xi$ approach to Case 3 may be problematic. In particular, if

$$\left\{V_{p\xi}(\hat{\theta}_C)\right\}^{-1} V_\xi \left\{E_p(\hat{\theta}_C)\right\}$$

is not trivial, then $f(\hat{\theta}_C, \hat{X}, \hat{\gamma})$ can seriously underestimate $V_{p\xi}(\hat{\theta}_C)$, and the resulting misspecification effect matrix,

$$\left\{f(\theta_\xi, X, \gamma)\right\}^{-1/2} V_{p\xi}(\hat{\theta}_C)\left\{f(\theta_\xi, X, \gamma)\right\}^{-1/2}$$

$$= I_k + \left\{f(\theta_\xi, X, \gamma)\right\}^{-1/2} V_\xi\left\{E_p(\hat{\theta}_C)\right\}\left\{f(\theta_\xi, X, \gamma)\right\}^{-1/2}$$

may have one or more eigenvalues substantially greater than unity. Thus, in this setting it is preferable to construct a generalized variance function model that produces a good approximation for $V_{p\xi}(\hat{\theta}_C)$, rather than for $V_p(\hat{\theta}_C)$.

2.2 Links with Univariate Generalized Variance Function Models

The preceding subsection considered the variability of the equation error at a fairly high level of generality, without specific ference to a particular parametric superpopulation model. The application in Section 3 obtains some model identification information through specific distributional assumptions for related univariate generalized variance functions.

For the current discussion, consider a set of J ordered quadruples $(\hat{\theta}_{Cj}, X_j, V_{pj}, \hat{V}_{pj})$, $j = 1, \ldots, J$ associated with J distinct estimands θ_j. In addition, assume that on a logarithmic scale,

$$\ln(\hat{V}_{pj}) = \ln(V_{pj}) + e_{pj} \qquad (2.9)$$

and

$$\ln(V_{pj}) = X_j \gamma + q_j \qquad (2.10)$$

where e_{pj} , $j = 1, \ldots, J$ are independent and identically distributed normal random variables with mean μ_e and variance σ_e^2; q_j , $j = 1, \ldots, J$ are independent and identically distributed normal random variables with mean 0 and variance σ_q^2; and the errors e_{pj} and q_j are mutually independent. Note that the lognormal assumption (2.9) is not consistent with the assumption (1.6) of an approximate Wishart distribution for $d_p\hat{V}_p$. In the current discussion, we will not make direct inferential use of the Wishart assumption (1.6), but we will assume that on the original scale, $d_p\hat{V}_{pj}/V_{pj}$ has the same first and second moments as a chi-square random variable on d_p degrees of freedom, i.e.,

$$E_p(\hat{V}_{pj}/V_{pj})=1 \qquad (2.11)$$

and

$$V_p(\hat{V}_{pj}/V_{pj})=2/d_p \qquad (2.12)$$

Routine results for the lognormal distribution (e.g., Casella and Berger, 1990, p. 628) and additional algebra show that

$$\mu_e = (-1/2)\ln(1+2/d_p)$$

and $\sigma_e^2 = \ln(1+2/d_p)$. In addition, under the assumption that the quadruples $(\hat{\theta}_{Cj}, X_j, V_{pj}, \hat{V}_{pj})$ are independent, ordinary least squares regression of $\ln(\hat{V}_{pj})$ on X_j leads to consistent estimators $\hat{\gamma}$ of γ and $(M\hat{S}E)$ of $\sigma_q^2 + \sigma_{ep}^2$. In addition,

$$\hat{\sigma}_q^2 = (M\hat{S}E) - \ln(1+2/d_p)$$

provides a consistent estimator of σ_q^2 under the assumptions listed above; and

$$V_{pj}^* = \exp(M\hat{S}E/2 + X_j\hat{\gamma})$$

is a consistent estimator of $f(\theta_\xi, X, \gamma) = E_\xi\{V_p(\hat{\theta}_C)\}$.

Also, additional algebra shows that ,

$$V_\xi\{V_{pj} - f(\theta_\xi, X_j, \gamma)\} = \{f(\theta_\xi, X_j, \gamma)\}^2 \exp(\sigma_q^2 - 1).$$

can be estimated by $\{V_{pj}^*\}^2 \exp(\hat{\sigma}_q^2 - 1)$. As noted above, we are using lognormal assumptions for both the sampling error e_{pj} and the equation error q_j. Nonetheless, analysts often summarize the stability of a variance estimator through a moment-based "degrees of freedom" calculation, which in this case would be,

$$\hat{d}_q = 2/\exp(\hat{\sigma}_q^2 - 1) \qquad (2.13)$$

In this case, expression (2.13) represents the uncertainty in V_{pj}^* as a predictor of the random variable V_{pj}, which is subject to the equation error q_j on the logarithmic scale. Finally, consider again Case 2, in which one has negligible equation error and nontrivial estimation error. Then additional routine algebra leads to the Satterthwaite-type "degrees of freedom" estimator,

$$\hat{d}_b = 2\left[\left\{(1/2, X_j)\hat{V}(\hat{\sigma}_q^2, \hat{\gamma})(1/2, X_j)'\right\}^{-1} - 1\right] \qquad (2.14)$$

Expression (2.14) represents the uncertainty in V_{pj}^* as an estimator of $f(\theta_\xi, X_j, \gamma) = V_{pj}$ under Case 2. Consequently, comparison of d_p, $\hat{d}_q$ and $\hat{d}_b$ can provide a rough indication of the relative magnitudes of, respectively, the sampling errors, equation error effects and estimation error effects considered in Cases 1 through 3.

2.2. Misspecification Effect Matrices and Rao-Scott Adjusted Test Statistics

Section 2.1 discussed misspecification effect matrices within the context of the equation error model (2.5). The original work by Rao and Scott (1981, 1984, 1987) on quadratic-form test statistics focused principal attention on approximations for the distribution of a test statistic under a null hypothesis associated with the parameter vector θ. We note, however, that to some degree, the adjusted matrices used in the Rao-Scott test statistics can be viewed as simple multivariate variance function estimators.

3. Application to Data from the U.S. Current Employment Survey

We applied the principal ideas of Section 2 to estimates of total employment for a large number of domains covered by the U.S. Current Employment Survey (CES). For some background on the Current Employment Statistics Survey, see American Statistical Association (1994), Werking (1997), Butani, Harter and Wolter (1997), Butani, Stamas and Brick (1997), West, Kratzke and Grden (1997). For the present discussion, four features are of principal interest. First, domains were defined by the intersection of three factors:

Industry $i = 1, \ldots, I$; $I = 6$ (Mining; mining and construction combined; construction; durables goods manufacturing; nondurable goods manufacturing; and wholesale trade. For some areas, mining and construction are combined, while for other areas, they are treated as distinct industries.)

Area $a = 1, \ldots, A$ ($A = 272$ metropolitan areas in the United States).

Month $t = 1, \ldots, 12$ (January through December, 2000).

To reflect the Industry$\times$Area$\times$Month structure used to define the domains of interest, the estimand subscript j used in Section 2 will be replaced by the triple subscript (i, a, t). For instance, for industry i, area a, and month t, we have

y_{iat} = True total employment;

$\hat{y}_{iat}$ = A direct (weighted link relative) estimator of y_{iat}, based only on data from industry i and area a.

$\hat{V}_{iat}$ = An estimator of the design variance of $\hat{y}_{iat}$, computed through standard fractionally weighted methods of balanced repeated replication.

4. Acknowledgements

The authors thank Larry Huff and Julie Gershunskaya for useful discussions of generalized variance functions; and for providing the data used in Section 3. The views expressed in this paper are those of the authors and do not necessarily represent the policies of the U.S. Bureau of Labor Statistics.

5. References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51** 279-292.

Butani, S., Harter, R., and Wolter, K. (1997). Estimation Procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 523-528.

Butani, S., Stamas, G. and Brick, M. (1997). Sample
Redesign for the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 517-522.

Casella, G. and Berger, R. L (1990). Statistical inference. Wadsworth.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhya C* **37**, 117- 132.

Johnson, E.G., and King, B.F. (1987). Generalized variance functions for a complex sample survey. *Journal of Official Statistics*, 3, 235-250.

Korn, E. L. andGraubard, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics. *The American Statistician* **44** 270-276

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research* **5,** 239- 261.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* **76** 221- 230

Rao, J. N. K. and Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics* **12**, 46-60.

Rao, J. N. K. and Scott, A. J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics* **15**, 385-397.

Skinner, C. J., Holt, D., and Smith, T. M. F., Eds. (1989). *Analysis of complex surveys.* New York: Wiley.

Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling. *Journal of the American Statistical Association* **82** 499-508.

Werking, G. (1997). Overview of the CES Redesign. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 512-516.

Wolter, K.M. (1985). Introduction toVariance Estimation. New York: Springer Verlag.

Table 1: Tests for Homogeneity of Growth Rates Across Areas, By Industry. Pennsylvania, June 2000

Industry	A	$\hat{d}_{bi}$	T^*_{M0i}	Critical value
Durables	11	20.8	7.1	48.6
Nondurables	9	27.4	2.4	26.2
Wholesale	4	56.2	2.7	8.6