EVALUATION OF DESCRIPTIVE ANALYSES OF SURVEY VARIANCES AND CONFIDENCE INTERVAL WIDTHS

D.S. Jang, Mathematica Policy Research and J.L. Eltinge, U.S. Bureau of Labor Statistics

Key words: Average design effect; Average standard error; Complex sample design; Generalized variance function; Parametric inference; Subpopulation.

Sample survey organizations often characterize the precision of point estimators through approximations based on, e.g., average design effects, average standard errors, generalized variance functions or average confidence interval widths. Practical evaluation of the adequacy of these approximations will depend on the whether one is interested in: (1) summary descriptions of a fixed set of variance estimates or confidence interval widths; or (2) formal inference (e.g., confidence intervals or test statistics) for related functions of finite-population or superpopulation parameters (e.g., design effects or specific coefficients of GVF models). Following a review of issues (1) and (2), this paper uses fixed-effect analysis of variance methods to develop some specific diagnostics for issue (1). Some of the proposed diagnostics are applied to data from the U.S. Third National Health and Nutrition Examination Survey (NHANES III).

I. Introduction

1.1 Summary descriptions of confidence interval widths

Statistical agencies frequently publish analysis results for a large number of relatively fine subpopulations formed by the intersection of several classification factors. For example, some agencies in the United States publish estimates for subpopulations defined by Age × Race × Sex classifications. In developing policies on the publication and interpretation of fine-subpopulation estimates, an important consideration is estimation uncertainty, as measured by confidence interval width.

At one extreme, one could account for estimation uncertainty through a simple rigid numerical rule, e.g., "For disease variable *X*, publish prevalence rates only for subpopulations with 95% confidence interval ha1f-widths less than 0.04." This extreme has the advantage of ensuring that each published estimate achieves a specified level of estimated precision. However, the resulting mixture of "published" and "unpublished" cells can produce a "swiss cheese" pattern of missing data that is likely to confuse or frustrate many data users. Also, to some degree the resulting publication pattern will be artefact of the sampling variability of the interval half-widths, and thus may be unstable across replications of the survey. At the other extreme, one could have a publication policy that does not include any explicit precision criteria, e.g., "Publish prevalence rates for all Age × Race, Age × Sex, and Race × Sex classifications, but not for any individual Age × Race × Sex cells." The resulting published tables will have a relatively simple structure, but may include some very imprecise estimates and omit other estimates that are more precise. Between these two extremes, one may make publication decisions for relatively simple groups of cells, where the decision for a given group is informed by summaries of the precision of point estimates for the member cells. For example, Section 3 will discuss confidence interval widths for a health survey; preliminary discussion of this survey anticipated that Age × Sex subpopulations for Mexican-Americans might have substantially wider intervals than Age × Sex subpopulations for other rare-ethnic groups. If this conjecture were supported by the data, then one might advise against publication of Age × Sex estimates for Mexican-Americans, or one might publish these estimates with special cautionary notes that emphasize limitations on precision.

Within the decision process, it is important to use relatively simple summaries of observed patterns of confidence interval widths, without missing salient features of these patterns. Consequently, three important questions are as follows.

(1) Within a given set of subpopulations, do confidence interval widths differ substantially (relative to mean width) across the levels of a given factor, e.g., race?

(2) Does the variability observed in (1) account for a substantial amount of the overall variability in confidence interval widths?

(3) To what extent are the answers for (1) and (2) sensitive to prospective changes in variance estimation?

If the answers to both (1) and (2) are "yes" for a given factor, then that factor may be useful in developing the abovementioned summary descriptions. Conversely, if a given factor does not satisfy both criteria (1) or (2), then summary

descriptions based on that factor may be relatively uninformative, and possibly misleading. Similarly, pronounced sensitivity, as defined by (3), indicates that variance estimation issues warrant further examination before the agency reaches to final decisions on publication of subpopulation estimates.

1.2 Outline of proposed exploratory methods and results

This paper shows that coefficient estimates and diagnostics from a fixed-main-effects analysis of variance (ANOVA) model can provide useful tools to address questions (1) and (2). This in turn allows one to identify specific variables and subpopulation factors for which it is useful to draw distinctions regarding confidence interval widths. Throughout this work, the models in question are used in an exploratory and descriptive manner, and not for purposes of formal inference.

Section 2 presents the main ideas used in this paper. Subsection 2.1 motivates our emphasis on confidence interval widths, rather than on the more commonly examined variance estimates. Subsection 2.2 presents the principal notation, models and methods used here. Subsection 2.3 discusses some diagnostics based on main-effect treatment mean squares and on R^2 statistics. Section 3 applies the proposed methods to Age × Race × Sex subpopulation confidence intervals computed with data from Phase I of the U.S. Third National Health and Nutrition Examination Survey (NHANES III). Subsection 3.1 describes the salient aspects of the NHANES III design and variables of interest. Subsection 3.2 uses the fixed-main-effect coefficient estimates, and related diagnostics, to describe variability of interval widths across groups of subpopulations. Subsection 3.3 examines some related two-factor interactions. Section 4 reviews the main ideas presented here.

2. Confidence Interval Widths and Related Fixed-Main-Effects Models

2.1 Motivation for modeling of confidence interval widths for $\hat{\theta}$

To begin the discussion, consider a subpopulation parameter 2, an associated point estimator $\hat{\theta}$, and a variance estimator $\hat{V}(\hat{\theta})$. We will examine estimation uncertainty through the widths of the nominal $(1-\alpha)100\%$ confidence intervals

$$\hat{\theta} \pm t_{\hat{d},\alpha/2}\{\hat{V}(\hat{\theta})\}^{1/2} \qquad (2.1)$$

where $\hat{d}$ is a "degrees of freedom" measure computed from the available data, and $t_{\hat{d},\alpha/2}$ is the associated upper $\alpha/2$ quantile of a *t* distribution on $\hat{d}$ degrees of freedom.

In previous work with descriptions of, or models for, uncertainty, the sample survey literature has tended to emphasize variance estimators $\hat{V}(\hat{\theta})$, rather than confidence interval widths as such. See, for example, the discussion of generalized variance functions in Wolter (1985, Chapter 5), Valliant (1987) and references cited therein. Consequently, it is important to motivate clearly our present emphasis on confidence interval widths. For estimation at a full-population level, the effective degrees of freedom term $\hat{d}$ tends to be relatively large. Thus, at this level, the $t_{\hat{d},\alpha/2}$ multipliers tend to display relatively little variability, and modeling of confidence interval widths is roughly equivalent to modeling of the standard errors $\{\hat{V}(\hat{\theta})\}^{1/2}$. Also, if one intends to use modeling results to produce improved measures of uncertainty, then one generally prefers to focus on modeling of the $\hat{V}(\hat{\theta})$, and to consider confidence interval construction only after one produces an improved variance estimator and an associated new "degrees of freedom" measure.

By contrast, in the present work we focus on subpopulation estimates, and we intend strictly to provide summary descriptions of estimation uncertainty. Since confidence intervals (or equivalent hypothesis tests) are the predominant method for reflecting uncertainty in formal statistical inference work with survey data, we prefer to focus our descriptive methods on confidence interval widths, rather than on the variance estimates $\hat{V}(\hat{\theta})$ by themselves. Moreover, for some variables studied in Section 3, the multipliers $t_{\hat{d},\alpha/2}$ varied substantially across subpopulations. Consequently, at the fine-subpopulation level, descriptive results for confidence interval widths are not necessarily equivalent to results for the standard errors $\{\hat{V}(\hat{\theta})\}^{1/2}$.

2.2 *Estimation and interpretation of fixed-main-effects models*

Now consider a set of subpopulations defined by several classifications. To keep notation relatively simple, we will restrict attention to three classification factors: F_1, with levels $i = 1,2,...,I$;

F_2, with levels $j = 1,2,...,J$; and F_3, with levels $k = 1,2,...,K$; extension to more factors is straightforward. For a given sub population defined by the triple (*i,j,k*), let θ_{ijk} be the subpopulation parameter of interest and let $w_{ijk} = 2t_{\hat{d},\alpha/2}\{\hat{V}(\hat{\theta}_{ijk})\}^{1/2}$ be the width of the customary confidence interval defined by (2.1). A fixed-main-effects model for w_{ijk} is

$$w_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \text{error} \qquad (2.2)$$

where μ, α_i, β_j and τ_k are fixed coefficients, and the estimability of these coefficients is ensured by the constraints $\sum_{i=1}^{I}\alpha_i = \sum_{j=1}^{J}\beta_j = \sum_{k=1}^{K}\tau_k = 0$. Due to balance across the three factors, standard least squares methods (e.g., Draper and Smith, 1981, Chapter 9) lead to estimates $\hat{\mu} = \bar{w}... = (IJK)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} w_{ijk}$, $\hat{\alpha}_i = \bar{w}_{i..} - \bar{w}...$, $\hat{\beta}_j = \bar{w}_{.j.} - \bar{w}...$ and $\hat{\tau}_k = \bar{w}_{..k} - \bar{w}...$, where $\bar{w}_{i..} = (JK)^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K} w_{ijk}$, and $\bar{w}_{.j.}$ and $\bar{w}_{..k}$ are defined similarly. Slightly more complex least-squares computations are used for unbalanced cases that exclude one or more (*i,j,k*) cells.

We emphasize that the motivating expression (2.2) and the resulting estimates are used here strictly to provide a descriptive summary of patterns of variability in subpopulation confidence interval widths, and not to draw formal statistical inferences regarding the conceptual parameters μ, α_i, β_j, or τ_k. To some degree, a similar interpretation arises in the generalized-variance-function literature, where one generally does not carry out formal inference for variance-model coefficients.

The estimates $\hat{\alpha}_i$, $\hat{\beta}_j$ and $\hat{\tau}_k$ give a simple indication of whether the interval widths differ substantially across groups defined by a given classification factor *i, j* or *k*. For example, suppose that for a given $i = 1,2,...,I$, the estimate $\hat{\alpha}_i$ is substantially larger than zero. This occurs if and only if the reported interval widths for subpopulations (*i,j,k*), averaged across the levels $j = 1,2,...,J$ and $k = 1,2,...,K$, are substantially larger than the corresponding average reported interval widths for subpopulations with classification factor F_1 not equal to *i*.

2.3 Two related sets of diagnostics

Two other ANOVA quantities are useful in describing the variability of confidence interval widths. First, consider the mean square $MS(F_1)$ associated with the factor F_1, define the ratio $\hat{\delta}_{F1} = \bar{w}_{...}^{-1}\{MS(F_1)\}^{1/2}$, and define $\hat{\delta}_{F2}$ and $\hat{\delta}_{F3}$ similarly. Within the context of question (1) from Section 1, a large value of $\hat{\delta}_{F1}$ (greater than one, say) indicates that the variability of the means $\bar{w}_{i..}$ across $i = 1,2,...,I$ is substantial, relative to the overall mean $\bar{w}_{...}$. This suggests that in reporting confidence intervals for (*i,j,k*)-level subpopulations, the statistical agency may need to consider separate publication and interpretation guidelines for different subpopulation groups formed according to $i = 1,2,...,I$.

Second, note that for the confidence-interval-width analysis, the customary ratio R^2 represents the proportion of observed variability in the widths w_{ijk} that is attributable to the main effects α_i, β_j and τ_k in expression (2.2). Similar comments apply to the related ratios R^2_{F1}, R^2_{F2}, and R^2_{F3}, where, e.g., $R^2_{F1} = (SSCT)^{-1}SS(F_1)$, $SSCT$ is the usual "corrected total sum of squares" and $SS(F_1)$ is the sum of squares associated with main effect F_1.

Within the context of question (2) from Section 1, a large value of R^2_{F1} (greater than 0.5, say) indicates that much of the variability of the w_{ijk} arises from differences across the main-effect groups defined by the factor F_1. In such cases, the analysis results from (2.2) may be useful in developing guidelines for publication of estimates. Conversely, relatively small values of R^2_{F1} (less than 0.2 or 0.3, say) indicate that summary descriptions of interval widths based on (2.2) may be relatively uninformative. Finally, if addition of some two-factor interactions to the main-effects expression (2.2) leads to a large absolute increase in R^2, then one may prefer to include these interactions in (2.2) when attempting to address questions (1) and (2). Inclusion of such interaction terms may be numerically important in some cases, but necessarily leads to greater complexity in the summary description of the observed patterns of confidence interval widths.

2.4 Two alternatives based on relative widths

For the application discussed in Section 3, the confidence interval widths w_{ijk} were of principal interest, so we have focused our analyses on these widths. However, in other applications there also may be interest in analyses for two related quantities. The general exploratory methods presented in this paper carry over to these two quantities.

First, analysis of the rescaled widths $\hat{\theta}_{ijk}^{-1} w_{ijk}$ would describe the variability of the interval widths w_{ijk}, relative to the associated point estimates $\hat{\theta}_{ijk}$. These rescaled widths are confidence interval analogues of the coefficient of variation, and would be important to a survey organization that emphasizes *relative* precision in the development of publication guidelines.

Second, one could apply the proposed methods to the ratios $t_{m-1,\alpha/2}^{-1} \hat{V}_{SRS}^{-1/2}(\hat{\theta}_{ijk}) w_{ijk}$, where $t_{m-1,\alpha/2}$ is the customary upper $\alpha/2$ quantile of a *t* distribution on $m-1$ degrees of freedom, *m* is the number of sampled elements, and $\hat{V}_{SRS}(\hat{\theta}_{ijk})$ is an estimator of the design variance of $\hat{\theta}_{ijk}$ derived under the (incorrect) assumption that our survey data were collected through of simple random sampling of elements. In this second case, the ratios of interest involve a form of "misspecification effect" for confidence interval construction. Application of expression (2.2) to this second set of ratios leads to an exploratory analysis of "confidence interval misspecification effects," e.g., identification of main-effect factors for which these misspecification effects are especially severe. For a general discussion of design effects and misspecification effects associated with variance estimation and related confidence interval issues, see, e.g., Skinner et al. (1989, Section 2.2).

3. Application to the U.S. Third National Health and Nutrition Examination Survey

3.1 NHANES III sample design, stratum collapse, estimation methods and survey items

The methods outlined in Section 2 were used to analyze the widths of confidence intervals for Age x Race x Sex subpopulations computed with data from Phase I of the U.S. Third National Health and Nutrition Examination Survey (NHANES III). National Center for Health Statistics (1996) provides a detailed description of NHANES III; for the present discussion, the following three points are especially important. First, the survey was based on a stratified multistage design, where the survey elements were individuals in the U S. civilian noninstitutionalized population.

Second, Phase I of the survey was carried out over a three-year period (1988-1991); the design can be viewed as involving a total 44 strata (large groups of counties), with one primary sample unit (generally an individual county) selected per stratum. Consequently, some stratum collapse was required to carry out customary design-based variance estimation and confidence interval construction. The primary collapse method considered here led to a total of 22 collapsed strata, with two selected primary units per stratum.

Third, standard design-based methods were used to compute point estimates and variance estimates for thirty-six subpopulations defined by the intersection of six age groups (20-29, 30-39, 40-49, 50-59, 60-69 and 70+), three race-ethnic groups (1: Mexican-American; 2: Black non-Hispanic; and 3: White/Other) and two sex groups (1: Male; and 2: Female). For each of the 12 interview and examination variables listed in Table 3.1, nominal 95% confidence intervals for the subpopulation means were constructed by expression (2.1). In this work, the degrees-of-freedom estimates $\hat{d}$ were computed through a modification of the Satterthwajte (1946) method discussed in Jang and Eltinge (1995). The standard "degrees of freedom" calculation (e.g., Frankel, 1971, p. 89; Skinner et *al.,* 1989, p. 57; or Korn and Graubard, 1995, p. 278), equal to (number of PSUs) – (number of strata), was considered too optimistic to use in this application, due to marked heterogeneity of variances across strata.

Table 3.1: Twelve NHANES III Variables

Var. #	Variable Name	Description
1	HDRESULT	HDL cholesterol
2	TCRESULT	Serum total cholesterol
3	LEAD	Blood lead
4	log(LEAD)	Natural log of blood lead
5	HAE2	Diagnosed hypertension
6	HAE7	Diagnosed high cholesterol
7	HAD1	Diagnosed diabetes
8	HAR3	Smoking now?
9	BMPHT	Height
10	BMPWT	Weight
11	BP1K1	Systolic blood pressure
12	BP1K5	Diastolic blood pressure

3.2 Comparison of estimated main effects

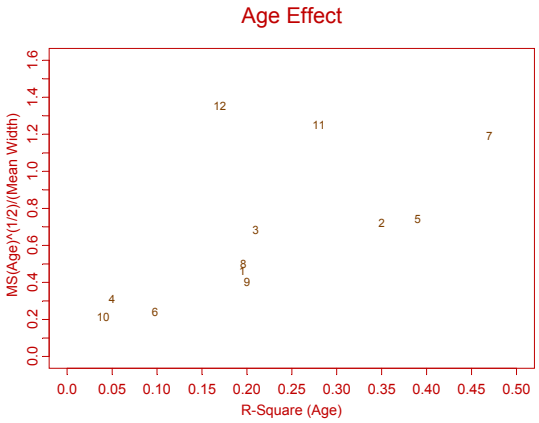
3.2.1 Age effects

As a preliminary screening device, Figure 3.1 presents a plot of $\hat{\delta}_{Age}$ against R^2_{Age} for the twelve NHANES variables. The plotting symbol is the variable number, as listed in Table 3.1. At one extreme are variables 4 (log(lead)), 6 (diagnosed high cholesterol) and 10 (weight), located in the lower left-hand corner of the plot. The low values of $\hat{\delta}_{Age}$ and R^2_{Age} indicate that mean observed confidence interval widths do not vary in a pronounced manner across age groups. Thus, administrative distinctions among age groups are relatively uninformative for these three variables.

At the other extreme is variable 7 (diagnosed diabetes), which has $\hat{\delta}_{Age} = 1.2$ and $R^2_{Age} = 0.47$. This variable has observed confidence interval widths that show a substantial amount of variability across age groups, relative to the mean width $\hat{\mu}$; and the age main effect accounts for almost one- half of the overall variability of the observed widths. Thus, for variable 7, the age main effect is fairly "interesting" in the sense defined by both questions (1) and (2) in Section 1.

Variables 11 (systolic blood pressure) and 12 (diastolic blood pressure) also have $\hat{\delta}_{Age} > 1$. However, their R^2_{Age} values are relatively small (0.28 and 0.17, respectively). Thus, the age main effects are fairly large relative to $\hat{\mu}$, but do not account for the bulk of the variability in widths for these two variables. This illustrates an important distinction between the dimensions of variability reflected by the diagnostics $\hat{\delta}_{Age}$ and R^2_{Age} respectively.

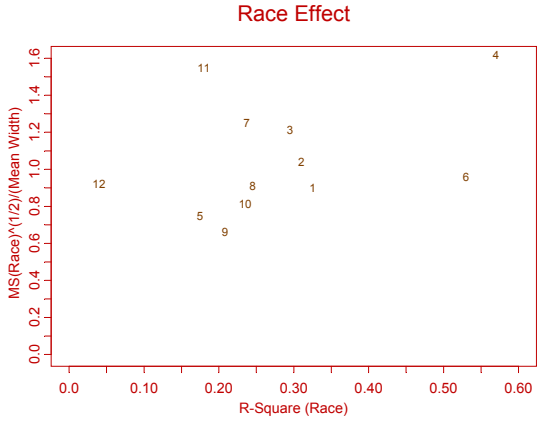
Figure 3.1: Plot of Relative Contribution of Age vs. R-Square (Age)



3.2.2 Race effects

Figure 3.2 presents a plot of $\hat{\delta}_{Race}$ against R^2_{Race}. Variables 2 (total cholesterol), 3 (lead), 4, 7 and 11 have $\hat{\delta}_{Race} > 1$, and variable 6 (diagnosed high cholesterol), has $\hat{\delta}_{Race} = 0.97$. In addition, variables 4 and 6 have $R^2_{Race} = 0.57$ and 0.53, respectively, while all other variables have $R^2_{Race} < 0.4$. Thus, for variables 4 and 6, there is some justification for drawing distinctions among race groups in summary description and interpretation of confidence interval widths.

Figure 3.2: Plot of Relative Contribution of Race vs. R-Square (Race)



3.2.3 Sex effects

Work that will not be detailed here indicated that, in the sense defined by questions (1) and (2), the sex main effects were somewhat less "interesting" than the age and race effects. For example, all twelve variables had $R^2_{Sex} < 0.2$.

3.3 Two-factor interactions

Figures 3.1 and 3.2 identified several combinations of variables and main effects *F* that had relatively large values of $\hat{\delta}_F$, but relatively small values of R^2_F. Some examples include variables 11 and 12 for both Age and Race main effects. For such cases, it is useful to study whether two-factor interactions account for a substantial amount of the remaining "unexplained" variability. Thus, we fit an expanded form of model (2.2) that included all main effects, and two-factor interactions associated with a given pairing of the age, race and sex factors. The resulting diagnostics δ and R^2 were computed for each variable and each set of two-factor interactions.

For the Age × Race interaction, variables 11 and 12 had $\hat{\delta}_{Age\times Race}$ equal to 1.05 and 1.72, respectively, with associated $R^2_{Age\times Race}$ equal to 0.39 and 0.54. Thus, for these two variables, it is not advisable to use a pure main-effects model in attempting to address questions (1) and (2). By contrast, no other variables had either $\hat{\delta}_{Age\times Race} > 1$ or $R^2_{Age\times Race} > 0.3$.

The estimated Age × Sex and Race × Sex interactions identified relatively few patterns of interest. For example, variables 11 and 12 had $\hat{\delta}_{Age\times Sex}$ equal to 0.57 and 1.22, respectively, but had fairly weak $R^2_{Age\times Sex}$ values, equal to 0.06 and 0.14. No other variable had either $\hat{\delta}_{Age\times Sex} > 0.5$ or $R^2_{Age\times Sex} > 0.2$. Also, no variable had either $\hat{\delta}_{Race\times Sex} > 0.5$ or $R^2_{Race\times Sex} > 0.1$.

4. Discussion

When a statistical agency works with a large number of subpopulations, there is a strong incentive to summarize confidence-interval-width patterns through simple descriptive statistics. This paper has shown that a fixed-main-effects analysis of variance (ANOVA) provides a simple framework for this summary description. Equally important, related diagnostics such as $\hat{\delta}_F$, R^2_F, and two-factor-interaction estimates, provide useful indications of the extent to which a given summary description is potentially informative, or potentially misleading.

We illustrated both ideas with an application of the proposed methods to data from Phase I of NHANES III. For example, the race main effects for variables 4 (log(blood lead)) and 6 (diagnosed high cholesterol) were strong enough that one might reasonably consider separate guidelines for interpretation of confidence interval widths for Age × Sex subpopulations in the three race groups. On the other hand, variables 9 (height) and 10 (weight) displayed relatively weak main effects. Thus, in a discussion of Age × Race × Sex sub population confidence interval widths, distinctions among simple age, race or sex groups would be relatively uninformative for variables 9 and 10. Finally, a balanced examination of $\hat{\delta}_F$, R^2_F, and two-factor interactions indicated a considerably more complex pattern of widths for the variables 11 (systolic blood pressure) and 12 (diastolic blood pressure); for these variables, direct interpretation of simple group mean widths could be misleading.

5. References

Draper, N. and Smith, H. (1981). *Applied Regression Analysis, Second Edition.* New York: Wiley.

Ezzati, TM., Massey, JT., Waksberg, J., Chu, A. and Maurer, KR. (1992). *Sample Design: Third National Health and Nutrition Examination Survey.* National Center for Health Statistics, Vital and Health Statistics Series, 2, Number 113. Hyattsville, MD.

Frankel, M.R. (1971). *Inference from Survey Samples: An Empirical Investigation.* Ann Arbor, Michigan: University of Michigan Institute for Social Research.

Jang, D.S. and Eltinge, JL. (1995), "Empirical Assessment of the Stability of Variance Estimators Based on a Two-Clusters-Per-Stratum Design." Technical Report #225, Department of Statistics, Texas A&M University. Submitted for publication.

Korn, E.I. and Graubard, B.G. (1995), "Analysis of Large Health Surveys: Accounting for the Sampling Design," *J. Royal Statist. Sac., Ser. A* 158, 263-295.

National Center for Health Statistics (1996), NHANES III Reference Manuals and Reports, CD-

ROM GPO, 017-022-1358-4. Washington, D.C.: United States Government Printing Office.

Rust, K. and Kalton, G. (1987), "Strategies for Collapsing Strata for Variance Estimation," *J. Off. Statist.* 3, 69-81.

Satterthwaite, F. E. (1946). "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2, 110-114.

Skinner, C.J., Holt, D. and Smith, T MF. (1989). *Analysis of Complez Surveys.* New York: Wiley.

Valliant, R. (1987), "Generalized Variance Functions in Stratified Two-Stage Sampling," *J. Amer. Statist. Assoc.* 82, 499-508.

Wolter, K. M. (1985), *Introduction to Variance Estimation,* New York: Springer-Verlag.