Journal of Official Statistics, Vol. 26, No. 4, 2010, pp. 585–602

Degrees of Freedom Approximations and Rules-of-Thumb

Richard Valliant[1] *and Keith F. Rust*[2]

In complex samples, *t*-distributions are used when performing hypothesis tests and constructing confidence intervals. Rules-of-thumb are typically used to approximate degrees of freedom for the *t*-distributions. The standard rule is to set the degrees of freedom equal to the number of primary sampling units minus the number of strata. We illustrate some circumstances where these rules can be poor. A simple estimate of degrees of freedom is presented that leads to improved confidence interval coverage.

Key words: Complex samples; kurtosis; linearization variance estimator; nonlinear estimators; skewness; variance of variance estimator.

1. Introduction

Analysts of survey data use approximate degrees of freedom (*DF*) to account for imprecision of variance estimates when computing confidence intervals and performing hypothesis tests. Rust (1984, 1985, 1986) and Rust and Rao (1996) review the problem of approximating degrees of freedom in a variety of situations, particularly for replication variance estimators. Rust (1984) and Rust and Kalton (1987) address the effect of collapsing strata on *DF* approximations. Eltinge and Jang (1996) cover the problem of gauging the stability of variance component estimators in complex designs. Korn and Graubard (1990; 1999, Section 5.2) discuss approximating degrees of freedom for analytic statistics.

Software packages typically use the rule-of-thumb that the *DF* for a variance estimator is the number (n) of primary sampling units (PSUs) minus the number of strata (H). One motivation for this rule-of-thumb is to suppose that a variance estimator has a chi-square distribution and to apply an approximation due to Satterthwaite (1946). As shown in Section 4, the Satterthwaite approximation specializes to $n - H$ under some restrictive conditions. Stata® (Stata Corporation 2005), SUDAAN® (Research Triangle Institute 2004), and WesVar® (Westat 2000), among other packages, use this approximation unless it is over-ridden by a user. This rule-of-thumb works under the assumptions described in Section 4, but can be poor for many variables. The fact that the rules-of-thumb can be faulty in some circumstances is known, or at least suspected, by practitioners. Johnson and

[1] Survey Research Center, University of Michigan and Joint Program in Survey Methodology, University of Maryland, 1218 Lefrak Hall, College Park, MD 20742, U.S.A. Email: rvalliant@survey.umd.edu

[2] Westat and Joint Program in Survey Methodology, University of Maryland, 1650 Research Boulevard, Rockville, MD 20850-3195, U.S.A. Email: rustk1@westat.com

Acknowledgments: The work of the first author was partially supported by grant SES-0416662 from the National Science Foundation (NSF). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We thank the Associate Editor and referees for their useful comments.

Rust (1992) and Johnson, Rust, and Hansen (1988) found that the effective *DF* for jackknife variance estimators depended on whether estimates were for the full population or subpopulations. Subpopulations that occurred in only some of the strata or PSUs had fewer *DF* than suggested by the standard rules-of-thumb. Graubard and Korn (1996) and Burns et al. (2003) proposed that the *DF* for domain estimates be modified to account for the possibility that a domain may not occur within all primary sampling units (PSUs) in a design. Their suggested rule-of-thumb was (number of PSUs with sampled members of the domain) minus (number of strata with sampled members of the domain). Although overriding the default in a software package is usually permitted, users may not have enough knowledge to do this with any degree of accuracy.

In this article, we present some theory for the usual rules-of-thumb and some illustrations of when they can be far from correct. Section 2 describes Satterthwaite approximations; Section 3 summarizes calculations of the variance of a variance estimator, which is a key ingredient in Satterthwaite. The fourth section covers some special cases that lead to the rules-of-thumb. In Section 5, we present some simulation results that show when the rules-of-thumb are acceptable and when they are poor. We also give an estimator of the *DF* that is simple to compute and is an improvement over the rules-of-thumb. The last section summarizes our results and recommendations.

2. Satterthwaite Approximations

Suppose a stratified probability sample s_h of n_h primary sampling units (PSUs) from a universe U_h of PSUs is selected from stratum h ($h = 1, \ldots, H$) with replacement. The total number of sample PSUs is $n = \sum_h n_h$. Throughout this article, we use the design-based approach to inference. We will refer to a plan in which PSUs are selected with varying probabilities with replacement as PPSWR. Although with-replacement (WR) sampling is rarely used in practice, the design-based theory for WR is more tractable and often more enlightening than for without-replacement sampling. Consequently, assuming WR sampling is a standard simplification for many theoretical analyses in sampling theory.

A sample of elementary units is selected within each PSU and $j = 1, \ldots, J$ variables are collected on each unit in such a way that a design-unbiased estimator of the PSU total of each variable can be constructed. The vector of population totals of the J variables is $\mathbf{t} = (t_1, \ldots, t_J)^T$. The estimator of the population total for variable j is $\hat{t}_j = \sum_h \hat{y}_{jh}$ where $\hat{y}_{jh} = n_h^{-1} \sum_{i \in s_h} y_{jhi}/p_{hi}$, y_{jhi} is the estimated total for variable j for units in PSU (hi), p_{hi} is the single-draw selection probability of PSU (hi). The $\hat{t}_j$'s are combined to create a nonlinear estimator $\hat{\theta} = g(\hat{\mathbf{t}})$ with $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_J)^T$ and g being some differentiable function. The standard linear approximation gives

$$\hat{\theta} - \theta \doteq \mathbf{d}^T(\hat{\mathbf{t}} - \mathbf{t}) \tag{1}$$

with $\theta = g(\mathbf{t})$ and $\mathbf{d} = (\partial g/\partial t_1, \ldots, \partial g/\partial t_J)^T$ is the vector of partial derivatives evaluated at $\mathbf{t}$.

By reversing the order of summation in (1) between variables and PSUs, the approximation can be written as

$$\hat{\theta} - \theta \doteq \sum_h \frac{1}{n_h} \sum_{i \in s_h} \frac{u'_{hi}}{p_{hi}} \equiv \sum_h \hat{u}_h \tag{2}$$

where $u'_{hi} = \sum_j d_j y_{jhi}$ and $d_j = \partial g/\partial t_j$. Expression (2) is in the form of a "pwr" estimator as discussed in Särndal, Swensson, and Wretman (1992, Section 2.9). In PPSWR sampling of PSUs, the u'_{hi}/p_{hi} are independent and have expected value $u_{Uh} = \sum_{i \in U_h} u_{Uhi}$ with $u_{Uhi} = \sum_{j=1}^{J} d_j t_{jhi}$ where $t_{jhi} = \sum_{k \in U_{hi}} y_{jhik}$ with y_{jhik} the value observed on variable j for unit (hik) and U_{hi} the universe of units in PSU hi. This assumes that y_{jhi} is a design-unbiased estimator of the total for PSU (hi). Thus, we also have $E(\hat{u}_h) = u_{Uh}$. Using the standard, conditional variance formula, $VE(\cdot|s_h) + EV(\cdot|s_h)$ with s_h the set of sample PSUs in Stratum h, the design-variance of $\hat{u}_h$ is $V(\hat{u}_h) = \sigma_h^2/n_h$ where

$$\sigma_h^2 = \sum_{U_h} p_{hi}(u_{Uhi}/p_{hi} - u_{Uh})^2 + \sum_{U_h} V(u'_{hi})/p_{hi}$$

and $V\left(u'_{hi}\right)$ is the variance of u'_{hi} with respect to whatever type of sampling is used within PSU hi. Thus, the approximate variance of $\hat{\theta}$ is $AV(\hat{\theta}) = \sum_h \sigma_h^2/n_h$. Note that σ_h^2 can be interpreted as the contribution to the variance of an estimated total for a single variable from a sample of size 1 in Stratum h. For example, in a single-stage, stratified simple random sample with replacement, $\sigma_h^2 \doteq N_h^2 S_h^2$ with $S_h^2 = \sum_{U_h}(y_{hi} - \bar{y}_{Uh})^2/(N_h - 1)$ (see Case 2 in Section 4). Notice in particular that in our notation σ_h^2 is not the unit variance S_h^2.

The standard estimator of the approximate variance is

$$\begin{aligned} v(\hat{\theta}) &= \sum_h \frac{1}{n_h(n_h - 1)} \sum_{s_h} \left(\frac{u'_{hi}}{p_{hi}} - \frac{1}{n_h} \sum_{s_h} \frac{u'_{hi}}{p_{hi}} \right)^2 \\ &= \sum_h \frac{n_h}{(n_h - 1)} \sum_{s_h} (u_{hi} - \bar{u}_h)^2 \end{aligned} \tag{3}$$

where $u_{hi} = u'_{hi}/n_h p_{hi}$. Expression (3) implicitly reflects the contribution of sampling within PSUs and is an example of what is known as the "linear substitute" method (Wolter 2007, Section 6.5).

A Satterthwaite approximation (Satterthwaite 1946) is based on assuming that a variance estimator has a chi-square distribution and solving for the implied degrees of freedom, using the method of moments. If $DFv(\hat{\theta})/AV(\hat{\theta}) \sim \chi^2_{DF}$ where DF is the degrees of freedom for $v(\hat{\theta})$ and χ^2_{DF} is the central chi-square distribution with DF degrees of freedom, then

$$\begin{aligned} \mathrm{Var}[DFv(\hat{\theta})/AV(\hat{\theta})] &= DF^2 \mathrm{Var}[v(\hat{\theta})]/[AV(\hat{\theta})]^2 \\ &= 2DF \end{aligned}$$

Solving for DF gives

$$DF = 2/\mathrm{relvar}[v(\hat{\theta})]. \tag{4}$$

where relvar $[v(\hat{\theta})] = \mathrm{Var}[v(\hat{\theta})]/[AV(\hat{\theta})]^2$ is the relvariance of the variance estimator. Somewhat more generally, (4) can be derived assuming only that the first two moments of $DFv(\hat{\theta})/AV(\hat{\theta})$ match those of a central chi-square distribution. The pivot

$t(\hat{\theta}) = (\hat{\theta} - \theta)/\sqrt{v(\hat{\theta})}$ is treated as having a central t-distribution with DF degrees of freedom. To formally justify this, we would need to show that $DFv(\hat{\theta})/AV(\hat{\theta})$ is approximately the sum of weighted independent chi-square random variables and that $\hat{\theta} - \theta$ and $v(\hat{\theta})$ are independent. But, we have no such design-based theorem. The model-based result, which can be found in many mathematical statistics books (e.g., Bickel and Doksum 1977), requires independent and identically distributed normal random variables. Expression (4), along with the t approximation for $t(\hat{\theta})$, is an *ad hoc,* but practical, fix-up to account for instability of the variance estimator. However, in cases where the data for individual units are far from normally distributed the variance of the variance estimator may be larger than expected (Cochran 1963, Section 2.14), making the assumption that $v(\hat{\theta})$ has a distribution proportional to a chi-square a poor one.

3. Variance of the Variance Estimator

The linearization variance estimator in (3) can be written as

$$v(\hat{\theta}) = \sum_h \frac{1}{n_h(n_h - 1)} \sum_{s_h} (z_{hi} - \bar{z}_h)^2$$

where $z_{hi} = u'_{hi}/p_{hi} - u_{Uh}$. Under PPSWR sampling of PSUs, the z_{hi} are independent with mean 0 and variance σ_h^2. Using the same steps as in Hansen, Hurwitz, and Madow (1953, pp. 99–101), calculation of the variance of $v(\hat{\theta})$ is tedious but straightforward. The Appendix gives some of the details. The result is

$$V[v(\hat{\theta})] = \sum_h \frac{1}{n_h^3}\left[\mu_{4h} - \sigma_h^4 \frac{n_h - 3}{n_h - 1}\right].$$

It follows that the approximate DF in (4) is

$$DF = \frac{2\left[\sum_h \sigma_h^2/n_h\right]^2}{\sum_h \frac{\sigma_h^4}{n_h^3}\left[\beta_h - \frac{n_h - 3}{n_h - 1}\right]} \tag{5}$$

with $\beta_h = \mu_{4h}/\sigma_h^4$. The expression $\beta_h - 3$ is Fisher's measure of kurtosis (e.g., see Cochran 1963, Section 2.14) and is sometimes called "excess" kurtosis since (when it is positive) $\beta_h - 3$ is the amount over and above the value of 3 for a normal distribution. Note that for a domain that is defined by a subset of strata, the summations in (5) would cover only the strata containing the domain.

4. Special Cases

Evaluating (5) in some special cases leads to the rules-of-thumb that practitioners and software packages often use.

Case 1. $n_h = 2$. Expression (5) reduces to

$$DF = \frac{4\left[\sum_h \sigma_h^2\right]^2}{\sum_h \sigma_h^4[\beta_h + 1]} \tag{6}$$

If, in addition, $\beta_h = 3$ as for a normal distribution and $\sigma_h^2 = \sigma^2$ in all strata, then (6) equals H. This is the rule-of-thumb that 1 degree of freedom is picked up per stratum when 2 PSUs are selected per stratum.

Case 2. $n_h > 2$. If $\beta_h = 3$ and $\sigma_h^2 = \sigma^2$, (5) becomes

$$DF = \frac{\left(\sum_h n_h^{-1}\right)^2}{\sum_h n_h^{-2}(n_h - 1)^{-1}}.$$

If the sample size is the same in each stratum, say $n_h = \bar{n}$, then $DF = H(\bar{n} - 1)$. This corresponds to the prescription that the DF should be (number of PSUs) – (number of strata). As noted in Fuller (1984), the assumption that $\beta_h = 3$ may be reasonable in multistage samples since u'_{hi} is a sum across a number of elementary units. In that case, failure of the rule-of-thumb may be due to the σ_h^2 being different across strata.

Case 3. Single-stage, stratified simple random sampling with replacement; $\hat{\theta} = \sum_h N_h\bar{y}_h$ with $\bar{y}_h = \sum_{s_h} y_{hi}/n_h$. In this case, $u'_{hi} = y_{hi}, p_{hi} = 1/N_h$, and $\sigma_h^2 \doteq N_h^2S_h^2$ if N_h is large. Assuming further that n_h is large, the DF approximation (5) is

$$DF = \frac{2\left(\sum_h N_h^2S_h^2/n_h\right)^2}{\sum_h \frac{N_h^4S_h^4}{n_h^3}(\beta_h - 1)} \tag{7}$$

where $\beta_h = S_h^{(4)}/S_h^4$ with $S_h^{(4)} = \sum_{U_h}(y_{hi} - \bar{y}_{Uh})^4/(N_h - 1)$ and $S_h^4 = \left(S_h^2\right)^2$ with $S_h^2 = \sum_{U_h}(y_{hi} - \bar{y}_{Uh})^2/(N_h - 1)$. Note that, when $\beta_h = 3$, n_h is large, and n_h/N_h is small, expression (5.16) in Cochran (1977) is a special case of (7). If $N_h \equiv \bar{N}$, $n_h \equiv \bar{n}$, $S_h^2 = S^2$, and $\beta_h \equiv \beta$, then (7) reduces to $2H\bar{n}/(\beta - 1)$. If y has heavier tails than the normal distribution ($\beta > 3$), DF can be much smaller than the rule-of-thumb value, $H(\bar{n} - 1) \doteq H\bar{n}$.

With Neyman allocation, $n_h = nN_hS_h/\sum_h N_hS_h$, and when $\beta_h \equiv \beta$, (7) reduces to $2n/(\beta - 1)$. With proportional allocation, $n_h = nN_h/N$, approximation (7) becomes

$$DF = \frac{2}{\beta - 1}\frac{n}{N}\frac{\left(\sum_h N_hS_h^2\right)^2}{\sum_h N_hS_h^4}. \tag{8}$$

Letting $x_h = \sqrt{N_h}$ and $w_h = \sqrt{N_hS_h^2}$, the Cauchy-Schwartz inequality, $\left(\sum x_hw_h\right)^2 \leq \sum x_h^2 \sum w_h^2$, implies that $N^{-1}\left(\sum_h N_hS_h^2\right)^2/\sum_h N_hS_h^4 \leq 1$. As a result, the DF with proportional allocation is less than $2n/(\beta - 1)$, the value for Neyman allocation with $\beta_h \equiv \beta$. As noted before, if β is much larger than three, the DF will be considerably less than the rule-of-thumb value of $H(\bar{n} - 1) \doteq n$.

In a single-stage design, it can often be the case that some or many of the β_h values are very much larger than three. Kurtosis for a binary variable is $(1 - 6p + 6p^2)/p(1 - p) + 3$ where p is the proportion with the characteristic. This becomes arbitrarily large as p becomes close to 0 or 1. (In the example population discussed in the next section, values as large as 2,000 are encountered.) In these cases the standard rule-of-thumb is completely inappropriate, as will be seen. However, provided that the stratum sample sizes, n_h, are for the most part not small, it is possible to use the sample data to estimate the values of β_h, and thus obtain an estimate of DF. However, note that having an improved DF estimate may still not yield CIs for rare characteristics that cover the population parameter at the desired rates. Kott and Liu (2009) present an alternative using an Edgeworth approximation that is preferable for small proportions.

As described in Joanes and Gill (1998), an estimator of the kurtosis in Stratum h, $G_{2h} = S_h^{(4)}/S_h^4 - 3$, in a simple random sample is

$$\hat{G}_{2h} = \frac{n_h - 1}{(n_h - 2)(n_h - 3)}\left[(n_h + 1)\frac{m_{4h}}{m_{2h}^2} - 3(n_h - 1)\right] \tag{9}$$

where $m_{4h} = n_h^{-1}\sum_{s_h}(y_{hi} - \bar{y}_{hs})^4$ is the fourth central moment among the sample observations, y_{hi}, in Stratum h and $m_{2h} = n_h^{-1}\sum_{s_h}(y_{hi} - \bar{y}_{hs})^2$. For nonlinear estimators, the linear element, u'_{hi}, is used in place of y_{hi} in m_{4h} and m_{2h}. In large stratum samples (9) is approximately $\hat{G}_{2h} \doteq m_{4h}/m_{2h}^2 - 3$. An estimator of $\beta_h = S_h^{(4)}/S_h^4$ is then $\hat{\beta}_h = \hat{G}_{2h} + 3$. Joanes and Gill (1998) discuss other estimators of G_{2h} and recommend (9) as having smaller mean squared error in highly skewed populations. Using (7), an estimator of the DF from a particular stratified simple random sample is then

$$\widehat{DF} = \frac{2\left(\sum_h N_h^2\hat{S}_h^2/n_h\right)^2}{\sum_h \frac{N_h^4}{n_h^3}\hat{S}_h^4(\hat{\beta}_h - 1)} \tag{10}$$

where $\hat{S}_h^2 = \sum_{i\in s_h}(y_{hi} - \bar{y}_{hs})^2/(n_h - 1)$. As for $\hat{G}_{2h}$, u'_{hi} is used instead of y_{hi} in $\hat{S}_h^2$ for nonlinear estimators.

The approximation to DF in (7) and the estimator in (10) can be quite sensitive to the sample allocation. The estimates of β_h in (10) are weighted together in proportion to $N_h^4\hat{S}_h^4/p_h^3$ where $p_h = n_h/n$. If a stratum with a large kurtosis receives a relatively small allocation, the DF can be much smaller than when, say, an equal allocation is used. The estimator in (10) may also be unstable since it involves fourth moments which are notoriously difficult to estimate—a point noted in Fuller (1984).

5. A Simulation Study

We examined the accuracy of the rule-of-thumb and more finely tuned approximations to DFs in a simulation study. The study involved a population and design of the type

represented in Case 3 of the previous section. The population consisted of 11,389 school districts in the 50 United States and the District of Columbia—a subset of the population used in Brick et al. (2005). To eliminate extreme observations, districts in the original data set were dropped that had values of 0 for the numbers of administrators, students, or teachers or had more than 50 administrators, 40,000 students, or 2,000 teachers. Having such outliers randomly enter a sample would affect variance estimates, but our goal is not to study the effect of individual extremes. The districts were categorized into twelve design strata based on size (four categories based on the number of students) crossed with percentage of students at or below the poverty line (three categories). The distribution of the population among the strata is shown in Table 1.

We studied estimates of the following six quantities: (i) total administrators, (ii) total administrators in districts with poverty level 1, (iii) student-teacher ratio (total students divided by total teachers), (iv) student-teacher ratio in districts with poverty level 1, (v) total number of districts with more than 15 administrators, and (vi) number of districts with poverty levels 1 and 2 with more than 15 administrators. As the final six columns of Table 1 show, these variables have kurtoses that are far from the normal distribution value of three. Estimand (vi) is for a rare characteristic since only 2.5% of districts with poverty levels 1 and 2 have more than 15 administrators. Estimands (ii), (iv), and (vi) are for domains that are contained in a subset of the strata.

Two sets of 10,000 stratified simple random samples without replacement (STSRS) were selected. In the first set the stratum sample sizes were 60, 100, 100, 140, 160, 160, 60, 60, 40, 40, 40, and 40 for a total of 1,000. This allocation is similar to proportional allocation, with some smoothing done to increase the allocation to strata 10–12. The second set of samples used an equal allocation of 10 districts per stratum for a total of 120. We also ran simulations for an equal allocation of 84 per stratum ($n = 1{,}008$) to explore the effects of an allocation different from the first one above but of about the same total sample size. Results were qualitatively similar to those reported below and are not recorded here.

In each sample, the six estimands above were calculated along with the variance estimates appropriate for STSRS without replacement. In particular, (3) was used with the addition of a stratum-specific finite population correction factor. In (3), $p_{hi} = 1/N_h$ and $u'_{hi} = y_{hi}$ in the case of estimated totals (estimands (i), (ii), (v), and (vi)) and $u'_{hi} = (y_{hi} - \hat{\theta}x_{hi})/\hat{t}_x$ for ratios (estimands (iii) and (iv)) with $\hat{\theta}$ being the estimated ratio for the population or subpopulation, y_{hi} the number of students in a district, x_{hi} the number of teachers in the district, and $\hat{t}_x$ the estimated number of teachers in the population or subpopulation. In each sample, we computed the estimated degrees of freedom in (10) separately for each of the six estimands described above.

Figures 1 and 2 are histograms of the variance estimators for the two sets of 10,000 samples. To draw the histograms, the variance estimates v are scaled to be $X = \text{sim}(DF)*v/\bar{v}$ where sim(DF) is the degrees of freedom estimated using (4) across the 10,000 samples, and $\bar{v}$ is the average variance estimate across the samples. If the variance estimator behaves as needed for the Satterthwaite approximation, then X will have an approximate chi-square distribution. Note that sim(DF) would not normally be available to a practitioner analyzing a single sample, although a single-sample analog might be constructed via bootstrapping. We use sim(DF) here as a standard against which to compare the estimated DFs from (10), which can be computed directly in each individual sample.

Table 1. Population distribution and sample allocation for a population of school districts stratified by size of student body (4) and poverty status (3)

				Kurtosis β_h					
Stratum	District size	Poverty status	Population size (N_h)	Admins	Admins in poverty level 1	Student/teacher ratio[a]	Student/teacher ratio in poverty level 1[b]	Districts with > 15 admins	Districts with > 15 admins in poverty levels 1,2
1	1	1	550	11.0	11.0	63.9	63.7	–	–
2	1	2	1,031	7.7	–	595.2	–	–	–
3	1	3	1,105	16.2	–	28.9	–	–	–
4	2	1	1,709	18.3	18.3	4.7	4.7	855.0	855.0
5	2	2	2,288	11.2	–	20.0	–	2,291.0	2,291.0
6	2	3	1,878	10.3	–	7.9	–	939.5	–
7	3	1	692	21.5	21.5	3.8	3.8	36.7	36.7
8	3	2	579	7.2	–	46.5	–	71.0	71.0
9	3	3	524	7.0	–	7.3	–	38.7	–
10	4	1	311	7.2	7.2	5.4	5.3	3.3	3.3
11	4	2	405	7.0	–	4.3	–	3.3	3.3
12	4	3	317	6.9	–	6.5	–	4.0	–
Total			11,389	27.0	27.3	46.7	116.8	45.1	42

[a] Kurtosis is computed among values of the linear substitute, $y_i - \theta x_i$, where y_i is the number of students in district i, x_i is the number of teachers, and θ is the population student-teacher ratio.

[b] The domain stratum kurtosis can differ from the full population kurtosis because the linear substitute is $y_i - \theta_d x_i$ for the domain with θ_d being the student teacher ratio for the domain.

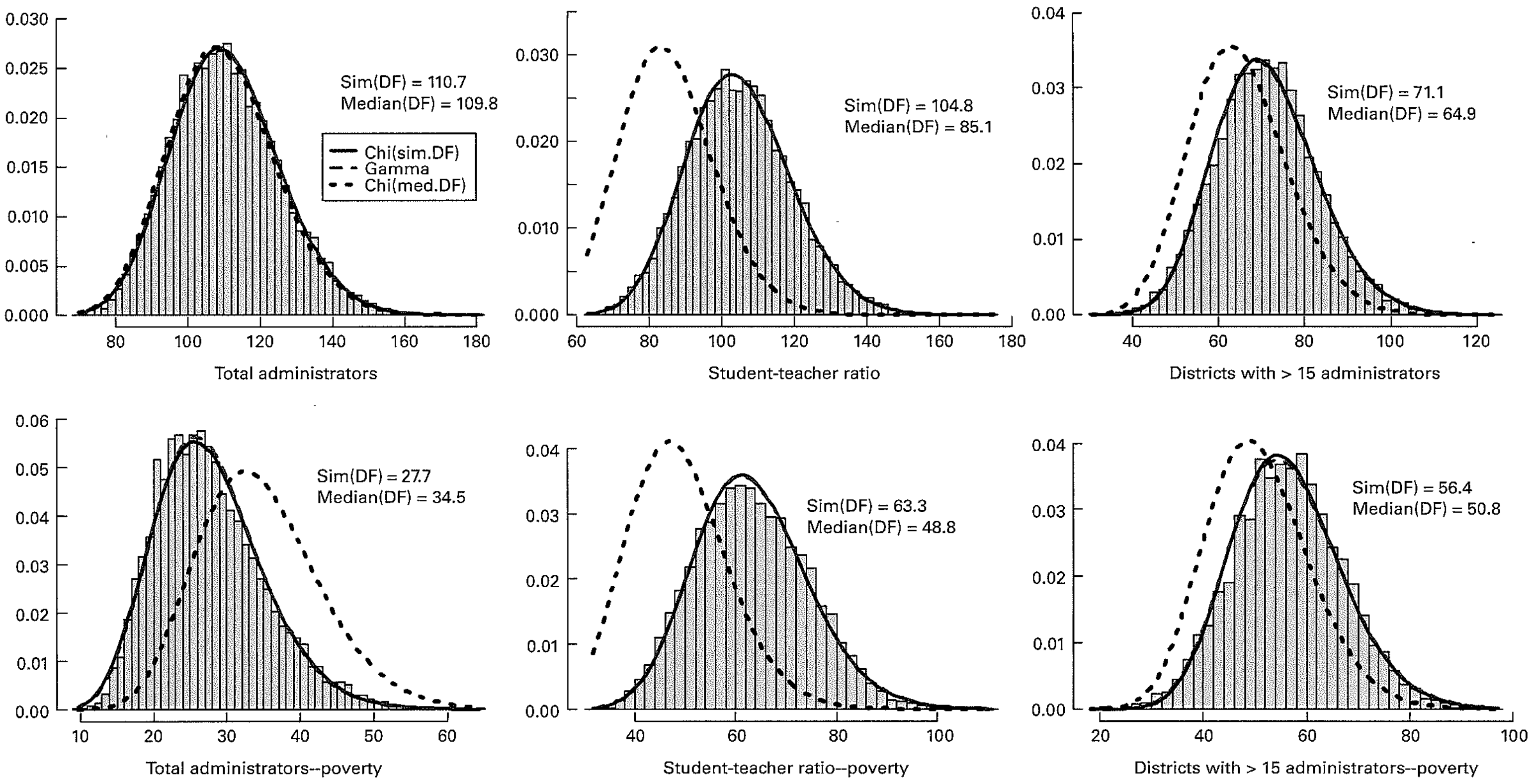

Fig. 1. Histograms of variance estimates for 10,000 samples of size 1,000. Sim(DF) is the DF estimated via expression (4) across all 10,000 samples; median(DF) is the median across the 10,000 samples of the DF's estimated via (10) from each sample. The solid black line is for a chi-square density with sim(DF) degrees of freedom. The short-dashed line (coincident with the black line here) is a gamma density (not constrained to be the special case of a chi-square). The dotted line is a chi-square density with median(DF) degrees of freedom

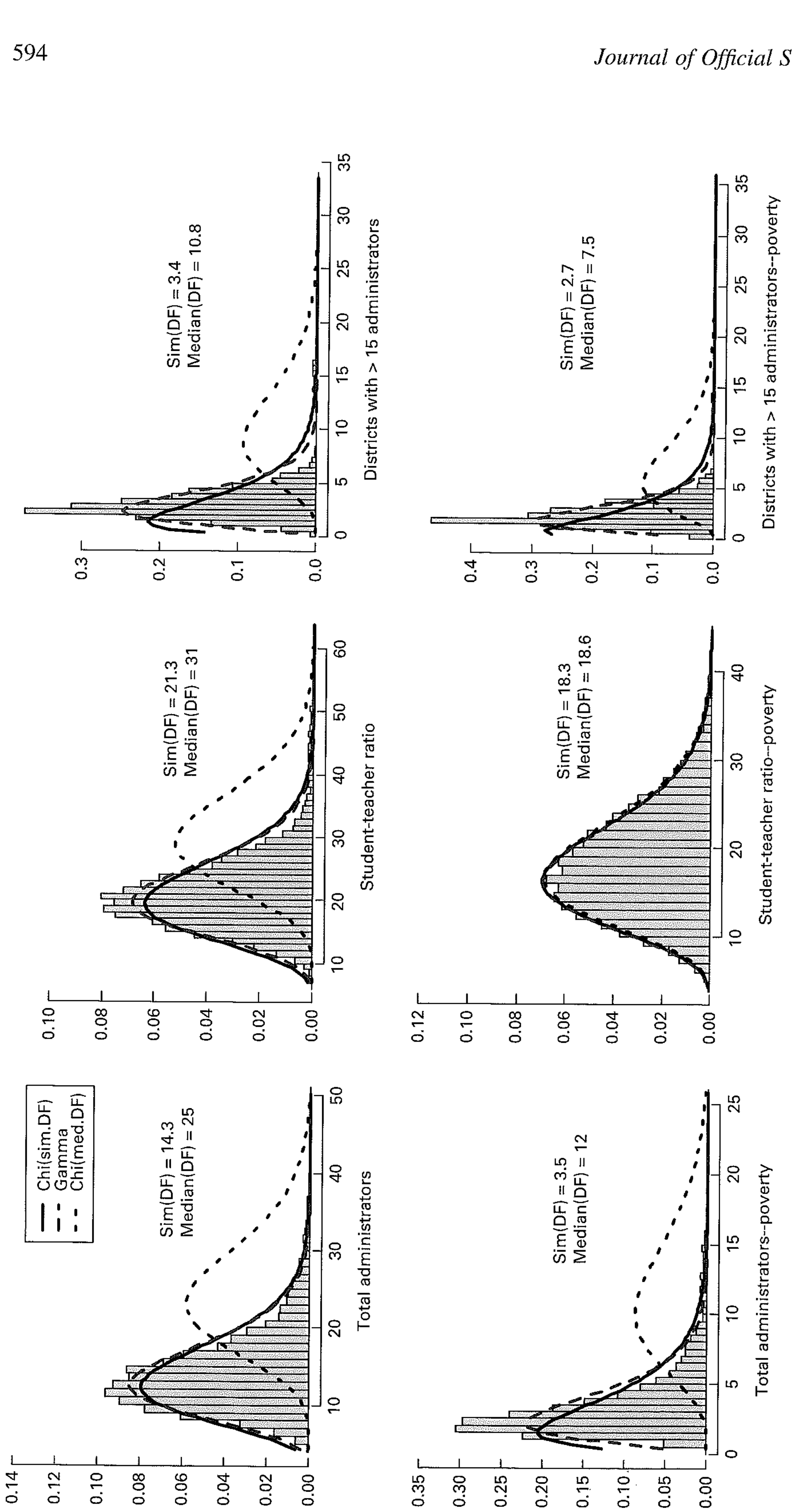

Fig. 2. Histograms of variance estimates for 10,000 samples of size 120. Sim(DF) is the DF estimated via expression (4) across all 10,000 samples; median(DF) is the median across the 10,000 samples of the DF's estimated via (10) from each sample. The solid black line is for a chi-square density with sim(DF) degrees of freedom. The short-dashed line is a gamma density (not constrained to be the special case of a chi-square). The dotted line is a chi-square density with median(DF) degrees of freedom

The solid line in each panel of Figures 1 and 2 is a chi-square density with sim(*DF*) degrees of freedom. The short-dashed line is the maximum likelihood estimate of a more general gamma density. When the chi-square and gamma are close in shape this provides evidence that the distribution of the scaled relative variance of the sample variance is close to chi-square, since the additional freedom for the parameters of the gamma distribution is not needed to improve the fit. The dotted line is a chi-square density with the *DF*, denoted median(*DF*), estimated by the median of (10) across the samples. The degrees of freedom estimated from (10) are extremely skewed across the samples so that the means are somewhat larger than the medians. If (10) is a useful alternative to the $n - H$ rule-of-thumb, then median(*DF*) and average(*DF*) should be near sim(*DF*) and should be considerably different from $n - H$. The rules-of-thumb are 998 for $n = 1{,}000$ and 12 strata and 108 for $n = 120$.

In Figure 1 the chi-square with sim(*DF*) is virtually identical to the gamma density, implying that the density of variance estimates is well-approximated by a chi-square. The chi-square density with median(*DF*) has too few *DF* to match the empirical densities in four of the six panels in Figure 1, but in all panels the median *DF* is 34 or more so that *t* confidence intervals will be similar to ones based on the normal approximation. In Figure 2 for the samples of $n = 120$, the chi-square densities with sim(*DF*) degrees of freedom are also close to the more general gamma densities. However, the chi-square densities with median(*DF*) have too many degrees of freedom to match the empirical densities well in five of six panels. In some cases, this would make a noticeable difference in the length of confidence intervals. For instance, median(*DF*) = 12 for total administrators in poverty level 1 districts while sim(*DF*) is 3.5. The 0.975 *t*-distribution values for these are 2.94 and 2.18, i.e., *t* intervals based on the empirical, simulation distribution will be 35% (2.94/2.18 − 1) longer than those based on the median estimated *DF* for the individual samples. However, these median estimated *DF*s are all much smaller than the rules-of-thumb *DF* of 988 for the samples in Figure 1 and 108 for the samples in Figure 2.

Tables 2 and 3 are numerical summaries of the simulations. For samples of $n = 1{,}000$ in Table 2, the range of estimated sample *DF*s from (10) is large. For example, the range for student-teacher ratio is (35.5, 227.9). In Table 2, the normal approximation *z*-intervals, i.e., ones in which $|\hat{\theta} - \theta|/\sqrt{v(\hat{\theta})}$ is compared to 1.96, are essentially the same as CIs computed using the rules-of-thumb for *DF*. Consequently, normal approximation 95% confidence intervals (CIs) cover about as well as *t*-intervals when $n = 1{,}000$. Ideally, about 95% of CIs constructed in the simulated samples should contain the population values. Faulty methods will typically have coverage rates noticeably less than 95%. The picture changes for $n = 120$ in Table 3. For all six estimands the *t*-intervals give closer to 95% coverage than do the normal approximation intervals. For example, the coverage for total districts with more than 15 administrators is 96.1% for *DF*s estimated from (10) but is 90.8% for normal approximation intervals. This increased coverage is obtained by making the CIs longer than those based on the normal approximation. For example, in Table 3 the median *DF*s range from 7.5 to 31. Thus, the ratios of the lengths of *t* intervals based on the median *DF*s to the normal approximation interval lengths range from 1.041 to 1.190. At the extreme, the minimum *DF* in Table 3 is 1.7 for which the *t*-multiplier for a 95% CI is 5.12. In that case, the *t* interval is 2.6 (5.12/1.96) times longer than the normal approximation interval.

Table 2. DF estimates and coverage of 95% confidence intervals over 10,000 samples of size 1,000

Statistic	Simulation *DF* from (4) across all samples	Sample *DF* from (10)				Coverage of 95% confidence intervals			
		Min	Max	Average *DF*	Median *DF*	Simulation *DF* from (4)	Estimated *DF* from (10)	Normal approx.	$\frac{Avg\left(v\left(\hat{\theta}\right)\right)}{mse\left(\hat{\theta}\right)}$
Total administrators	110.7	36	345.8	113.0	109.8	95.5	95.4	95.1	1.019
Total administrators in poverty level 1 districts	27.7	6.6	172.1	38.3	34.5	95.1	94.9	94.2	1.000
Student-teacher ratio	104.8	35.5	227.9	88.6	85.1	95.2	95.3	95.0	1.018
Student-teacher ratio in poverty level 1 districts	63.3	13.6	202.6	56.7	48.8	95.2	95.3	94.8	1.004
Total districts with > 15 administrators	71.1	22.2	335.5	67.9	64.9	94.7	94.8	94.3	1.001
Total districts in poverty levels 1 and 2 with > 15 administrators	56.4	11	336.8	54.9	50.8	89.4	89.6	88.5	0.784

Table 3. DF estimates and coverage of 95% confidence intervals over 10,000 samples of size 120

Statistic	Simulation *DF* from (4) across all samples	Sample *DF* from (10)				Coverage of 95% confidence intervals			
		Min	Max	Average *DF*	Median *DF*	Simulation *DF* from (4)	Estimated *DF* from (10)	Normal approx.	$\frac{Avg\left(v\left(\hat{\theta}\right)\right)}{mse\left(\hat{\theta}\right)}$
Total administrators	14.3	2.9	195.8	28.8	25	95.2	94.4	93.3	0.994
Total administrators in poverty level 1 districts	3.5	1.8	581.7	16.2	12	97.2	92.6	90.7	0.978
Student-teacher ratio	21.3	3.1	261.0	34.6	31	94.6	94.2	93.4	0.969
Student-teacher ratio in poverty level 1 districts	18.3	2.7	332.6	21.7	18.6	94.5	94.5	92.7	0.985
Total districts with > 15 administrators	3.4	1.7	72.6	13.3	10.8	97.3	96.1	90.8	1.020
Total districts in poverty levels 1 and 2 with > 15 administrators	2.7	1.7	48.4	10.9	7.5	96.1	94.5	85.5	1.012

Notice that the variance estimators themselves are approximately unbiased, as shown in the last column of Tables 2 and 3, which gives the ratio of the average variance estimate to the empirical mean squared error of the estimate in each row. The one exception is estimand (vi), which is the rare characteristic. Constructing CIs in such cases is well-known to have special problems and to require special solutions (Korn and Graubard 1998; Kott and Liu 2009).

Although the chi-square density based on the median of (10) was not a good match for the actual density of the variance estimates when $n = 120$, estimating the DF separately in each sample was a substantial improvement over the rule-of-thumb of $n - H$, which is standard practice.

Figure 3 illustrates how coverage can vary depending on the estimated DF for the variables total administrators and total administrators in poverty level 1. The 10,000 samples of size 120 were sorted by estimated DF and divided into 40 groups of 250 samples each. Coverage rates and average values of the estimated DF were computed within each group for the t-intervals and z-intervals and are shown in the panels in the first row. Two nonparametric smoothers are shown in each first-row panel along with the individual coverages. The second row shows the differences in the coverage rates using the estimated DF and the rule-of-thumb (which is the normal approximation in this case). A smoother is also shown in the second-row plots. The t-intervals uniformly give a few points higher coverage at every value of average DF. For small estimated DF this is actually a disadvantage since the t-intervals over-cover there. Nonetheless, through most of the range the coverage of the t-intervals is closer to the nominal percentage of 95, leading to the better overall coverage shown in Table 3.

6. Conclusion

We provide theory that covers the approximation of degrees of freedom in stratified multistage samples and investigate the empirical properties of an improved degrees of freedom estimator in the case of stratified simple random sampling. The standard rule-of-thumb for degrees of freedom of a linearization variance estimator is (number of PSUs) − (number of strata). The accuracy of this rule hinges on several assumptions. At least the first two moments of the distribution of a variance estimator have to be well-approximated by those of a chi-square distribution. The point estimate to which the variance estimate applies must be for the full population or for a domain that is spread across all strata and PSUs. Stratum variance components and first-stage sample sizes must be equal or nearly so.

As illustrated here, the rule-of-thumb DF can be a substantial over-estimate if the assumptions that justify it are violated. The faults are more likely to occur when the sample size is not large and may be especially severe in the types of single-stage samples used in establishment or institutional populations. However, there is a sample estimator of the actual DF that is simple to compute. Although imperfect, the estimated DF leads to t confidence intervals that have coverage probabilities that are much closer to nominal values, at least for stratified single-stage samples with moderate to large sample sizes in each stratum.

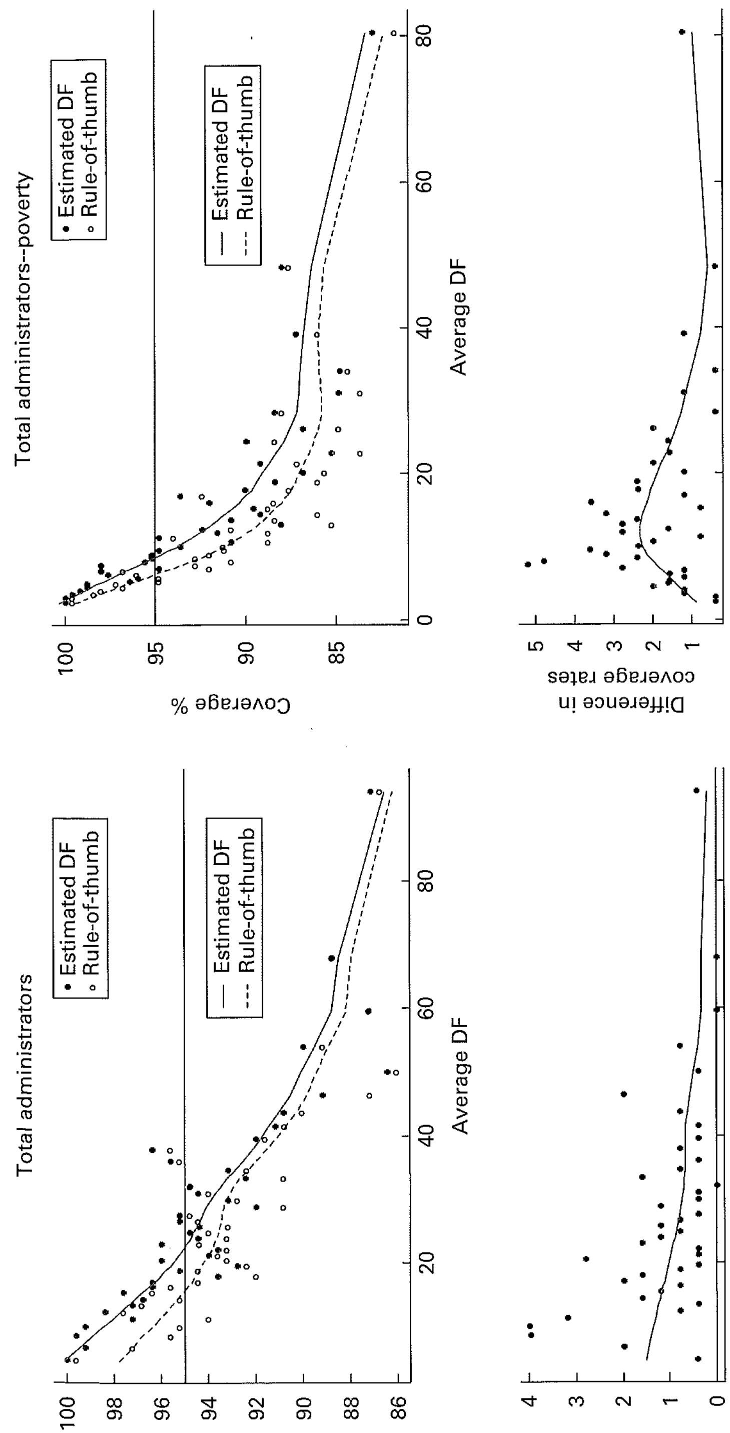

Fig. 3. Plot of empirical coverage rates of 95% confidence intervals versus estimated DF for 10,000 samples of size 120 for total administrators and total administrators in poverty level 1 districts. In the first row black dots are coverage rates based on t-intervals with DF estimated from (10). Open circles are rates for normal approximation z-intervals. Solid line is a nonparametric smoother for coverage of t-intervals; dashed line is a smoother for coverage of z-intervals. The differences in coverage rates are plotted in the second row

In this article, we addressed only exact or linearization variance estimators. A separate set of rules-of-thumb is used for replication variances, particularly ones that involve collapsing of strata or PSUs. The *DF* appropriate for replication variances depends on how the replication method is implemented for a particular survey. Collapsing of strata or PSUs will typically lose *DF* compared to the maximum that is available. Collapsing is sometimes used for linearization variance estimators, but our experience is that it is more common when setting up replication weighting systems. We plan to focus on those applications in another report.

Appendix. Variance of the Variance Estimator

This appendix sketches some of the details needed to derive $V[v(\hat{\theta})]$ given in Section 3. By definition, $V[v(\hat{\theta})] = E[v(\hat{\theta})^2] - [Ev(\hat{\theta})]^2$. The variance estimator is

$$v(\hat{\theta}) = \sum_h \frac{1}{n_h(n_h-1)} \sum_{s_h} (z_{hi} - \bar{z}_h)^2$$

where $z_{hi} = u'_{hi}/p_{hi} - u_{Uh}$. In PPSWR sampling, the z_{hi} are independent random variables with mean 0 and variance σ_h^2, as defined in Section 2. Since u'_{hi}/p_{hi} is a 1-PSU estimator of the stratum total and u_{Uh} is itself a constant, the variance of z_{hi} is calculated as

$$\begin{aligned} V(z_{hi}) &= VE(u'_{hi}/p_{hi}|s_h) + EV(u'_{hi}/p_{hi}|s_h) \\ &= V(u_{Uhi}/p_{hi}) + E\left[p_{hi}^{-2}V(u'_{hi}|s_h)\right] \\ &= \sum_{U_h} p_{hi}(u_{Uhi}/p_{hi} - u_{Uh})^2 + \sum_{U_h} V(u'_{hi})/p_{hi} \\ &\equiv \sigma_h^2. \end{aligned}$$

It follows that $E\sum_{s_h}(z_{hi} - \bar{z}_h)^2 = (n_h - 1)\sigma_h^2$ and

$$Ev(\hat{\theta}) = \sum_h \sigma_h^2/n_h. \tag{A.1}$$

Next, the expectation of the square of the variance estimator is

$$\begin{aligned} E[v(\hat{\theta})^2] = E\Bigg\{ &\sum_h \frac{1}{[n_h(n_h-1)]^2} \left[\sum_{s_h}(z_{hi} - \bar{z}_h)^2\right]^2 \\ &+ \sum_h \sum_{h' \neq h} \frac{1}{n_h(n_h-1)} \frac{1}{n_{h'}(n_{h'}-1)} \sum_{s_h}(z_{hi} - \bar{z}_h)^2 \sum_{s_{h'}}(z_{h'i} - \bar{z}_{h'})^2 \Bigg\} \end{aligned} \tag{A.2}$$

To evaluate (A.2), we need the expectation of

$$\left[\sum_{s_h}(z_{hi} - \bar{z}_h)^2\right]^2 = \left(\sum_{s_h} z_{hi}^2\right)^2 - 2n_h\bar{z}_h^2 \sum_{s_h} z_{hi}^2 + n_h^2\bar{z}_h^4. \tag{A.3}$$

By squaring out and collecting terms, the expectation of the first term in (A.3) is $E\left[\left(\sum_{s_h} z_{hi}^2\right)^2\right] = n_h\mu_{4h} + n_h(n_h - 1)\sigma_h^4$ where $\mu_{4h} = E\left(u'_{hi}/p_{hi} - u_{Uh}\right)^4$. As an example of how to compute the expectation of the remaining terms in (A.3), consider $\bar{z}_h^2 \sum_{s_h} z_{hi}^2 = n_h^{-2}\left[\sum_{s_h} z_{hi}^4 + 2\sum_{i \neq j \in s_h} z_{hi}^3 z_{hj} + \sum_{i \neq j \in s_h} z_{hi}^2 z_{hj}^2 + \sum_{i \neq j \neq k \in s_h} z_{hi}^2 z_{hj} z_{hk}\right]$.

By independence of the z_{hi}, $E\left(z_{hi}^2 z_{hj}^2\right) = \sigma_h^4$ *and* $E\left(z_{hi}^3 z_{hj}\right) = E\left(z_{hi}^2 z_{hj} z_{hk}\right) = 0$. Thus, $E\left(\bar{z}_h^2 \sum_{s_h} z_{hi}^2\right) = n_h^{-1}\left[\mu_{4h} + (n_h - 1)\sigma_h^4\right]$. Similar computations lead to $E(\bar{z}_h^4) = n_h^{-3}\left[\mu_{4h} + 3(n_h - 1)\sigma_h^4\right]$. Substitution of these results gives

$$E\left[\sum_{s_h}(z_{hi} - \bar{z}_h)^2\right]^2 = (n_h - 1)n_h^{-1}\left[\mu_{4h}(n_h - 1) + \sigma_h^4\left(n_h^2 - 2n_h + 3\right)\right]. \tag{A.4}$$

Expression (A.2) is then

$$E[v(\hat{\theta})^2] = \sum_h \frac{1}{n_h^3(n_h - 1)}\left[\mu_{4h}(n_h - 1) + \sigma_h^4\left(n_h^2 - 2n_h + 3\right)\right] + \sum_h \sum_{h' \neq h} \frac{\sigma_h^2}{n_h}\frac{\sigma_{h'}^2}{n_{h'}}.$$

Noting that $[Ev(\hat{\theta})]^2 = \sum_h \sigma_h^4/n_h^2 + \sum_h \sum_{h' \neq h}\left(\sigma_h^2/n_h\right)\left(\sigma_{h'}^2/n_{h'}\right)$, we have

$$\begin{aligned} V[v(\hat{\theta})] &= E[v(\hat{\theta})^2] - [Ev(\hat{\theta})]^2 \\ &= \sum_h \frac{1}{n_h^3(n_h - 1)}\left[\mu_{4h}(n_h - 1) + \sigma_h^4\left(n_h^2 - 2n_h + 3\right)\right] \\ &\quad + \textstyle\sum_h \sum_{h' \neq h} \frac{\sigma_h^2}{n_h}\frac{\sigma_{h'}^2}{n_{h'}} - \sum_h \frac{\sigma_h^4}{n_h^2} - \sum_h \sum_{h' \neq h}\left(\frac{\sigma_h^2}{n_h}\right)\left(\frac{\sigma_{h'}^2}{n_{h'}}\right) \end{aligned}$$

which after some algebra simplifies to $V[v(\hat{\theta})] = \sum_h \frac{1}{n_h^3}\left[\mu_{4h} - \sigma_h^4 \frac{n_h - 3}{n_h - 1}\right]$.

7. References

Bickel, P. and Doksum, K. (1977). Mathematical Statistics: Basic Ideas and Selected Topics. San Francisco: Holden-Day.

Brick, J.M., Jones, M.E., Kalton, G., and Valliant, R. (2005). Variance Estimation with Hot Deck Imputation: A Simulation Study of Three Methods. Survey Methodology, 31, 151–159.

Burns, A., Morris, R., Liu, J., and Byron, M. (2003). Estimating Degrees of Freedom for Data from Complex Surveys. Proceedings of the American Statistical Association, Section on Survey Research Methods, 727–729.

Cochran, W.G. (1963, 1977). Sampling Techniques, (Second and Third editions). New York: John Wiley.

Eltinge, J.L. and Jang, D.S. (1996). Stability Measures for Variance Component Estimators under a Stratified Multistage Design. Survey Methodology, 22, 157–165.

Fuller, W. (1984). Least Squares and Related Analyses for Complex Survey Designs. Survey Methodology, 10, 97–114.

Graubard, B.I. and Korn, E.L. (1996). Survey Inference for Subpopulations. American Journal of Epidemiology, 144, 102–106.

Hansen, M.H., Hurwitz, W.H., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vol. II. New York: Wiley.

Joanes, D.N. and Gill, C.A. (1998). Comparing Measures of Sample Skewness and Kurtosis. The Statistician, 47, 183–189.

Johnson, E. and Rust, K. (1992). Population Inferences and Variance Estimation for NAEP Data. Journal of Educational Statistics, 17, 175–190.

Johnson, E.G., Rust, K.F., and Hansen, M.H. (1988). Weighting Procedures and Estimation of Sampling Variance. Chapter 8 in E.G. Johnson and R. Zwick, Focusing the New Design: The NAEP 1988 Technical Report. Washington DC: U.S. Department of Education.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous Testing of Regression Coefficients with Complex Survey Data: Use of Bonferroni *t* Statistics. The American Statistician, 44, 270–276.

Korn, E.L. and Graubard, B.I. (1998). Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data. Survey Methodology, 24, 193–201.

Korn, E.L. and Graubard, B.I. (1999). Analysis of Health Surveys. New York: John Wiley.

Kott, P.S. and Liu, Y. (2009). One-Sided Coverage Intervals for a Proportion Estimated from a Stratified Simple Random Sample. International Statistical Review, 77, 251–265.

Research Triangle Institute (2004). SUDAAN User's Manual, Release 9.0. Research Triangle Park, NC: Research Triangle Institute.

Rust, K.F. (1984). Techniques for Estimating Variances for Sample Surveys, unpublished Ph.D. dissertation, Ann Arbor MI: University of Michigan.

Rust, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. Journal of Official Statistics, 1, 381–397.

Rust, K. (1986). Efficient Replicated Variance Estimation. Proceedings of the American Statistical Association, Section on Survey Research Methods, 81–87.

Rust, K. and Kalton, G. (1987). Strategies for Collapsing Strata for Variance Estimation. Journal of Official Statistics, 3, 69–81.

Rust, K.F. and Rao, J.N.K. (1996). Variance Estimation for Complex Estimators in Sample Surveys. Statistical Methods in Medical Research, 5, 381–397.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Satterthwaite, F.W. (1946). An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin, 2, 110–114.

Stata Corporation (2005). Survey Data Reference Manual, Release 9. College Station: Stata Press.

Westat (2000). WesVar 4.0 User's Guide, available at www.westat.com/wesvar. Rockville MD: Westat.

Wolter, K.M. (2007). Introduction to Variance Estimation, (Second edition). New York: Springer-Verlag.

Received October 2008
Revised January 2010