CONFIDENCE INTERVALS FOR SUB-DOMAIN PARAMETERS WHEN THE SUB-DOMAIN SAMPLE SIZE IS RANDOM

ROBERT J. CASADY, ALAN H. DORFMAN[1], and SUOJIN WANG[2]

ABSTRACT

Let A be a population sub-domain of interest and assume that the elements of A cannot be identified on the sampling frame and the number of elements in A is not known. Further assume that a sample of fixed size (say n) is selected from the entire frame and the resulting sub-domain sample size (say n_A) is random. The problem addressed is the construction of a confidence interval for a sub-domain parameter such as the sub-domain aggregate $T_A = \sum_{i \in A} x_i$. The usual approach to this problem is to redefine x_i, by setting $x_i = 0$ if $i \notin A$. Thus, the construction of a confidence interval for the sub-domain total is recast as the construction of a confidence interval for a population total which can be addressed (at least asymptotically in n) by normal theory. As an alternative, we condition on n_A and construct confidence intervals which have approximately nominal coverage under certain assumptions regarding the sub-domain population. We evaluate the new approach empirically using data from the Bureau of Labor Statistics (BLS) Occupational Compensation Survey.

KEY WORDS: Bayes Method, Conditioning, Establishment Surveys, Simple Random Sampling, Stratification, Survey Methods

1. INTRODUCTION

Let x_i be the value of the characteristic of interest for the i^{th} $(i = 1, 2, \ldots, N)$ element of the population and let A be a sub-domain of interest. The elements of A cannot be identified on the frame and the number of elements in A (say N_A) is not known; however, it is assumed that any element of A included in a sample can be identified. The problem is to construct a confidence interval for the sub-domain total, $T_A = \sum_{i \in A} x_i$, based on a sample of n elements selected from the entire frame.

The usual approach to this problem is to redefine x_i, by setting $x_i = 0$ if $i \notin A$, which forces the population total $T = \sum_{i=1}^{N} x_i$ to be equal to T_A. Thus, the construction of a

[1] Robert J. Casady and Alan H. Dorfman, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E.,Washington D.C., 20212-0001.
[2] Suojin Wang, Department of Statistics, Texas A&M University, College Station, TX, 77843.

confidence interval for the sub-domain total is recast as the construction of a confidence interval for a population total. In what follows it is assumed that the x_i's have been redefined as above. An overview of domain estimation can be found in Chapter 10 of Sarndal et.al. (1992). If we assume a simple random sample with replacement, the standard approach to this problem is along the following line:

Define the additional population parameters,

$\overline{X} = T/N$ = population mean,

$S^2 = \sum_{i=1}^{N} \left(x_i - \overline{X}\right)^2 \big/ N$ = population variance, and

$p_A = N_A/N$ = proportion of population in *A*.

Then

(1) $\hat{T} = (N/n)\sum_{i=1}^{n} x_i$, $\bar{x} = \sum_{i=1}^{n} x_i \big/ n = \hat{T}\big/N$, $s^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \big/ (n-1)$, and $\hat{p}_A = n_A/n$ (where n_A is the number of sample elements in *A*) are unbiased for the corresponding population parameters,

(2) $E\left(\hat{T}\right) = T = T_A$, so, we define the sub-domain estimator $\hat{T}_A \equiv \hat{T}$,

(3) $\operatorname{var}\left(\hat{T}_A\right) = N^2\, S^2/n$,

(4) $\sqrt{n}\left(\hat{T}_A - T_A\right) \big/ (NS) \xrightarrow{d} N(0,1)$, and

(5) s^2 is consistent for S^2.

It follows that $\sqrt{n}\left(\hat{T}_A - T_A\right) \big/ (Ns) \xrightarrow{d} N(0,1)$, so, when *n* is "sufficiently large", appropriate values from the normal distribution can be used to construct confidence intervals for T_A.

However, the proportion of the x_i's that are equal to zero is at least 1-p_A, therefore, when p_A is small and the values of the x_i's for $i \in A$ are concentrated away from zero, the convergence in distribution in (4) can be extremely slow. Consequently, the distribution of $\dfrac{\sqrt{n}\left(\hat{T}_A - T_A\right)}{Ns}$ can be far from normal even for what are usually considered to be moderate to large values of *n*. Dorfman and Valliant (1993) noted this problem in their study of wage distributions for sub-domains consisting of workers in specific occupational groups. Preliminary empirical work by the authors indicated that supposed 95% confidence intervals for total workers and total wages for occupation based sub-domains typically provided only 75% to 85% coverage even for a large total sample size (*n*=353 establishments). Furthermore, this work indicated that the distribution of $\hat{T}_A - T_A$ was strongly dependent on the realized value of n_A, which suggested that some type of "conditional" confidence interval should be considered. Thus, the formal goal of our research was to establish methodology for the construction of conditional (on n_A or equivalently $\hat{p}_A$) confidence intervals for T_A, which provide nominal, or near nominal, coverage regardless of the realized value of the sub-domain sample size.

In this paper we propose several methods of conditional confidence interval construction. These methods result from a Bayes based analysis of the conditional distribution of a random variable of the form $\hat{\theta} = \left(\hat{T}_A - T_A\right) \big/ s_{\hat{T}_A}$, where $s_{\hat{T}_A}$ is a standardizing random variable. The cases of simple random sampling and stratified random sampling are considered in Sections 2 and 3, respectively. The results of an empirical evaluation of the methods are discussed in Section 4. Section 5 provides a summary and concluding remarks.

2. THE CASE OF SIMPLE RANDOM SAMPLING

2.1 Definitions and Notation

We define the following parameters and estimators:

Sub-domain parameters:

$\mu_A = T_A / N_A$ = sub-domain mean,

$\sigma_A^2 = \sum_{i \in A} \left(x_i - \mu_A\right)^2 \big/ N_A$ = variance of population elements in *A*.

Sub-domain estimators:

$\hat{N}_A = \hat{p}_A N$,

$\hat{\mu}_A = \sum_{i=1}^{n_A} x_i \Big/ n_A = \hat{T}_A \Big/ \hat{N}_A$ (only defined for $n_A \geq 1$), and

$\hat{\sigma}_A^2 = \sum_{i=1}^{n_A} \left(x_i - \hat{\mu}_A\right)^2 \Big/ \left(n_A - 1\right)$ (only defined for $n_A \geq 2$).

In what follows it is understood that $n_A \geq 2$ (or equivalently $\hat{p}_A \geq 2/n$) unless specifically stated otherwise. The relationships given below follow directly from the definitions:

$T_A = N p_A \mu_A$ and $\hat{T}_A = N \hat{p}_A \hat{\mu}_A$,

$\overline{X} = p_A \mu_A$ and $\bar{x} = \hat{p}_A \hat{\mu}_A$,

$S^2 = p_A\left(1 - p_A\right)\mu_A^2 + p_A \sigma_A^2$

and

$$s^2 = \frac{n}{n-1}\hat{p}_A\left(1-\hat{p}_A\right)\hat{\mu}_A^2 + \frac{n\hat{p}_A - 1}{n-1}\hat{\sigma}_A^2. \qquad (1)$$

Also, it is straightforward to verify that

$$\begin{aligned}\left(\sqrt{n}/N\right)\left(\hat{T}_A - T_A\right) &= \sqrt{n}\left(\mu_A\left(\hat{p}_A - p_A\right) + \hat{p}_A\left(\hat{\mu}_A - \mu_A\right)\right) \\ &= \mu_A\sqrt{p_A\left(1-p_A\right)}\frac{\sqrt{n}\left(\hat{p}_A - p_A\right)}{\sqrt{p_A\left(1-p_A\right)}} + \sqrt{\hat{p}_A}\sigma_A\frac{\sqrt{n\hat{p}_A}\left(\hat{\mu}_A - \mu_A\right)}{\sigma_A} \qquad (2) \\ &= \mu_A\sqrt{p_A\left(1-p_A\right)}Z_1 + \sqrt{\hat{p}_A}\sigma_A Z_2,\end{aligned}$$

where $Z_1 = \dfrac{\sqrt{n}\left(\hat{p}_A - p_A\right)}{\sqrt{p_A\left(1-p_A\right)}}$, $Z_2 = \dfrac{\sqrt{n\hat{p}_A}\left(\hat{\mu}_A - \mu_A\right)}{\sigma_A}$

and

$$\frac{(n-1)s^2}{\sigma_A^2} = \left(1-\hat{p}_A\right)\left[Z_2 + \sqrt{n\hat{p}_A}\gamma_A\right]^2 + \left(n\hat{p}_A - 1\right)\frac{\hat{\sigma}_A^2}{\sigma_A^2}, \text{ where } \gamma_A = \mu_A/\sigma_A.$$

2.2 General Methodology for Confidence Intervals

Let $\hat{\theta} = \dfrac{\left(\hat{T}_A - T_A\right)}{s_{\hat{T}_A}}$ and assume that the conditional (on $\hat{p}_A$) distribution function of $\hat{\theta}$, say $H\left(\cdot \,|\hat{p}_A; p_A, \mu_A, \sigma_A^2\right)$, is known. In order to construct a conditional equal tailed $(1-\alpha)\times 100\%$ CI for T_A, we define an upper critical value

$$c_u \equiv c_u\left(\alpha, \hat{p}_A, p_A\right) = -\inf\left\{x \middle| H\left(x|\hat{p}_A; p_A\right) \geq \alpha/2\right\} = -H^{-1}\left(\alpha/2, \hat{p}_A; p_A\right)$$

where p_A is considered fixed and the dependence on μ_A and σ_A^2 is temporarily suppressed; a lower critical value, say c_ℓ, is defined in a similar manner. A conditional, equal tailed $(1-\alpha)\times 100\%$ CI for T_A is then given by $CI(1-\alpha) = (\ell, u)$, where

$$\begin{aligned} u &= \hat{T}_A + c_u s_{\hat{T}_A} \text{ and} \\ \ell &= \hat{T}_A + c_\ell s_{\hat{T}_A}. \end{aligned} \tag{3}$$

At this point the obvious practical problem is that the critical values c_u and c_ℓ depend not only on $\hat{p}_A$ but also on the unknown parameter p_A. One approach to this problem is to take a Bayesian tack and assume the parameter p_A is the realization of a random variable. Adjusting the notation for this assumption, we have $H\left(x|\hat{p}_A, p_A\right) \equiv H\left(x|\hat{p}_A; p_A\right)$, and it follows that

$$\begin{aligned} \Pr\left\{\hat{\theta} \leq x|\hat{p}_A\right\} &= F\left(x|\hat{p}_A\right) \\ &= \frac{1}{h\left(\hat{p}_A\right)} \int H\left(x|\hat{p}_A, p_A\right) f\left(\hat{p}_A|p_A\right) g\left(p_A\right) dp_A, \end{aligned} \tag{4}$$

where $h\left(\hat{p}_A\right) = \int f\left(\hat{p}_A|p_A\right) g\left(p_A\right) dp_A$ and $g\left(p_A\right)$ is the pdf of p_A. It should be noted that as a consequence of our sampling scheme the distribution of $n\hat{p}_A$, conditional on p_A, is

Binomial (n, p_A) so that $f(\hat{p}_A | p_A)$ is known. Under the Bayesian approach, the critical values are $c_u^* \equiv c_u^*(\alpha, \hat{p}_A) = -F^{-1}(\alpha/2|\hat{p}_A)$ and $c_\ell^* \equiv c_\ell^*(\alpha, \hat{p}_A) = -F^{-1}(1-\alpha/2|\hat{p}_A)$ so the upper and lower limits for a conditional $(1-\alpha)\times 100\%$ CI for T_A are

$$\begin{aligned} &u = \hat{T}_A + c_u^* s_{\hat{T}_A} \text{ and} \\ &\ell = \hat{T}_A + c_\ell^* s_{\hat{T}_A}. \end{aligned} \tag{5}$$

For the purposes of our current research, we are assuming that the prior distribution $g(p_A)$ is $N(\mu_{p_A}, \sigma_{p_A}^2)$ with μ_{p_A} and $\sigma_{p_A}^2$ to be specified. For an empirical Bayes approach, we used $\mu_{p_A} = \hat{p}_A$ and considered several possible alternatives for $\sigma_{p_A}^2$ which we discuss in detail below. Our experience indicates that the normality assumption is not crucial, rather, it is primarily a matter of convenience. On the other hand, the choice of values for the mean and variance is relatively more important. This will also be discussed in more detail at later point.

2.3 Confidence Intervals Under Normal Assumptions

Assume that within the sub-domain *A* the x_i are distributed $N(\mu_A, \sigma_A^2)$. Then

(a) $\left[\sqrt{n}(\hat{T}_A - T_A)/N | \hat{p}_A, p_A\right]$ is distributed $N(\sqrt{n}\mu_A(\hat{p}_A - p_A), \hat{p}_A\sigma_A^2)$,

(b) $\left[(n\hat{p}_A - 1)\dfrac{\hat{\sigma}_A^2}{\sigma_A^2} | \hat{p}_A, p_A\right]$ is distributed $\chi^2(n\hat{p}_A - 1)$,

(c) $\left[\left(Z_2 + \sqrt{n\hat{p}_A}\gamma_A\right)^2 | \hat{p}_A, p_A\right]$ is distributed non-central $\chi^2(1; \lambda^*)$ with $\lambda^* = n\hat{p}_A\gamma_A^2$, and

(d) the conditional random variable in (b) is stochastically independent of the conditional random variables in (a) and (c).

Consider $\hat{\theta}_1 = \frac{\left(\hat{T}_A - T_A\right)}{\left(N\hat{\sigma}_A\sqrt{\hat{p}_A}\big/\sqrt{n}\right)}$ which utilizes the conditional variance of $\hat{T}_A$ as the standardizing term. It follows immediately from (a), (b) and (d) that, conditional on $\left(\hat{p}_A, p_A\right)$, the random variable $\hat{\theta}_1$ is distributed as a non-central *t* with $n\hat{p}_A - 1$ degrees of freedom and non-centrality parameter $\lambda = \sqrt{n}\gamma_A \frac{\left(\hat{p}_A - p_A\right)}{\sqrt{\hat{p}_A}}$. Thus, we have specified the conditional distribution function $H\left(\ \cdot\ \middle| \hat{p}_A, p_A\right)$ of $\hat{\theta}_1$. As $f\left(\hat{p}_A \middle| p_A\right)$ and $g\left(p_A\right)$ have been previously specified, it follows that $F\left(\cdot\middle|\hat{p}_A\right)$ in (4) is well-defined although extremely cumbersome to calculate. The dependence on μ_A and σ_A^2, through γ_A, should be noted.

Although $F\left(\cdot\middle|\hat{p}_A\right)$ as given above can be used to determine the critical values, they are extremely difficult to calculate. A relatively simple approach, given in the next paragraph, provides a close approximation to the critical values. We have verified the closeness of the approximation by computing the exact values for selected cases using large scale simulations.

Under the assumption that p_A is distributed as a $N\left(\hat{p}_A, \sigma_{p_A}^2\right)$, it follows from Appendix A that $\left[\lambda\middle|\hat{p}_A\right]$ is distributed approximately as a normal with mean zero and variance $\frac{\gamma_A^2\left(1-\hat{p}_A\right)}{1+\psi_A}$, where $\psi_A = \frac{\hat{p}_A\left(1-\hat{p}_A\right)/n}{\sigma_{p_A}^2}$. It then follows from the result in Appendix B that, conditional on $\hat{p}_A$,

$$\frac{\left(\hat{T}_A - T_A\right)}{\frac{N\hat{\sigma}_A\sqrt{\hat{p}_A}}{\sqrt{n}}\sqrt{\frac{\gamma_A^2\left(1-\hat{p}_A\right)}{1+\psi_A}+1}}$$

is distributed as a central *t* with $n\hat{p}_A - 1$ degrees of freedom. The upper confidence limit

u, defined in (5), is given (approximately) by

$$u = \hat{T}_A + \frac{N\hat{\sigma}_A\sqrt{\hat{p}_A}}{\sqrt{n}} \frac{\left(\gamma_A^2\left(1-\hat{p}_A\right)+1+\psi_A\right)^{1/2}}{\left(1+\psi_A\right)^{1/2}} t_{1-\alpha/2,n_A-1}. \tag{6}$$

As $\hat{\sigma}_A^2$ is conditionally unbiased for σ_A^2 and $\hat{\mu}_A^2 - \hat{\sigma}_A^2 / n_A$ is conditionally unbiased for μ_A^2, we use $\hat{\gamma}_A^2 = \left(\hat{\mu}_A^2 - \hat{\sigma}_A^2 / n_A\right) / \hat{\sigma}_A^2$ to estimate γ_A^2 and approximate u with

$$\begin{aligned}\tilde{u} &= \hat{T}_A + \frac{N\hat{\sigma}_A\sqrt{\hat{p}_A}}{\sqrt{n}} \frac{\left(\hat{\gamma}_A^2\left(1-\hat{p}_A\right)+1+\psi_A\right)^{1/2}}{\left(1+\psi_A\right)^{1/2}} t_{1-\alpha/2,n_A-1} \\ &= \hat{T}_A + \frac{Ns}{\sqrt{n}}\left(\frac{n-1}{n}\right)^{1/2} \frac{\left(1+\left(\frac{n}{n-1}\right)\frac{\hat{p}_A\hat{\sigma}_A^2\left(\psi_A+1/n\right)}{s^2}\right)^{1/2}}{\left(1+\psi_A\right)^{1/2}} t_{1-\alpha/2,n_A-1}\end{aligned}.$$

A necessary condition for the results in Appendix A to hold is that n be large enough to ignore the terms with order $O\left(n^{-1}\right)$, in which case

$$\tilde{u} \cong \hat{T}_A + \frac{Ns}{\sqrt{n}}\left(\left(1+\frac{\hat{p}_A\hat{\sigma}_A^2\psi_A}{s^2}\right) \Big/ \left(1+\psi_A\right)\right)^{1/2} t_{1-\alpha/2,n_A-1}. \tag{7}$$

It should be noted that $\tilde{u}$ is strictly decreasing as ψ_A increases and

$$\begin{aligned}&\tilde{u} \to \hat{T}_A + \frac{Ns}{\sqrt{n}} t_{1-\alpha/2,n_A-1} = \tilde{u}_1 \text{ as } \psi_A \text{ becomes small,} \\ &\tilde{u} \cong \hat{T}_A + \frac{Ns}{\sqrt{n}}\left(1-\frac{4}{4+n_A}\left(\frac{\hat{p}_A\left(1-\hat{p}_A\right)\hat{\mu}_A^2}{s^2}\right)\right)^{1/2} t_{1-\alpha/2,n_A-1} = \tilde{u}_2, \text{ for } \psi_A = \frac{4}{n_A} \\ &\tilde{u} = \hat{T}_A + \frac{Ns}{\sqrt{n}}\left(\frac{1+\hat{p}_A\hat{\sigma}_A^2/s^2}{2}\right)^{1/2} t_{1-\alpha/2,n_A-1} = \tilde{u}_3 \text{ for } \psi_A = 1 \text{, and} \\ &\tilde{u} \to \hat{T}_A + \frac{Ns}{\sqrt{n}}\left(\frac{\sqrt{\hat{p}_A}\hat{\sigma}_A}{s}\right) t_{1-\alpha/2,n_A-1} = \tilde{u}_4 \text{ as } \psi_A \text{ becomes large.}\end{aligned} \tag{8}$$

In each case the lower critical value can be dealt with in an analogous manner resulting in four competing confidence intervals; namely, $CI_i(1-\alpha)=(\tilde{\ell}_i,\tilde{u}_i),\ i=1,\cdots,4$, with $\tilde{\ell}_i$ defined similarly to $\tilde{u}_i$ in (8) with $t_{1-\alpha/2,n_A-1}$ replaced by $t_{\alpha/2,n_A-1}$. In general the competing confidence intervals are labeled in order by decreasing length, except that the length of CI_3 is longer than CI_2 for n_A=2 or 3.

The first case is equivalent to assuming that $\sigma^2_{p_A}$ is “large” relative to $\operatorname{var}(\hat{p}_A)$ and leads to using the usual unconditional variance but with degrees of freedom equal to n_A-1. In most practical problems this seems reasonable since $\sigma^2_{p_A}$ is an unknown constant and $\operatorname{var}(\hat{p}_A)$ is $O(p_A/n)$. It is interesting to note that the heuristic development in Appendix C also yields CIs with critical values dependent on *s*. The second case is motivated by empirical evidence for its advantage at n_A=2 and 3, and the consistency to the standard normal method as n_A becomes large. For this case, $\sigma^2_{p_A}=\frac{n_A}{4}\frac{\hat{p}_A(1-\hat{p}_A)}{n}=\frac{\hat{p}_A{}^2}{4}(1-\hat{p}_A)$. The third interval is based on the standard empirical Bayes assumption regarding the prior variance, namely, $\sigma^2_{p_A}=\hat{p}_A(1-\hat{p}_A)/n$. The last confidence interval is based on the assumption that p_A is essentially degenerate at $\hat{p}_A$.

A small empirical study, using an artificial population, suggested that

1. Standard confidence intervals using the usual variance estimate and normal quantiles can give very low coverage. The worst cases occurred when γ was a half, for several values of *p*.

2. The strictly conditional intervals (i.e., CI_4) using the conditional variance can give abominable coverage, especially when γ is large. That is, confidence intervals based on “large” values of ψ_A gave very poor results.

3. The use of the standard variance estimate with degrees of freedom based on the number of sample units in the domain (i.e., CI_1) give conservative coverage.

4. The Empirical Bayes estimates corresponding to $\psi_A = 1$ and $\psi_A = \frac{4}{n_A}$ give conservative coverage with narrower mean interval length than the CI_1. However, the differences were not very great and in proceeding to the more complex stratified random sampling case, we focus our attention on variants of CI_1.

3. THE CASE OF STRATIFIED RANDOM SAMPLING

3.1 Definitions and Notation

Assume there are K strata and, where appropriate, terms previously defined have corresponding stratum level definitions. For example, n_k is the sample size and n_{Ak} is the number of sample elements in A for the k^{th} stratum. Thus, a natural estimator for the sub-domain total $T_A = \sum_{k=1}^{K} \sum_{i \in A} x_{ki} = \sum_{k=1}^{K} N_k p_{Ak} \mu_{Ak}$ is

$$\hat{T}_A = \sum_{k=1}^{K} \hat{T}_{Ak} = \sum_{k=1}^{K} N_k \hat{p}_{Ak} \hat{\mu}_{Ak}. \tag{9}$$

It is straightforward to verify that

$$E\left[\left(\hat{T}_A - T_A\right) \middle| \hat{\mathbf{p}}_A, \mathbf{p}_A\right] = \sum_{k=1}^{K} N_k \left(\hat{p}_{Ak} - p_{Ak}\right) \mu_{Ak} \equiv \tilde{\mu}_A \text{ and}$$
$$\operatorname{var}\left[\left(\hat{T}_A - T_A\right) \middle| \hat{\mathbf{p}}_A, \mathbf{p}_A\right] = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak}^2 \frac{\sigma_{Ak}^2}{n_{Ak}} = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} \frac{\sigma_{Ak}^2}{n_k} \equiv \tilde{\sigma}_A^2,$$

where $\hat{\mathbf{p}}_A = \left[\hat{p}_{A1}\, \hat{p}_{A2} \cdots \hat{p}_{AK}\right]$, $\mathbf{p}_A = \left[p_{A1}\, p_{A2} \cdots p_{AK}\right]$ and $B_1 = \left\{k \middle| n_{Ak} \geq 1 \text{ and } 1 \leq k \leq K\right\}$.

3.2 General Methodology for Confidence Intervals

As before, $\hat{\theta} = \frac{\hat{T}_A - T_A}{s_{\hat{T}_A}}$ has a distribution function $H\left(\cdot|\hat{\mathbf{p}}_A;\mathbf{p}_A,\underline{\mu}_A,\underline{\sigma}^2_A\right)$, where $\hat{\mathbf{p}}_A$ is the known realization of a random vector and $\mathbf{p}_A$, $\underline{\mu}_A$ and $\underline{\sigma}^2_A$ are unknown vectors of parameters. The upper critical value required to construct a conditional equal tailed $(1-\alpha)\cdot 100\%$ CI for T_A, is denoted by $c_u \equiv c_u(\alpha,\hat{\mathbf{p}}_A,\mathbf{p}_A)$ and the lower critical value is denoted by $c_\ell \equiv c_\ell(\alpha,\hat{\mathbf{p}}_A,\mathbf{p}_A)$; the dependence on $\underline{\mu}_A$ and $\underline{\sigma}^2_A$ being suppressed. As in the previous section, the problem is that the critical values depend on the unknown vector of parameters $\mathbf{p}_A$ so the situation is basically the same as for simple random sampling, although as we shall note, there are complications of a new sort in dealing with the stratified case.

Analogous to the approach in Section 2.2, if it is assumed that $\mathbf{p}_A$ is the realization of a random vector, then we can write $H(x|\hat{\mathbf{p}}_A,\mathbf{p}_A) = \Pr(\hat{\theta} \le x|\hat{\mathbf{p}}_A,\mathbf{p}_A)$ and $f(\hat{\mathbf{p}}_A;\mathbf{p}_A) = f(\hat{\mathbf{p}}_A|\mathbf{p}_A)$. It follows that

$$\begin{aligned}\Pr(\hat{\theta} \le x|\hat{\mathbf{p}}_A) &= F(x,\hat{\mathbf{p}}_A)\\ &= \frac{1}{h(\hat{\mathbf{p}}_A)}\int\cdots\int H(x,\hat{\mathbf{p}}_A,\mathbf{p}_A)\, f(\hat{\mathbf{p}}_A|\mathbf{p}_A)\, g(\mathbf{p}_A)\, d\mathbf{p}_A,\end{aligned}$$

where $h(\hat{\mathbf{p}}_A) = \int\cdots\int f(\hat{\mathbf{p}}_A|\mathbf{p}_A)\, g(\mathbf{p}_A)\, d\mathbf{p}_A$ and $g(\mathbf{p}_A)$ is the joint pdf of $\mathbf{p}_A$. The upper and lower critical values for an equal tailed $(1-\alpha)\cdot 100\%$ conditional (on $\hat{\mathbf{p}}_A$) CI for T_A are $c^*_u \equiv c^*_u(\alpha,\hat{\mathbf{p}}_A) = -F^{-1}(\alpha/2|\hat{\mathbf{p}}_A)$ and $c^*_\ell \equiv c^*_\ell(\alpha,\hat{\mathbf{p}}_A) = -F^{-1}(1-\alpha/2|\hat{\mathbf{p}}_A)$.

Because the samples are selected independently from each stratum we have $f(\hat{\mathbf{p}}_A|\mathbf{p}_A) = \prod_{k=1}^{K} f_k(\hat{p}_{Ak}|p_{Ak})$ and, as a consequence of our within stratum sampling scheme, $n_k\hat{p}_{Ak}$ has a binomial distribution $B(n_k, p_{Ak})$. It is reasonable to assume that the $\{p_{Ak}|1 \le k \le K\}$ are jointly independent so that $g(\mathbf{p}_A) = \prod_{k=1}^{K} g_k(p_{Ak})$ which implies $f(\hat{\mathbf{p}}_A|\mathbf{p}_A)\, g(\mathbf{p}_A) = \prod_{k=1}^{K} f_k(\hat{p}_{Ak}|p_{Ak})\, g_k(p_{Ak})$ and $h(\hat{\mathbf{p}}_A) = \prod_{k=1}^{K} \int f_k(\hat{p}_{Ak}|p_{Ak})\, g_k(p_{Ak})\, dp_{Ak}$.

In what follows, we assume that the prior distribution of p_{Ak} is $N\left(\mu_{p_{Ak}}, \sigma^2_{p_{Ak}}\right)$ and for the empirical Bayes approach, we used $\mu_{p_{Ak}} = \hat{p}_{Ak}$ and, analogously to the case of simple random sampling, we define $\psi_{Ak} = \frac{\hat{p}_{Ak}\left(1-\hat{p}_{Ak}\right)/n_k}{\sigma^2_{p_{Ak}}}$.

3.3 Confidence Intervals Under Normal Assumptions

Assume that within sub-domain A for the k^{th} stratum the x_{ki} are distributed $N\left(\mu_{Ak}, \sigma^2_{Ak}\right)$ and the x_{ki} are distributed independently from stratum to stratum, then $\left[\left(\tilde{T}_A - T_A\right)|\hat{\mathbf{p}}_A, \mathbf{p}_A\right]$ is distributed $N\left(\tilde{\mu}_A, \tilde{\sigma}^2_A\right)$. It follows that $\left[\left(\tilde{T}_A - T_A\right)/\tilde{\sigma}_A|\hat{\mathbf{p}}_A, \mathbf{p}_A\right]$ is distributed $N\left(\tilde{\mu}_A/\tilde{\sigma}_A, 1\right)$. Furthermore, based on the empirical Bayes approach specified in the previous section, it is straightforward to extend the result in Appendix A to the case of stratified random sampling and it then follows that $\left[\tilde{\mu}_A/\tilde{\sigma}_A|\hat{\mathbf{p}}_A\right]$ is distributed $N\left(0, \operatorname{var}\left(\tilde{\mu}_A|\hat{\mathbf{p}}_A\right)/\tilde{\sigma}^2_A\right)$, where $\operatorname{var}\left(\tilde{\mu}_A|\hat{\mathbf{p}}_A\right) = \sum_{k\in B_1} N_k^2 \mu_{Ak}^2 \frac{\hat{p}_{Ak}\left(1-\hat{p}_{Ak}\right)}{n_k\left(1+\psi_{Ak}\right)}$.

Let $B_2 = \left\{k|n_{Ak} \geq 2 \text{ and } 1 \leq k \leq K\right\}$ and assume that $B_2 \neq \emptyset$. Then, for $k \in B_2$, the terms $\left\{\sum_{i=1}^{n_{Ak}}\left(x_{ki} - \hat{\mu}_{Ak}\right)^2/\sigma^2_{Ak}|\hat{\mathbf{p}}_A, \mathbf{p}_A\right\}$ are distributed independently as $\chi^2\left(n_{Ak}-1\right)$. Next, let $\left\{w_k|k \in B_2\right\}$ be non-negative constants with $\sum_{k\in B_2} w_k > 0$ and $\hat{\sigma}^2_{Ak} = \frac{\sum_{i=1}^{n_{Ak}}\left(x_{ki}-\hat{\mu}_{Ak}\right)^2}{\left(n_{Ak}-1\right)}$.

Then, based on the usual Satterthwaite (1946) two moment approximation, the conditional random variable

$$\left\{\frac{\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)\hat{\sigma}^2_{Ak}/\sigma^2_{Ak}}{c}\middle|\hat{\mathbf{p}}_A, \mathbf{p}_A\right\}$$

is distributed approximately as a $\chi^2(\nu)$, where $c = \sum_{k\in B_2} w_k^2\left(n_{Ak}-1\right) / \sum_{k\in B_2} w_k\left(n_{Ak}-1\right)$ and $\nu = \left[\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)\right]^2 / \sum_{k\in B_2} w_k^2\left(n_{Ak}-1\right)$. It follows that, conditional on $\hat{\mathbf{p}}_A$, the random variable

$$\hat{\theta} = \frac{\left(\hat{T}_A - T_A\right)\Big/\sqrt{\operatorname{var}\left(\tilde{\mu}_A \middle| \hat{p}_A\right)+\tilde{\sigma}^2_A}}{\sqrt{\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)\left(\hat{\sigma}^2_{Ak}/\sigma^2_{Ak}\right)\Big/\nu}} = \frac{\left(\hat{T}_A - T_A\right)\Big/\sqrt{\operatorname{var}\left(\tilde{\mu}_A \middle| \hat{p}_A\right)+\tilde{\sigma}^2_A}}{\sqrt{\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)\left(\hat{\sigma}^2_{Ak}/\sigma^2_{Ak}\right)\Big/\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)}}$$

is distributed as a central t with ν degrees of freedom. Analogous to the conditional upper confidence limit defined in (6), we define

$$u\left(\mathbf{w}\right) = \hat{T}_A + \left[\frac{\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)\left(\hat{\sigma}^2_{Ak}/\sigma^2_{Ak}\right)\sum_{k\in B_1}\frac{N_k^2\hat{p}_{Ak}\sigma^2_{Ak}}{n_k}\left[\frac{\gamma^2_{Ak}\left(1-\hat{p}_{Ak}\right)+1+\psi_{Ak}}{\left(1+\psi_{Ak}\right)}\right]}{\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)}\right]^{1/2} t_{1-\alpha/2,\nu} \qquad (10)$$

with the lower confidence limit defined in a similar manner. When the ψ_{Ak} are near zero we have (approximately)

$$u\left(\mathbf{w}\right) \cong \hat{T}_A + \left[\frac{\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)\left(\hat{\sigma}^2_{Ak}/\sigma^2_{Ak}\right)\sum_{k\in B_1}\frac{N_k^2\hat{p}_{Ak}\sigma^2_{Ak}}{n_k}\left(\gamma^2_{Ak}\left(1-\hat{p}_{Ak}\right)+1\right)}{\sum_{k\in B_2} w_k\left(n_{Ak}-1\right)}\right]^{1/2} t_{1-\alpha/2,\nu} \qquad (11)$$

We temporarily let $\Theta = \sum_{k\in B_1}\frac{N_k^2\hat{p}_{Ak}\sigma^2_{Ak}}{n_k}\left(\gamma^2_{Ak}\left(1-\hat{p}_{Ak}\right)+1\right)$ and consider two alternatives for specifying the w_k:

<u>Alternative 1</u> Let $w_k = \frac{N_k^2\hat{p}_{Ak}\sigma^2_{Ak}}{n_k\left(n_{Ak}-1\right)}\left(\gamma^2_{Ak}\left(1-\hat{p}_{Ak}\right)+1\right) = \frac{N_k^2\hat{p}_{Ak}}{n_k\left(n_{Ak}-1\right)}\left(\mu^2_{Ak}\left(1-\hat{p}_{Ak}\right)+\sigma^2_{Ak}\right)$.

Then, from equation (11),

$$u_1\left(\mathbf{w}\right) = \hat{T}_A + \Theta^{1/2} t_{1-\alpha/2,\nu_1}. \qquad (12)$$

where the parameter ν_1 is referred to as the weighted degrees of freedom.

For $k \in B_2$, we use $\hat{S}^2_{Ak}$ to estimate σ^2_{Ak}, $\hat{\mu}^2_{Ak} - \hat{S}^2_{Ak}/n_{Ak}$ to estimate μ^2_{Ak} and $\hat{w}_k = \frac{N_k^2 \hat{P}_{Ak}}{n_k\left(n_{Ak}-1\right)}\left(\left(\hat{\mu}^2_{Ak} - \hat{S}^2_{Ak}/n_{Ak}\right)\left(1-\hat{P}_{Ak}\right)+\hat{S}^2_{Ak}\right)$ to estimate w_k. We can then estimate the weighted degrees of freedom with $\hat{\nu}_1 = \left[\sum_{k\in B_2} \hat{w}_k\left(n_{Ak}-1\right)\right]^2 \Big/ \sum_{k\in B_2} \hat{w}_k^2\left(n_{Ak}-1\right)$. We will delay dealing with Θ until after specifying the second alternative.

Alternative 2 Let $w_k = 1$ then from equation (11)

$$u_2\left(\mathbf{w}\right) = \hat{T}_A + \Theta^{1/2}\left[\frac{\sum_{k\in B_2}\left(n_{Ak}-1\right)\left(\hat{S}^2_{Ak}/\sigma^2_{Ak}\right)}{\sum_{k\in B_2}\left(n_{Ak}-1\right)}\right]^{1/2} t_{1-\alpha/2,\nu_2} \tag{13}$$

with unweighted degrees of freedom $\nu_2 = \sum_{k\in B_2}\left(n_{Ak}-1\right)$. If we estimate σ^2_{Ak} with $\hat{S}^2_{Ak}$ then equation (13) simplifies to

$$u_2\left(\mathbf{w}\right) = \hat{T}_A + \Theta^{1/2} t_{1-\alpha/2,\nu_2} \tag{14}$$

It is straightforward to show that $\nu_2 \geq \hat{\nu}_1$, hence, for any specified value of Θ, the length of the confidence intervals under alternative 2 is less than or equal the length under alternative 1. Using a different approach, Kott (1994) also recommends lowering degrees of freedom using the Satterthwaite approximation. Similarly, Johnson and Rust (1993) use the Satterthwaite approximation to get degrees of freedom corresponding to a resampling variance estimator.

Addressing the problem of estimating Θ, we have

$$\begin{aligned}\Theta &= \sum_{k\in B_1} \frac{N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2}{n_k}\left(\gamma_{Ak}^2\left(1-\hat{p}_{Ak}\right)+1\right)\\ &= \sum_{k\in B_1} \frac{N_k^2 \hat{p}_{Ak}}{n_k}\left(\mu_{Ak}^2\left(1-\hat{p}_{Ak}\right)+\sigma_{Ak}^2\right)\\ &= \sum_{k\in B_1-B_2} \frac{N_k^2 \hat{p}_{Ak}}{n_k}\left(\mu_{Ak}^2\left(1-\hat{p}_{Ak}\right)+\sigma_{Ak}^2\right)+\sum_{k\in B_2} \frac{N_k^2 \hat{p}_{Ak}}{n_k}\left(\mu_{Ak}^2\left(1-\hat{p}_{Ak}\right)+\sigma_{Ak}^2\right).\end{aligned}$$

For $k \in B_1 - B_2$ the estimator $\hat{\sigma}_{Ak}^2$ is not defined, however, it is straightforward to verify that $\left(1-\hat{p}_{Ak}\right)E\left[\hat{\mu}_{Ak}^2 \middle| n_{Ak}\right] \le \sigma_{Ak}^2 + \mu_{Ak}^2\left(1-\hat{p}_{Ak}\right) \le E\left[\hat{\mu}_{Ak}^2 \middle| n_{Ak}\right]$. Therefore,

$$\begin{aligned}s_a^2 &= \sum_{k\in B_1-B_2} \frac{N_k^2 \hat{p}_{Ak}\left(1-\hat{p}_{Ak}\right)\hat{\mu}_{Ak}^2}{n_k} + \sum_{k\in B_2} \frac{N_k^2 \hat{p}_{Ak}}{n_k}\left\{\hat{\sigma}_{Ak}^2\left[1-\left(1-\hat{p}_{Ak}\right)/n_{Ak}\right]+\hat{\mu}_{Ak}^2\left(1-\hat{p}_{Ak}\right)\right\}\\ &= \sum_{k\in B_1} \frac{N_k^2 \hat{p}_{Ak}}{n_k}\left(1-\hat{p}_{Ak}\right)\hat{\mu}_{Ak}^2 + \sum_{k\in B_2} \frac{N_k^2 \hat{p}_{Ak}}{n_k}\hat{\sigma}_{Ak}^2\left(1+1/n_k-1/n_{Ak}\right)\end{aligned} \qquad (15)$$

is an "under-estimate" for Θ and

$$s_b^2 = \sum_{k\in B_1-B_2} \frac{N_k^2 \hat{p}_{Ak}^2}{n_k}\hat{\mu}_{Ak}^2 + \sum_{k\in B_1} \frac{N_k^2 \hat{p}_{Ak}}{n_k}\left(1-\hat{p}_{Ak}\right)\hat{\mu}_{Ak}^2 + \sum_{k\in B_2} \frac{N_k^2 \hat{p}_{Ak}}{n_k}\hat{\sigma}_{Ak}^2\left(1+1/n_k-1/n_{Ak}\right) \qquad (16)$$

is an "over-estimate". Clearly, $s_a^2 \le s_b^2$ with equality only when $B_1 = B_2$.

It can also be verified that in the case of stratified sampling, the standard variance estimator for estimated population totals is

$$\begin{aligned}s_{std}^2 &= \sum_{k\in B_1} N_k^2 \frac{s_k^2}{n_k}\\ &= \sum_{k\in B_1} \frac{N_k^2 \hat{p}_{Ak}}{n_k-1}\left(1-\hat{p}_{Ak}\right)\hat{\mu}_{Ak}^2 + \sum_{k\in B_2} \frac{N_k^2 \hat{p}_{Ak}}{n_k-1}\hat{\sigma}_{Ak}^2\left(1-1/n_{Ak}\right).\end{aligned} \qquad (17)$$

Note that s_{std}^2 and s_a^2 are equal to terms of order n_k^{-1}, however, s_a^2 will tend to be smaller in most practical situations.

These results imply that CIs of the form $\left(\hat{T}_A \pm s_b t_{1-\alpha/2,\hat{\nu}_1}\right)$ will provide the highest level of coverage while CIs of the form $\left(\hat{T}_A \pm s_a t_{1-\alpha/2,\nu_2}\right)$ will provide the lowest level. Also, CIs of the form $\left(\hat{T}_A \pm s_{std} t_{1-\alpha/2,\hat{\nu}_1}\right)$ and (especially) $\left(\hat{T}_A \pm s_{std} t_{1-\alpha/2,\nu_2}\right)$ have obvious computational advantages. Several of these competing forms of CI were evaluated in an empirical study which is reported in detail in Section 3.5.

3.4 The Case of Population Means and Medians

The results in the preceding sections can easily be extended to ratio estimators by the standard linearization approach. By way of example, suppose we are interested in estimating the average wage for workers in a particular occupation via a sample of business establishments; here we let *A* consist of establishments with employees in the occupation of interest. Let $\hat{T}_A(W)$ and $\hat{T}_A(M)$ be estimators of total wages (*W*) for employees in occupation *A* and total number of employees (*M*) in occupation *A* as in equation (9). We can write these estimators as $\hat{T}_A(W) = \sum_{i=1}^{n_A} c_i y_i$ and $\hat{T}_A(M) = \sum_{i=1}^{n_A} c_i m_i$, where for the i^{th} establishment, c_i is the sampling "weight", m_i is the total number of employees in occupation *A* and y_i is total wages for employees in occupation *A*. The ratio estimator $\overline{w}_A = \hat{T}_A(W)/\hat{T}_A(M)$ is used to estimate the average wage ($\overline{W}_A = W/M$) for employees in occupation *A*. It is straightforward to verify that the usual linear approximation for the difference $\left(\overline{w}_A - \overline{W}_A\right)$ is given by

$$\lambda\left(\overline{w}_A - \overline{W}_A\right) = \sum_{i=1}^{n_A} c_i z_i\,, \qquad (18)$$

where $z_i = M^{-1}\left(y_i - \overline{W}_A x_i\right)$. For stratified random sampling, we can re-label the sample establishments to reflect stratification and then equation (18) can be written as

$$\lambda\left(\overline{w}_A - \overline{W}_A\right) = \sum_{k=1}^{K} \left(N_k / n_k\right) \sum_{i=1}^{n_{Ak}} z_{ki} = \sum_{k=1}^{K} \hat{Z}_{Ak} \; .$$

This is of the same form as equation (9), so the results of Section 3.3 apply to $\lambda(\overline{w}_A - \overline{W}_A)$ under the appropriate normality assumptions on the z_{ki} .

Preliminary analytical investigation indicates that these results can also be extended to the construction of CIs for population medians, or other percentiles, by the use of either the Woodruff (1952) or the Francisco and Fuller (1991) techniques. Detailed investigation of this extension is beyond the scope of the current paper.

3.5 The Empirical Study For Stratified Random Sampling

Results on coverage and mean interval length, from two simulation studies, both on populations derived from a test sample of the Occupational Compensation Survey Program (OCSP) conducted in 1991, are included in Tables 1-4. One population (the "Small Population") took the sample itself as the population, with six non-certainty strata, and one certainty stratum of 12 establishments. Repeated samples were taken from this population at sizes n=36 and 60, corresponding to the choices n_k =4 and n_k =8. The second population (the "Large Population") was constructed by expanding the sample data through replication of establishments to achieve a population the size of the original population; again there were six noncertainty and 1 certainty strata; for each stratum samples were of the size of the actual sample. Domains are defined by the different occupations of interest; only a fraction of establishments have workers in a particular occupation, and lie in the corresponding domain.

In both cases sampling was without replacement, so finite population correction factors were included (as appropriate) in the construction of the CIs. Also, the study was limited to a concern with 95% coverage.

SMALL POPULATION: Table 1 for total wages and Table 2 for mean wages give coverage and interval length, at two sample sizes n_k =4 and n_k =8, for 8 occupations, and 4 variance-degrees of freedom combinations: the standard variance estimator, s_{std}^2, with the standard normal *z*-quantile, and with the unweighted and weighted degrees of freedom. Results are based on 500 runs. Occupations are ordered by increasing values of the average value over runs of ν_2. We note:

(1) Almost universally, coverage using the standard variance estimator and the standard normal quantiles (infinite *df*) is poor.

(2) Coverage for the other interval types is far more satisfactory, in the main matching nominal or conservative for the weighted degrees of freedom; as expected the unweighted degrees of freedom tends to yield coverage a few points below the weighted degrees of freedom coverage.

(3) Confidence intervals for means are better behaved on the whole than for totals. Two occupations (1122, 4021) yield seriously low coverage for totals even with the improved procedures; only 4021 gives poor coverage for means.

Interval lengths are taken relative to $2 \cdot z_{.975} \approx 4$ times the root mean square error of $\hat{T}_A$ calculated over runs; when the distribution of $\hat{T}$ is actually normal this ratio is 1.

(4) The relative interval length of the standard interval tends to be too small, that is, less than 1.

(5) Interval length among the other variance-degrees of freedom combinations is largest for s_{std}^2 with $\hat{\nu}_1$, and smallest for s_{std}^2 with ν_2. These differences can be appreciable; there is a tradeoff between coverage and interval size.

(6) For a given interval type, the relative interval length tends to 1 as ν_2 increases.

LARGE POPULATION. Tables 3-4 give coverage and interval length for totals workers and mean wage for the same four interval types, and a wider range of occupations, ordered by average n_2. Results are based on 5000 runs. Here the interval lengths are taken relative to the median interval length for the standard normal confidence interval. We note:

(1) Results are consistent with those on the Small Population, in terms of the relative coverage and interval sizes of the several interval types. The standard normal is unsatisfactory for many occupations.

(2) Coverage using $\hat{I}_1$ is less than 90% only in a small fraction of cases.

(3) There can be marked differences in interval length for the different interval types; however, all ratios of interval length to 4 · root mean square error tend to 1, as n_2 gets large.

(4) There are some differences in problem occupations from the Small Population Study; for example, the coverage for 4021 is much improved, but 2911 has poor coverage, especially for the mean, even with the non-standard intervals. These differences are probably due to some differences in the way the populations were structured; in particular, all certainty establishments in the original sample were treated as certainties in the Large Population; this was not the case in the Small Population.

(5) In the main, coverage is better for means than for totals, but there are some obvious exceptions, especially at low values of n_2.

4. SUMMARY AND CONCLUSIONS

From our theoretical investigation and the two simulation studies relying on OCSP data, we draw the following conclusions:

1. Standard 95% confidence intervals for domain means or totals, when based on the standard normal distribution and standard methods of variance estimation, tend to yield less than actual 95% coverage. The extent of the deviance will vary with domain

(occupation in the simulation study), but can be quite considerable even when the sample is large.

2. New nonstandard methods offer a sharp improvement, giving intervals with better coverage, typically at or close to the nominal 95% coverage. These intervals tend to be longer than the standard intervals. The increase in length will vary with domain, and will depend on the particular method for CI construction that is adopted among those we have considered. "Asymptotically", that is for "large sample domains", there will be little difference from standard intervals.

3. The basic ideas behind these intervals are (1) conditioning on the amount of information on the particular occupation, which, roughly speaking, is measured in terms of the number of units in the sample that belong to the domain, and (2) An important unknown is the fraction within each stratum of such units, and to handle this we put a prior distribution on this unknown, reflective of the degree of our ignorance of it, an idea we borrow from the Bayesians. However, the bottom line here is coverage probabilities.

4. The principal effect of these ideas is the abandonment, for purposes of CI construction, of the standard normal quantiles (±1.96 for 95% coverage). These are replaced by quantiles from the Student's *t*-distribution, with degrees of freedom determined from the sample and varying with domain. If because of publication requirements or for other reasons, there is need to report standard deviations rather than confidence intervals, then we recommend reporting an *effective* standard deviation given by the length of the 95% interval divided by twice 1.96.

5. The most likely candidate for estimate of variance, accompanying the new *t*-quantile, is the standard estimate of variance. In most instances this should be quite satisfactory, so

that the only change will be in the introduction of the new degrees of freedom methodology. However, we have considered alternatives to the standard variance estimator, which in some instances may improve coverage or reduce the length of confidence intervals.

6. An open question concerns what degree and type of collapsing of strata (if any) should be used in the estimation of variances and of the degrees of freedom for the purpose of confidence interval construction. In general, there will be a tradeoff: as strata are reduced in number, the estimate of variance will tend to increase, but so will the degrees of freedom (reducing the size of t_{n_2} or $t_{\hat{n}_1}$.) The answer to this question may well be population specific, and experience of the population from past surveys useful.

REFERENCES

DORFMAN, A. and VALLIANT, R. (1993). Quantile Variance Estimators in Complex Surveys, *American Statistical Association 1993 Proceedings of the Section of Survey Research Methods*, to appear.

FRANCISCO, C.A. and FULLER, W. A. (1991). Quantile Estimation with a Complex Survey Design, *Annals of Statistics*, 39, 454-469.

JOHNSON, E.G. and RUST, K.F. (1993). Effective Degrees of Freedom for Variance Estimates from a Complex Sample Survey, *American Statistical Association 1993 Proceedings of the Section of Survey Research Methods*, to appear.

KOTT, P. S. (1994). A Hypothesis Test of Linear Regression Coefficients with Survey Data, *Survey Methodology*, to appear.

SARNDAL, C-E, SWENSSON, B. and WRETMAN, J. (1992). Model Assisted Survey Sampling, New York: Springer-Verlag.

SATTERTHWAITE, F. (1946). An approximate Distribution of Estimates of Variance Components, *Biometrics*, 2, 110-114.

WOODRUFF, R.S. (1952). Confidence Intervals for Medians and other Position Measures, *Journal of the American Statistical Association*, 47, 635-646.

Appendix A

From the discussion in Section 2.2 we know that $n\hat{p}_A$ has a binomial distribution $Bi(n, p_A)$, hence, for $\hat{p}_A = 0,\ 1/n,\ 2/n, \ldots,\ 1$

$$
\begin{aligned}
f\left(\hat{p}_A \middle| p_A\right) &= \frac{\Gamma(n+1)}{\Gamma\left(n\hat{p}_A+1\right)\Gamma\left(n\left(1-\hat{p}_A\right)+1\right)} p_A^{n\hat{p}_A}\left(1-p_A\right)^{n\left(1-\hat{p}_A\right)} \\
&= \frac{\Gamma(n+1)}{\Gamma(n+2)} \frac{\Gamma(n+2)}{\Gamma\left(n\hat{p}_A+1\right)\Gamma\left(n\left(1-\hat{p}_A\right)+1\right)} p_A^{\left(n\hat{p}_A+1\right)-1}\left(1-p_A\right)^{\left(n\left(1-\hat{p}_A\right)+1\right)-1} \\
&= \frac{1}{n+1} k_{\hat{p}_A}\left(p_A\right).
\end{aligned}
$$

For each (fixed) value of $\hat{p}_A$, the function $k_{\hat{p}_A}(p_A)$ is the pdf of a Beta distribution with parameters $\omega_1 = n\hat{p}_A + 1$ and $\omega_2 = n(1-\hat{p}_A)+1$. As both ω_1 and ω_2 will be larger than unity with high probability (at least in most real world situations), it is reasonable to approximate $k_{\hat{p}_A}(p_A)$ with a normal pdf having equivalent mean and variance. For the Beta distribution in question the mean and variance are

$$\frac{n\hat{p}_A + 1}{n+2} \approx \hat{p}_A \text{ and}$$

$$\frac{\left(n\hat{p}_A+1\right)\left(n\left(1-\hat{p}_A\right)+1\right)}{(n+2)^2(n+3)} \approx \frac{\hat{p}_A\left(1-\hat{p}_A\right)}{n}.$$

Thus, the approximation is $k_{\hat{p}_A}\left(p_A\right) \approx \dfrac{1}{\sqrt{2\pi}\sqrt{\hat{p}_A\left(1-\hat{p}_A\right)/n}}\, e^{-\frac{1}{2}\frac{\left(p_A-\hat{p}_A\right)^2}{\hat{p}_A\left(1-\hat{p}_A\right)/n}}$.

Assuming that p_A is distributed $N(\mu, \sigma^2)$ it follows from the Bayes formula that the posterior distribution is

$$h(p_A|\hat{p}_A) = f(\hat{p}_A|p_A)\,g(p_A) \Big/ \int_0^1 f(\hat{p}_A|p_A)\,g(p_A)\,dp_A$$
$$\cong c e^{-\frac{1}{2}\left[\frac{(p_A-\hat{p}_A)^2}{\hat{p}_A(1-\hat{p}_A)/n}+\frac{(p_A-\mu)^2}{\sigma^2}\right]},$$

where $c = \left[\int_0^1 e^{-\frac{1}{2}\left[\frac{(p_A-\hat{p}_A)^2}{\hat{p}_A(1-\hat{p}_A)/n}+\frac{(p_A-\mu)^2}{\sigma^2}\right]} dp_A\right]^{-1} \cong \frac{\sqrt{(\hat{p}_A(1-\hat{p}_A)/n)+\sigma^2}}{\sigma\sqrt{2\pi}\sqrt{\hat{p}_A(1-\hat{p}_A)/n}}\, e^{\frac{1}{2}\left[\frac{(\hat{p}_A-\mu)^2}{(\hat{p}_A(1-\hat{p}_A)/n)+\sigma^2}\right]}$ is the normalizing constant.

Under the "empirical Bayes" assumption that $\mu = \hat{p}_A$ and $\sigma^2 = \hat{p}_A(1-\hat{p}_A)/n$ we have

$$h(p_A|\hat{p}_A) \cong \frac{1}{\sqrt{2\pi}\sqrt{\hat{p}_A(1-\hat{p}_A)/2n}}\, e^{-\frac{1}{2}\left[\frac{(p_A-\hat{p}_A)^2}{\hat{p}_A(1-\hat{p}_A)/2n}\right]}.$$

If we drop the specific assumption regarding σ^2, and let $\psi = (\hat{p}_A(1-\hat{p}_A)/n)/\sigma^2$ then $[p_A|\hat{p}_A]$ is distributed normal with mean equal $\hat{p}_A$ and variance $\frac{\hat{p}_A(1-\hat{p}_A)}{(1+\psi)n}$. Under the empirical Bayes variance assumption we have $\psi = 1$.

Appendix B

Result: Assume W is distributed $N(0, c^2)$ and, conditional on $W = w$, the random variable T is distributed as a non-central *t* with ν degrees of freedom and non-centrality parameter *w*. Then, the unconditional distribution of $T/\sqrt{c^2+1}$ is central *t* with ν degrees of freedom.

Proof: First notice that *T* can be written as

$$T = \frac{X + W}{\sqrt{S^2/\nu}},$$

where X is distributed as $N(0,1)$, S^2 is distributed as χ^2_ν, and X, W, and S^2 are mutually independent. Therefore, $X' = (X + W)/\sqrt{1+c^2}$ is distributed as $N(0,1)$. As X' and S^2 are independent, it follows by definition that

$$T' = \frac{T}{\sqrt{1+c^2}} = \frac{X'}{\sqrt{S^2/\nu}}$$

is distributed as t_ν.

Appendix C

We here make some observations on the use of $n_a - 1$ degrees of freedom with the standard statistic $\frac{n^{1/2}\left(\hat{T}_A - T_A\right)}{Ns}$, corresponding to the empirical Bayes format with $\Psi_A = 0$, from a frequentist standpoint.

The question is why the *t* distribution with n_A - 1 degrees of freedom, which would seem appropriate for use with a conditional distribution based on the domain size n_A and the within domain variance estimate $\hat{\sigma}^2_A$, yields generally sound confidence intervals in conjunction with the above unconditional statistic based on sample size *n* and sample variance s^2. We focus on 95% coverage and let $t^* = t\left(n_A - 1; .975\right)$. Also, assume μ_A is positive and, to abbreviate notation, let $\mu \equiv \mu_A$, $\sigma^2 \equiv \sigma^2_A$, $\gamma \equiv \gamma_A$, $p \equiv p_A$ and $\hat{p} \equiv \hat{p}_A$.

Consider Figure 1, which, for 500 samples of size 300 selected with replacement from a population of size 3000 having *p*=0.03 and γ = 9, graphs the ratio *R*= $(\hat{T}_A - T_A)/\left(Nst^*/n^{1/2}\right)$ against n_A; if the ratio *R* is less than 1, then T_A is in the corresponding interval estimate. We note (i) the intervals are well behaved, in fact conservative in that more than 95% of *R*-values are less than 1, (ii) that as n_A increases, *R* increases from relatively large negative to relatively large positive, seeming to reflect changes in the bias of $\hat{T}_A - T_A$ as n_A changes. Furthermore, there is a "within n_A" variation, which in the main seems to increase as n_A decreases.

We proceed to analyze *R* in light of this figure, attempting to get a handle on the bias and spread, for each $n_A \geq 2$ (or, equivalently, the "across n_A" and "within n_A" variation respectively.). Conditional on n_A, the bias of the numerator $E\left(\hat{T}_A - T_A\right) = N\left(\hat{p} - p\right)\mu \equiv b_{\hat{p}}$. Then *R*=$\frac{\hat{T}_A - T_A - b_{\hat{p}}}{t^* N\hat{p}^{1/2}\hat{\sigma}/n^{1/2}}\frac{\hat{p}^{1/2}\hat{\sigma}}{s} + \frac{b_{\hat{p}}}{t^* Ns/n^{1/2}}$. We suggest that, for given n_A, the first term reflects the spread of *R* , the second the bias.

For moderate or large n, we have by (1) that $s^2 = \hat{p}\left(1-\hat{p}\right)\hat{\mu}^2 + \frac{n_A - 1}{n}\hat{\sigma}^2$, and substituting this in both terms, we derive $R = \frac{\hat{T}_A - T_A - b_{\hat{p}}}{t^* N\hat{p}^{1/2}\hat{\sigma}/n^{1/2}} f + \frac{n_A - pn}{t^* n_A^{1/2}} g$, where $f = \left[\hat{p}\left(1-\hat{p}\right)\hat{\gamma}^2 + \left(1 - 1/n_A\right)\right]^{1/2}$ and $g = \left\{\frac{\tilde{\gamma}^2}{\left[\hat{p}\left(1-\hat{p}\right)\hat{\gamma}^2 + 1 - 1/n_A\right]}\right\}^{1/2}$, for $\tilde{\gamma} = \mu/\hat{\sigma}$ and $\hat{\gamma} = \hat{\mu}/\hat{\sigma}$.

Consider the first factor in each term. By Section 2.3, $\frac{\hat{T}_A - T_A - b_{\hat{p}}}{N\hat{p}^{1/2}\hat{\sigma}/n^{1/2}}$ is a standard studentized statistic and so, divided by t^*, will lie between -1 and 1, 95% of the time (this would not be so if we used the conventional z-statistic, in place of t^* or t with larger degrees of freedom). The first factor of the second term is fixed for given n_A. For given n,p we ask how often n_A will be such that this value will be large. The expression $\frac{n_A - pn}{n_A^{1/2}} = \left(\hat{p} - p\right)\Big/\sqrt{\frac{\hat{p}}{n}}$ is bounded above by the conventional binomial statistic, which for np moderate has an approximate z-distribution. But in the situation we are concerned with, np is typically small, and the division by t^* instead of $z_{.975}$ is an important safety factor. Values of $\frac{n_A - pn}{t^* n_A^{1/2}}$ and $\frac{n_A - pn}{z_{.975} n_A^{1/2}}$ and the probabilities with which they arise, were tabulated for n = 50, 100, 500, and 1000 and for a range of values of p, for example, for n=50, p= 0.01, .02, .04, .06, .12, .18. It was found that large (absolute) values of $\frac{n_A - pn}{t^* n_A^{1/2}}$ (say bigger than 0.8) have extremely small probability, with the most vulnerable situation being the possibility of getting n_A=3, 4, or 5, when np is about 12 or 15. In particular, the case when n_A=2 is not worrisome because of the large value of t*. By contrast, large values of $\frac{n_A - pn}{z_{.975} n_A^{1/2}}$ were not so improbable.

These considerations suggest the bias will be less than g (or $0.8g$) with extremely high probability, and the within n_A variation will be less than f with probability 95%. Note that for given n_A, f and g are functions of $\hat{\gamma}, \tilde{\gamma}$ both estimating γ. For the moment consider them as functions of γ itself, ignoring their stochastic variation. Then it is easy to see that f is monotonic down, with a maximum at γ=0 of $\left(1 - 1/n_A\right)^{1/2}$ and asymptoting to zero as γ increases [for example at γ=3 (the approximate practical lower bound we found in wage data), for n_A=2 (worst case) and $\hat{\rho}$ negligibly small, *f*=0.32]. On the other hand, *g* is monotonic increasing, with a value of 0 at γ=0 and asymptoting to $\left(1-\hat{\rho}\right)^{1/2} \approx 1$. Thus the larger γ, the more pronounced the across n_A variation and the less the within n_A variation, and vice versa. At γ=0, the bias term is zero. Clearly *f*+*g* is bounded by a small number, achieved when γ is between 0 and 1. In fact, squaring *f*+*g* and taking the derivative, one finds that to maximize *f*+*g,* we have $\gamma \approx 1 - 1/n_A$. Table 1 gives values of γ yielding maximum value of *f*+*g*, and the values so yielded, for small values of n_A. Table C-1 suggests the worst case for coverage occurs for γ about 0.65.

Table C-1. Maximum Values of *f*+*g*, for given γ

n_A	2	3	4	5	6	7	8	9	10
γ	.5	.67	.75	.8	.83	.86	.88	.89	.9
f	1.15	.95	.87	.83	.81	.79	.78	.77	.76
g	.58	.63	.65	.67	.67	.68	.68	.69	.69
f+*g*	1.73	1.58	1.53	1.50	1.48	1.47	1.46	1.46	1.45

Note: As n_A increases, both f and g approach $\sqrt{2}/2$

This discussion ignored the fact that not γ, but $\hat{\gamma}$ and $\tilde{\gamma}$ appear in *f* and *g*. Note that were it not for the fact that there are *two* estimators of γ, the above argument goes through, with same bounds on *f*+*g*, etc., with $\hat{\gamma}$ for example replacing γ. However, $\tilde{\gamma}$ and $\hat{\gamma}$ should typically be close. In particular, it is easy to see that $\tilde{\gamma}-\hat{\gamma}$ is distributed as $t / n_A^{1/2}$, where t has a t-distribution with n_A-1 degrees of freedom.

An incidental observation is that for large γ the within n_A distribution is clearly skewed downward (see Figure 2). The distribution of the ratio $\hat{\gamma}$ is skewed positive, so that f is skewed negatively, accounting for the effect.