CONFIDENCE INTERVALS FOR SUB-DOMAIN MEANS AND TOTALS

Robert J. Casady, Alan Dorfman, Bureau of Labor Statistics
Suojin Wang, Texas A&M University
Robert J. Casady, BLS, 2 Massachusetts Ave NE Washington, DC 20212

Key Words: Chi-square distribution, degrees of freedom, sampling

1. Introduction.

Let P be a population of Primary Sampling Units (PSU's) labeled hi, where $h=1,\ldots,H$ and $i=1,\ldots,N_h$. Let $A \subset P$ be a domain of interest of unknown size. We are interested in inference on A based on a sample s from P. An overview of domain estimation is given in (Sarndal et. al. 1992, Chapter 10). That there can be a serious problem in the construction of sound confidence intervals for domain quantities is noted in (Dorfman and Valliant 1993) in the context of quantile estimation. The present work verifies that this problem extends to means and totals, and offers a remedy.

We proceed by way of considering an establishment survey, taking wages earned as the variate of interest. This will help fix ideas, and leads into our simulation study, which was based on such a survey.

Suppose the PSU's are establishments and the domain A are those establishments having workers in a particular occupation **O**. Consider the problem of estimating the total wages for workers in **O**. (The problem of estimating the mean wage is basically a mild extension of this.) Without loss of generality re-order establishments within each stratum h so that "contributors" (establishments with one or more workers in **O**) are listed first. Then, if N_{Ah} is the number of contributors in the hth stratum, we define Y_{hi}=total wages paid by establishment hi to workers in **O**, for $i=1,\ldots,N_{Ah}$, and $Y_{hi}=0$, for $i=N_{Ah+1},\ldots,N_h$. The goal is to estimate $T \equiv \sum_{h=1}^{H}\sum_{i=1}^{N_{Ah}} Y_{hi} = \sum_{h=1}^{H}\sum_{i=1}^{N_h} Y_{hi}$.

Taking a stratified random sample without replacement, we re-order establishments so that sampled establishments are listed first, and, within sampled establishments, contributors first. Then the standard estimator $\hat{T} = \sum_{h=1}^{H} N_h n_h^{-1} \sum_{i=1}^{n_h} Y_{hi}$, where n_h= number of sampled establishments in stratum h, can be written $\hat{T} = \sum_{h=1}^{H} N_h \hat{p}_{Ah}\hat{\mu}_{Ah}$, where n_{Ah} is the number of contributors in sample of hth stratum, $\hat{p}_{Ah} = n_{Ah}/n_h$ estimates $p_{Ah} = N_{Ah}/N_h$, the proportion of contributors in h, and $\hat{\mu}_{Ah} = \sum_{i=1}^{n_{Ah}} Y_{hi} \Big/ n_{Ah}$ estimates $\mu_{Ah} = \sum_{i=1}^{N_{Ah}} Y_{hi} \Big/ N_{Ah}$, the mean wage in h, that is, the mean of the total amount Y_{hi} establishments pay to their workers in occupation **O**.

The standard variance estimator used to measure the error $\hat{T} - T$ is $s^2 \equiv \hat{\mathrm{var}}\left(\hat{T} - T\right) = \sum_{h=1}^{N_h} N_h^2 n_h^{-1}\left(1 - n_h/N_h\right)s_h^2$, where $s_h^2 = \sum_{i=1}^{n_h}\left(Y_{hi} - \bar{Y}_h\right)^2 \Big/ \left(n_h - 1\right)$, with $\bar{Y}_h = \sum_{i=1}^{n_h} Y_{hi} \Big/ n_h$; note that these last expressions contain the zero Y_{hi}'s. Typically, we use Normal Quantiles to construct confidence intervals, for example, $z_{.975} = 1.96 \approx 2$ for 95% confidence intervals. We can think of this as using degrees of freedom equal to infinity ($df=\infty$).

Note that s_h^2, the key ingredient in the above variance estimate, can be written in terms of $\hat{\mu}_{Ah}$, $\hat{\sigma}_{Ah}^2 = \sum_{i=1}^{n_{Ah}}\left(Y_{hi} - \hat{\mu}_{Ah}\right)^2 \Big/ \left(n_{Ah} - 1\right)$ (the sample domain variance) and $\hat{p}_{Ah}$:

$$s_h^2 = \frac{n_h}{n_h - 1}\hat{p}_{Ah}\left(1 - \hat{p}_{Ah}\right)\hat{\mu}_{Ah}^2 + \frac{n_h\hat{p}_{Ah} - 1}{n_h - 1}\hat{\sigma}_{Ah}^2.$$

2. New variance estimators and degrees of freedom

Alternative variance estimators and alternative expresssions for degrees of freedom can be derived by:

(i) Conditioning on $\hat{p}_{Ah} = n_{Ah}/n_h$; the attempt here is to embody the recognition that in cases where n_{Ah} is small, we have less information contributing to the estimator $\hat{T}$.

(ii) Setting a prior distribution on p_{Ah}, the unknown proportion of contributors in h. Specifically, we assume p_{Ah} normally distributed with mean $\hat{p}_{Ah}$

and variance $\{\hat{p}_{Ah}(1-\hat{p}_{Ah})/n_h\}/\psi_{Ah}$ where the expression in curly brackets is the standard ("empirical Bayes") prior variance of p_{Ah}, and ψ_{Ah} can be selected in different ways to yield alternative variance estimates. The idea is to capture our uncertainty about p_{Ah}. This step also makes tractable the conditional bias

$E(\hat{T}-T|\hat{p}_{Ah}) \approx \sum_{h=1}^{H} N_h \mu_{Ah}(\hat{p}_{Ah}-p_{Ah})$. Roughly speaking, by using this prior, the bias can be regarded as *zero*, and a component is added to the conditional variance, yielding a new variance estimate (or, rather family of estimates corresponding to choice of ψ_{Ah}):

$$\hat{\text{var}}(\hat{T}-T|p_{Ah}) \equiv s_{\psi}^2 \equiv$$

$$\sum_{h=1}^{H} N_h^2 n_h^{-1}(1-n_h/N_h)\left\{\hat{p}_{Ah}\left[\frac{1-\hat{p}_{Ah}}{1+\psi_{Ah}}\gamma_{Ah}^2+1\right]\hat{\sigma}_{Ah}^2\right\}$$

where $\gamma_{Ah}^2 = \mu_{Ah}^2/\sigma_{Ah}^2$ is the reciprocal of the squared coefficent of variation in stratum h. Note that γ_{Ah}^2 will be estimated from the sample data. When ψ_{Ah}=0, the factor in curly brackets is approximately s_h^2, the corresponding component in s^2, the standard variance estimator above. So $s_0^2 \approx s^2$; under some necessary adjustment (see (iv) below), we find $s_0^2 > s^2$.

(iii) Getting degrees of freedom. A simple unweighted expression for df is $df_{uw} = \sum_{h=1}^{H}(n_{Ah}-1)$. A "weighted df" is derived from a chi-square approximation to the distribution of s_{ψ}^2 using the approach of Satterthwaite (1946). We write

$$s_{\psi}^2 = \sum_{h=1}^{H} w_h \frac{(n_{Ah}-1)\hat{\sigma}_{Ah}^2}{\sigma_{Ah}^2}, \text{ where}$$

$$w_h = N_h^2 n_h^{-1}(1-n_h/N_h)\times$$

$$\left\{\hat{p}_{Ah}\left[\frac{1+\hat{p}_{Ah}}{1+\psi_{Ah}}\gamma_{Ah}^2+1\right]\sigma_{Ah}^2\right\}(n_{Ah}-1)^{-1}$$ and note that $\frac{(n_{Ah}-1)\hat{\sigma}_{Ah}^2}{\sigma_{Ah}^2}$ has a $\chi_{n_{Ah}-1}^2$ distribution so that s_{ψ}^2 is a mixture of chi-squares. Such a mixture can be approximated by a factor times a simple chi-square having degrees of freedom

$$df_w = \left\{\sum_{h=1}^{H} w_h(n_{Ah}-1)\right\}^2 \Big/ \sum_{h=1}^{H} w_h^2(n_{Ah}-1).$$ It can be shown that $df_w \le df_{uw}$. We substitute $\hat{\gamma}_{Ah}^2 = \hat{\mu}_{Ah}^2/\hat{\sigma}_{Ah}^2$ and $\hat{\sigma}_{Ah}^2$ in w_h where needed. We note that, in the context of hypothesis testing in complex surveys, Kott (1994) also recommends lowering degrees of freedom using the Satterthwaite approximation, using a somewhat different approach. Likewise, Johnson and Rust (1993) use the Satterthwaite approximation to get degrees of freedom corresponding to a resampling variance estimator.

(iv) The $n_{Ah} < 2$ problem. In case $n_{Ah} = 1$, $\hat{\sigma}_{Ah}^2$ and terms in df_w cannot be calculated. The basic strategy is to substitute a global estimate $\hat{\sigma}_A^2$ for unavailable $\hat{\sigma}_{Ah}^2$, and recalculate df_w based on the modified s_{ψ}^2, numerical examples suggest there will be very little difference from estimating the degrees of freedom by restricting the original expression for df_w to $n_{Ah} \ge 2$. Strata with $n_{Ah} = 0$ are omitted from calculation.

3. Simulation Studies

Simulation studies were carried out, on two populations both derived from data arising in the Bureau of Labor Statistics Occupational Compensation Survey Program (OCSP). One population (the "Small Population") took the sample itself as the population, and sampled from 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_h=4 and n_h=8. The second population (the "Large Population") was constructed by expanding the available data through replication of establishments to achieve a population the size of the original population; again there were six noncertainty and 1 certainty strata; samples were of the size of the the actual sample.

Results on coverage and mean interval length are included in Tables 1-4.

SMALL POPULATION: Table 1 for total wages and Table 2 for mean wages give coverage and interval length, at two sample sizes n_h=4 and n_h=8, for 8 occupations, and 3 variance-df combinations: the standard variance estimator with the standard normal z-quantile, and with the unweighted and weighted degrees of freedom, referred to respectively as Normal, t_{simple} and $t_{complex}$. The combination of s_{ψ}^2 and df_w with ψ_{Ah} =0 was generally more conservative than

$t_{complex}$ with wider intervals, but is here omitted to conserve space.

Results are based on 500 runs; however not all runs are "viable" for all estimators; for example, if n_{Ah} is less than 2 over all strata, then s^2 may be calculable but s_{ψ}^2 is not. For a given "interval type" (variance-*df* combination), coverage was calculated over viable runs. Occupations are ordered by increasing values of the average value over runs of df_{uw}. We note:

* Coverage using the standard variance estimator and the standard normal quantiles (infinite *df*) is poor almost universally.
* Coverage for the other interval types is far more satisfactory, in the main matching nominal or being conservative for the weighted degrees of freedom: the unweighted degrees of freedom tends to yield coverage a few points below the weighted *df* coverage.
* Confidence intervals for means are better behaved on the whole than for totals. Two occupations (Secretary I, GuardsI) yield seriously low coverage for totals even with the improved procedures; only Guards I gives poor coverage for means.

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

* Relative interval length of the standard interval tends to be too small, that is, less than 1.
* Interval length among the other variance-*df* combinations is largest for s_{ψ}^2 with df_w, next largest for s^2 with df_w, and smallest for s^2 with df_{uw}. These differences can be appreciable; there is definitely a tradeoff between coverage and interval size.
* For a given interval type, the relative interval length tends to 1 as n_h 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, numerically labelled for simplicity and ordered by average df_{uw}. Results are based on 5000 runs. As before, coverage was calculated over viable runs; because of the larger sample size there were typically at most a handful of non-viable runs for any occupation.. We note:

*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.

*Coverage using df_w is less than 90% only in a small fraction of cases.

*There can be marked differences in interval length for the different interval types; all ratios tend to 1, as df_{uw} gets large.

*There are some differences in problem occupations from the Small Population Study; for example, Guards I (4021) does fine, but Computer System Analyst I (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 OCSP sample were treated as certainties in the Large Population; this was not the case in the Small Population.

*In the main, coverage is better for means than for totals, but there are some obvious exceptions, especially at low df_{uw}.

4. Conclusions

The following points are evident from our results:

1. Standard 95% confidence intervals for mean or total wages (or total workers) based on the standard normal distribution and standard methods of variance estimation almost always yield less than actual 95% coverage. The extent of the undercoverage will vary with occupation, but can be quite considerable. Basically, this arises because of the "domain problem". Although the samples we take are quite sizeable in terms of number of PSU's (establishments) selected, these numbers can shrink drastically when restricted to a particular domain (occupation).

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 have to do with the notion of conditioning on the amount of information on the particular domain, which, roughly speaking, is measured in terms of the number of units in the sample that actually belong to the domain. An important unknown is the fraction within each stratum of such units, and one idea which we have exploited is

that of putting 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. (Note: frequently for purposes of publication, it is desirable to make available standard deviations, with the understanding that these translate to a confidence interval of a specified level by multiplying by a suitable normal distribution quantile. We need not abandon this practice, but we do need to make available the *effective* standard deviation, that is, that number such that the normal interval based on it equals the sounder t-interval.)

5. The most likely candidate for estimate of variance, accompanying the new t-quantile, is the standard estimate of variance. In most instances, this will be quite satisfactory, so that the only innovation is the new degrees of freedom methodology. However, we have considered alternatives to the standard variance estimator, for the sake both of increasing coverage in problem occupations and of narrower CI's, and it may in the long run prove desirable to use one or other of the alternatives, possibly for a limited set of domains, for example, in the case of occupations, where an occupation is sharply split along union- non-union lines, so that the within domain distribution is bimodal or sharply skewed.)

6. Preliminary work suggests that coverage for quantiles can be improved by applying the present approach to get confidence intervals for the distribution function prior to application of the Woodruff (1952) transformation.

7. We have some question about what degree and type of collapsing of strata should be used in the estimation of variances and of the degrees of freedom. Some degree of collapsing might be desirable: 1) if the original strata are designed primarily to give a reasonable hope of capturing the variety domains rather than because of any anticipated differences in values of the variate of interest across strata; it can be expected that these strata can be joined into still homogeneous mega-strata; 2) using the original strata for purposes of variance estimation will often, we suspect, give a quite low estimate of degrees of freedom; 3) as strata are collapsed the estimate of variance will increase, but the estimated t-quantile will be reduced; thus there may be an optimal tradeoff here, assuring covera ge with minimal cost in interval length. This issue was not addressed in the simulation studies we did if only because the populations we considered were not large enough or inhomogeneous enough to get useful information on this point.

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

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 Postion Measures, *Journal of the American Statistical Association* **47**, 635-646

Table 1. Coverage and average degrees of freedom for **Total Wages** for a stratified random sample

$n = 36$ ($n_h = 4$)

Occupation	Coverage			interval length			df	
	Normal	t_{simple}	$t_{complex}$	Normal	t_{simple}	$t_{complex}$	t_{simple}	$t_{complex}$
Guards I	.47	.89	.92	0.53	2.65	3.30	1.5	1.3
Secretary I	.69	.92	.93	0.75	3.67	4.32	1.6	1.3
Key Entry II	.51	.93	.95	0.59	2.80	3.19	1.6	1.4
Gen Main Wrks	.75	.99	.99	0.70	2.60	3.40	2.0	1.5
Com. Sys. Al. I	.73	.95	.96	0.74	2.20	3.08	2.3	1.7
Secretary II	.85	.96	.96	0.85	1.98	3.06	2.8	1.9
Switchbrd Oper	.89	.97	.98	0.90	1.50	2.70	4.3	2.3
Accountant III	.87	.92	.95	0.88	1.14	1.58	6.1	3.5

$n = 60$ ($n_h = 8$).

Occupation	Coverage			interval length			df	
	Normal	t_{simple}	$t_{complex}$	Normal	t_{simple}	$t_{complex}$	t_{simple}	$t_{complex}$
Guards I	.49	.65	.74	0.63	1.09	2.40	3.8	2.3
Secretary I	.74	.87	.91	0.87	1.63	3.08	3.7	2.0
Key Entry II	.65	.75	.80	0.66	1.13	2.38	3.9	2.3
Gen Main Wrks	.79	.89	.94	0.80	1.10	2.00	5.6	3.1
Com. Sys. Al. I	.78	.86	.89	0.83	1.10	1.74	6.0	3.5
Secretary II	.86	.90	.95	0.88	1.06	1.38	8.0	4.3
Switchbrd Oper	.88	.90	.96	0.92	1.02	1.38	12.3	5.4
Accountant III	.92	.94	.96	0.96	1.04	1.13	16.6	9.7

Table 2. Coverage and average degrees of freedom for **Average Wage** for a stratified random sample

$n = 36$ ($n_h = 4$).

Occupation	Coverage			interval length			df	
	Normal	t_{simple}	$t_{complex}$	Normal	t_{simple}	$t_{complex}$	t_{simple}	$t_{complex}$
Guards I	.19	.79	.85	0.22	1.11	1.20	1.5	1.3
Secretary I	.68	.97	.98	0.57	2.80	3.08	1.6	1.3
Key Entry II	.65	.96	.97	0.46	2.09	2.30	1.6	1.4
Gen Main Wrks	.76	.97	.99	0.62	2.40	2.95	2.0	1.5
Com. Sys. Al. I	.75	.96	.98	0.76	2.37	2.88	2.3	1.8
Secretary II	.77	.96	.98	0.69	1.62	2.33	2.8	1.9
Switchbrd Oper	.79	.93	.97	0.69	1.19	2.00	4.3	2.4
Accountant III	.87	.94	.97	0.86	1.11	1.67	6.1	3.5

$n = 60$ ($n_h = 8$).

Occupation	Coverage			interval length			df	
	Normal	t_{simple}	$t_{complex}$	Normal	t_{simple}	$t_{complex}$	t_{simple}	$t_{complex}$
Guards I	.28	.49	.62	0.27	0.54	0.81	3.8	2.4
Secretary I	.81	.93	.94	0.70	1.50	2.20	3.7	2.3
Key Entry II	.75	.96	.98	0.57	1.04	1.47	3.9	2.6
Gen Main Wrks	.82	.92	.97	0.69	1.01	1.65	5.6	2.9
Com. Sys. Al. I	.80	.88	.91	0.82	1.12	1.46	6.0	3.9
Secretary II	.84	.90	.95	0.81	0.97	1.34	8.0	4.3
Switchbrd Oper	.88	.91	.95	0.77	0.86	1.21	12.3	5.6
Accountant III	.90	.93	.94	0.88	0.96	1.09	16.6	9.0

Table 3. Coverage and average degrees of freedom for **Total Workers**; N=3062, n=354.

Occupation	Coverage			interval length			df	
	Normal	t_{simple}	$t_{complex}$	Normal	t_{simple}	$t_{complex}$	t_{simple}	$t_{complex}$
1718	.89	.96	.97	0.99	2.14	2.34	2.97	2.67
1604	.60	.83	.89	0.78	1.47	2.27	3.45	2.34
1802	.85	.94	.94	0.92	1.40	2.48	4.44	2.35
1716	.87	.89	.92	0.97	1.08	1.37	11.9	5.97
2911	.87	.88	.90	0.95	1.06	1.39	12.4	5.90
2052	.89	.91	.97	0.96	1.06	1.60	13.1	4.25
1332	.93	.95	.96	0.99	1.08	1.13	15.3	11.4
1141	.93	.95	.96	0.98	1.06	1.18	16.9	9.00
4021	.89	.91	.91	0.96	1.04	1.34	16.8	6.32
1232	.92	.94	.94	0.97	1.04	1.05	17.3	15.5
2853	.92	.94	.94	0.98	1.04	1.13	20.6	13.5
3020	.92	.93	.95	0.98	1.03	1.18	24.9	10.4
1122	.88	.88	.89	0.95	0.99	1.04	28.0	15.2
1142	.89	.90	.91	0.96	1.00	1.20	28.6	9.67
1714	.85	.86	.86	0.93	0.98	1.04	29.1	15.3
1514	.93	.93	.94	0.98	1.01	1.04	34.8	18.0
3180	.92	.92	.93	1.00	1.03	1.06	41.5	25.2
4030	.81	.81	.83	0.91	0.93	1.07	59.9	14.3
1063	.94	.95	.95	1.00	1.01	1.10	77.6	27.4
1403	.94	.94	.95	1.00	1.01	1.05	77.9	28.5
1180	.95	.95	.95	1.01	1.02	1.02	128	90.0

Table 4. Coverage and average degrees of freedom for **Average Wage**; N=3062, n=354.

Occupation	Coverage			interval length			df	
	Normal	t_{simple}	$t_{complex}$	Normal	t_{simple}	$t_{complex}$	t_{simple}	$t_{complex}$
1718	.64	.87	.91	0.82	1.89	3.02	3.00	2.08
1604	.63	.89	.92	0.59	1.24	1.87	3.37	2.08
1802	.87	.97	1.0	0.89	1.36	2.71	4.47	2.32
1716	.88	.90	.92	0.92	1.04	1.32	12.1	5.98
2911	.77	.80	.83	0.81	0.90	1.32	12.2	4.98
2052	.87	.90	.95	0.87	0.96	1.48	13.2	4.31
1332	.92	.94	.95	0.94	1.03	1.11	15.4	9.76
1141	.87	.89	.91	0.92	0.99	1.01	16.8	7.96
4021	.95	.96	.97	0.92	1.00	1.16	17.0	9.90
1232	.91	.93	.94	0.68	0.74	0.87	17.3	14.2
2853	.93	.94	.95	0.97	1.04	1.08	20.6	13.9
3020	.92	.93	.95	0.95	1.00	1.10	24.8	11.8
1122	.93	.94	.95	0.85	0.89	1.17	28.0	15.5
1142	.91	.93	.96	0.87	0.91	1.11	28.8	8.42
1714	.90	.91	.92	0.94	0.98	1.02	29.0	17.6
1514	.90	.90	.91	0.93	0.96	1.01	34.9	15.8
3180	.93	.93	.94	0.96	0.98	1.00	41.5	30.7
4030	.88	.89	.92	0.63	0.65	0.72	59.7	17.2
1063	.94	.94	.94	0.98	0.99	1.00	77.4	54.2
1403	.93	.94	.94	0.97	0.99	1.00	77.7	40.2
1180	.94	.95	.95	0.99	1.00	1.00	128	90.3