QUANTILE VARIANCE ESTIMATORS IN COMPLEX SURVEYS

Alan Dorfman, Richard Valliant, Bureau of Labor Statistics
Alan Dorfman, 2 Massachusetts Avenue, N.E., Washington DC 20212

KEY WORDS: Density estimation, distribution function, simulation study, stratified cluster sampling

This paper summarizes some results on a project to study methods of variance estimation and confidence interval construction for quantiles of wages over a variety of occupations in the Bureau of Labor Statistics' Occupational Compensation Survey. Our work consists mainly of simulations from a population consisting of a single sample from one Metropolitan Statistical Area. Four methods are reported on here: (1) Woodruff, (2) Francisco-Fuller, (3) balanced repeated replication (BRR), and (4) balanced repeated replication with grouping of establishments (BRRg). Also some incomplete results are noted for an approach involving direct estimation of the density in the expression for the asymptotic variance of the quantiles. Our findings are that either of the BRR methods outperforms Woodruff and Francisco-Fuller. Woodruff and Francisco-Fuller are about on a par, with Woodruff having a great edge in computational ease. For some occupations, *all* methods are abysmal; some analysis and conjecture are offered to account for this.

1. POPULATION AND METHODS

The population dataset is from a test sample of the Occupational Compensation Survey conducted in 1991. The population consists of 353 establishments which we divided into 7 strata, based on industrial type (Manufacturing versus Services) and size class based on number of workers employed (whether in occupations of interest or not); see Table 1. Additionally, we included one certainty stratum consisting of the 12 largest establishments. The goal was to create a population from which we could sample in a manner that mimicked the sampling process from the original MSA population.

Two establishment sample sizes were used in the simulations: (1) $n = 36$ ($n_h = 4$ in strata 1-6 plus 12 certainties), and (2) $n = 60$ ($n_h = 8$ in strata 1-6 plus 12 certainties). For each sample size, one hundred one-stage cluster samples were selected. In each non certainty stratum, each sample was selected by simple random sampling without replacement. All workers in a sample establishment were enumerated. In each sample, estimates of the 25^{th}, 50^{th}, and 75^{th} percentiles of wages were computed for the 9 occupations listed in Table 2. The column labeled $cv(\hat{M})$ gives the coefficient of variation of the estimated total number of workers in the occupation for the $n_h = 4$ sample design. This measure is related to the degree to which an occupation is clustered in the establishments and was used to identify occupations that would represent a range of difficulty for estimation of the variance of quantiles, and reflects an important difference between this population and other populations for which quantile variance estimation has been studied (e.g. Francisco & Fuller (1991), Wheeless and Shah (1988)) in that for a given occupation the number of units (workers) can vary widely across clusters (establishments), even within a given stratum.

The γ^{th} quantile of wages for a particular occupation is defined as

$$q_\gamma = \min_{(hij)\in U}\left\{y_{hij} : F(y_{hij}) \geq \gamma\right\}$$

where h is a stratum, i is an establishment, and j is a worker, U is the universe of all workers, y_{hij} is the wage of worker hij, and F is the cumulative distribution function (CDF) for wages in the given occupation defined by

$$F(y) = \sum_h \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} I\{y_{hij} \leq y\} \Big/ M,$$

where N_h is the number of establishments in stratum h, M_{hi} is number of workers in establishment (hi) in the particular occupation, I is the indicator function defined as $I\{y_{hij} \leq y\} = 1$ if $y_{hij} \leq y$ and 0 otherwise, and M is the number of workers in the given occupation. As an example, the 50^{th} percentile (the median) is the smallest wage value in the population such that at least 50 percent of the workers make that value or less. The sample estimate of the CDF at some fixed wage value y is

$$\hat{F}(y) = \sum_h \sum_{i\in s_h} \sum_{j=1}^{M_{hi}} w_{hij} I\{y_{hij} \leq y\} \Big/ \hat{M}$$

where s_h is the sample of establishments in stratum h, and, for $w_{hij} = N_h/n_h$ the sample weight of

worker (hij), $\hat{M}=\sum_h \sum_{i\in s_h} \sum_{j=1}^{M_{hi}} w_{hij}$ is an estimator of the total number of workers in the occupation. The sample estimate of the γ^{th} quantile is then

$$\hat{q}_\gamma = \min\left\{y: \hat{F}(y) \geq \gamma\right\}. \quad (1)$$

In practice, the sample estimator of q_γ is found by first sorting the sample workers in order by the wage paid to each and then cumulating their normalized sample weights $w_{hij}/\hat{M}$ until γ is first exceeded.

For each of the three quantiles listed above, we applied the methods of variance estimation of Woodruff, Francisco-Fuller, BRR, and grouped BRR. The principal difference between the first of these two and the BRR algorithms is that the former involve preliminary calculations of the estimated distribution function and of its standard deviation, which is usually estimated by the linearization method.

The Woodruff method (Woodruff 1952) takes several steps: (1) estimation of the variance s_y^2 of the estimated distribution function $\hat{F}(y)$, at $y=\hat{q}_\gamma$; the usual design-based estimator is

$$s_y^2 = \sum_h (1-f_h)\frac{n_h}{n_h-1}\sum_{i\in s_h}\left(d_{hi}-\bar{d}_h\right)^2$$

where $f_h = n_h/N_h$ $d_{hi} = \hat{M}^{-1}\sum_{j=1}^{M_{hi}} w_{hij}\left[I\left\{y_{hij}\leq y\right\}-F(y)\right]$ with $y=q_\gamma$; (2) construction of a (1-α)100% confidence interval $\hat{F}(y)\pm t_\alpha s_y \approx \gamma \pm t_\alpha s_y$ for $F(y)$, with $y=\hat{q}_\gamma$ regarded as fixed; (3) determination of the quantiles corresponding to the endpoints of this interval to give $\mathbf{I}_W(q_\gamma)=\left[q_{\gamma-t_\alpha s_\gamma}, q_{\gamma+t_\alpha s_\gamma}\right]$ as the basic (1-α)100% Woodruff confidence interval for the population quantile q_γ. Note that this interval is not typically symmetric about the quantile estimate itself. (4) The standard deviation for the quantile is usually taken as the length of this interval divided by twice the standard normal percentile, that is, as $s_{q_\gamma}=\left(q_{\gamma+t_\alpha s_\gamma}-q_{\gamma-t_\alpha s_\gamma}\right)/2t_\alpha$. Finally, a symmetric confidence interval is just $\hat{q}_\gamma \pm t_\alpha s_{q_\gamma}$. An alternative which we have not seen mentioned in the literature is to take $s_{q_\gamma}=\max\left(q_{\gamma+t_\alpha s_\gamma}-q_\gamma, q_\gamma - q_{\gamma-t_\alpha s_\gamma}\right)/t_\alpha$; this yields a symmetric confidence interval with coverage guaranteed as large as the basic Woodruff interval. In any case, the estimate of the standard deviation depends in part on the choice of α; the literature generally suggests use of α=0.05, corresponding to 95% intervals.

It may be noted that in simulation studies using simple random sampling, carried out by the authors of the software package SUDAAN (Wheeless and Shah 1988), the Woodruff estimator performed somewhat better than the estimator finally chosen for that package; there is reason to believe that the gap between the estimators would widen in the context of complex surveys, especially with data less well behaved than that from the standard normal and lognormal distributions used by Wheeless and Shah (1988). The Woodruff procedure is a good deal faster to compute than the other methods we are considering.

The Francisco-Fuller (FF) or "test-inversion" method (Francisco and Fuller 1991) is closely related to Woodruff. Whereas the Woodruff bases the CI for the quantile on the CI for the distribution function *at* the estimated quantile, the FF method relies on confidence bounds of the distribution function at values *around* the quantile. In particular, the FF CI is the set

$$\Gamma = \left\{y: \hat{F}(y)+t_\alpha s_y > \gamma \text{ and } \hat{F}(y)-t_\alpha s_y < \gamma\right\}.$$

The literature tends to suggest FF will outperform Woodruff, although computationally it is far more intensive, requiring estimation of $\hat{F}(y)$ and s_y over an array of y-values, as well as various transformations and smoothing operations. To our knowledge it has not hitherto been tested in the context of complex sampling of the sort arising in the OCS. In particular, the case of unequal-sized clusters has not been dealt with.

The BRR method consists of dividing the establishment sample into half-samples in a prescribed way, estimating the desired quantile from each half-sample, and computing the variance among the half-sample estimates. For the test population, we treated each sample as if a two-per-stratum design had been used. For n_h=4, for instance, we treated the first two randomly drawn establishments in a stratum as a pair and the third and fourth units as a second pair. For the n_h=4 design, 6×2=12 pairs were created in the noncertainty strata. The minimal number of balanced half-samples is 16 in this case. The noncertainty units in a particular half-sample were then combined with the 12 certainties and a sample quantile was computed using (1). For the n_h=8 design, 6×4=24 pairs were formed with the minimal number of balanced half-samples being 28. The BRR estimate of variance of $\hat{q}_\gamma$ is then

$$v_{BRR}(\hat{q}_\gamma) = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{q}_\gamma^{(r)}-\hat{q}_\gamma\right]^2$$

where $\hat{q}_\gamma^{(r)}$ is the estimated quantile based on half-sample r and R is the total number of replicates (half-samples).

The grouped BRR method is similar to the full BRR but reduces the number of replicates. In each stratum, the sample units were put into two groups, and the groups were assigned to half-samples. For n_h=4, two groups of two establishments each were formed. For n_h=8, two groups of four establishments each were formed. Establishments were assigned to groups in each sample in the order that they were selected. For example, when n_h=4, the first and second units selected in a sample were put in group 1 while the third and fourth were put into group 2. With the grouped method, two groups are formed in a design stratum regardless of the number of sample establishments. The number of replicates in a balanced set with 6 strata is 8, rather than 16 for the ungrouped case when n_h=4 or 28 when n_h=8. Although popular as a means of reducing the number of replicates and therefore computation, the grouped BRR estimator will typically be less stable than the full BRR and will produce CI's with inferior coverage properties. When the number of strata is small and n_h is large, the grouped BRR estimator may, in fact, be inconsistent (Rao and Shao 1993).

2.RESULTS FOR VARIANCE ESTIMATORS

As a measure of the bias of one of the four standard error estimators, we computed the ratio of its average to the empirical root mean square error over the 100 samples. The ratio was defined as $\overline{v^{1/2}}/rmse(\hat{q})$ where v is one of the four variance estimators, $\overline{v^{1/2}} = \sum_{k=1}^{100} v_k^{1/2}/100$, and $rmse(\hat{q}) = \sqrt{\sum_{k=1}^{100}(\hat{q}_k - q)^2/100}$ with $\hat{q}_k$ being the estimated quantile for sample k and q being the population quantile. The ratios are plotted in Figures 1(a) and 1(b) for the 50th percentile for both n_h=4 and 8. A few notable features are:

(a) The order of performance, as measured by bias, from best to worst is BRR, grouped BRR, Woodruff, and FF. The BRR and grouped BRR tend to be fairly close. Some evidence suggests that the weak point in the Woodruff and FF estimator is in s_y^2, the linearization estimate of variance of the distribution function.

(b) Underestimates are more common than overestimates. Although not shown in Figure 1, the degree of underestimation is greatest for the 25th percentile and smallest for the 75th.

(c) The negative bias can be substantial at either sample for all variance estimators for occupations with larger values of $cv(\hat{M})$. In particular, performance is poor for occupations 5,6,8, and 9. However, there are some anomalies; BRR and BRRg are positively biased for occupation 7 having high $cv(\hat{M})$, and the estimators do universally poorly for 2 having small $cv(\hat{M})$. This last case is explained by the fact that occupation 2 occurs almost exclusively in certainty establishments, so that the variance estimate is frequently *zero* for all estimators.

3. CONFIDENCE INTERVALS

Coverages of 95% confidence intervals across the 100 samples are plotted in Figures 1(c) and 1(d) for each occupation and variance estimator. For the Woodruff and FF methods, CI's were computed as described previously. For BRR and BRRg, normal approximation CI's were calculated as $\hat{q}_\gamma \pm 1.96v^{1/2}$ where v is one of the BRR estimators. Some general observations are:

(a) The order of performance, as measured by CI coverage, is the same as for variance estimation: BRR, grouped BRR, Woodruff, and FF.

(b) Coverage percentages are typically less than the nominal 95%.

(c) The occupations where variances were underestimated the most have extremely poor coverage.

Security guard (9) is the worst case having coverages for the median between 20% and 30% using Woodruff or FF. BRR and BRRg are better (40%-50%) but far from acceptable.

4. DISCUSSION

It follows from the Bahadur representation of quantiles $\hat{q}_\gamma \approx q_\gamma + \left[1/f(q_\gamma)\right]\left(F(q_\gamma) - \hat{F}(q_\gamma)\right)$ (see for example Francisco and Fuller 1991) that $\mathrm{var}(\hat{q}_\gamma) \approx \left[1/f(y_\gamma)\right]^2 \mathrm{var}\left(\hat{F}(y_\gamma)\right)$ where $f(y)$ is the density associated with the CDF $F(y)$. Thus, variance estimators of the quantile implicitly or explicitly must perform two tasks: (1) estimate the variance of the distribution function, and (2) estimate the density function $f(y)$. BRR and BRRg do both implicitly; Woodruff and FF rely on explicit estimation of the distribution function and use inversion of confidence intervals to estimate $f(y)$ implicitly.

An alternative is to estimate $f(y)$ directly using non-parametric density estimation, a subject

which has seen a great deal of research in recent years; Silverman (1986) is a good introduction. Hall and Sheather (1988) discuss one simple version of this approach first proposed by Bloch and Gastwirth (1968), viz. $\hat{f}(\gamma_p) = (2m/n)/(X_{r+m} - X_{r-m})$ where $X_r = \hat{\gamma}_p$ and m is an integer to be selected by the analyst.

Wheeless and Shah (1988) propose another simple estimator, namely $\hat{f}(\gamma_p) = \left(\hat{F}(x'_{j+1}) - \hat{F}(x'_j)\right)/(x'_{j+1} - x'_j)$, where $\{x'_v\}$ forms an evenly spaced grid on the interval covered by the sample values, and $[x'_j, x'_{j+1}]$ contains $\hat{\gamma}_p$. This estimator has been incorporated in SUDAAN software; we hazard it will be vulnerable to unevenness in the data, especially if the number of bins $[x'_j, x'_{j+1}]$ is selected arbitrarily, since at most one data point is guaranteed to be in a particular bin.

More sophisticated methods of density estimation exist, for example, the simple kernel estimator

$$\hat{f}(\gamma_p) = (nb)^{-1} \sum_s K\left([X_i - \hat{\gamma}_p]/b\right),$$

where $K()$ is a density function symmetric about 0, and b is a bandwidth selected by the analyst. Figures 1(c) and 1(d) show coverage yielded by variance estimation based on an adaptive kernel estimator (Silverman, 1986, Section 5.3); it outperforms Woodruff and FF at the median, with uneven results at the other quartiles. This is encouraging, since no serious attempt was made to choose a bandwidth geared for sound confidence intervals, as in Hall and Sheather (1988).

For the Woodruff, FF, and direct density methods, one can also seek improvement in estimation of $\text{var}\left(\hat{F}(y_\gamma)\right)$ We ran simulations in which the true (population) variance (estimated by the empirical variance of $\hat{F}(y_\gamma)$ over all runs) of the distribution function is employed in the Woodruff method; Figures 1(c) and 1(d) show that the resulting empirical coverage of confidence intervals for the quantile tends to closely match the nominal. This suggests that an improved estimator of $\text{var}\left(\hat{F}(y_\gamma)\right)$ would lead to improved coverage.

To get further light on this, consider the working model which posits that, conditional on the number of employees in the given occupation in an establishment M_{hi}, the wages Y_{hij} are realizations of random variables of the form $Y_{hij} = \mu_h + e_{hi} + e_{hij}, h = 1,\ldots,H; i = 1,\ldots,N_h; j = 1,\ldots,M_{hi}$ with the between and within establishment errors e_{hi} and e_{hij} mutually independent. The M_{hi} themselves are taken as independent with mean and variance m_h, Φ_h respectively; also let $m_h^2 = E\left(M_{hi}^2\right)$ and $K_h = M^{-2} n_h^{-1} N_h^2 (1 - f_h)$. Then it can be shown, using the usual linearization of a ratio, that $\text{var}\left(\hat{F}(t) - F(t)\right) \approx \sum_h K_h \left\{A_h(t)\Phi_h + B_h(t) m_h + C_h(t) m_h^2\right\}$ where $A_h(t), B_h(t), C_h(t)$ are positive functions given by $A_h(t) = \left(H_h(t) - F(t)\right)^2$, $B_h(t) = H_h(t) - L_{hi}(t)$, and $C_h(t) = L_{hi}(t) - H_h^2(t)$, with $H_h(t) = P\left(Y_{hij} \le t\right) = \int G_{hi}(t - \mu_h - x)\, dG_h(x)$ and $L_{hi}(t) = \int G_{hi}^2(t - \mu_h - x)\, dG_h(x)$, for G_h and G_{hi} distribution functions for e_{hi}, e_{hij} respectively. One can show that the standard variance estimator $s_y^2 = \sum_h K_h \{\hat{A}_h(t)\hat{\Phi}_h + \hat{\text{var}}(M_{hi}\Delta_{hi}(t)) + 2\left(\hat{H}_h(t) - \hat{F}(t)\right)\hat{\text{cov}}(M_{hi}, \Delta_{hi}(t) M_{hi})$, where $\hat{A}_h(t) = \left(\hat{H}_h(t) - \hat{F}(t)\right)^2$ with $\hat{H}_h(t) = \sum_{i,j} I\left(Y_{hij} \le t\right)/\sum_i M_{hi}$, and $\hat{\Phi}_h = (n_h - 1)^{-1} \sum_i \left(M_{hi} - \overline{M}_h\right)^2$. Except for the first term, this does not match $\text{var}\left(\hat{F}(t) - F(t)\right)$, so that there is some hope that a model-based estimator might improve things.

However, this hope is not too strong. If only a few strata contribute actual workers for a given occupation, and within the contributing strata the distribution of the M_{hi} is positively skewed, with, say, mostly zeroes, a scattering of small values, and one or two large values, then when the few large values are missing from the sample, Φ_h, m_h, m_h^2 will all tend to be underestimated and so, consequently, will $\hat{\text{var}}\left(\hat{F} - F\right)$, whether we use the standard estimator s_y^2 or a more elaborate model-based estimator. This also suggests why the index $CV\left(\hat{M}\right)$ foretells as well as it does when coverage is low. Since $\hat{F}$ is a ratio estimator, this result would seem to be anticipated in (Hansen, Hurwitz, and Madow 1953, Ch. 4, Sec. 12), which discusses the relationship between the coefficient of variation of the estimator of the denominator of the ratio, and the linearized estimate of variance of the ratio estimator. However, a parallel study carried out on confidence intervals for total wages, yielded coverage very similar to that for quantiles, and an explanation in terms of underestimation of the parameters Φ_h, m_h, m_h^2 is possible for totals also.

Thus improvements in coverage will not be easy. We see several avenues for exploration: (1) estimating the variance of the distribution

function by BRR; this should certainly improve the linearization estimator and may have certain advantages over the direct application of BRR to the quantiles in terms of computation time and incorporation of sampling fractions into the variance structure; (2) incorporating information on occupation from other regions or time periods or from related occupations into the variance estimates; this amounts to using small area estimation for variance estimation; (3) using the bootstrap rather than normal confidence intervals; and (4) adjusting the degrees of freedom associated with the variance estimate, to take into account the number of clusters (establishments) actually contributing data on a given occupation.

REFERENCES

Bloch, D.A. and Gastwirth, J.L. (1968) "On a Simple estimate of the Reciprocal of the Density Function," *Annals of Statististics,* **39,** 1083-1085.

Francisco, C.A. and Fuller, W. A. (1991), "Quantile Estimation with a Complex Survey Design," *Annals of Statististics,* **19,** 454-469.

Hall, P. and Sheather, S.J. (1988) "On the Distribution of a Studentized Quantile," *Journal of the Royal Statistical Society* B, **50,** 381-391.

Hansen, Hurwitz &Madow (1953), *Sample Survey Methods and Theory*, Vol. II Theory, John Wiley and Sons, New York

Rao, J.N.K. and Shao, J. (1993), "On balanced Half-sample Variance Estimation in Stratified Sampling," preprint.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis,* Chapman and Hall: London.

Wheeless, S.C. and Shah, B.V. (1988), "Results of a Simulation for Comparing Two Methods for Estimating Quantiles and their Variances for Data from a Sample Survey," *American Statistical Association 1988 Proceedings of the Section on Survey Research Methods*, 722-727.

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

Table 1. Description of Population

Stratum number		Size classes (Number of Employees)	No. of estabs. in population N_h	No. of sample estabs n_h
1	Manufacturing	<250	35	4 or 8
2	Manufacturing	250-999	35	4 or 8
3	Manufacturing	≥1000	33	4 or 8
4	Services	<250	136	4 or 8
5	Services	250-999	66	4 or 8
6	Services	≥1000	36	4 or 8
7	Certain	12 largest	12	12

Table 2. Occupations in Study

Occ. no.	Description	$cv(\hat{M})$
1	Accountant III	.052
2	Engineering technician I	.055
3	Secretary II	.160
4	Switchboard operator-receptionist	.174
5	Computer systems analyst I	.177
6	General maintenance worker	.438
7	Secretary I	.666
8	Key entry operator II	1.163
9	Guard I.	2.271

Figure 1. Ratios of standard error estimates to empirical root mean square errors and empirical coverage of 95% confidence intervals for median wages