Economics Bulletin

Volume 33, Issue 3

Weighted empirical likelihood-based inference for quantiles under stratified random sampling

Tahsin Mehdi
University of Guelph

Abstract

A growing body of literature is emerging on empirical likelihood methods for complex surveys. These works largely focus on the population mean. We propose a weighted empirical likelihood approach as a method of inference for quantiles under stratified random sampling, which is one of the most popular complex survey designs. A simulation study substantiates our proposed methodology.

Citation: Tahsin Mehdi, (2013) "Weighted empirical likelihood-based inference for quantiles under stratified random sampling", *Economics Bulletin*, Vol. 33 No. 3 pp. 2437-2442.
Contact: Tahsin Mehdi - tmehdi@uoguelph.ca.
Submitted: July 26, 2013. **Published:** September 19, 2013.

1 Introduction

Ever since the pioneering work of Owen (1988, 1990), there has been a proliferation of literature on empirical likelihood (EL), which is a powerful nonparametric statistical tool. An advantage of this method is that one does not need to assume anything about the underlying distribution of the data. The EL ratio is entirely data driven. The main focus of Owen (1988) was to construct confidence intervals for a population mean given a single sample of independent and identically distributed (iid) observations. For a detailed overview of EL, see Owen (2001).

EL methods for complex surveys are yet to be investigated to their full extent. Chen and Sitter (1999) and Zhong and Rao (2000) were among the first to consider EL in the context of complex survey designs. Complex survey designs pose additional difficulties for the conventional EL approach. Asymptotic results from conventional EL are not directly applicable to complex surveys as special types of constraints may need to be imposed, depending on the survey design. Generally, the EL ratio in such cases will not have the same calibration as in the case of simple random sampling. Also, existing computational procedures may not be readily applicable. To alleviate such problems, Fu et al. (2008) introduced a weighted empirical likelihood method and developed an unified approach for making inferences on population means in the presence of multiple samples. One of the cases they consider is stratified random sampling where the focal point of interest is on the overall population mean. Their approach relies on the augmentation of the special types of constraints induced by stratified samples.

Though a large body of literature exists regarding inference on population means, quantiles have received relatively less attention. This is especially true in the case of complex survey designs. Such designs could conceivably give rise to multiple distribution functions (as is the case with stratified random sampling) instead of just one. Thus, deriving asymptotic expressions for quantiles can get quite tedious and sometimes may not even be possible. For simple random sampling, (Owen, 2001, Ch. 3.6) provides a good introduction to EL methods for quantiles.

The promising results of Fu et al. (2008) warrant further research into the weighted empirical likelihood approach. Drawing upon their work, we propose a weighted empirical likelihood-based inference method for quantiles under stratified random sampling. Our results rely on very similar asymptotic expansions.

The rest of this article is organized as follows. In Section 2, we introduce our proposed methodology and establish some asymptotic results under stratified random sampling. In Section 3, we present a Monte Carlo study which assesses the accuracy of the confidence interval obtained from our method. Section 4 concludes.

2 Inference for a quantile

Suppose that a population is divided into k mutually exclusive strata of known sizes $N_1, ..., N_k$. The weight associated with the ith stratum is $w_i = N_i/N$, where $N = \sum_{i=1}^{k} N_i$ is the overall population size. Let $\{Y_{ij}, j = 1, ..., n_i\}$, $i = 1, ..., k$, be k independent samples of size n_i extracted from the strata and let $n = \sum_{i=1}^{k} n_i$ be the pooled sample size. Assume the strata sampling fraction n_i/N_i is negligible so that $\{Y_{ij}, j = 1, ..., n_i\}$ is regarded as an iid sample generated by the continuous random variable Y_i with distribution function F_i. The overall distribution function is then given by $F(y) = \sum_{i=1}^{k} w_i F_i(y)$. Let Q^α denote the α-quantile of F. This quantile is implicitly characterized by $F(Q^\alpha) = \alpha$. Our focal point of interest is on constructing confidence intervals for Q^α given α.

The weighted empirical log-likelihood (WEL) function of Fu et al. (2008) is given by

$$l_w(F_1, ..., F_k) = \sum_{i=1}^{k} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \log(p_{ij}), \tag{1}$$

where p_{ij} is the probability associated with Y_{ij}. The formulation of (1) was motivated using the argument of Chen and Sitter (1999). See Fu et al. (2008) for more on this.

An advantage of using the WEL function is that the usual large sample properties of EL can be established under the special type of constraints induced by stratified samples. If the constraints are reformulated in a suitable way, computational procedures are also readily available.

To construct a confidence interval for Q^α, we maximize (1) subject to $p_{ij} > 0$ and

$$\sum_{j=1}^{n_i} p_{ij} = 1, \quad i = 1, ..., k, \tag{2}$$

$$\sum_{i=1}^{k} w_i \sum_{j=1}^{n_i} p_{ij} 1_{Y_{ij} \leq Q^\alpha} = \alpha, \tag{3}$$

where $1_{(\cdot)}$ is an indicator function which evaluates to one if the argument $(\cdot)$ is true, and zero otherwise. Constraint (3) identifies the quantile Q^α and its use can be justified by arguments similar to (Owen, 2001, Ch. 3.6). Since $E_i(1_{Y_i \leq Q^\alpha}) = F_i(Q^\alpha)$, where E_i denotes the expectation under distribution F_i, constraint (3) indeed identifies Q^α.

To construct confidence intervals for Q^α, we require the asymptotic distribution of the WEL ratio which Fu et al. (2008) defines as

$$r_w(Q^\alpha) = \sum_{i=1}^{k} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \log(n\hat{p}_{ij}),$$

where $\hat{p}_{ij}$ given by (7) solve the maximization problem. Assume $n_i/n \to x \neq 0$, so that it is unnecessary to distinguish between $O(n^{-1/2})$ and $O(n_i^{-1/2})$, and between $o(n^{-1/2})$ and $o(n_i^{-1/2})$. The following theorem establishes the asymptotic distribution of $r_w(Q^\alpha)$ at $Q^\alpha = Q_0^\alpha$.

Theorem 2.1. *Suppose $\{Y_{ij}, j = 1, ..., n_i\}$ is an iid sample, with finite variance, from $F_i, i = 1, ..., k$, and the k samples are independent of each other. If Q_0^α is the α-quantile of the overall distribution function F, then $-2r_w(Q_0^\alpha)/c \xrightarrow{d} \chi^2_{(1)}$, where the scaling constant c is given by* (12).

Proof. Our proof follows very closely the proof of Fu et al. (2008) for stratified sampling. For ease of notation and without loss of generality, consider $k = 3$. Constraints (2) and (3) can be reformulated as

$$\sum_{i=1}^{3} w_i \sum_{j=1}^{n_i} p_{ij} = 1, \tag{4}$$

$$\sum_{i=1}^{3} w_i \sum_{j=1}^{n_i} p_{ij} \boldsymbol{Z}_{ij} = \boldsymbol{\eta}, \tag{5}$$

where the vector-valued variables $\boldsymbol{Z}_{ij}$ and $\boldsymbol{\eta}$ are given by

$$\begin{aligned} \boldsymbol{Z}_{1i} &= (1, 0, 1_{Y_{1i} \leq Q^\alpha})', \\ \boldsymbol{Z}_{2i} &= (0, 1, 1_{Y_{2i} \leq Q^\alpha})', \\ \boldsymbol{Z}_{3i} &= (0, 0, 1_{Y_{3i} \leq Q^\alpha})', \end{aligned}$$

and

$$\boldsymbol{\eta} = (w_1, w_2, \alpha)'.$$

Equation (5) can be re-written as

$$\sum_{i=1}^{3} w_i \sum_{j=1}^{n_i} p_{ij} \boldsymbol{u}_{ij} = \boldsymbol{0}, \tag{6}$$

where $\boldsymbol{u}_{ij} = \boldsymbol{Z}_{ij} - \boldsymbol{\eta}$. The reformulation of constraints (2) and (3) ensure that the probabilities in each of the stratum sum to unity. The maximization of (1) subject to (4) and (6) can be carried out using the Lagrange multiplier technique. For a given Q^α, it can be shown that the optimized probabilities are

$$\hat{p}_{ij}(Q^\alpha) = \frac{1}{n_i(1 + \boldsymbol{\lambda}'\boldsymbol{u}_{ij})}, \tag{7}$$

where the vector-valued Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$\sum_{i=1}^{3} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \frac{\boldsymbol{u}_{ij}}{1+\boldsymbol{\lambda}'\boldsymbol{u}_{ij}} = \mathbf{0}. \tag{8}$$

The above equation can be solved using the algorithm described in Wu (2004). Rewriting the numerator $\boldsymbol{u}_{ij}$ in (8) as $\boldsymbol{u}_{ij}[(1+\boldsymbol{\lambda}'\boldsymbol{u}_{ij}) - \boldsymbol{u}_{ij}'\boldsymbol{\lambda}]$, equation (8) can be expressed as

$$\left[\sum_{i=1}^{3} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \frac{\boldsymbol{u}_{ij}\boldsymbol{u}_{ij}'}{1+\boldsymbol{\lambda}'\boldsymbol{u}_{ij}}\right] \boldsymbol{\lambda} = \sum_{i=1}^{3} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \boldsymbol{u}_{ij}. \tag{9}$$

Noting that $\sum_{j=1}^{n_i}[n_i(1+\boldsymbol{\lambda}'\boldsymbol{u}_{ij})]^{-1} = 1$, for $i = 1, 2, 3$, the order of $\boldsymbol{\lambda}$ is related to the right-hand side of (9), which can be written as

$$\boldsymbol{U} = \sum_{i=1}^{3} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \boldsymbol{u}_{ij} = \left(0, 0, \hat{F}(Q^\alpha) - \alpha\right)', \tag{10}$$

where (for $k = 3$) $\hat{F}(Q^\alpha) = \sum_{i=1}^{3}(w_i/n_i)\sum_{j=1}^{n_i} 1_{Y_{ij} \le Q^\alpha}$. Since $\alpha = F(Q^\alpha)$, it immediately follows that $\boldsymbol{U} = O_p(n^{-1/2})$ (component-wise) when $Q^\alpha = Q_0^\alpha$.

Letting $\boldsymbol{D} = \sum_{i=1}^{3}(w_i/n_i)\sum_{j=1}^{n_i} \boldsymbol{u}_{ij}\boldsymbol{u}_{ij}'$ and noting that it is $O_p(1)$, from (9) we have that $\boldsymbol{\lambda} = O_p(n^{-1/2})$. The finite variance assumption allows us to have $\max_{ij}|\boldsymbol{u}_{ij}| = o_p(n^{1/2})$ and $\boldsymbol{\lambda}'\boldsymbol{u}_{ij} = o_p(1)$ uniformly over all i and j (see Owen, 2001, Ch. 11.1). An asymptotic expression for the Lagrange multiplier is obtained as

$$\boldsymbol{\lambda} = \boldsymbol{D}^{-1}\boldsymbol{U} + o_p(n^{-1/2}). \tag{11}$$

The WEL ratio function at Q_0^α is

$$r_w(Q_0^\alpha) = -\sum_{i=1}^{3} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \log(1+\boldsymbol{\lambda}'\boldsymbol{u}_{ij}).$$

Using a second order Taylor expansion on $\log(\cdot)$, we obtain the following asymptotic expansion of the WEL ratio,

$$\begin{aligned}
-2r_w(Q_0^\alpha) &= 2\sum_{i=1}^{3} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \log(1+\boldsymbol{\lambda}'\boldsymbol{u}_{ij}) \\
&= 2\sum_{i=1}^{3} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \log\left(\boldsymbol{\lambda}'\boldsymbol{u}_{ij} - \frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{u}_{ij}\boldsymbol{u}_{ij}'\boldsymbol{\lambda}\right) + o_p(n^{-1}) \\
&= \boldsymbol{U}'\boldsymbol{D}^{-1}\boldsymbol{U} + o_p(n^{-1}) \\
&= d_{33}(\hat{F}(Q_0^\alpha) - \alpha)^2 + o_p(n^{-1}),
\end{aligned}$$

where the last step is a consequence of (10) and d_{33} is the last (third for $k = 3$) diagonal element of $\boldsymbol{D}^{-1}$. If we let

$$c = d_{33}\sum_{i=1}^{3} \frac{w_i^2}{n_i - 1} F_i(Q_0^\alpha)(1 - F_i(Q_0^\alpha)), \tag{12}$$

it immediately follows that $-2r_w(Q_0^\alpha)/c$ will have a limiting χ^2 distribution with one degree of freedom. □

The scaling constant c involves the true distribution function F_i and quantile Q_0^α. Replacing Q_0^α by its weighted sample quantile $\hat{Q}^\alpha = \hat{F}^{-1}(\alpha)$ and F_i by its empirical counterpart $\hat{F}_i(\hat{Q}^\alpha) = n_i^{-1}\sum_{j=1}^{n_i} 1_{Y_{ij} \le \hat{Q}^\alpha}$ will not affect the limiting distribution of the test statistic.

Under the WEL approach, a $100(1-\rho)\%$ confidence interval for Q_0^α can be constructed as $\{Q^\alpha | -2r_w(Q^\alpha)/c < \chi_{(1)}^{2,\rho}\}$, where $\chi_{(1)}^{2,\rho}$ is the ρ-quantile from the χ^2 distribution with one degree of freedom. The ratio $r_w(Q^\alpha)$ is computable for any Q^α such that Q^α is in the convex hull formed by the overall sample. A bootstrap calibration of the confidence interval is also a possibility. See Fu et al. (2008) for details.

3 Simulation study

To assess the finite sample performance of our proposed methodology, we now present the results of some Monte Carlo simulations. As a benchmark, we consider the approach of Woodruff (1952) who basically suggested constructing confidence intervals for quantiles of complex surveys by inverting the confidence intervals of the distribution function. Sitter and Wu (2001) found this method to be quite reliable even in the moderate to extreme tail regions of distributions.

We consider a population divided into three strata with weights $0.50, 0.30$, and 0.20. The samples for the strata are independently generated from three lognormal distribution functions with means and standard deviations $(1.5, 0.3)$, $(2, 0.4)$, and $(2.1, 0.4)$. We use pooled sample sizes of $n = 50, n = 100$, $n = 200$, and construct 95% confidence intervals for seven different quantiles. For each specification, we conduct $5,000$ simulations. Table 1 reports the simulated coverage probability (CP), lower tail error rates (L), upper tail error rates (U), and the average length (AL) of the intervals. With the exception of the case where $n = 50$ and $\alpha = 0.05$, both confidence intervals seem to have excellent coverage rates even in the tails of the distribution. Interestingly, the tail error rates of the WEL interval seem to be much more balanced than Woodruff's. In the moderate to extreme tail regions (i.e., $\alpha = 0.05, 0.10, 0.90, 0.95$), WEL tends to slightly outperform Woodruff as WEL's coverage probabilities are closer to the nominal level of 95%. This is not true for all instances but the "overall picture" gives WEL a slight advantage. The quantiles towards the center of the distribution do not pose much problems (which is expected). The WEL intervals are roughly on par with Woodruff's.

Table 1: Simulated coverage and tail error rates for 95% confidence intervals

(n_1, n_2, n_3)		$Q^{0.05}$	$Q^{0.10}$	$Q^{0.25}$	$Q^{0.50}$	$Q^{0.75}$	$Q^{0.90}$	$Q^{0.95}$
Woodruff Confidence Interval for Quantile Q^α								
$(20, 20, 10)$	CP	82.94	94.58	95.48	95.52	95.42	96.24	94.86
	L	11.34	0.78	1.08	1.52	1.56	1.92	1.02
	U	5.72	4.64	3.44	2.96	3.02	1.84	4.12
	AL	0.91	1.34	1.28	1.56	2.66	5.30	7.51
$(40, 40, 20)$	CP	93.90	94.72	95.04	95.42	95.12	95.98	96.64
	L	0.98	0.68	1.38	1.84	1.96	1.72	1.72
	U	5.12	4.60	3.58	2.74	2.92	2.30	1.64
	AL	1.01	0.93	0.88	1.10	1.84	3.39	5.85
$(80, 80, 40)$	CP	94.62	94.36	94.92	95.26	94.70	95.60	95.60
	L	0.98	1.62	1.74	1.78	2.22	2.06	2.16
	U	4.40	4.02	3.34	2.96	3.08	2.34	2.24
	AL	0.73	0.62	0.62	0.77	1.28	2.31	3.59
WEL Confidence Interval for Quantile Q^α								
$(20, 20, 10)$	CP	87.34	94.38	95.56	94.56	94.74	94.50	94.00
	L	11.34	3.06	2.38	2.38	2.20	2.52	1.88
	U	1.32	2.56	2.06	3.06	3.06	2.98	4.12
	AL	1.09	1.30	1.28	1.57	2.65	5.08	7.36
$(40, 40, 20)$	CP	96.08	95.10	94.80	95.18	94.78	95.26	94.88
	L	1.84	2.78	2.64	2.32	2.56	1.94	2.18
	U	2.08	2.12	2.56	2.50	2.66	2.80	2.94
	AL	1.06	0.88	0.87	1.10	1.84	3.37	5.43
$(80, 80, 40)$	CP	94.64	94.78	95.10	95.18	94.48	95.22	94.68
	L	3.18	3.04	2.40	2.24	2.56	2.32	2.52
	U	2.18	2.18	2.50	2.58	2.96	2.46	2.80
	AL	0.66	0.61	0.62	0.77	1.28	2.28	3.44

4 Conclusion

Following up on the work of Fu et al. (2008), we proposed a weighted empirical likelihood-based inference method for quantiles in the presence of a stratified random sampling design. Our method is very easy to implement as computational routines are readily available. Through simulations, we were able to show that the confidence intervals obtained from our method perform just as well (and slightly better in some cases) as the popular method of Woodruff (1952). Thus, the WEL approach is a perfectly reliable method of inference for quantiles arising from stratified random samples.

So far, we have limited ourselves to inferences on a single measure (i.e., one quantile or one mean). But one may be interested in making simultaneous inference on multiple quantiles or means. The nature of complex surveys make the asymptotics much more difficult in such cases. Our work along with the work of Fu et al. (2008) provide partial guidance for future research into inference for a vector of measures.

References

Chen, J. and Sitter, R. R. (1999) "A pseudo empirical likelihood approach to the effective use of auxilary information in complex surveys" *Statistica Sinica* **9**, 385-406.

Cochran, W. G. (1977) *Sampling Techniques*, 3rd ed., John Wiley and Sons Inc.

Francisco, C. A. and Fuller, W. A. (1991) "Quantile estimation with a complex suvey design" *The Annals of Statistics* **19**, 454-469.

Fu, Y., Wang, X. and Wu, C. (2008) "Weighted empirical likelihood inference for multiple samples" *Journal of Statistical Planning and Inference* **139**, 1462-1473.

Gross, T. (1980) "Median estimation in sample surveys" in *Proceedings of Section on Survey Research Methods, American Statistical Association*, 181-184

Owen, A. B. (1988) "Empirical likelihood ratio confidence intervals for a single functional" *Biometrika* **75**, 237-249.

Owen, A. B. (1990) "Empirical likelihood ratio confidence regions" *The Annals of Statistics* **18**, 90-120.

Owen, A. B. (2001) *Empirical Likelihood*, Chapman and Hall/CRC.

Qin, J. and Lawless, J. (1994) "Empirical likelihood and general estimating equations" *The Annals of Statistics* **22**, 300-325.

Sitter, R. R. and Wu, C. (2001) "A note on Woodruff confidence intervals for quantiles" *Statistics & Probability Letters* **52**, 353-358

Woodruff, R. (1952) "Confidence intervals for medians and other position measures" *Journal of the American Statistical Association* **47**, 635-646.

Wu, C. (2004) "Some algorithmic aspects of the empirical likelihood method in survey sampling" *Statistica Sinica* **14**, 1057-1069.

Wu, C. (2005) "Algorithms and R codes for the pseudo empirical likelihood method in survey sampling" *Survey Methodology* **31**, 239-243.

Wu, C. and Rao, J. N. K. (2006) "Pseudo-empirical likelihood ratio confidence intervals for complex surveys" *The Canadian Journal of Statistics* **34**, 359-375.

Zhong, B. and Rao, J. N. K. (2000) "Empirical likelihood inference under stratified random sampling using auxilary population information" *Biometrika* **87**, 929-938.