行政院國家科學委員會專題研究計畫　成果報告

經驗概度於多零資料的應用(2/2)

計畫類別： 個別型計畫
計畫編號： NSC91-2118-M-032-003-
執行期間： 91 年 08 月 01 日至 92 年 07 月 31 日
執行單位： 淡江大學數學系

計畫主持人： 陳順益

報告類型： 完整報告

處理方式： 本計畫可公開查詢

中　華　民　國 93 年 2 月 13 日

Empirical likelihood with application to data sets with large proportion of zeroes

Shun-Yi CHEN (with Jiahua CHEN and J. N. K. RAO)

Abstract: If a population contains many zero values and the sample size is not very large, the traditional normal approximation based confidence intervals for the population mean may have poor coverage probabilities. This problem is substantially reduced by constructing parametric likelihood ratio intervals when an appropriate mixture model can be found. However, in the context of survey sampling, a general preference is to make minimal assumption about the population. The authors have therefore investigated the coverage properties of nonparametric empirical likelihood confidence intervals for the population mean. Under a variety of hypothetical populations, the empirical likelihood intervals often outperformed parametric likelihood intervals by having more balanced coverage rates and larger lower bounds. The have also used data from the Canadian Labour Force Survey-2000 to illustrate the empirical likelihood method.

1. INTRODUCTION

We discuss the problem described in Kvanli, Shen & Deng (1998, hereafter KSD). In accounting practice, a sample of about 100 claims is often obtained for re-counting. Most of the claims will be found legitimate, but a small portion of claims may be excessive. There is hence a need to construct a confidence interval (CI) for the average (or total) amount of excessive claims. The lower confidence bound is often used, for example, by the government to compute the amount of money owed to the government. The accuracy of the lower bound is of particular importance.

Often, a CI of a parameter can be constructed based on the asymptotic normality of the corresponding point estimator. Thus, the true coverage of the CI hinges on the precision of the normal approximation. When the population distribution is highly skewed, the normal approximation becomes unreliable unless the sample size is very large. The non-coverage below the lower bound and above the upper bound could be totally unbalanced. In this situation, one may instead construct a CI via the likelihood function. The coverage probability of a likelihood interval is usually based on a chi-square approximation to the distribution of the likelihood ratio statistic. Likelihood based inferences are known to have many optimal properties under some regularity conditions. However, the optimal properties naturally depend on the appropriateness of the parametric model and the precision of the chi-square approximation.

When a population contains many zero values, commonly used parametric models such as normal, Poisson or Gamma distribution, have to be modified. KSD discussed the use of mixtures of 0 and one of the commonly used parametric models in this situation. They demonstrated that with appropriate choice of the mixture models, the resulting CIs lead to more accurate coverage probabilities compared to the ones obtained by the traditional, normal approximation based intervals. They also observed that the accuracy is dependent on the correctness of the parametric model chosen.

As pointed out in Cochran (1977), the preference in survey sampling is to make, at most, limited assumptions about the population distribution, without specifying a parametric form. In this spirit, we explore the possibility of constructing accurate CIs without making explicit parametric assumptions. We use the nonparametric empirical likelihood (EL) method to construct such intervals.

The EL methodology was first discussed in the context of survey sampling by Hartley & Rao (1968). Its application to survey sampling has been further explored in Chen & Qin (1993), Chen & Sitter (1999), Zhong & Rao (2001), and Wu & Sitter (2001). Some important work on empirical likelihood can be found in Owen (1988, 1990, 1991), and Qin & Lawless (1994, 1995). The recent book by Owen (2001) contains a comprehensive account of recent developments in EL methodology.

The EL function has properties similar to the parametric likelihood. Most notably, under very mild conditions, the EL ratio statistic has a chi-square limiting distribution. At the same time, it does not require a parametric model assumption. The CI has a data-driven shape, and remains inside the natural bounds of the parameter space.

In this paper, we give a brief account of the EL method in general, and its application to the problem discussed in KSD in particular. The EL method is straightforward, but its finite sample performance in the intended application can only be predicted by theory and determined through simulation. Therefore, we examine the performance of the EL interval under a variety of scenarios and also using data from the Canadian Labour Force Survey-2000. In general, we found that the EL intervals have overall coverage properties similar to parametric likelihood (PL) intervals based on the normal mixture model. However, the EL gives the best lower bounds relative to competing methods. Its non-coverage probabilities below the lower bound are consistently closer to the target value (2.5% for the usual 95% CI). In addition, the EL lower bound is often the largest one on average. Thus, it is particularly advantageous to use the lower bound of the EL interval. The comparison of the upper bounds of the PL intervals and EL intervals is mixed. However, the differences, if any, are usually small.

We introduce the EL method and the KSD method in Sections 2 and 3 respectively. Extension of the EL method to stratified random sampling is given in Section 4. Simulation results are presented in Section 5. We also use these methods in Section 5 to analyze data from the Canadian Labour Force Survey-2000. Some concluding remarks are given in Section 6.

2. EMPIRICAL LIKELIHOOD METHOD FOR THE MEAN

Let $Y_1, Y_2, \ldots, Y_n$ be a sample of independent and identically distributed (iid) random variables with the common distribution $F(y)$. Let p_i be the probability of observing Y_i. The empirical log-likelihood function is defined as

$$el_n(F) = \sum_{i=1}^{n} \log p_i; \quad 0 \le p_i \le 1; \quad \sum_{i=1}^{n} p_i = 1. \tag{1}$$

Without further restrictions on the choice of $F(y)$, the log-likelihood (1) is maximized when $p_i = n^{-1}$ and the resulting maximum empirical likelihood estimate of $F(y)$ is the well known empirical distribution function $F_n(y) = n^{-1}\sum_{i=1}^{n} I(Y_i \le y)$ where $I(\cdot)$ is the indicator function. Let $\tau = \tau(F) = E(Y)$, the mean of Y. Then the maximum empirical likelihood estimate of τ is $\bar{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i$.

We can also construct a "profile" empirical log-likelihood function of τ. Let us maximize $el_n(F)$ under (1) and an additional restriction

$$\sum_{i=1}^{n} p_i Y_i = \tau$$

for each given τ. The maximization can be done with surprising simplicity. The maximum is attained when

$$p_i = [n\{1 + \lambda(Y_i - \tau)\}]^{-1}, \tag{2}$$

where λ is the Lagrange multiplier that solves the equation

$$\sum_{i=1}^{n} \frac{Y_i - \tau}{1 + \lambda(Y_i - \tau)} = 0. \tag{3}$$

The solution to (3) exists whenever τ falls inside the convex hull generated by $Y_1, \dots, Y_n$. Hence, substituting (2) in (1), we get the profile empirical log-likelihood

$$el_n(\tau) = -\sum_{i=1}^{n} \log\{1 + \lambda(Y_i - \tau)\} - n \log n. \tag{4}$$

In this paper, we use the same notation, *el*, for the empirical log-likelihood (1) and the resulting profile empirical log-likelihood (4). However, its meaning should be clear from the associated argument(s). The profile EL ratio function is defined as

$$er_n(\tau) = 2\sum_{i=1}^{n} \log\{1 + \lambda(Y_i - \tau)\}. \tag{5}$$

Letting τ_0 be the true mean, and assuming that the third moment of Y exists, Owen (1990) showed that

$$er_n(\tau_0) \rightarrow \chi_1^2$$

in distribution as $n \rightarrow \infty$, where χ_1^2 denotes the chi-square variable with one degree of freedom (df). Hence, an approximate $100(1-\alpha)\%$ CI for τ is given by

$$\{\tau : er_n(\tau) \leq \chi_{1-\alpha,1}^2\} \tag{6}$$

where $\chi_{1-\alpha,1}^2$ is the $(1-\alpha)$-th quantile of the chi-square distribution with one df. The interval (6) is the $100(1-\alpha)\%$ empirical likelihood CI for the mean $\tau_0 = E(Y)$. Appendix A.1 gives a simple algorithm for calculating the empirical likelihood CI (6).

When a population contains a large proportion of zero values, the EL method described above can be applied directly. In the context of a finite population, we assume that the current population is a member of a sequence of evolving finite populations such that the proportion of zeroes in the sequence is fixed and hence does not change with the sample size n. This condition is equivalent to saying that the chi-square approximation works better in practice when the proportion of non-zero values is not too small. We also assume that the sample fraction is small so that the iid assumption on the sample $y_1, \dots, y_n$ is approximately valid.

KSD did not explicitly examine finite population situations, but the assumption of fixed proportion of non-zero values is implicit in KSD as they required iid structure. Unlike the mixture models suggested in KSD, the chi-square limiting distribution of the empirical likelihood ratio statistic remains valid without any parametric model assumption. Hence, the method provides another useful solution to the problem discussed in their paper.

3. MIXTURE MODELS

The mixture models discussed in KSD have density functions of the form

$$f(y; \mu, \theta, p) = pf_1(y; \mu, \theta)I(y \neq 0) + (1-p)I(y = 0),$$

where p is the population error rate and $f_1(y; \mu, \theta)$ is a parametric density function with conditional mean μ and nuisance parameters θ. The unconditional mean of Y is therefore $\tau = p\mu$. The problem of interest is to construct a CI for the mean τ.

Suppose $Y_1, \ldots, Y_n$ are iid random variables with common density function $f(y; \mu, \theta, p)$. The log-likelihood function is then given by

$$l_n(\mu, \theta, p) = \sum_{i=1}^{n} \log f(y_i; \mu, \theta, p).$$

The likelihood ratio function for testing $\tau = \tau_0$ is defined as

$$r_n(\tau_0) = 2\{\sup_{\mu,\theta,p} l_n(\mu, \theta, p) - \sup_{\mu,\theta,p:\tau=\tau_0} l_n(\mu, \theta, p)\}. \tag{7}$$

When the density function $f_1(y; \mu, \theta, p)$ satisfies some regularity conditions, which is the case for most commonly used models, the likelihood ratio statistic has χ_1^2 limiting distribution based on the result of Wilks (1938). Accordingly, a two-sided approximate $100(1-\alpha)\%$ CI for τ is given by

$$\{\tau : r_n(\tau) \le \chi^2_{1-\alpha,1}\}. \tag{8}$$

Note that we need to assume that p is not equal to 0 or 1.

Despite the perceived simplicity of parametric mixture models, computations associated with (8) are more complex than these for EL. When f_1 is a normal or an exponential distribution, KSD provided detailed instructions for the related numerical solution. We give a very brief summary of the numerical solution here. Under the normal mixture model, let k be the number of zero values in the sample of size n and $\bar{y}_n$ be the sample mean. The computation starts with an initial value $\tau^{(0)}$ less than $\bar{y}_n$. Given $\tau = \tau^{(0)}$, the likelihood is maximized when $\mu = \mu^{(0)}$ solves the cubic equation

$$\mu^3 - A\mu^2 + B\mu - C = 0,$$

where

$$A = \frac{(2n-k)\tau^{(0)} + 3T}{2(n-k)}, \quad B = \frac{S(n-k) + (3n-k)T\tau^{(0)}}{2(n-k)^2}, \quad C = \frac{nS\tau^{(0)}}{2(n-k)^2}$$

with $T = \sum_{i=1}^n y_i$ and $S = \sum_{i=1}^n y_i^2$. The likelihood ratio function (7) can then be computed at $\tau^{(0)}$ and $\mu^{(0)}$. If it is smaller than $\chi^2_{1-\alpha,1}$, the lower bound of the CI is larger than τ. The lower bound is obtained iteratively by increasing $\tau^{(0)}$ and so on until (7) is equal to $\chi^2_{1-\alpha,1}$. The upper bound is obtained in the same way. To properly estimate the variance of the normal distribution component, we further require $n - k \ge 2$.

Under the exponential model, we also start with an initial value $\tau^{(0)}$. Given $\tau^{(0)}$, the likelihood is maximized when $\mu = \mu^{(0)}$, where

$$\mu^{(0)} = \frac{A + \sqrt{A^2 - 4B}}{2}$$

with $A = (2n\tau^{(0)} + T - k\tau^{(0)})/\{2(n-k)\}$ and $B = T\tau^{(0)}/\{2(n-k)\}$. The rest of the computations are similar to these for the normal mixture model.

The likelihood ratio function (7) for the parametric mixture method can be extended to empirical likelihood by changing $f_1(y_i; \mu, \theta)$ to $\tilde{p}_i$, $i = 1, \ldots, n-k$, where $\tilde{p}_i \ge 0$, $\sum_{i=1}^{n-k} \tilde{p}_i = 1$ and $\sum_{i=1}^{n-k} \tilde{p}_i y_i = \mu = \tau/p$. It can be shown that the resulting EL ratio function is identical to $er_n(\tau)$ given by (5). Therefore, the mixture model formulation of EL is not providing any additional information.

4. EXTENSION TO STRATIFIED RANDOM SAMPLING

Simple random sampling is an important building block for more complex sampling designs. However, it is rarely applied directly. Often, the population is divided into several strata, and independent simple random samples are drawn within strata. Interestingly, the EL method can be applied to the stratified case; see Zhong & Rao (2000) and Chen & Sitter (1999) for details.

To simplify the discussion, we consider populations with two strata, denoted by 1 and 2, with weights W_1 and W_2. Further, we assume that the sampling fractions within strata are small so that the dependence between the sampled values can be ignored. The following development can be applied to more than two strata without additional technical difficulties. Generalization to more complex designs is possible, but may not be straightforward. However, such generalizations are not the focus of this paper.

Suppose that the sample sizes in strata 1 and 2 are m and n with observations $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$, respectively. The log-empirical likelihood function can then be written as

$$el_{m,n}(p_1, \ldots, p_m, q_1, \ldots, q_n) = \sum_{i=1}^{m} \log p_i + \sum_{j=1}^{n} \log q_j. \tag{9}$$

The corresponding restrictions are given by

$$\sum_{i=1}^{m} p_i = 1; \quad \sum_{j=1}^{n} q_j = 1. \tag{10}$$

Of course, we also require $p_i, q_j > 0$. The empirical likelihood is maximized when $p_i = m^{-1}$ and $q_j = n^{-1}$.

To obtain the profile empirical log-likelihood for the overall population mean τ, we need only maximize (9) subject to (10) and an additional constraint

$$W_1 \sum_{i=1}^{m} p_i x_i + W_2 \sum_{m=1}^{n} q_j y_j = \tau.$$

The profile empirical log-likelihood (of τ) is maximized at

$$\tau = W_1 \bar{x}_m + W_2 \bar{y}_n,$$

where $\bar{x}_m$ and $\bar{y}_n$ are the sample means. In this case, $p_i = m^{-1}$ and $q_j = n^{-1}$ for all i, j.

It appears that the computational problem associated with the profile log-likelihood of τ is very complex. However, a very simple solution in fact exists. Starting with an appropriate value t, we solve

$$\sum_{i=1}^{m} \frac{x_i - \tau_1}{1 + m^{-1} W_1 t (x_i - \tau_1)} = 0; \quad \sum_{j=1}^{n} \frac{y_j - \tau_2}{1 + n^{-1} W_2 t (y_j - \tau_2)} = 0$$

to obtain $\tau_1(t)$ and $\tau_2(t)$. Since the functions in the above equations are monotone in τ_1 and τ_2 respectively, the computation can be done easily. The profile empirical log-likelihood at $\tau(t) = W_1 \tau_1(t) + W_2 \tau_2(t)$ is then given by

$$\begin{aligned} el_{m,n}\{\tau(t)\} &= -\sum_{i=1}^{m} \log\{1 + m^{-1} W_1 t (x_i - \mu_1)\} - \sum_{j=1}^{n} \log\{1 + n^{-1} W_2 t (y_j - \mu_2)\} \\ &\quad - m \log m - n \log n. \end{aligned}$$

The corresponding profile EL ratio function is

$$er_{m,n}\{\tau(t)\} = 2 \sum_{i=1}^{m} \log\{1 + m^{-1} W_1 t (x_i - \mu_1)\} + 2 \sum_{j=1}^{n} \log\{1 + n^{-1} W_2 t (y_j - \mu_2)\}. \tag{11}$$

To compute its value at a given point τ_0, we need only solve the equation $\tau(t) = \tau_0$ for the corresponding value of t. The function $\tau(t)$ is monotone in t. Hence, a linear search can be used to find the solution. Appendix A.2 gives details of the above algorithm.

We have shown in Appendix A.3 that the profile EL ratio function at the true parameter point τ_0 converges in distribution to χ_1^2. More precisely, we have the following theorem.

THEOREM 1. *Suppose $x_i, i = 1, \ldots, m$ and $y_i, j = 1, \ldots, n$ are two sets of iid random variables and $m/n \to \rho \in (0, 1)$ as $n \to \infty$. Assume $E(|X|^3) < \infty$ and $E(|Y|^3) < \infty$. Let $\tau_0 = W_1E(X) + W_2E(Y)$. Then $er_{m,n}(\tau_0)$ as defined by (11) has chi-square limiting distribution with one degree of freedom.*

Using Theorem 1, an asymptotic $100(1-\alpha)\%$ CI for $\tau = W_1E(X) + W_2E(Y)$ can constructed as

$$\{\tau : er_{m,n}(\tau) \leq \chi^2_{1-\alpha,1}\}.$$

Due to the monotonicity between t and τ discussed earlier, this set is indeed an interval. Its coverage properties are investigated by simulation in Section 5.3.

5. SIMULATION STUDY AND AN APPLICATION

We first consider the simplest case when the sample consists of essentially iid observations. In this case, we compare the traditional normal approximation based method, the KSD method, and the EL method through simulated data (Section 5.1). Second, we compare the KSD method and the EL method using data from the Canadian Labour Force Survey-2000 (Section 5.2). Third, we consider the situation when the population is stratified. In this case, the KSD method needs substantial extension and is not included in the simulation. Instead, a simple method, proposed by Godambe & Thompson (1999), is compared to the proposed EL method.

5.1. Simple random sampling

The simulations in this section mimic the simulations done in KSD. We generated data from mixture models with $\mu = 5$, $\sigma^2 = 16$ and error rate parameter p varying from 0.05 to 0.25 in increments of 0.05. The sample size is n=100. The nominal confidence level is taken as 95%. For each p and the parametric mixture model specified, 10,000 simulation runs were generated. Thus, the simulation error associated with the coverage probability is less than 0.5%. We recorded the proportion of runs with the true mean smaller than the lower bound, and the proportion of runs with the true mean larger than the upper bound. We also computed the average lower bound and the average upper bound for each method.

When the sample size increases to $n = 200$, all methods improved, but the conclusions are not changed. Therefore, we have not reported our results for $n = 200$.

Due to the requirement $n-k \geq 2$ for the normal mixture model, we threw away simulation runs with fewer than 2 non-zero observations. This is not a problem in practice. In addition, we only allowed non-negative observations to fit the context of the targeted application. These restrictions made the true population mean of the sample inflated, but we adjusted the population mean when computing the coverage probabilities. We used an Splus function to compute the inflated mean for each configuration of n, p, μ, and σ^2. For example, when $n = 100$, $p = 0.05$, $\mu = 5$ and $\sigma^2 = 16$, the population mean becomes 0.3001629 instead of $p\mu = 0.25$. Based on 10,000 simulation runs, the overall sample mean is 0.30096 which is a close match. The inflation is not as severe when n or p is larger.

In Table 1, we report the results for the four methods of finding CIs, based on simulation runs generated from normal mixture models. Both non-coverage rates and average bounds are reported. The average length of a CI is the difference between the average upper bound and the average lower bound.

It follows from Table 1 that the CI constructed by the traditional (normal approximation) method has poor coverage probability in all situations. The coverage probability improves gradually as the number of non-zero observations increases. When the exponential mixture model is applied to normal mixture data, the problem of over-coverage becomes severe. Also, it gives longer than necessary CIs. In applications, if the lower bound is used as the standard for refund, the amount of refund is seriously lower than it should be. Using the correct normal mixture model or the EL method produce very similar results. The overall coverage of using the normal mixture model is slightly better when the number of non-zero values is low. The comparison is reversed when the number of non-zero values increases. In both cases, the difference is not significant. However, we note that the EL method produces intervals with a more balanced coverage. Its average lower bounds are larger in general. However, the difference is not statistically significant.

Why is the method based on the exponential model consistently producing intervals with higher than the nominal coverage rate? This phenomenon is also evident in the simulation results of KSD. By assuming the exponential model for the non-zero part of the data, we have in effect assumed that the population variance is equal to the square of its mean. In the above model, the true mean μ and variance σ^2 are 5 and 16 respectively. Using the exponential model is equivalent to assuming that the mean and variance are equal to 5 and 25 respectively. Thus, the assumed model gives larger variance than the true variance which in turn gives wider CIs. As a result the intervals lead to over-coverage.

Next, we generated simulation runs from the exponential mixture model. In this case, all the observations are non-negative. However, we still require at least two non-zero observations in each simulation run. Otherwise, the method based on the normal mixture model does not work. The simulation results are reported in Table 2.

We note from Table 2 that the exponential mixture model based confidence intervals have close to nominal coverage rates. Hence, it is evident that the appropriateness of the model assumption is very important in using the KSD approach. When the lower non-coverage rate is of primary concern, the EL method works the best. The EL interval has non-coverage rates below lower bound closer to the target value of 2.5%. In addition, its average lower bounds are consistently larger than others. This is of particular importance in applications, as noted before.

The lower bound from the EL method is highly correlated to the desired lower bound from the exponential model. Their correlation is about 0.9 for all cases. On the other hand, the correlation between the bounds from the normal mixture model and the exponential mixture model is about 0.8. We also note that the upper tail of the exponential distribution is much heavier than the normal distribution. Thus, the normal mixture model based method consistently under-estimates the variance in the upper tail. This is clearly the reason behind the poor coverage probabilities above the upper bound. On the other hand, it over-estimates the variance in the lower tail, resulting in low non-coverage below the lower bound. The EL method picks the shape of the distribution from the sample, so it is expected to work reasonably well in all situations. However, when there are too few non-zero observations, the information in the upper tail of the distribution may not be represented in the sample. As a result, the EL method will still under-estimate the variance in the upper tail simply by chance, unless some prior information about the upper tail is known.

Next, we consider the situation when none of the parametric model assumptions are true. In Table 3, we report the simulation results when the runs are generated from a mixture of 0 and a Gamma distribution with shape parameter 3 and scale parameter 5/3 such that the mean $\mu = 5$. Clearly, in this situation, the exponential model performs poorly. The normal mixture model and the EL method have similar overall coverage probabilities, and they are close to the target of 95%. It is also clear that the EL intervals have more balanced non-coverage rates as well as larger average lower bounds. Thus, we conclude that the EL method is better, albeit the advantage is not as large as we would like to see.

5.2. Analysis of data from Canadian Labour Force Survey-2000

The simulations in Section 5.1 are based on artificial populations. It would be important to study the performance of the methods on a real data set. For this purpose, we down-loaded a data set provided by Statistics Canada through the TriUniversity Data Resources. This data set is

Table 1: Results on 95% confidence intervals under a normal mixture model

Lower and upper non-coverage rates								
p	Normal Approximation		Exponential Mixture		Normal Mixture		Empirical Likelihood	
0.05	0.19	10.46	0.63	0.55	1.51	4.35	1.57	4.77
0.10	0.58	8.65	0.87	0.55	2.08	4.14	2.21	4.26
0.15	0.86	6.60	0.82	0.48	2.05	3.63	2.12	3.53
0.20	1.11	5.59	0.74	0.71	2.13	3.07	2.20	3.04
0.25	1.37	5.10	0.78	0.57	2.52	3.03	2.61	2.87
Average lower and upper bounds								
p	Traditional Approximation		Exponential Mixture		Normal Mixture		Empirical Likelihood	
0.05	0.023	0.579	0.092	1.060	0.100	0.693	0.107	0.687
0.10	0.192	0.968	0.248	1.386	0.275	1.073	0.280	1.074
0.15	0.398	1.337	0.436	1.743	0.479	1.433	0.484	1.437
0.20	0.628	1.697	0.646	2.105	0.705	1.785	0.710	1.790
0.25	0.870	2.042	0.868	2.455	0.943	2.121	0.948	2.128

Table 2: Results on 95% confidence intervals under exponential mixture model

Lower and upper non-coverage rates								
p	Normal Approximation		Exponential Mixture		Normal Mixture		Empirical Likelihood	
0.05	0.09	17.62	1.69	3.42	1.07	10.32	1.88	11.14
0.10	0.23	12.64	1.79	3.02	1.29	7.40	2.02	7.13
0.15	0.47	10.99	2.05	3.32	1.29	6.80	2.16	6.22
0.20	0.51	9.09	2.04	3.04	1.35	5.67	2.21	5.06
0.25	0.63	8.49	1.92	3.11	1.32	5.61	2.00	4.77
Average lower and upper bounds								
p	Traditional Approximation		Exponential Mixture		Normal Mixture		Empirical Likelihood	
0.05	-0.020	0.526	0.079	0.909	0.053	0.665	0.081	0.664
0.10	0.103	0.893	0.213	1.188	0.177	1.121	0.215	1.048
0.15	0.257	1.230	0.374	1.493	0.335	1.349	0.376	1.392
0.20	0.430	1.552	0.551	1.797	0.508	1.664	0.553	1.718
0.25	0.618	1.864	0.739	2.095	0.694	1.968	0.741	2.029

from the Canadian Labour Force Survey-2000. The website address is http://tdr.uoguelph.ca, but there are some restrictions to data access. We took a 10% random sample of the data from the province of Ontario for the purpose of illustration. We used the number of extra hours worked as response variable y. Among the 17,415 sampled observations, we have 3,677 non-zero values. The proportion of non-zero observations is 21%. The mean of the non-zero observations is 92.15 and the corresponding standard deviation is 74. The histogram (not shown here) indicated that the exponential mixture model of KSD may be appropriate. The mean of the entire data set is 19.46.

Due to the common practice of rounding off, there are a large number of tied observations in the data set. If we sample 100 units from it, we may obtain several runs with all observed non-zero values equal. A sample like this will disable the KSD method under the normal mixture model, as the sample variance of non-zero values will be zero. To avoid this technical problem, we added a uniform error to the non-zero values. We generated 10,000 random samples, each of size 100, from this data set. Since the sampling fraction is less than 1%, we regarded each sample as a set of independent observations.

The results of the simulation are summarized in Table 4. We note that the traditional normal approximation interval is again the worst. Using exponential mixture model results in serious over-coverage. As a result, its lower bound is the smallest, and its use will imply gross under-estimation of the average extra-hours Canadian worked in a year. The intervals based on normal mixture model and the empirical likelihood have very close overall coverage rates. However, the EL intervals are more balanced than the normal mixture model intervals. Consequently, their non-coverages below the lower bound are closer to the target of 2.5% for 95% CIs.

5.3. Stratified random sampling

In this section, we consider the situation when the population is divided into two strata, and a simple random sample is obtained from each stratum. Further, we assume that the strata sample fractions are negligible so that the observations may be regarded as iid within strata.

We generated samples of size $m = 60$ and $n = 140$ from two distributions. The percentage of non-zero observations ranged from 0.05 to 0.15. In general, we assumed that the second stratum has higher percentage of zeroes which warranted larger sample size. With two distributions to be selected, the number of choices is large. We considered only two distributions: mixture of 0 and normal, and mixture of 0 and exponential. We also allowed different distributions in the two strata. Thus, we considered three different types of populations. The strata weights are chosen as $W_1 = 0.4$ and $W_2 = 0.6$.

We studied the EL method and the method of Godambe & Thompson (1999) for stratified sampling. The pivotal quantity, proposed by Godambe and Thompson, reduces to

$$\frac{\hat{\tau} - \tau}{\sqrt{W_1^2 s_x^2/m + W_2^2 s_y^2/n + (W_1^2/m + W_2^2/n)(\hat{\tau} - \tau)^2}} \qquad (12)$$

in the case of two strata, where $\hat{\tau} = W_1\bar{x}_m + W_2\bar{y}_n$. The CI is based on the asymptotic normality of the pivotal quantity (12). The resulting CI is simply an inflated traditional normal approximation based interval (see Section 6).

Simulation results are given in Table 5 for the EL and the Godambe and Thompson (GT) methods. Table 5 shows that the GT intervals have lower coverage probabilities in all cases compared to EL intervals. In addition, EL intervals have more balanced coverages. The average lengths of the two intervals are close, but EL intervals have larger lower bounds on the average.

6. CONCLUDING REMARKS

For simple random sampling, we considered four methods of constructing confidence intervals for the population mean when the population contains many zero values. Intervals are based on the traditional normal approximation, the parametric likelihood (PL) ratios of normal mixture model and the exponential mixture model and the empirical likelihood (EL) ratio. Godambe & Thompson (1999) proposed another method based on the asymptotic normality of the pivotal quantity

$$\frac{\bar{X} - \tau}{\sqrt{\sum_{i=1}^{n}(X_i - \tau)^2}}.$$

The resulting interval is still centered at $\bar{X}$, but its length is increased by a factor of $1+\chi^2_{1-\alpha,1}/(2n)$ relative to the length of the normal approximation interval. Thus, it has higher coverage probability but it may not be useful when the population is highly skewed.

Many re-sampling methods aimed at improving the accuracy of the confidence intervals have also been proposed. In general, re-sampling methods have the ability to eliminate the effect of skewness which can be demonstrated through the Edgeworth expansion (see Hall, 1988). However, when the sample contains only a few non-zero observations, re-sampling methods are not expected to be helpful.

Our simulation results on the first three methods are consistent with Kvanli *et al.* (1998) in general. The method of Kvanli *et al.* (1998) has much better coverage properties compared to the traditional method based on the normal approximation. We also found that the EL method has some additional advantages: (1) EL method produces more balanced coverage probabilities. (2) EL method gives a larger average lower bound while maintaining the the non-coverage rate below lower bound close to the target value of 2.5%. (3) Unlike the normal mixture model, the method works whether the non-zero observations are tied or not. Appropriateness of the parametric model assumptions is important for the method of Kvanli *et al.* (1998), whereas the EL method is nonparametric.

ACKNOWLEDGMENTS

This research is partially supported by the Natural Science and Engineering Research Council of Canada. We are thankful to the referees for constructive suggestions.

APPENDIX

In this section, we give a brief description of the algorithm for computing the CIs given by (6) and a detailed description of the algorithm for computing the CIs under stratified simple random sampling. Further, we sketch a proof of Theorem 1.

A.1 Algorithm for EL interval: iid case.

It is easy to verify that $er_n(\tau)$ is a concave function in τ and is maximized when $\tau = \bar{y}_n$. Let $Y_{(n)}$ be the largest observation of Y_i's. A simple algorithm for the upper bound is as follows:

1. Let $t_1 = \bar{y}_n$, $t_2 = Y_{(n)}$
2. If $|t_2 - t_1|$ is small, stop. Otherwise, let $\tau = (t_1 + t_2)/2$.
3. Solve (3) by linear search and use (5) to get $el_n(\tau)$.
4. If $er_n(\tau) > \chi^2_{1-\alpha,1}$, set $t_2 = \tau$; otherwise $t_1 = \tau$. Go to step 2.

The algorithm for the lower bound is similar.

A.2 Algorithm for EL interval: stratified random sampling.

For any given τ_1 and τ_2 within the feasible range, we could compute the profile log-likelihood by maximizing the empirical log-likelihood with the restrictions

$$\sum_{i=1}^{m} p_i x_i = \tau_1; \quad \sum_{j=1}^{n} q_j y_j = \tau_2$$

in addition to

$$\sum_{i=1}^{m} p_i = 1; \quad \sum_{j=1}^{n} q_j = 1; \quad p_i > 0; \quad q_j > 0.$$

Using the Lagrange multiplier method, we find that

$$el_{m,n}(\tau_1, \tau_2) = -\sum_{i=1}^{m} \log\{1 + \lambda_1(x_i - \tau_1)\} - \sum_{j=1}^{n} \log\{1 + \lambda_2(y_j - \tau_2)\} - m \log m - n \log n$$

with λ_1 and λ_2 being the solutions of

$$\sum_{i=1}^{m} \frac{x_i - \tau_1}{1 + \lambda_1(x_i - \tau_1)} = 0; \quad \sum_{j=1}^{n} \frac{y_j - \tau_2}{1 + \lambda_2(y_j - \tau_2)} = 0. \tag{13}$$

We can then link the profile log-likelihood of τ and $el_{m,n}(\tau_1, \tau_2)$ by the following simple relationship:

$$el_{m,n}(\tau) = \sup\{el_{m,n}(\tau_1, \tau_2) : W_1\tau_1 + W_2\tau_2 = \tau\}.$$

To get a more explicit relationship between $el_{m,n}(\tau)$ and $el_{m,n}(\tau_1, \tau_2)$, we further utilize the Lagrange multiplier method. Let

$$g(\tau_1, \tau_2, t) = el_{m,n}(\tau_1, \tau_2) - t(W_1\tau_1 + W_2\tau_2 - \tau),$$

where t is the Lagrange multiplier. We first take the derivatives of $g(\tau_1, \tau_2, t)$ with respect to τ_1, τ_2 and t and set them equal to zero. Note that λ_1 and λ_2 are functions of τ_1 and τ_2 in (13). The first equation $\partial g(\tau_1, \tau_2, t)/\partial \tau_1 = 0$ reduces to

$$-\sum_{i=1}^{m} \frac{\lambda_1'(x_i - \tau_1) - \lambda_1}{1 + \lambda_1(x_i - \tau_1)} - W_1 t = 0,$$

where $\lambda_1' = \partial \lambda_1 / \partial \tau_1$. Since

$$\sum_{i=1}^{m} \frac{x_i - \tau_1}{1 + \lambda_1(x_i - \tau_1)} = 0$$

and

$$\sum_{i=1}^{m} \frac{1}{1 + \lambda_1(x_i - \tau_1)} = m,$$

we get $\lambda_1 = W_1 t/m$. Similarly, we get $\lambda_2 = W_2 t/n$ from the second equation $\partial g(\tau_1, \tau_2)/\partial \tau_2 = 0$. Consequently,

$$\begin{aligned} el_{m,n}\{\tau(t)\} &= el_{m,n}\{\tau_1(t), \tau_2(t)\} \\ &= -\sum_{i=1}^{m} \log[1 + \lambda_1\{x_i - \tau_1(t)\}] - \sum_{j=1}^{n} \log[1 + \lambda_2\{y_j - \tau_2(t)\}] - m \log m - n \log n, \end{aligned}$$

with $\tau_1(t), \tau_2(t)$ defined by (13) by letting $\lambda_1 = W_1 t/m$ and $\lambda_2 = W_2 t/n$.

According to the above analysis, we can compute the empirical profile log-likelihood function of τ by following a few simple steps. Suppose we want to compute $el_{m,n}(\tau_0)$ for a given τ_0.

1. Choose an initial value of $t = 0$.
2. Let $\lambda_1 = W_1 t/m$ and $\lambda_2 = W_2 t/n$.

3. Solve (13) for τ_1 and τ_2.

4. Compute the value of $W_1\tau_1 + W_2\tau_2$. If it is close to τ_0, then $el_{m,n}(\tau_0) = el_{m,n}(\tau_1, \tau_2)$. Otherwise, update the value of t and go to step 2.

Note that the values of τ_1 and τ_2 are determined by t. Hence, they are functions of t. Their derivatives are

$$\frac{d\tau_1}{dt} = -\frac{W_1}{m^2}\sum_{i=1}^{m}\frac{(x_i - \tau_1)^2}{\{1 + W_1(x_i - \tau_1)t/m\}^2}$$

and

$$\frac{d\tau_2}{dt} = -\frac{W_2}{n^2}\sum_{i=1}^{n}\frac{(y_i - \tau_2)^2}{\{1 + W_2(y_i - \tau_2)t/n\}^2}.$$

Consequently, $\tau(t) = W_1\tau_1(t) + W_2\tau_2(t)$ is monotone in t. Hence, the updating step can be easily done.

If we are only interested in using the empirical profile likelihood of τ to construct CIs, the algorithm can be simplified. We need only solve $el_{m,n}(\tau_1(t), \tau_2(t)) = \chi^2_{1-\alpha,1}$ for two values of t. The corresponding values of $\tau(t)$ will then be our lower and upper bounds of the CI. Again, it can be easily verified that $el_{m,n}\{\tau_1(t), \tau_2(t)\}$ is monotone in t. Solving this equation numerically is simple.

A.3 Proof of Theorem 1:

The proof of Theorem 1 given here is not completely rigorous. We assume, without proof, that

$$\tau_1(t) = \bar{x}_m + O_p(m^{-1/2}), \quad \tau_2(t) = \bar{y}_n + O_p(n^{-1/2}),$$

where $\bar{x}_m$ and $\bar{y}_n$ are the sample means, and t solves $W_1\tau_1(t) + W_2\tau_2(t) = \tau_0$ and $O_p(\cdot)$ denotes order in probability. Under this assumption, and the finiteness of the third moments of X and Y, we can easily show that $t/n = O_p(n^{-1/2})$. Note that under the assumption that m/n has a non-zero limit, n and m are of the same order. Hence, we need not distinguish between $O(n)$ and $O(m)$. For notational simplicity, we put $\tau_1 = \tau_1(t)$ and $\tau_2 = \tau_2(t)$. Consequently,

$$\begin{aligned} 0 &= \sum_{i=1}^{m}\frac{x_i - \tau_1}{1 + m^{-1}W_1t(x_i - \tau_1)} \\ &= \sum_{i=1}^{m}(x_i - \tau_1) - m^{-1}W_1t\sum_{i=1}^{m}(x_i - \tau_1)^2 + O_p(1). \end{aligned}$$

Therefore, we get

$$\begin{aligned} \tau_1 &= \bar{x}_m + \frac{W_1t}{m^2}\sum_{i=1}^{m}(x_i - \bar{x}_m)^2 + O_p(m^{-1}) \\ &= \bar{x}_m + \frac{W_1t}{m}s_x^2 + O_p(m^{-1}), \end{aligned}$$

where $s_x^2 = m^{-1}\sum_{i=1}^{m}(x_i - \bar{x}_m)^2$. Using similar notation, we also get

$$\tau_2 = \bar{y}_n + \frac{W_2t}{n}s_y^2 + O_p(n^{-1}).$$

Setting $W_1\tau_1 + W_2\tau_2 = \tau_0 = W_1E(X) + W_2E(Y)$, we get

$$t = \frac{W_1\{\bar{x}_m - E(X)\} + W_2\{\bar{y}_n - E(Y)\}}{W_1s_x^2/m + W_2s_y^2/n} + O_p(1).$$

Hence

$$\begin{aligned}
er_{m,n}\{\tau(t)\} &= 2\sum_{i=1}^{m}\log\{1+W_1t(x_i-\tau_1)/m\}+2\sum_{j=1}^{n}\log\{1+W_2t(y_j-\tau_2)/n\} \\
&= \frac{2W_1t}{m}\sum_{i=1}^{m}(x_i-\tau_1)-\frac{W_1^2t^2}{m^2}\sum_{i=1}^{m}(x_i-\tau_1)^2 \\
&\quad +\frac{2W_2t}{n}\sum_{j=1}^{n}(y_j-\tau_2)-\frac{W_2^2t^2}{n^2}\sum_{j=1}^{n}(y_j-\tau_2)^2+o_p(1) \\
&= \frac{[W_1\{\bar{x}_m-E(X)\}+W_2\{\bar{y}_n-E(Y)\}]^2}{W_1^2s_x^2/m+W_2^2s_y^2/n}+o_p(1),
\end{aligned}$$

where $o_p(1)$ denotes that the term goes to zero in probability as $n \to \infty$.

The conclusion then follows from the fact that $W_1\{\bar{x}_m - E(X)\} + W_2\{\bar{y}_n - E(Y)\}$ is asymptotically normal with mean 0 and variance $W_1^2m^{-1}\sigma_x^2 + W_2^2n^{-1}\sigma_y^2$.

REFERENCES

J. Chen & J. Qin (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

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

W. G. Cochran (1977). *Sampling Techniques*, 3rd Edition. Wiley, New York.

V. P. Godambe & M. E. Thompson (1999). A new look at confidence intervals in survey sampling. *Survey Methodology*, 25, 161-174.

P. Hall (1988). Theoretical comparisons of bootstrap confidence intervals (with discussion). *The Annals of Statistics*, 16, 927-985.

H. O. Hartley, & J. N. K. Rao (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

A. H. Kvanli, Y. K. Shen & L. Y. Deng (1998). Construction of confidence intervals for the mean of a population containing many zero values. *Journal of Business & Economic Statistics*, 16, 362-368.

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

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

A. B. Owen (1991). Empirical likelihood for linear models. *The Annals of Statistics* 19, 1725-47.

A. B. Owen (2001). *Empirical likelihood.* Chapman & Hall, New York.

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

J. Qin & J. F. Lawless (1995). Estimating equations, empirical likelihood and constraints on parameters. *The Canadian Journal of Statistics*, 23, 145-159.

S. S. Wilks (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9, 60-62.

C. Wu & R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-93.

C. X. B. Zhong & J. N. K. Rao (2000). Empirical likelihood inference under stratified random sampling using auxiliary information. *Biometrika*, 87, 929-38.

Table 3: Results on 95% confidence intervals under Gamma mixture model

Lower and upper non-coverage rates								
p	Normal Approximation		Exponential Mixture		Normal Mixture		Empirical Likelihood	
0.05	0.40	9.92	0.83	0.79	1.94	4.27	1.95	4.58
0.10	0.77	7.86	0.98	0.58	2.25	3.23	2.26	3.40
0.15	0.99	7.41	0.91	0.38	2.09	4.17	2.26	4.06
0.20	0.84	6.27	0.68	0.48	2.10	3.80	2.25	3.58
0.25	1.32	5.48	0.91	0.53	2.35	3.49	2.55	3.29
Lower and upper bounds								
p	Traditional Approximation		Exponential Mixture		Normal Mixture		Empirical Likelihood	
0.05	0.021	0.496	0.079	0.913	0.087	0.593	0.093	0.591
0.10	0.166	0.827	0.214	1.195	0.236	0.926	0.239	0.922
0.15	0.072	0.876	0.376	1.500	0.414	1.230	0.420	1.241
0.20	0.598	1.401	0.556	1.809	0.608	1.531	0.614	1.544
0.25	0.749	1.753	0.745	2.108	0.810	1.820	0.818	1.835

Table 4: Results on 95% confidence intervals based on data from the Canadian Labour Force Survey-2000

p	Normal Approximation		Exponential Mixture		Normal Mixture		Empirical Likelihood	
Lower and upper non-coverage rates								
0.21	0.70	7.25	1.30	1.21	1.63	4.56	2.11	4.11
Average lower and upper bounds								
0.21	9.824	28.996	10.958	34.577	11.189	30.691	11.621	31.386

Table 5: Results on 95% empirical likelihood and Godambe and Thompson confidence intervals under stratified random sampling

p_1, p_2: proportions of zero values in strata 1 and 2
p_l, p_u: lower and upper non-coverage rates
C_l, C_u: lower and upper confidence bounds

		Empirical likelihood interval					Godambe and Thompson interval				
p_1	p_2	p_l	p_u	C_l	C_u	length	p_l	p_u	C_l	C_u	length
Normal, Normal											
0.05	0.05	1.81	3.16	0.185	0.703	0.518	0.53	6.89	0.131	0.635	0.504
0.10	0.05	2.33	3.39	0.241	0.797	0.556	0.68	6.77	0.188	0.731	0.543
0.15	0.05	2.43	2.92	0.308	0.900	0.592	0.98	6.84	0.255	0.836	0.581
0.10	0.10	2.34	3.15	0.439	1.143	0.704	1.02	6.17	0.382	1.073	0.691
0.15	0.10	3.24	3.24	0.511	1.241	0.730	1.08	5.69	0.455	1.178	0.723
Exponential, Exponential											
0.05	0.05	1.66	6.81	0.145	0.712	0.567	0.28	12.30	0.071	0.592	0.521
0.10	0.05	2.24	5.97	0.191	0.824	0.633	0.34	10.92	0.117	0.688	0.571
0.15	0.05	1.96	6.25	0.245	0.937	0.692	0.29	10.44	0.170	0.785	0.615
0.10	0.10	2.31	5.70	0.353	1.162	0.809	0.58	9.43	0.270	1.011	0.741
0.15	0.10	3.02	5.36	0.415	1.276	0.861	0.57	8.32	0.331	1.111	0.780
Exponential, Normal											
0.05	0.05	2.04	5.15	0.159	0.717	0.558	0.18	10.13	0.089	0.614	0.524
0.10	0.05	2.16	4.70	0.212	0.804	0.592	0.36	9.39	0.145	0.705	0.560
0.15	0.05	2.53	4.86	0.279	0.909	0.630	0.59	8.62	0.214	0.813	0.600
0.10	0.10	2.50	4.52	0.375	1.147	0.772	0.54	8.33	0.298	1.031	0.733
0.15	0.10	3.23	4.29	0.450	1.249	0.799	0.58	7.83	0.372	1.136	0.763