Empirical Likelihood Confidence Intervals for a Population Containing Many Zero Values

Jiahua Chen

Department of Statistics & Actuarial Science

University of Waterloo

Revised for a presentation at McMaster University

Jan 22, 2003

This presentation is based on the joint work with Shuiyi Chen and J.N.K. Rao

It can be downloaded from

www.uwaterloo.ca/stats.uwaterloo.ca

Research partially supported by NSERC.

Outline of the talk

- Description of the Problem;
- Literature review;
- Empirical likelihood approach;
- Simulation study;
- Generalizations.

Description of the Problem

We consider a general problem in survey sampling.

Assume that there is a finite population consists of N sampling units.

Each unit has some characteristics of interest.

The problem of survey is to make inference on the finite population parameters such as:

$$\text{Population mean}: \quad \bar{Y} = N^{-1} \sum_{i=1}^{N} y_i;$$

$$\text{Population CDF}: \quad F_N(y) = N^{-1} \sum_{i=1}^{N} I(y_i \leq y)$$

with $I(y_i \leq y)$ being an indicator function.

Typically, the inference is done through a survey sampling.

First, we obtain a random sample from the finite population according to a sampling plan.

Second, we obtain measurements on sampled units.

Finally, we analysis the data and make inferences on the finite population parameters.

More often than not, we are not satisfied with merely giving a point estimate. An associated confidence interval is desirable.

In the case of population mean, it is usually true that the sample mean is asymptotically normal.

Hence, a typical 95% confidence interval has the form

$$\bar{y}_n - 1.96 n^{-1/2} \widehat{s}_n, \bar{y}_n + 1.96 n^{-1/2} \widehat{s}_n,$$

where $\bar{y}_n$ is the sample mean and s_n^2 is the sample variance (assuming $n << N$).

Under some conditions, the coverage rate of this CI converges to 95% when n, N go to infinity.

However, when n, N are finite, the coverage can be very different from 95%.

A classical example is when the population is severely skewed.

The special case we consider is when the population contains a large proportion of zero values.

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.

Thus, the amount of the excessive claim for most sampling unit is zero, with some non-zeroes.

An accurate confidence interval can be used by the government to compute the amount of money the firm owes.

The classical central limit theorem based CI is obviously not ideal in this case.

1. The coverage rate might be far from 95%;

2. The lower bound can be smaller than zero.

Kvanli, Shen and Deng (KSD, 1998) considered a method using mixture models.

Literature review

KSD suggested that perhaps an appropriate parametric model can be found for non-zero values in the population.

If so, such a population can be described by the following density function:

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

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.

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

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

We can therefore define the likelihood ratio function for testing $\tau = \tau_0 (= p\mu)$ as

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

Accordingly, a two-sided approximate $100(1-\alpha)\%$ CI for τ is given by

$$\{\tau : r_n(\tau) \leq \chi^2_{1-\alpha,1}\}. \qquad (1)$$

The theory behinds this procedure is the famous Wilks(1938) result of chisquare limiting distribution of the likelihood ratio statistic.

Even though this method also relies on asymptotic results, in this case the coverage rates are usually much better as the likelihood is tailor made just for such populations.

KSD provided simulation results when f is normal and exponential density functions.

They also discussed computational problems related to this procedure.

Motivation

In survey, we try to avoid parametric model assumptions whenever possible.

For example, no models are needed when using sample mean, ratio estimator or regression estimator.

Is there any possibility of achieving the same precision without using the mixture model?

The empirical likelihood method comes into picture as it is totally non-parametric.

What is it and how can it be applied here?

Empirical Likelihood

Let $Y_1, Y_2, \ldots, Y_n$ be a set of independent and identically distributed (iid) random variables.

Assume their common distribution is given by $F(y)$.

Let p_i be the probability of observing Y_i.

The empirical log-likelihood function is defined as

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

Without further restrictions on the choice of F, the log-likelihood is maximized when $p_i = n^{-1}$.

The resulting maximum empirical likelihood estimate of $F(y)$ is the well known empirical distribution function $F_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \leq y)$.

Let $\tau = \tau(F) = E(Y_1)$. Then the maximum likelihood estimate of τ is $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$.

This likelihood can also be used for constructing confidence intervals for τ.

Let us maximize $el_n(F)$ under an additional restriction

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

for each given τ.

The result is:

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

where λ is the Lagrange multiplier that solves the equation

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

We get the "profile" empirical log-likelihood

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

Under mild conditions,

$$er_n(\tau_0) \to \chi^2_1$$

in distribution as $n \to \infty$.

Hence, an approximate $100(1-\alpha)\%$ CI for τ is given by

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

Note that this procedure does not assume a parametric model, nor is designed for the population with a large number of zero values.

It is advantages to be totally non-parametric.

It may not work as perfectly as we would like.

After all, we do not change anything to fit our specific problem.

How good is it? Simulation!

One can easily observe that the performance of KSD method depends on the appropriate choice of models.

We cannot choose a model to meet the characteristic of the population for empirical likelihood. The coverage is better when the data is bell shaped compared to exponentially shaped.

One advantage of empirical likelihood is that it provides more balanced coverage; its lower bound is rightfully larger (compared to normal mixture method).

Generalization

If the data is collected via a stratified simple random sample design, can we still use these methods?

The finite mixture model approach might still work. However, it will meet quite a bit challenge in computation and in theory.

The empirical likelihood method works just as simple as before.

Let use present the case when there are two strata in the population.

The likelihood looks like

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

An additional constraint regarding to the population mean is give by

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

where W_1, W_2 are stratum weights.

A large sample result can be established.

Theorem 4.1 *Suppose* $x_i, i = 1, \ldots, x_m$ *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* $E[|X_1|^3] < \infty$ *and* $E[|Y_1|^3] < \infty$. *Let* $\tau_0 = W_1 E(X_1) + W_2 E(Y_1)$. *Then* $er_{m,n}(\tau_0)$ *as defined earlier has chisquare limiting distribution with one degree of freedom.*

Let us try a rought proof.

We assume, without proof, that

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

where $\bar{x}_m$ and $\bar{y}_n$ are the sample means, t solves $W_1\tau_1(t) + W_2\tau_2(t) = \tau_0$ and $O_p(\cdot)$ denotes order in probability.

It can then be shown that $t/n = O_p(n^{-1/2})$.

Put $\tau_1 = \tau_1(t)$ and $\tau_2 = \tau_2(t)$. We have

$$\begin{aligned} 0 &= \sum_{i=1}^{m} \frac{x_i - \tau_1}{1 + m^{-1} W_1 t (x_i - \tau_1)} \\ &= \sum_{i=1}^{m} (x_i - \tau_1) - \\ &\qquad m^{-1} W_1 t \sum_{i=1}^{m} (x_i - \tau_1)^2 + O_p(1). \end{aligned}$$

Therefore, we get

$$\begin{aligned}\tau_1 &= \bar{x}_m + \frac{W_1 t}{m^2}\sum_{i=1}^{m}(x_i - \bar{x}_m)^2 + O_p(m^{-1}) \\ &= \bar{x}_m + \frac{W_1 t}{m} s_x^2 + O_p(m^{-1}),\end{aligned}$$

where $s_x^2 = m^{-1}\sum_{i=1}^{m}(x_i - \bar{x}_m)^2$.

Similarly,

$$\tau_2 = \bar{y}_n + \frac{W_2 t}{n} s_y^2 + O_p(n^{-1}).$$

Setting

$$W_1\tau_1 + W_2\tau_2 = \tau_0 = W_1 E(X) + W_2 E(Y),$$

we get

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

Hence

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

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

$$W_1^2 m^{-1}\sigma_x^2 + W_2^2 n^{-1}\sigma_y^2.$$

We skip the details of the computational issue.

However, we have an algorithm which allows us to do a linear search on a convex function.

Hence, the computation is fast and the computational convergence is guaranteed.

The simulation indicates that the coverage properties of empirical likelihood method are very good in this situations too.