Pseudo Empirical Likelihood Confidence Intervals for Complex Sample Survey Data

N. GLENN GRIESINGER
Texas Southern University
Department of Mathematics
3100 Cleburne Street, Houston, Texas
UNITED STATES
Nancy.GlennGriesinger@tsu.edu

ANDREA J. SHELTON
Texas Southern University
Department of Health Sciences
3100 Cleburne Street, Houston, Texas
UNITED STATES
andrea.shelton@tsu.edu

KIRAN CHILAKAMARRI†
Texas Southern University
Department of Mathematics
Houston, Texas 77004
UNITED STATES

DEMETRIOS KAZAKOS
Texas Southern University
Department of Mathematics
3100 Cleburne Street, Houston, Texas
UNITED STATES
demetrios.kazakos@tsu.edu

Abstract: Complex sample survey data are obtained through multistage sampling designs that involve clustering, stratification, and non–responce adjustments. Standard statistical methods such as empirical likelihood are typically not applicable to complex samples because independent, identically distributed observations seldom result from such data. Hence, we derive pseudo empirical likelihood confidence intervals for stratified single–stage and stratified multistage sampling designs. Use of such designs include national health data sets.

Key–Words: Complex sample, survey data, empirical likelihood.

1 Introduction

Complex sample surveys which include a large representative sample of various demographic groups provide excellent sources of data for accessing various health measures. Data used to demonstrate these methods are typically found in large data sets such as national health surveys. Using standard statistical methods in this context induces a nonstandard covariance structure among sample quantities as patterns in the covariance matrix are nonstandard [4].

2 Statistical Sampling

We introduce a new sampling method and evaluate the method by determining its design effect. Estimators may have a loss or gain of efficiency when simple random sampling is not used. A sequence of estimators is asymptotically efficient for a parameter if the asymptotic variance of the estimator achieves the Cramer–Rao lower bound [1]. The design effect is a measure of the loss or gain or efficiency when sampling methods other than simple random sampling are used. The design effect is computed by determining two quantities: (1) the design specific estimate of the sampling variance, and (2) the sampling variance under the simple random sampling assumption, then determining the ratio $(1) : (2)$.

3 Complex Sample Survey Data

Complex survey data are obtained by stratification, cluster sampling, or unequal probability sampling. Since these procedures do not produce random samples, standard statistical methods are not appropriate for such data sets. Complex survey data analysis instead uses more suitable statistical methods. Such methodologies include using weights to assign greater or lesser importance to sample elements in order to accurately represent the population. An element's weight is determined by taking the inverse of its inclusion probability which is the probability of an item in the population becoming part of the sample during the drawing of a single sample.

One example of a method appropriate for complex survey data is pseudo–empirical likelihood, developed by [2] as a more suitable suitable methodology for complex survey data. The empirical likelihood

†Deceased

methodology was initially introduced as a sample survey method known as the scale load approach developed by [3].

4 Empirical Likelihood

Empirical likelihood has since been extended to other data types such as biased data, incomplete data, dependent data, spatial data, and complex survey data [7], [5], [2]. In the case of independent and identically distributed data, the parametric likelihood is a function of the parameter θ which takes values in the space Θ.

Definition 4.1 For a random sample $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ of size n, the *parametric likelihood* is

$$\begin{aligned} L(\theta) &= L(\theta; X_1, X_2, ..., X_n), \theta \in \Theta \quad (1) \\ &= \prod_{i=1}^{n} f(X_i; \theta), \quad (2) \end{aligned}$$

if $\{X_1, X_2, \ldots, X_n\}$ are $independent.$

When the density function $f(\mathbf{X}; \theta)$ is unknown one can use the empirical likelihood to determine the likelihood of a parameter [6], [7].

Definition 4.2 Suppose $\{X_1, X_2, \ldots, X_n\}$ is a random sample from an unknown distribution. Let the parameter θ denote the mean. Suppose p_i is the probability mass placed on X_i, $\sum_{i=1}^{n} p_i = 1$, $p_i \geq 0$. Let $t(\mathbf{p}) = \sum_{i=1}^{n} p_i X_i$ denote the value t assumes at $\mathbf{p}$. The *empirical likelihood [6], [7] for* θ *is defined as*

$$L(\theta) = \max_{\mathbf{p}, t(\mathbf{p})=\theta} \prod_{i=1}^{n} p_i. \quad (3)$$

For all θ in the convex hull of $\mathbf{X}$,

$$\begin{aligned} \max_{\mathbf{p}, t(\mathbf{p})=\theta} \prod_{i=1}^{n} p_i &\leq \max_{\mathbf{p}} \prod_{i=1}^{n} p_i \quad (4) \\ &\leq \prod_{i=1}^{n} \frac{1}{n} \\ &= L(\hat{\theta}), \end{aligned}$$

where $\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n}$. Therefore, $L(\hat{\theta}) = n^{-n} = max_\theta L(\theta)$. The corresponding *empirical log–likelihood ratio* is

$$l(\theta) = \log(\mathcal{T}(\theta)), \quad (5)$$

where

$$\mathcal{T}(\theta) = L(\theta)/L(\hat{\theta}). \quad (6)$$

5 Pseudo Empirical Likelihood

Chen and Sitter developed a pseudo empirical likelihood that extended to complex surveys (Chen and Sitter, 1999). Wu and Rao later developed pseudo–empirical likelihood ratio confidence intervals for complex surveys [8]. The pseudo empirical likelihood, developed as a more feasible alternative for analyzing survey data, is a design unbiased estimate of the log empirical likelihood function

$$\hat{l}(\mathbf{p}) = \sum_{i \in s} d_i \log(p_i), \quad (7)$$

where s is the set of units selected using a complex survey design, $d_i = \pi_i^{-1}$ are design weights, π_i are inclusion probabilities, s is the sample, and p_i is the probability mass placed on the data. Since pseudo empirical likelihood is design unbiased, the estimator is unbiased under the design protocol.

There is a modification in the estimate depending on whether there is stratified or nonstratified, single–stage or multi–stage sampling. The purpose of stratifying is to separate the population into overlapping, homogeneous subpopulations called strata. Independent samples are then drawn from each stratum. Stratified sampling can be done in either single or double stage manner.

In single-stage sampling, samples are drawn independently from each stratum. All samples are not necessarily the same size. In double–stage sampling, data are clustered within a given stratum, then at least two clusters are drawn from the stratum. Lastly, subsamples are drawn from each cluster [2].

To determine empirical likelihood confidence interval for the mean for complex survey data, we use the pseudo likelihood to expand the random sample case to complex samples. We first discuss random samples.

6 Random Sample Confidence Interval Derivation

In the case of a random sample, the empirical likelihood confidence interval for the mean θ, is derived by determining upper and lower limits θ_+ and θ_- for which $\mathcal{T}(\theta_\pm) = t_0 \in (0, 1)$, where t_0 is a threshold value based on the confidence level; $\mathcal{T}(\theta_\pm)$ is defined in Equation (6). Since empirical likelihood is defined on the support, upper and lower limits are bounded by the 1^{st} and n^{th} order statistic for a sample of size n,

$$X_{(1)} \leq \theta_- \leq \bar{X} \leq \theta_+ \leq X_{(n)}. \quad (8)$$

Inequality (8) can be used in two separate safeguarded searches for the limits. However, a faster approach is

to reformulate the problem as two optimization problems [7]:

$$\max_{\mathbf{p}} \sum_{i=1}^{n} p_i X_i \tag{9}$$

and

$$\min_{\mathbf{p}} \sum_{i=1}^{n} p_i X_i \tag{10}$$

subject to

$$p_i \geq 0 \tag{11}$$

$$\sum_{i=1}^{n} p_i = 1 \tag{12}$$

$$\sum_{i=1}^{n} \log(np_i) = \log(t_0). \tag{13}$$

Equation (9) subject to constraints given in Equations (11) through (13) leads to θ_+. Similarly, Equation (10) subject to the same constraints leads to θ_-. The optimal values are where the empirical likelihood curve, which is strictly concave, and the horizontal line $y = t_0$ intercept.

The threshold value t_0 for constructing a $100(1-\alpha)\%$ confidence interval is based on limiting values of a $\chi^2_\alpha(df)$ distribution; df represents degrees of freedom. Since the $\alpha = .05$ quantile for establishing a 95% confidence interval is $\chi^2_{.05}(1) \approx 3.8415$,

$$-2\log \mathcal{T}(\theta) \approx 3.8415. \tag{14}$$

Solving Equation (14) for $\mathcal{T}(\theta)$ establishes the threshold value for establishing a 95% confidence interval, $\mathcal{T}(\theta) \approx .147 = t_0$.

Since the optimization problems Equation (9) and Equation (10) are nonlinear, the optimality conditions are the Karush Kuhn Tucker conditions. To derive the confidence interval limits, determine the Lagargian:

$$\begin{aligned} G_1 &= \sum_{i=1}^{n} p_i X_i && (15)\\ &\quad -\gamma_1 \left(\sum_{i=1}^{n} \log(np_i) - \log(t_0) \right) && (16)\\ &\quad -\gamma_2 \left(\sum_{i=1}^{n} p_i - 1 \right), && (17)\\ & && (18) \end{aligned}$$

where γ_1 and γ_2 are Lagrange multipliers. Suppose $\mathbf{X}^*$ is a local solution of Equation (15). A first order necessary condition is that the gradient of the Lagrangian equals zero,

$$\nabla G_1(\mathbf{p}^*, \gamma_1^*, \gamma_2^* | \mathbf{X}, t_0) = 0, \tag{19}$$

where $\mathbf{p}^*$ is the optimal probability vector, and $(\gamma_1^*, \gamma_2^*) = \gamma^*$ are Lagrange multipliers such that the Karush Kuhn Tucker conditions are satisfied at $(\mathbf{X}^*, \gamma^*)$.

Determining the form of the optimal probability vector $\mathbf{p}^*$ entails taking the partial derivative of the Lagrangian Equation (15) with respect to p_k, where $k = 1, 2, \ldots, n$, then also taking the partial derivative of Equation (15) with respect to γ_1:

$$\begin{aligned} \frac{\partial G_1}{\partial p_k} &= X_k - \frac{n\gamma_1}{np_k} - \gamma_2 \\ &= X_k - \frac{\gamma_1}{p_k} - \gamma_2 \\ &= 0, \end{aligned} \tag{20}$$

and

$$\frac{\partial G_1}{\partial \gamma_1} = \sum_{i=1}^{n} p_i - 1 = 0. \tag{21}$$

From Equation (20),

$$p_k = \frac{\gamma_1}{X_k - \gamma_2}. \tag{22}$$

From Equation (22), $\sum_{i=1}^{n} p_i = 1$ implies

$$\sum_{k=1}^{n} \frac{\gamma_1}{X_k - \gamma_2} = 1, \tag{23}$$

which implies

$$\gamma_1 = \frac{1}{\sum_{k=1}^{n} \frac{1}{X_k - \gamma_2}}. \tag{24}$$

From Equation (24), the k^{th} element of the optimal probability vector, a function of γ_2, is:

$$\begin{aligned} p_k(\gamma_2) &= \frac{\gamma_1}{X_k - \gamma_2} && (25)\\ &= \frac{(X_k - \gamma_2)^{-1}}{\sum_{k=1}^{n} (X_k - \gamma_2)^{-1}} && (26) \end{aligned}$$

The Lagrange multiplier γ_2 is determined in such a manner that the constraints in Equations (11) through (13) are satisfied. This can be accomplished through a grid search.

7 Complex Sample Confidence Interval Derivation

Pseudo–empirical likelihood ratio confidence interval derivation for complex surveys is analogous to random sample interval derivation. However, pseudo–empirical likelihood is separated into two cases, nonstratified and stratified sampling.

For nonstratified sampling, the upper and lower confidence interval bounds are determined through the following optimization problem:

$$\max_{\mathbf{p}} \sum_{i=1}^{n} p_i X_i \tag{27}$$

and

$$\min_{\mathbf{p}} \sum_{i=1}^{n} p_i X_i \tag{28}$$

subject to

$$\begin{aligned} p_i &\geq 0 && (29)\\ \sum_{i=1}^{n} p_i &= 1 && (30)\\ n\sum_{i\in s} \tilde{d}_i(s)\log(p_i) &= \log(t_0), && (31) \end{aligned}$$

where s represents the set of units selected using a complex survey design, $\tilde{d}_i(s)$ are normalized design weights, $\tilde{d}_i(s) = \frac{d_i}{\sum_{i\in s} d_i}$, $d_i = \pi_i^{-1}$, π_i represents the inclusion probability, t_0 represents the threshold value for constructing a $100(1-\alpha)\%$ confidence interval based on limiting values of a $\chi^2_\alpha(df)$ distribution.

Equation (27) subject to constraints given in Equations (29) through (31) leads to θ_+. Similarly, Equation (28) subject to the same constraints leads to θ_-. The optimal values are where the pseudo–empirical likelihood curve and the horizontal line $y = t_0$ intercept.

As in the random sample case, we have nonlinear optimization problems. The optimality conditions are the Karush Kuhn Tucker conditions and the Lagargian, G_2 is:

$$G_2 = \sum_{i=1}^{n} p_i X_i \tag{32}$$

$$-\gamma_1\left(\sum_{i=1}^{n} \tilde{d}_i(s)\log(np_i) - \log(t_0)\right) - \gamma_2\left(\sum_{i=1}^{n} p_i - 1\right). \tag{33}$$

Determining the form of the optimal probability vector $\mathbf{p}^*$ entails taking the partial derivative of the Lagrangian Equation (32) with respect to p_k, where $k = 1, 2, \ldots, n$, then also taking the partial derivative of Equation (32)

$$\frac{\partial G_2}{\partial p_i} = X_i - \gamma_1 \frac{n\tilde{d}_i(s)}{p_i} - \gamma_2 = 0, \tag{34}$$

for $i = 1, 2, \ldots, n$. Therefore,

$$p_k = \frac{\gamma_1}{X_k - \gamma_2}. \tag{35}$$

For stratified sampling, the upper and lower confidence interval bounds are determined through an optimization problem similar to the nonstratified case. The only difference is that instead of the constraint given in Equation (31), the constraint in the stratified case is:

$$n\sum_{h=1}^{L} W_h \sum_{i\in s_h} \tilde{d}_{hi}(s_h)\log(p_{hi}) = \log(t_0), \tag{36}$$

where n represents the stratum size for stratum h, L is the strata, W_h is the stratum weight which is obtained by dividing the stratum size by the population size, $\tilde{d}_{hi}(s_h)$ are the normalized design weights within stratum h, s_h is the set of sample points in stratum h, and p_{hi} is the probability mass placed on the sample points in stratum h. Equation (36) is the constraint that represents the pseudo–empirical log–likelihood for the stratified unistage design.

8 Conclusion

Standard methods for empirical likelihood typically use data from a simple random sample of the population. However, complex sample survey data require specialized procedures designed for such data. Standard statistical software procedures do not allow analysts to take properties of survey data into account. A failure to use more specialized procedures designed for survey data analysis can impact both point and interval estimation of parameter s. We derive pseudo empirical likelihood confidence intervals for stratified single–stage and stratified multistage sampling designs.

References:

[1] G. Casella & R. L. Berger, (2002), *Statistical Inference*, Duxbury.

[2] 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.

[3] H. O. Hartley and J. N. K. Rao (1968), *A new estimation theory for sample survey*, *Biometrika*, **55**(**3**), 547 – 557.

[4] J. R. Landis, J. M. Lepkowski, S. A. Eklund, S. A. Stehouwer, (1982), a statistical methodology for analyzing data from a complex survey, the First National Health and Nutrition Examination Survey, National Center for Health Statistics, *Vital Health Statistics*, **2**(**92**).

[5] D. J. Nordman, (2008), A blockwise empirical likelihood for spatial lattice data, *Satistica Sinica* **18**: 1111 – 1129.

[6] A. Owen (1988), Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, **75**, 237 – 49.

[7] A. Owen (2001), *Empirical Likelihood*, Chapman and Hall / CRC Press.

[8] C. Wu and J. N. K. Rao (2006), Pseudo–empirical likelihood ratio confidence intervals for complex surveys, *The Canadian Journal of Statistics*, **34**(3), 359 – 375.