7-15-2013

Jackknife Empirical Likelihood Inference for the Absolute Mean Deviation

xueping meng

JACKKNIFE EMPIRICAL LIKELIHOOD INFERENCE FOR THE ABSOLUTE MEAN DEVIATION

by

XUEPING MENG

Under the Direction of Dr. Yichuan Zhao

ABSTRACT

In statistics it is of interest to find a better interval estimator of the absolute mean deviation. In this thesis, we focus on using the jackknife, the adjusted and the extended jackknife empirical likelihood methods to construct confidence intervals for the mean absolute deviation θ of a random variable. The empirical log-likelihood ratio statistic is derived whose asymptotic distribution is a standard chi-square distribution. The results of simulation study show the comparison of the average length and coverage probability by using jackknife empirical likelihood methods and normal approximation method. The proposed adjusted and extended jackknife empirical likelihood methods perform better than other methods for symmetric and skewed distributions. We use real data sets to illustrate the proposed jackknife empirical likelihood methods.

INDEX WORDS: Confidence interval, Coverage probability, Jackknife empirical likelihood, Adjusted jackknife empirical likelihood, Extended jackknife empirical likelihood.

JACKKNIFE EMPIRICAL LIKELIHOOD INFERENCE FOR THE ABSOLUTE MEAN DEVIATION

by

XUEPING MENG

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2013

JACKKNIFE EMPIRICAL LIKELIHOOD INFERENCE FOR THE ABSOLUTE MEAN DEVIATION

by

XUEPING MENG

Committee Chair: Dr. Yichuan Zhao

Committee: Dr. Yi Jiang

Dr. Remus Osan

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2013

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Yichuan Zhao for the persistent support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me all the time in researches and life. He always encourages me to insist when I have difficulties and intent to give up. Without his strong support, I have no chance to finish this thesis. And the working attitude I learnt from him will be my priceless wealth in my future life.

I also would like to thank other thesis committee: Dr. Yi Jiang, Dr. Remus Osan for their encouragement, insightful comments, and helpful questions.

My sincere thanks also go to my friends Xue Yu and Jing Wang, for enlightening me the first glance of research. I could not forget Jing's help to understand many basic concepts and Xue's effort to check error of my codes. The time we worked and studied together in GSU will be my precious memories.

Finally, I would like to thank my family: my parents, husband and lovely daughter Erin Li, for supporting me spiritually throughout my life. I love you all forever.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In this chapter, we will introduce some basic concept and methods we used in the thesis research. In a data set, for an element, the absolute mean deviation is the difference between that element and a given point [see wikipedia]. Large sample theory is introduced by the central limit theory. We elaborate the concept of absolute mean deviation which is the main research target of this thesis. Also we need to use Newton-Raphson and bisection methods when we solve out the key step nonlinear equations. In addition to the empirical likelihood method, we also introduce several jackknife empirical likelihood related methods to compare with the normal approximation based method in terms of coverage probability and average length of confidence intervals.

1.1 Central limit theory

In probability theory, the central limit theorem (CLT) states that, given certain conditions, the mean of a sufficiently large number of independent random variables, each with a well-defined mean and well-defined variance, will be approximately normally distributed.

Let $\{X_1, ..., X_n\}$ be a random sample of size n—that is, a sequence of independent and identically distributed (iid) random variables drawn from distributions of expected values given by μ and finite variance given by σ^2. Suppose we are interested in the sample mean $S_n = \frac{X_1+X_2+...+X_n}{n}$ of these random variables. By the law of large numbers, the sample means converge in probability to μ as n goes to infinity. The classical central limit theorem describes the distributional form of the stochastic fluctuations around μ during this convergence.

1.2 Large sample theory

Large sample theory (LST), also called asymptotic theory, is used to approximate the distribution of an estimator when the sample size is large. This theory demonstrates the advantage if the sampling distribution of the estimator is complicated or unknown. Before using this theory, one must determine which estimator is used. The rate of convergence, asymptotic distribution, and sample size must be assumed for the approximation. Moreover, if the estimator is to be useful for inference, the asymptotic standard error (SE), an estimator of the asymptotic standard deviation, must be computable.

We can express the idea as follow: If *F(X)* is a cumulative distribution function, and $X_1, X_2 \ldots X_n$ form a sequence of independent identically distributed (iid) random variables with the mean μ and variance σ^2. One measure of "spread "of a cumulative distribution function, *F(x)* is the absolute mean deviation proposed by Gastwirth (1974) as follows:

$$\theta = \int_{-\infty}^{\infty} |\mathrm{x}-\mu| \, \mathrm{dF(x)} = E\left|\mathrm{X} - \mathrm{E(X)}\right|. \tag{1.1}$$

Since we have [see Gastwirth (1974)] :

$$\hat{\theta} = \mathrm{n}^{-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} \left|\mathrm{X_i}-\overline{\mathrm{X}}\right| = \mathrm{n}^{-1} \sum_{\mathrm{X_i}<\overline{\mathrm{X}}} \left|\overline{\mathrm{X}}-\mathrm{X_i}\right| + \mathrm{n}^{-1} \sum_{\mathrm{X_i}>\overline{\mathrm{X}}} \left|\overline{\mathrm{X}}-\mathrm{X_i}\right|, \tag{1.2}$$

where $\overline{\mathrm{X}} = \mathrm{n}^{-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}$. Using the same methods as those in Gastwirth (1974), denote N as the number of the observations which are less than $\overline{\mathrm{X}}$, $\mathrm{X_i} < \overline{\mathrm{X}}$. Gastwirth (1974) shows:

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \left|\overline{\mathrm{X}}-\mathrm{X_i}\right| = 2[\mathrm{N}\overline{\mathrm{X}} - \sum_{\mathrm{X_i}<\overline{\mathrm{X}}} \mathrm{X_i}]. \tag{1.3}$$

We have the following result given by Gastwirth (1974):

$$\mathrm{v}^2 = 4 * \left\{ \mathrm{p}^2 \int_{\mu}^{\infty} (\mathrm{x}-\mu)^2 \mathrm{dF(x)} + (1-\mathrm{p})^2 \int_{-\infty}^{\mu} (\mathrm{x}-\mu)^2 \mathrm{dF}\,(\mathrm{x}) - \frac{\theta^2}{4} \right\}, \tag{1.4}$$

and

$$\mathrm{p} = \mathrm{F}(\mu). \tag{1.5}$$

The asymptotic normality of $\hat{\theta}$ is given by Gastwirth (1974).

$$\mathrm{n}^{\frac{1}{2}}\left(\hat{\theta}-\theta\right) \rightarrow \mathrm{N}\left(0, \mathrm{v}^{2}\right). \tag{1.6}$$

Then we construct a $100(1-\alpha)\%$ normal approximation based confidence interval for θ:

$$R=\left\{\theta : \hat{\theta} \pm Z_{a/2} * v/\sqrt{n}\right\}. \tag{1.7}$$

1.3 Empirical likelihood

Empirical likelihood (EL) method was first introduced by Owen (1988). It is the method to construct confidence regions for the mean of a random vector. This nonparametric inference method is based on a data-driven likelihood ratio function, rather than an assumption that the entire data from a known distribution. The empirical likelihood can be thought of as a bootstrap that does not resample, and as "likelihood without parametric assumptions " [see Owen (2001)]. It also has better asymptotic power properties and small sample performance compared to other methods.

Since Owen (1988) derived the asymptotic χ^2 distribution of empirical likelihood ratio statistic for the mean μ, there have been many important contributions to the development of the EL method in mainstream statistics. This is an evidence from Owen (2001) on empirical likelihood. Among other results, Qin and Lawless (1994), which showed that side information in the form of a set of estimating equations can be used to improve the maximum EL estimators and the EL ratio confidence intervals, is particularly appealing for inference from survey data in the presence of auxiliary information. Hall (1990) and DiCiccio et al. (1991) have developed the empirical likelihood regions. Qin and Lawless (1994) proposed an empirical likelihood for a parameter solved by general estimating equations, which established the Wilks theorem. Ren (2008) and Keziou and Leoni-Aubin (2008) worked on the two-sample problem. Recent censored linear regression models have been extensively discussed by Zhao (2011) and Zhou and Li (2008), etc. And Tsao (2013) proposed the extended

empirical likelihood for general estimating equations.

Empirical likelihood has been widely utilized in many settings, when data subjects to constraints are linear. However, there exist a lot of computational difficulties when applied to complicated statistics, such as nonlinear functional.

To overcome the computational difficulties, a modified empirical likelihood method was proposed by Jing et al. (2009) and Wang (2010), which was called jackknife empirical likelihood (JEL). This method combines two of the popular nonparametric approaches: the jackknife and the empirical likelihood. The main idea of the JEL is to " turn the statistic of interest into a sample mean based on jackknife pseudo-values" [see Quenouille (1956)]. If we can prove that these pseudo-values are asymptotically independent, Owen's [see Owen (1988), Owen (1990)] empirical likelihood should be applied for the mean of the jackknife pseudo-values.

As a new approach, jackknife empirical likelihood method has the most brilliant feature - simplicity, and it is a simple application of empirical likelihood to simplify the computation to complicated statistics. Also, some other new methods from jackknife empirical likelihood method having better performance in terms of coverage probability and average length, are adjusted jackknife empirical likelihood proposed by Chen et al. (2008) and extended jackknife empirical likelihood proposed by Tsao (2013). Our main contribution in this thesis is to develop new jackknife empirical likelihood methods for the absolute mean deviation to achieve better small sample performance.

1.4 Structure

We develop the jackknife empirical likelihood (JEL), adjusted jackknife empirical likelihood (AJEL), extended jackknife empirical likelihood (EJEL) method for the absolute mean deviation in chapter 2.

In chapter 3, we will report that the results of simulation studies on the finite sample performance in terms of coverage probability, average length of standard method, jackknife

empirical likelihood, adjusted jackknife empirical likelihood and extended jackknife empirical likelihood based confidence interval on the absolute mean deviation θ. We will further apply these methods to four real data with different sample sizes to check the performance in chapter 4. In chapter 5, we make the conclusion, and propose some ideas for the future work.

CHAPTER 2

INFERENCE PROCEDURE

2.1 Jackknife empirical likelihood

We plug the estimator (1.3) in JEL method:

$$\hat{\theta}_{\mathrm{n-1}}^{(-\mathrm{i})}=\frac{1}{(\mathrm{n}-1)}\sum_{\mathrm{j}\neq\mathrm{i}}^{\mathrm{n}}\left|\mathrm{X_j}-\bar{X}_{(-i)}\right|, \tag{2.1}$$

where $\bar{\mathrm{X}}_{(-\mathrm{i})}=\frac{1}{\mathrm{n-1}}\sum_{\mathrm{j=i}}^{\mathrm{n}}\mathrm{X_j}$.

This equation means that we estimate the estimator by removing the i-th item, where $\hat{\theta}_{\mathrm{n-1}}^{(-\mathrm{i})}=\hat{\theta}(\mathrm{X}_1, \mathrm{X}_2, \ldots \mathrm{X}_{i-1}, \mathrm{X}_{i+1}, \ldots \mathrm{X}_n)$. We define our jackknife pseudo-values by:

$$\widehat{\mathrm{V}}_{\mathrm{i}}=\mathrm{n}\hat{\theta}-(\mathrm{n}-1)\,\hat{\theta}_{\mathrm{n-1}}^{(-\mathrm{i})}. \tag{2.2}$$

The jackknife estimator of θ is the average of preudo-values defined as follows:

$$\widehat{\hat{\theta}_{\mathrm{n,jack}}}=\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\widehat{\mathrm{V}}_{\mathrm{i}}. \tag{2.3}$$

Since we know that Owen's empirical likelihood is particularly easy to apply for the sample mean, we will proceed as follows: Let $\mathrm{P}=(\mathrm{P}_1, \mathrm{P}_2, \ldots \mathrm{P}_{\mathrm{n}})$ be the probability vector, we have that: $\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{P_i}=1$ and $\mathrm{P_i}>0$ for $1\leq \mathrm{i}\leq \mathrm{n}$. Then following Owen (1988, 1990) and Qin and Lawless (1994), we have

$$\mathrm{L}(\theta)=max\left\{\prod_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{P_i}:\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{P_i}=1,\ \sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{P_i}(\hat{V}_i-\theta)=0\ ,\ \ P_i>0\right\}. \tag{2.4}$$

Note: $\prod_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{P_i}$ reaches its maximum when $P_i=1/n$. Next, we can define the jackknife empirical likelihood ratio at θ by:

$$\mathrm{R}\left(\theta\right)=\max\left\{\prod_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{nP_i}\right):\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{P_i}=1,\ \sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{P_i}(\hat{V}_i-\theta)=0\ ,\quad P_i>0\right\}. \tag{2.5}$$

and

$$\mathrm{logR}(\theta)=\max\left\{\prod_{\mathrm{i}=1}^{\mathrm{n}}\log(\mathrm{nP_i}):\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{P_i}=1,\ \sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{P_i}(\hat{V}_i-\theta)=0\ ,\quad P_i>0\right\}. \tag{2.6}$$

By using the Lagrange multipliers method, we have:

$$\mathrm{P_i}=\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\widehat{\mathrm{V}}_\mathrm{i}-\theta}{1+\lambda(\widehat{\mathrm{V}}_\mathrm{i}-\theta)}, \tag{2.7}$$

where λ satisfies:

$$\mathrm{f}\left(\lambda\right)=\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\widehat{\mathrm{V}}_\mathrm{i}-\theta}{1+\lambda(\widehat{\mathrm{V}}_\mathrm{i}-\theta)}=0. \tag{2.8}$$

Next, we plug equation (2.7) into equation (2.5), then take log:

$$\log\ \mathrm{R}(\theta)=-\sum_{\mathrm{i}=1}^{\mathrm{n}}\log\left\{1+\lambda\left(\widehat{\mathrm{V}}_\mathrm{i}-\theta\right)\right\}. \tag{2.9}$$

We plug equation (2.7) into equation (2.6), we have:

$$\mathrm{l}\left(\theta\right)=-2\mathrm{logR}\left(\theta\right)=-2\sum_{\mathrm{i}=1}^{\mathrm{n}}\log\left(\mathrm{nP_i}\right)\ . \tag{2.10}$$

Thus, we establish the following Wilk's theorem and show how the result can be used to construct confidence interval for θ. Let θ_0 be the true value of θ.

Theorem 1: Under the above conditions, $\mathrm{l}(\theta_0)$ converges in distribution to χ^2, where χ^2 is a chi-square random variable with 1 degree of freedom.

An asymptotic 100(1-α) % JEL confidence interval can be constructed with the above theorem:

$$\mathrm{R_c}=\{\theta\colon -2\mathrm{logR}\,(\theta)\leq \mathrm{C}\}\,, \tag{2.11}$$

where C is chosen to satisfy $\mathrm{P}\,(\chi^2\leq \mathrm{C})=1-\alpha$.

2.2 Adjusted jackknife empirical likelihood

When the sample size is not significantly large, the coverage probability could be deviated significantly from the corresponding nominal level. Chen et al. (2008) developed an adjusted empirical likelihood method. This method significantly improves the performance of the empirical likelihood method. They showed that " the first-order asymptotic properties of the adjusted empirical likelihood remains the same while the error of coverage probability could be reduced significantly when the sample size is small for the first-order asymptotic properties of the adjusted empirical likelihood under the population mean case"[see Chen et al. (2008)].

Moreover, this method could efficiently avoid convex hull restriction and guarantees a sensible value of the empirical likelihood when the parameter value varies. So it will be very easy for the algorithm of the standard empirical likelihood to be extended to the adjusted method. We adapt their approach to the jackknife empirical likelihood for θ.

The adjusted jackknife empirical likelihood function for fixed θ is defined to be as Chen et al. (2008) did:

$$\mathrm{L}\,(\theta)=\max\{\ \prod_{\mathrm{i=1}}^{\mathrm{n+1}}\mathrm{p_i},\ \text{subject to}\ \sum_{\mathrm{i=1}}^{\mathrm{n+1}}\mathrm{p_i g_i^{ad}}\,(\theta)=0,\quad \sum_{\mathrm{i=1}}^{\mathrm{n+1}}\mathrm{p_i}=1,\ \ P_i>0\}, \tag{2.12}$$

where: $\mathrm{i}=1,\ 2,\ \ldots,\mathrm{n}$, $\mathrm{g_i^{ad}}\,(\theta)=\left(\widehat{V}_i-\theta\right)$, and $\mathrm{g_{n+1}^{ad}}\,(\theta)=-\mathrm{a_n}\overline{\mathrm{g}}_{\mathrm{n}}(\theta)$. Here $a_n=max(1,\ log(n)/2)$, which is recommended by Chen et al. (2008),

$$\overline{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta). \tag{2.13}$$

The resulting adjusted jackknife empirical likelihood is:

$$R^{ad}(\theta) = \prod_{i=1}^{n+1} \left\{(n+1)p_i^{ad}(\theta)\right\}, \tag{2.14}$$

where

$$p_i^{ad}(\theta) = \frac{1}{n+1}\frac{1}{1+\lambda g_i^{ad}(\theta)}, \tag{2.15}$$

i= 1, 2, 3, …, n+1, and λ satisfies

$$f(\lambda) = \sum_{i=1}^{n+1} \frac{g_i^{ad}(\theta)}{1+\lambda g_i^{ad}(\theta)} = 0. \tag{2.16}$$

Next, we plug equation (2.15) into equation (2.14), then take log:

$$\log R^{ad}(\theta) = -\sum_{i=1}^{n+1} \log(1+\lambda g_i^{ad}(\theta)). \tag{2.17}$$

For the adjusted jackknife empirical likelihood method, we can combine Chen et al. (2008) and the Jing et al. (2009) to get the following Wilk's theorem.

Theorem 2: Under the above conditions, $-2\log R^{ad}(\theta_0)$ converges in distribution to χ^2.

For the adjusted jackknife empirical likelihood (AJEL) method, an asymptotic 100(1-α) % confidence interval for θ can be constructed with the above theorem:

$$R^{ad} = \left\{\theta: -2\log R^{ad}(\theta) \le C\right\}, \tag{2.18}$$

where C is chosen by $P(\chi^2 \le C) = 1-\alpha$.

2.3 Extended jackknife empirical likelihood

Tsao and Wu (2013) proposed a new empirical likelihood by extending the empirical likelihood domain expansion. They extended the empirical likelihood beyond its domain Θ_n by expanding its contours nested inside the domain with a similarity transformation. At the same time, the extended empirical likelihood achieves two objectives. The first one is to escape the "convex hull constrain" on the empirical likelihood. The second one is to improve the coverage accuracy of the empirical likelihood ratio confidence region to $O(1/\mathrm{n}^2)$ [see Tsao and Wu (2013)].

The extended empirical likelihood [see Tsao and Wu (2013)] achieved the second objective through a special transformation. The extended EL confidence region not only retains the shape of the EL confidence region but also works efficiently on the small sample size [see Tsao (2013) and Tsao and Wu (2013)].

Following Tsao and Wu (2013), we define $\mathrm{h}_\mathrm{n}^\mathrm{c}$ by using JEL $l(\theta)$ at θ:

$$\mathrm{h}_\mathrm{n}^\mathrm{c}(\theta) = \hat{\theta} + \gamma(\mathrm{n}, \mathrm{l}(\theta))(\theta - \hat{\theta}), \tag{2.19}$$

where $\gamma(\mathrm{n}, \mathrm{l}(\theta))$ is the expansion factor given as

$$\gamma(\mathrm{n}, \mathrm{l}(\theta)) = 1 + \frac{\mathrm{l}(\theta)}{2\mathrm{n}}. \tag{2.20}$$

Applying the method of Lagrange multipliers, we have extended jackknife EL ratio as follows:

$$\begin{aligned} -2logR^{ed}(\theta) &= l\left\{\hat{\theta} + \left(1 + \frac{\mathrm{l}(\theta)}{2\mathrm{n}}\right)\left(\theta - \hat{\theta}\right)\right\} \qquad (2.21) \\ &= 2\sum_{\mathrm{i}=1}^{\mathrm{n}} \log\left\{1 + \left(1 + \frac{\mathrm{l}(\theta)}{2\mathrm{n}}\right)\left(\theta - \hat{\theta}\right)\right\} \end{aligned}$$

.

Theorem 3: Under the above conditions, $-2\log R^{ad}(\theta_0)$ converges in distribution to χ^2.

Thus, the 100(1-α)% EJEL confidence interval for θ is

$$R^{ed} = \left\{\theta: -2\log R^{ed}(\theta) \leq C\right\}, \tag{2.23}$$

where C is defined as before.

CHAPTER 3

NUMERICAL STUDIES

Based on the result in the proposed inference procedure, two groups of simulation studies are conducted to explore the performance of standard normal approximation method, JEL, adjusted JEL and extended JEL methods for the absolute mean deviation with different sample sizes. In this chapter, we simulate the data from normal distribution and exponential distribution, then compare the performance of 4 different methods.

In the simulation study, we simulate a group of data with standard normal distribution with mean 0 and standard deviation 1. Also we simulate another group of data with exponential distribution with λ =1.

For the standard normal approximation (NA) based method, in order to find the cumulative p, we compare it with the mean for each individual X. If the X is greater than $\overline{\mathrm{X}}$, we count once. Then we plug p into the variance function to find each of the upper and lower bounds. We compare the true value of the absolute mean deviation with the upper and lower bounds. If it is inside, we count once.

For the JEL method, we need to check whether *-2logR(θ_0)* is less or equal to $\chi_1^2(\alpha)$ to calculate the coverage probability. Here we choose α to be 0.1, 0.05 and 0.01 to define three nominal levels 90%, 95% and 99%. For example, we check if *-2logR(θ_0)* $\leq 1.96^2$, when α is 0.05. The length of the confidence interval is also very important because the shorter length means higher accuracy in obtaining the true value of an absolute mean deviation. We choose our sample size from smaller to larger: 30, 50, 100, 200 and 300. For each different sample size, the repetition is 5000 times.

3.1 Simulation for normal distribution

We simulate a group of data in the standard normal distribution, with mean μ=0 and standard deviation σ=1. From Table 3.1, we can find the following results.

The coverage probability is not satisfied on the normal approximation methods (NA) when the sample size is small and moderate. However, JEL, AJEL and EJEL have much better coverage probability for the same sample size. For example, when the sample size n=30, and nominal level =95%, coverage probability of NA method is 81.64%, coverage probability of JEL method is 93.02%, coverage probability of AJEL method is 94.62% and coverage probability of EJEL method is 95.12%. From the results, we can see AJEL and EJEL are very close to nominal level 95%. Thus we say AJEL and EJEL have better performance than JEL on the small sample size.

When the sample size is large, NA, JEL, AJEL and EJEL have similar performance in terms of coverage probability. For example when the sample size n=300, and nominal level =95%, coverage probability of NA method is 93.82%, coverage probability of JEL confidence interval is 95.12%, coverage probability of AJEL confidence interval is 95.40% and coverage probability of EJEL confidence interval is 95.20%.

For all the methods, the length of confidence interval becomes shorter when the sample size becomes larger. When the sample size is from moderate to large, the length of confidence interval for all the methods is very close. When the sample size is smaller, the length of the NA method is slight shorter than other three methods due to serious under coverage problem for the NA method.

Table 3.1 : Coverage probability under normal distribution

n	$Nominal$ $Level$	NA	JEL	$AJEL$	$EJEL$
30	99%	85.18%	98.16%	98.64%	99.64%
	95%	81.64%	93.02%	94.62%	95.12%
	90%	79.02%	87.42%	89.74%	89.48%
50	99%	90.44%	98.72%	98.98%	99.98%
	95%	86.44%	94.28%	95.60%	95.10%
	90%	83.06%	88.84%	90.44%	89.60%
100	99%	94.70%	98.72%	99.14%	98.78%
	95%	90.48%	94.78%	95.12%	94.94%
	90%	86.12%	89.80%	90.54%	90.20%
200	99%	96.88%	99.02%	99.10%	99.06%
	95%	93.00%	95.28%	95.80%	95.32%
	90%	88.74%	90.10%	90.56%	90.22%
300	99%	97.56%	99.00%	99.04%	99.04%
	95%	93.82%	95.12%	95.40%	95.20%
	90%	88.88%	90.60%	91.24%	90.70%

Note:

NA: Normal approximation method

JEL: Jackknife empirical likelihood

AJEL: Adjusted Jackknife empirical likelihood

EJEL: Extended Jackknife empirical likelihood

Table 3.2 : Average length under normal distribution

n	$Nominal$ $Level$	NA	JEL	$AJEL$	$EJEL$
30	99%	0.507	0.583	0.585	0.584
	95%	0.386	0.446	0.442	0.448
	90%	0.324	0.367	0.366	0.371
50	99%	0.411	0.456	0.446	0.447
	95%	0.313	0.340	0.340	0.339
	90%	0.263	0.281	0.283	0.283
100	99%	0.303	0.316	0.313	0.316
	95%	0.230	0.239	0.237	0.238
	90%	0.193	0.200	0.198	0.199
200	99%	0.217	0.222	0.222	0.221
	95%	0.165	0.169	0.168	0.168
	90%	0.138	0.141	0.140	0.141
300	99%	0.178	0.181	0.180	0.181
	95%	0.135	0.137	0.137	0.137
	90%	0.114	0.115	0.115	0.115

Note:

NA: Normal approximation method

JEL: Jackknife empirical likelihood

AJEL: Adjusted Jackknife empirical likelihood

EJEL: Extended Jackknife empirical likelihood

3.2 Simulation for exponential distribution

We simulate a group of data in exponential distribution, with λ=1. From Table 3.2, we have the following findings:

The coverage probability is not satisfied with the NA method at all sample sizes we tried. However, JEL, AJEL and EJEL methods have much better coverage probability even when the sample size is small. Also, AJEL and EJEL methods are slightly better than JEL method when we compare these three methods on small sample size. For example, when the sample size n=30, and the nominal level =95%, coverage probability of NA confidence interval is 69.68%, coverage probability of JEL confidence interval is 92.00%, coverage probability of AJEL confidence interval is 94.30% and coverage probability of EJEL confidence interval is 95.04%. From the simulation results, we can see that the AJEL and EJEL methods are very close to the nominal level 95%. Thus we can say that AJEL and EJEL method have better performance than JEL and NA methods for the small sample sizes.

When the sample size is large, JEL, AJEL and EJEL methods have similar performance in terms of coverage probability. For example when the sample size n=300, and nominal level =95%, the coverage probability of NA confidence interval is 84.88%, the coverage probability of JEL confidence interval is 93.64%, the coverage probability of AJEL confidence interval is 93.98% and the coverage probability of EJEL confidence interval is 93.72%. We can see all the jackknife methods have better coverage probability than the normal approximation method. The results are close to our expectation for all sample sizes.

For all the methods, the length becomes shorter when the sample size becomes larger. When the sample size changes from smaller to larger, the length of JEL, AJEL and EJEL methods are very close and the length converges faster than normal distribution. When the sample size is small, the length of NA method is shorter than other methods, but when the sample size is large, the length of NA method is longer than other methods.

Table 3.3 : Coverage probability under exponential distribution

n	$Nominal$ $Level$	NA	JEL	$AJEL$	$EJEL$
30	99%	73.52%	98.20%	99.50%	98.94%
	95%	69.68%	92.00%	94.30%	95.04%
	90%	66.76%	86.00%	89.20%	87.82%
50	99%	80.60%	97.46%	98.20%	98.70%
	95%	76.46%	92.44%	93.84%	93.58%
	90%	72.18%	87.08%	88.98%	87.08%
100	99%	85.62%	98.00%	98.44%	98.42%
	95%	80.54%	93.02%	94.08%	93.14%
	90%	75.12%	87.60%	88.60%	87.52%
200	99%	90.94%	98.34%	98.64%	98.40%
	95%	84.56%	94.16%	94.82%	94.22%
	90%	78.26%	88.98%	89.84%	88.98%
300	99%	91.76%	98.46%	98.70%	98.42%
	95%	84.88%	93.64%	93.98%	93.72%
	90%	78.18%	87.88%	88.70%	88.06%

Note:

NA: Normal approximation method

JEL: Jackknife empirical likelihood

AJEL: Adjusted Jackknife empirical likelihood

EJEL: Extended Jackknife empirical likelihood

Table 3.4 : Average length under exponential distribution

n	$Nominal$ $Level$	NA	JEL	$AJEL$	$EJEL$
30	99%	0.447	0.783	0.787	0.770
	95%	0.333	0.619	0.565	0.595
	90%	0.362	0.502	0.491	0.490
50	99%	0.343	0.635	0.617	0.624
	95%	0.260	0.488	0.485	0.474
	90%	0.221	0.395	0.398	0.392
100	99%	0.342	0.470	0.466	0.469
	95%	0.273	0.350	0.349	0.345
	90%	0.220	0.291	0.289	0.290
200	99%	0.343	0.331	0.330	0.332
	95%	0.262	0.247	0.247	0.246
	90%	0.221	0.206	0.206	0.205
300	99%	0.342	0.269	0.269	0.268
	95%	0.262	0.202	0.202	0.201
	90%	0.225	0.169	0.168	0.168

Note:

NA: Normal approximation method

JEL: Jackknife empirical likelihood

AJEL: Adjusted Jackknife empirical likelihood

EJEL: Extended Jackknife empirical likelihood

CHAPTER 4

REAL DATA ANALYSIS

In this chapter, we studied four real data sets with the sample size small, moderate and large to illustrate the proposed methods in chapter 3.

The first data set named "pottery" has 26 observations and the second data set named "hotdogs" has 54 observations. These two data sets were obtained from the Data and Story Library (DASL) at Carnegie Mellon University. The third data set named "discoveries " has 114 observations and the last data set named "faithful " has 272 observations. These two data sets were obtained from R dataset package in R program.

In order to compare the results with the simulation study, we check the normality of each dataset, the normality test called Shapiro-Wilk test has been conducted. The null hypothesis of Shapiro-Wilk test is that sample data distribution is normal distribution. We check the p-value to reject or accept the null hypothesis. If the p-value is smaller than the nominal level, we reject null hypothesis which means the sample data is from non-normal distribution. Otherwise we can treat the sample data are from normal distribution. Thus we can compare the result with the normal distribution result.

4.1 Pottery data analysis

For the data set "pottery", 26 observations are 26 samples of Romano-British pottery which were found at four different kiln sites in Wales, Gwent and the New Forest. The 6 variables are the percentage of oxides of various metals measured by atomic absorption spectrophotometry. The data were collected in order to see if different sites contained pottery of different chemical compositions.

The size 26 is similar to small sample size that we simulated in chapter 3, therefore we can use this data set to illustrate the proposed methods. Among the 6 variables, we only choose one variable which is the percentage of aluminum oxide in sample to illustrate our methods.

We obtain the lower bound, upper bound and length by using the NA, JEL, AJEL and EJEL methods. From the results, we can see that the lengths of JEL, AJEL, EJEL methods are very close and are clearly longer than one of the normal approximation method.

After using the Shapiro-Wilk test, the calculated p-value is 0.09447 which has very weak evidence to support the data is from a normal distribution. Thus we need to check the histogram of the data and it shows the distribution is close to a exponential distribution. We also can see our result is coherent to the simulation result of exponential distribution.

Table 4.1 : Length of confidence intervals of pottery data set

Nominal		*NA*		*JEL*		*AJEL*		*EJEL*	
Level		UB	LB	UB	LB	UB	LB	UB	LB
		3.234	1.795	3.578	1.761	3.571	1.810	3.598	1.747
99%	Length	1.739		1.817		1.761		1.851	
		3.062	1.967	3.303	1.943	3.306	1.989	3.318	1.932
95%	Length	1.095		1.3560		1.317		1.386	
		2.974	2.055	3.171	2.038	3.179	2.083	3.183	2.030
90%	Length	0.919		1.133		1.096		1.153	

Note:

NA: Normal approximation method

JEL: Jackknife empirical likelihood

AJEL: Adjusted Jackknife empirical likelihood

EJEL: Extended Jackknife empirical likelihood

UB: Upper bound

LB: lower bound

4.2 Hot dogs data analysis

The data set "hot dogs" is about the results of a laboratory analysis of calories and sodium content of major hot dog brands. Researchers for Consumer Reports analyzed three types of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat).

There are 54 observations in this dataset which are similar to small sample size. We have two variables calories and sodium in this data set, but we only choose sodium in sample to analysis it.

Similar to the data set "pottery", we also find the lower bound, upper bound and length by using the NA, JEL, AJEL and EJEL methods. From the results, we can see the lengths of all the methods are almost the same.

Regarding to Shapiro-Wilk test, the calculated p-value is 0.4836. We fail to reject null hypothesis, which means this sample data is from a normal distribution. Also the histogram of the data is shown colse to a normal distribution. Our result is coherent to the simulation result with the normal distribution.

Table 4.2 : Length of confidence intervals of hot dogs data set

Nominal		*NA*		*JEL*		*AJEL*		*EJEL*	
Level		UB	LB	UB	LB	UB	LB	UB	LB
		96.815	58.469	99.392	60.101	100.476	61.365	99.283	60.189
99%	Length	38.346		39.291		39.111		39.094	
		92.230	63.053	93.411	63.941	94.605	65.238	93.331	64.011
95%	Length	29.177		29.470		29.367		29.320	
		89.885	65.399	90.535	65.956	91.773	67.268	90.470	66.015
90%	Length	24.486		24.579		24.505		24.455	

Note:

NA: Normal approximation method

JEL: Jackknife empirical likelihood

AJEL: Adjusted Jackknife empirical likelihood

EJEL: Extended Jackknife empirical likelihood

UB: Upper bound

LB: lower bound

4.3 Discoveries data analysis

The data set "discoveries" is a group of time series data. The number is the "great" inventions and scientific discoveries in each year from 1860 to 1959.

There are 114 observations in this dataset which are similar to moderate sample size. In this data set, we have only one variable which means the number of discoveries were found in each year.

Similar to the data set in sections 4.1 and 4.2, we also find the lower bound, upper bound and length by using the NA, JEL, AJEL and EJEL methods.

We can see the lengths of JEL, AJEL and EJEL are longer than one of the NA method from the results. Thus we can get a conclusion there is no big difference among the lengthes of JEL, AJEL and EJEL methods.

According to Shapiro-Wilk test, the calculated p-value is 0.000001524. Since the p-value is very small, so we reject null hypothesis, which means this sample data is from a non-normal distribution. Also the histogram of the data is shown very skewed. We can compare our result with the simulation result of exponential distribution. Under the same sample size, the length of NA method is slightly shorter than those of JEL, AJEL and EJEL methods. These two results are also coherent.

Table 4.3 : Length of confidence intervals of discoveries data set

Nominal		*NA*		*JEL*		*AJEL*		*EJEL*	
Level		UB	LB	UB	LB	UB	LB	UB	LB
		2.007	1.341	2.352	1.268	2.343	1.270	2.355	1.267
99%	Length	0.666		1.084		1.073		1.088	
		1.928	1.421	2.163	1.356	2.156	1.357	2.164	1.355
95%	Length	0.507		0.807		0.799		0.809	
		1.887	1.461	2.074	1.403	2.068	1.404	2.075	1.402
90%	Length	0.426		0.671		0.664		0.673	

Note:

NA: Normal approximation method

JEL: Jackknife empirical likelihood

AJEL: Adjusted Jackknife empirical likelihood

EJEL: Extended Jackknife empirical likelihood

UB: Upper bound

LB: lower bound

4.4 Faithful data analysis

The data set "faithful "is also from R dataset in R program. This data set has two variables. One is waiting time between eruptions and another one is the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

There are 272 observations in this dataset which are similar to large sample size. We choose the duration of the eruption as our parameter to study in this data set.

We also check the normality of this dataset by Shapiro-Wilk test. The calculated p-value is 9.036e-16 which is close to 0. We reject null hypothesis, and it means this sample data are from a non-normal distribution. The histogram of the data is shown skewed and it looks like a mixed normal distribution. Thus we can not compare the result with the simulation results of exponential distribution or normal distribution.

By each of the method, we also find the lower bound, upper bound and length at nominal level 99%, 95% and 90%.

We can see the lengths of JEL, AJEL and EJEL are almost same and are much longer than one of the NA method in the table. We should try the simulation study of mixed normal distribution and compare with this result in the future.

Table 4.4 : Length of confidence intervals of faithful data set

Nominal		*NA*		*JEL*		*AJEL*		*EJEL*	
Level		UB	LB	UB	LB	UB	LB	UB	LB
		1.106	0.979	1.148	0.949	1.150	0.951	1.148	0.949
99%	Length	0.127		0.199		0.199		0.199	
		1.090	0.994	1.123	0.972	1.125	0.974	1.123	0.972
95%	Length	0.096		0.151		0.151		0.151	
		1.083	1.002	1.110	0.984	1.113	0.986	1.111	0.983
90%	Length	0.081		0.126		0.127		0.128	

Note:

NA: Normal approximation method

JEL: Jackknife empirical likelihood

AJEL: Adjusted Jackknife empirical likelihood

EJEL: Extended Jackknife empirical likelihood

UB: Upper bound

LB: lower bound

CHAPTER 5

SUMMARY AND FUTURE WORK

5.1 Summary

In this thesis, we used three types of JEL methods to construct confidence interval for the absolute mean deviation.

According to the simulation study, we can easily conclude that JEL, AJEL and EJEL methods have much better performance than the standard normal approximation method in terms of coverage probability when the sample size is small. Lengths for all the JEL, AJEL and EJEL methods are very close. Therefore it is hard to say which method is better. Especially under exponential assumption, the coverage probability of the standard normal approximation method is far away from our expectation when the sample size is small. However, the coverage probability of JEL, AJEL and EJEL methods is very close to nominal level 95%.

For the real data analysis part, we calculated the interval lengths for each data set, but the result is not very satisfied since the length of all the JEL methods we used are similar, and we could not choose which one is better. We also check the normality of data, the dataset "pottery" and dataset "hotdog" follows a normal distribution, and the results are coherent with the simulation results of normal distribution. The datasets "discoveries"and "faithful" do not follow normal distribution. Thus we compare the result with exponential distribution simulation. These two results are also comparable.

Therefore, we conclude that JEL, AJEL and EJEL methods perform better than standard normal approximation based method in terms of coverage probability when the sample size is small. In practice, we recommend AJEL and EJEL methods. From computational issue, we find the AJEL method is easy and shares the very good small sample performance.

5.2 Future work

From the result, we can see the JEL and AJEL methods have very good performance no matter the sample size is small or large. Theoritically, the EJEL method should have better performance than other methods when the sample size is small [see Tsao and Wu (2013)]. However, the result of EJEL method is comparable with AJEL method when the smaller sample size is small. In order to overcome this drawback, we need to improve this method in the future.

In addition to normal and exponential distributions, we also can try simulation of other distributions, such as mixed normal distributions.

In addition, we also can try other empirical likelihood methods, such as bootstrap method to explore the accuracy by jackknife empirical likelihood method.

Bibliography

Chen, J., Variyath, A., and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *J Comput Graph Stat*, 17:426–443.

DiCiccio, T., Hall, P., and Romano, J. (1991). Empirical likelihood is bartlett correctable. *The Annals of Statistics*, 19:1053–1061.

Gastwirth, J. L. (1974). Large sample theory of some measures of income inequality. *Econometrica*, 42:191–196.

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *Annals of Statistics*, 75(2):121–140.

Jing, B., Yuan, Q., and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104(487):1224–1232.

Keziou, A. and Leoni-Aubin, S. (2008). On empirical likelihood for semi-parametric two sample density ratio models. *J. Statist. Plann and Inference*, 138:915–928.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika Trust*, 75(2):237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

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

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistis*, 22:300–325.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika*, 10:353–360.

Ren, J. (2008). Weighted empirical likelihood in some two-sample semi-parametric models with various types of censored data. *Annals of Statistics*, 36:147–166.

Tsao, M. (2013). Extending the empirical likelihood by domain expansion. *The Canadian Journal of Statistics*, 40:1–18.

Tsao, M. and Wu, F. (2013). Empirical likelihood on the full parameter space. *Annals of Statistics*.

Wang, X. (2010). Empirical likelihoood with applications. *Ph.D Thesis in National University of Singapore*.

Zhao, Y. (2011). Empirical likelihood inference for the accelerated failure time model. *Statistics and Probability Letters*, 81:603–610.

Zhou, M. and Li, G. (2008). Empirical likelihood analysis of the buckley-james estimator. *Journal of Multivariate Analysis*, 99:1069–1112.