Spring 5-1-2012

Confidence Interval Estimation for Coefficient of Variation

Shuang Liu

CONFIDENCE INTERVAL ESTIMATION FOR COEFFICIENT OF VARIATION

by

SHUANG LIU

Under the Direction of Dr. Gengsheng Qin

ABSTRACT

The coefficient of variation (CV) is a helpful quantity to describe the variation in evaluating results from different populations. There are many papers discussing methods of constructing confidence intervals for a single CV, such as exact method and approximation methods for CV when the underlying distribution is a normal distribution. However, the exact method is computationally cumbersome, and approximation methods can't be applied when the underlying distribution is unknown. In this thesis, we propose the generalized confidence interval for CV when the underlying distribution is normal and three empirical likelihood-based non-parametric intervals for CV when the underlying distribution is unknown. Simulation studies are conducted to compare the relative performances of these intervals based on the coverage probability and average interval length. Finally, the application of the proposed methods is demonstrated by using some real examples.

INDEX WORDS: Approximation method, Boostrapping, Coefficient of variation, Empirical likelihood, Generalized pivotal quantity, Jackknife

CONFIDENCE INTERVAL ESTIMATION FOR COEFFICIENT OF VARIATION

by

SHUANG LIU

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

Master of Science

in the College of Arts and Sciences

Georgia State University

2012

CONFIDENCE INTERVAL ESTIMATION FOR COEFFICIENT OF VARIATION

by

SHUANG LIU

Committee Chair: Dr. Gengsheng Qin

Committee: Dr. Xin Qi

Dr. Xu Zhang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2012

ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my advisor, Dr. Gengsheng Qin, for all his help and guidance. Without his support, advice and patience, this thesis would not have been possible to accomplish. I feel very lucky to be one of Professor Qin's student.

And I would like to thank other thesis committee members, Dr. Xin Qi and Dr. Xu Zhang, for taking their precious time out from their busiest and important moment and giving helpful suggestion to my thesis work. I would also like to show my gratitude to some of my classmates studying in similar field, Binhuan Wang, Haochuan Zhou and Aekyung Jung, for their unselfish help and support. In addition, I would like to thank those who have been traveling in the statistical world with me, my other classmates Hanfang Yang, Shan Luo, Hongwei Wang, Zhu Zi for their encouragement. And I also want thank Ruya Zhao who has been always supporting and understanding.

Lastly, I owe my deepest gratitude to my parents who have been supporting and encouraging me with all their hearts in all the ways. I couldn't have achieved any success without their love.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

The coefficient of variation (CV) is an important quantity to describe the variation. It provides an alternative index besides the most commonly used measurements of variation such as variance or standard deviation, which come across with difficulty in comparing the variations from different populations with different units. Take the data from students physical examination as an example. If one wants to analyze variations of both height and weight, it is not appropriate and logically incorrect to directly compare the variances or standard deviation, as the units from both populations are not matched. Another practical example in which commonly used measurements of variation might fail to work properly is to describe the relationship between the variation of average individual income and the variation of tax income of States. The different scale of data makes the direct comparison invalid or misleading. Nevertheless, With the property of standardization and unitlessness, CV makes the comparison possible. It is a good measure of the reliability of the experiment, which is, the smaller the CV values, the higher is the reliability of experiment (Gomez and Gomez, 1984; Steel and Torrie, 1980). Hence, CV is frequently provided by researchers, especially those in agricultural fields, in major publications. It is as well widely applied to describe the variation related to average in different engineering research.

Let X be a random variable with mean μ and variance σ^2. The population coefficient of variation is defined as follows:

$$k = \frac{\sigma}{\mu}$$

Assuming that observations X_i , i=1, 2, ..., n, are the independent identically distributed sample from N(μ ,σ) . The sample mean $\bar{X}$ and sample variance S^2 are the unbiased point estimates of μ and σ^2, respectively. An estimator of parameter k is

$$K = \frac{S}{\bar{X}},$$

where $\bar{X} = \frac{1}{n}\sum X_i$, and $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$.

1.2 Parametric inference on Coefficient of Variation

CV does not appear to be the top mentioned index to measure the variation. Therefore, not those like standard deviation or variance which has no secret of anything at all, there aren't many papers published about the inference of CV. However, the helpful application of the index in the agricultural research and engineering field still drew some attention from statisticians. And some authors have already done some impressive work for the inference of CV. When the underlying distribution of X is normally distributed, the sampling distribution

of CV is given by Hendricks and Robey (1936) as follows,

$$dF_{cv} = \frac{2}{\pi^{\frac{1}{2}}\Gamma(\frac{n-1}{2})} e^{-\frac{n}{2}\frac{cv^2}{\sigma^2}\frac{cv^2}{1+cv^2}} \frac{cv^{n-2}}{(1+cv^2)^{\frac{n}{2}}} \sum_{i=0}^{n-1} \frac{(n-1)!\Gamma(\frac{n-i}{2})}{(n-1-i)!i!} \frac{n^{\frac{i}{2}}}{2^{\frac{i}{2}}(\frac{\sigma}{\mu})^i} \frac{1}{(1+cv^2)^{\frac{i}{2}}} dcv.$$

Lehmann (1986) also derived the sample distribution of CV in order to give an exact method for the construction of a confidence interval for CV. Suppose X_i are identically and independently distributed from a normal distribution with mean μ and variance σ^2, then

$$\frac{\bar{X}}{\frac{S}{\sqrt{n}}} \sim NCT_{n-1}(\frac{\mu\sqrt{n}}{\sigma}),$$

where $NCT_{n-1}(\frac{\mu\sqrt{n}}{\sigma})$ denotes a noncentral t-distribution with n-1 degrees of freedom and non-centrality parameter $\frac{\mu\sqrt{n}}{\sigma}$. With this distribution function, the confidence interval of CV can be developed with standard procedure, yet with cumbersome calculation.

Routinely constructing an exact confidence interval for CV appears to be with difficulty thanks to the complexity of distribution function. Thus, statistician chose to sacrifice certain level of accuracy and proposed some simpler but relatively accurate approximation methods. McKay (1932) first published the approximation method for constructing confidence interval for CV under normal assumption. Based on the McKay's work, David (1949) proposed a modified approximation method. Since the most important part of the approximation

method is selecting an appropriate pivot quantity, he proved that by selecting a very simple or "naive" approximate pivot, confidence interval for CV can be still obtained with acceptable accuracy. Later, based on analysis of the distribution of a class of approximate pivotal quantities, Vangel (1996) modified McKay's method. He compared David's (1949) approximation with Mckay's (1932) and pointed out that the "naive" method resulted less accuracy than McKay's method. He then proposed his own approximation by extending McKay's (1932) method and obtained the result with satisfactory accuracy. The form of approximate pivotal quantities are presenting in the later chapter and the method of constructing confidence interval for CV is given in the same chapter too.

Besides all the approximate pivotal quantity methods, Barndorff-Nielsen (1986, 1991), Pierce and Peters (1992) and Reid (1996) proposed various likelihood based inference procedures which can be as well used to construct confidence interval for CV.

Let $\theta = (k, \mu)$, where k and μ are population CV and mean respectively, and let $l(\theta) = l(k, \mu)$ be the log likelihood function of θ. The Signed log-likelihood ratio statistics for k is

$$r(k) = sgnd(\hat{k} - k)2[l(\hat{\theta}) - l(\hat{\theta_k})]^{\frac{1}{2}},$$

where $\hat{\theta} = (\hat{k}, \hat{\mu})$ is overall maximum likelihood estimate of $\theta = (k, \mu)$ and $\hat{\theta_k} = (k, \hat{\mu_k})$ is a constrained maximum likelihood estimate of θ for a given k . Then it can be proved that $r(k)$ is asymptotically following the standard normal distribution. Thus, the confidence interval for k can be constructed. However, Pierce and Peters(1992) pointed out that this asymptotic method has accuracy of order $O(n^{-\frac{1}{2}})$ and performs unsatisfactorily with small

sample size. A modified Signed log-likelihood ratio statistic was proposed by Barndorff-Nielsen(1986,1991) as follows:

$$r^*(k) = r(k) + r(k)^{-1} log\{\frac{u(k)}{r(k)}\}.$$

And he also showed that r^* is asymptotically distributed with $N(0,1)$ with accuracy of $O(n^{-\frac{3}{2}})$. In order to obtain the $100(1-\alpha)\%$ confidence interval, a simplified signed log-likelihood ratio statistics is also given as

$$r(k) = sgnd(\hat{k}-k)\{2nlog(\frac{k\hat{\mu}_k}{\hat{k}\hat{\mu}}) + \frac{n}{k^2\hat{\mu}_k^2}[\hat{k}^2\hat{\mu}^2 + (\hat{\mu}-\hat{\mu}_k)^2] - n\}^{\frac{1}{2}},$$

where

$$\hat{\mu}_k = \frac{-\bar{y} + \sqrt{\bar{y}^2 + 4k^2(S^2+\bar{y}^2)}}{2k^2}, \quad \hat{\mu} = \bar{y}$$

and

$$u(k) = \sqrt{n}\{\frac{1}{2} + \frac{1}{2}(\frac{\hat{k}\hat{\mu}}{k\hat{\mu}_k})^2 - \frac{\hat{\mu}}{\hat{\mu}_k}\}(\frac{\hat{k}\hat{\mu}}{k\hat{\mu}_k})(1 + k^2 - \frac{\hat{\mu}}{2\hat{\mu}_k})^{-\frac{1}{2}}.$$

Hence, with all the formulas above, r^* can be easily obtained. Therefore, a $100(1-\alpha)\%$ confidence interval for k is

$$CI = \{k : |r^*(k)| \leq z_{\frac{\alpha}{2}}\}.$$

1.3 The purpose of the present study

The goal of this thesis is to propose three new methods of constructing confidence interval for CV. Comparison will be made with existing methods according to the coverage probability and length of confidence interval. Methodology for different interval estimates will be presented in detail in chapter 2 and adequate simulation work will be given to show the performance of different interval estimates in Chapter 3. In Chapter 4 conclusion of this thesis will be made with discussion and some future work will be raised.

CHAPTER 2

METHODOLOGY

2.1 Existing method

As mentioned in Chapter 1, the complication of distribution function for CV rises the difficulty of routinely constructing exact confidence interval for CV with standard process. In order to avoid the cumbersome calculation, many authors proposed relatively accurate approximation methods. McKay (1932) first proposed the approximation method by using the approximate pivotal quantity. And later David (1949) and Vangel (1996) gave the modified methods based on different pivot quantity selections. Since the sufficient part of the approximation method is selecting the appropriate pivot quantity, the detailed mathematical procedure of Vangel's (1996) work will be presented to illustrate the method.

Let Y_v be a random variable following the chi-square distribution with degree of freedom $v = n - 1$, and define $W_v = \frac{Y_v}{v}$. For any $\alpha \in (0, 1)$, let $\chi^2_{v,\alpha}$ denote the 100α percentile of the distribution of Y_v , then $t = \frac{\chi^2_{v,\alpha}}{v}$ is the corresponding quantile for W_v . Vangel (1996) defined the random variable as follows:

$$Q = \frac{K^2(1 + k^2)}{(1 + \theta K^2)k^2},$$

where $\theta = \theta(k, \alpha)$ is a known function, $K = \frac{S}{\bar{Y}}$ is the sample CV where S^2 and $\bar{Y}$ is the sample variance and sample mean respectively. Select a proper θ such that

$$Pr(Q < t) \approx Pr(W_v \leq t).$$

The selection of θ is the most important part for the approximation method. The approximation methods mentioned above basically were all trying to select an appropriate θ. They are relatively simple but maintain relatively satisfactory accuracy. McKay (1932) claims that the selection of $\theta = \frac{v}{v+1}$ would make a good approximation for Q. And David (1949) gave an even simpler selection of θ which set $\theta = 1$. Another "naive" approximation was proposed by setting $\theta = \frac{1}{t}$. Later Vangel (1996) proved that the distribution of W_v is known and is free of K, and he proposed a modified θ with the expression of

$$\theta = \frac{v}{v+1}[\frac{2}{\chi^2_{v,\alpha}} + 1].$$

In general, with the selected θ, the $100(1-\alpha)\%$ approximate confidence interval for k is given as follows by using the approximate pivot:

$$CI = (\frac{K}{\sqrt{t_1(\theta_1 K^2+1)-K^2}}, \frac{K}{\sqrt{t_2(\theta_2 K^2+1)-K^2}}),$$

where $t_1 = \frac{\chi^2_{v,1-\frac{\alpha}{2}}}{v}$, $t_2 = \frac{\chi^2_{v,\frac{\alpha}{2}}}{v}$, and $\theta_i = \frac{2}{(v+1)t_i} + \frac{v}{k+1},\ i = 1, 2$. So the McKay and Vangel's interval estimates are

$$CI_1 = \{K[(\frac{u_1}{v+1}-1)K^2 + \frac{u_1}{v}]^{-\frac{1}{2}}, K[(\frac{u_2}{v+1}-1)K^2 + \frac{u_2}{v}]^{-\frac{1}{2}}\},$$

and

$$CI_2 = \{K[(\frac{u_1+2}{v+1}-1)K^2 + \frac{u_1}{v}]^{-\frac{1}{2}}, K[(\frac{u_2}{v+1}-1)K^2 + \frac{u_2+2}{v}]^{-\frac{1}{2}}\},$$

respectively, where $u_i \equiv vt_i$, for $i = 1, 2$.

Although, by using the approximate pivot, one can obtain a confidence interval for desired index with very satisfactory accuracy for some true values of k, the McKay and Vangel's methods come across some inevitable problem. With some simple calculation, one can prove that the pivotal quantities may be a complex value when selecting a large k. That directly limited the application of these methods. It is clearly suggested by Mckay himself

that his methods could only be applied under the condition of $k < 0.33$ (McKay, 1932). The same drawback appears for Vangel's method too. These unavoidable problems inspired us for seeking some new methods introduced in next section.

2.2 New Methods

2.2.1 *Generalized Confidence Interval*

Inspiring by McKay(1932) and Vangel's(1996) approximation methods which introduced some pivotal quantities to construct the confidence interval, we are thinking about if it is possible to propose a different type of pivotal quantity. Since lots of works about Generalized Confidence Interval have been done based on Generalized Pivotal Quantity (GPQ) in the literature, in this section, we proposed a method to develop a GPQ-based confidence interval for CV. The definition of GPQ is as follows, cited from Weerahandi (1993):

Let X be an observable random vector with the cdf $F(x|v)$, where $v = (\theta, \delta)$ is a vector of unknown parameters, θ is the parameter of interest, and δ is a vector of nuisance parameters. Let χ be the sample space of possible values of X and let Θ be the parameter space of θ. An observation from X is denoted by x, where $x \in \chi$. Let $R = r(X; x, v)$ be a function of X, x, v (but not necessarily a function of all), where $v = (\theta, \delta)$. Then it is said to be a generalized pivotal quantity if property A: R has a probability distribution free of unknown parameters. Property B: The observed pivotal, defined as $r_{obs} = r(x; x, v)$, does not depend on the nuisance parameter δ .

Then a two-sided $100(1 - \alpha/2)\%$ confidence interval for parameter θ is $(R_{\alpha/2}, R_{1-\alpha/2})$, where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $100(\alpha/2) - th$ percentile and $100(1 - \alpha/2) - th$ percentile of

distribution R respectively. More detailed introduction of GPQ-based confidence interval is referred to Weerahandi (1993) and Hanning et al. (2006).

Now let observations $X_1, ..., X_m$ be the random sample from a normal distribution with mean μ and standard deviation σ . Then the sufficient estimator for (μ, σ) is

$$(\hat{\mu}, \hat{\sigma}) = (\bar{X}, S)$$

where $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$, and $S^2 = \frac{\sum_{i=1}^{m}(X_i - \bar{X})}{m-1}$.

Under normality assumption, the parameter of interest D is a function of parameters (μ, σ) which is

$$(\mu, \sigma^2, k) = (\mu, \sigma^2, \frac{\sigma}{\mu}).$$

So the parameter can be estimated by

$$(\hat{\mu}, \hat{\sigma}^2, \hat{k}) = (\bar{X}, S^2, \frac{S}{\bar{X}})$$

Let U ,Z be the quantities defined as follow:

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

$$Z = (\frac{\sigma^2}{n})^{-1/2}(\bar{X} - \mu) \sim N(0, 1).$$

And we can easily verify that both properties A and B are satisfied in these quantities. Therefore, defined quantities are the pivotal quantities corresponding to estimators.

Let $\bar{x}$ and s^2 be the observation of $\bar{X}$, S^2 respectively. Then the GPQ of σ^2,and μ is in expression of

$$R_{\sigma^2} = \frac{(n-1)s^2}{U}, \tag{2.1}$$

and

$$R_\mu = \bar{x} - (\frac{R_{\sigma^2}}{n})^{1/2}Z. \tag{2.2}$$

Based on this, the GPQ for the parameter of interest k is expressed as

$$R_k = \frac{\sqrt{R_{\sigma^2}}}{R_\mu}. \tag{2.3}$$

To construct a GPQ-based confidence interval with given observations $x_1, ..., x_m$ from $N(\mu, \sigma^2)$, we propose the following Monte-Carlo algorithm:

STEP 1: Compute the sample mean and sample standard deviation using original observation samples.

STEP 2: Randomly generate one U from corresponding chi-squared distribution with $n-1$ degree of freedom and one Z from corresponding normal distribution.

STEP 3: Calculate R_{σ^2},R_{μ} by using (2.1) and (2.2).

STEP 4: Calculate R_k by using (2.3).

STEP 5: Repeat STEP2-STEP4 H times (10000 times is recommended) to obtain H values of R_k .

Consequently, the $100(1-\alpha)\%$ GPQ-based confidence interval for CV can be obtained as $(R_{k,\alpha/2}, R_{k,1-\alpha/2})$, where $R_{k,\alpha/2}$ and $R_{k,1-\alpha/2}$ are the $100(\alpha/2)-th$ and $100(1-\alpha/2)-th$ percentiles of R_k's.

2.2.2 *Empirical likelihood based method*

Parametric methods of obtaining confidence interval for CV have been discussed and improved by many authors over the past several decades as we mentioned above. When the underlying distribution is a normal distribution, one is well-equipped of several methods to construct a confidence interval for CV. Nevertheless, in practice, the normality assumption may not be easily guaranteed. In these cases, which happen quite frequently, those methods are performing poorly or even invalid. Then it is really crucial to introduce a non-parametric method to compensate the incompletion of existing methods.

When addressing the non-parametric method for constructing confidence intervals, Owen (1988, 1990) is the one who must be mentioned, because he proposed a very powerful non-parametric method for constructing confidence interval for parameters of interest, which is well-known as Empirical likelihood (EL). With a lot of advantages, EL method was in the center of attention once the method purposed and has been widely applied to many different contexts. We summarize the advantages of as follows: 1. it does not require a piv-

otal statistic; 2. No prior constraints of the shape of confidence region is needed; 3. it may confesses a Bartlett correction that tolerates low coverage error. More detailed information on EL method can be refereed to Hall and La Scala (1990). However, the EL-based method for CV has not been well-developed. So, in the following paragraph, we attempt to apply the empirical method to the construction of confidence intervals for CV.

Recall that

$$k = \frac{\sqrt{E(X^2) - (E(X)^2)}}{E(X)}$$

is a smooth function of the mean $m = (E(X), E(X^2))$. Owen (1988) showed that the limiting distribution of the empirical log-likelihood ratio for m is a chi-square distribution with 2 degree of freedom. Hence, an EL-based joint confidence region for m can be obtained from the chi-square distribution, and an EL-based confidence interval for the CV can be found based on this confidence region. This is the original idea from Owen (1988) and an indirect way to construct confidence interval for the CV. However, this method involves in a 2-dimensional confidence region and it is somewhat uncomfortable in implementation. Thus, here we proposed a plugging-in EL method for constructing confidence interval of the CV.

We observe that the coefficient of variation k satisfies the following equation:

$$E(\sigma - kX) = 0.$$

Let $P = (p_1, ..., p_n)$ be a probability vector. Then the EL for k can be defined as follows:

$$L(k) = \sup_P \{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i W_i = 0\},$$

where $W_i = \sigma - kX_i,\ i = 1, \cdots, n$. Since parameter σ is unknown in practice, we use sample standard deviation S to estimate it. Then, the profile EL for k can be defined as follows:

$$\hat{L}(k) = \sup_P \{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{W}_i = 0\},$$

where $\hat{W}_i = S - kX_i$.

With Lagrange multiplier method, we can obtain the expression for p_i , which is

$$p_i = \frac{1}{n}\{1 + \lambda \hat{W}_i\}^{-1}, i = 1, ..., n,$$

where λ is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{W}_i}{1 + \lambda \hat{W}_i} = 0.$$

Therefore, the profile EL ratio for k is:

$$R(k) = \prod_{i=1}^{n}(np_i) = \prod_{i=1}^{n}\{1 + \lambda \hat{W}_i\}^{-1}.$$

Then the corresponding empirical log-likelihood ratio for k is

$$l(k) = 2\sum_{i=1}^{n} log\{1 + \lambda \hat{W}_i\}, \tag{2.4}$$

where λ is the solution of the equation $\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{W}_i}{1+\lambda\hat{W}_i} = 0$. Following Theorem 2.1 in Hjort *et al.* (2009), we find out that $l(k)$ actually asymptotically follows a scaled chi-square distribution.

Theorem: If k is the true value of the coefficient of variation, then the limiting distribution of $l(k)$, defined by (3), is a scaled chi-square distribution with degree of freedom one. i.e. ,

$$cl(k) \xrightarrow{d} \chi_1^2,$$

where the scale constant is $c \approx \frac{1}{E(l(k))}$.

Since c is an unknown constant, in order to construct confidence interval for the CV based on this theorem, we have to estimate c. Here we propose the following bootstrap procedure to estimate the scale constant.

Step 1: generate bootstrap sample $X_1^*, \cdots, X_n^*$ from the original sample $X_1, \cdots, X_n$.

Step 2: find the bootstrap estimate $l^*(\hat{k})$ of $l(k)$:

$$l^*(\hat{k}) = 2\sum_{i=1}^{n} log\{1 + \lambda^* \hat{W}_i^*\}$$

where $\hat{W}_i^* = S^* - \hat{k}X_i^*$ is the bootstrap version of $\hat{W}_i$, $\hat{k}$ is the sample coefficient of variation from the original sample, and λ^* is the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{W}_i^*}{1 + \lambda^* \hat{W}_i^*} = 0.$$

Step 3: repeat steps 1-2 B times ($B \geq 200$ is recommended) to obtain B bootstrap copies of $l(k)$:

$$\{l_1^*(\hat{k}), l_2^*(\hat{k}), ..., l_B^*(\hat{k})\}.$$

Step 4: Estimate the constant c as follows:

$$c^* = [\frac{1}{B}\sum_{i=1}^{n} l_i^*(\hat{k})]^{-1}$$

Based on Theorem and the estimated constant c^*, we can construct the bootstrap EL-based confidence interval (BEL interval) for k as follows:

$$\{k : c^* l(k) \leq \chi_1^2(1-\alpha)\}$$

where $\chi_1^2(1-\alpha)$ is the $(1-\alpha)-th$ quantile of χ_1^2 .

2.2.3 *Jackknife Empirical Likelihood Method*

In previous section, we introduced the bootstrap EL-based confidence interval (BEL interval). Despite of all the advantages the method has, serious computational difficulty is inevitably occurred, as well as one has to estimate the constant coefficient c for the scaled chi-square distribution, which may introduce extra bias. Recently, Jing, Yuan and Zhou (2009) raised a new approach called jackknife empirical likelihood (JEL) method. The logarithm of the Jackknife empirical likelihood ratio asymptotically follows the standard Chi-square distribution under certain conditions. Thus, constructing the JEL-based confidence interval

is extremely simple in calculation. In following context, we will develop the JEL-based method to construct confidence interval for CV.

Let $X_i,\ i = 1, ..., n$ be a random sample from a population with unknown underlying distribution. The sample coefficient of variation can be calculated as follows:

$$\hat{k} = \frac{S}{\bar{X}}$$

The corresponding jackknife pseudo values are

$$\hat{W}_i = n\hat{k} - (n-1)k_i, i = 1, ..., n$$

where k_i is the sample coefficient of variation computed with $(n-1)$ sample observations after deleting $i-th$ observation from the original sample.

With similar argument in Tukey (1958), $\hat{W}_i$'s are expected to be asymptotically independent. Then standard empirical likelihood methods could be applied to these jackknife samples for constructing empirical likelihood confidence interval for CV. Let $P = (p_1, ..., p_n)$

be a probability vector. The jackknife empirical likelihood for CV can be defined as follows:

$$L_J(k) = \sup_P \{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i(W_i - k) = 0\}.$$

And just as previous section, the Lagrange multiplier method provides the Jackknife empirical log-likelihood ratio:

$$l_J(k) = 2\sum_{i=1}^{n} log\{1 + \lambda_J(W_i - k)\},$$

where λ_J is the solution of the equation:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{W_i - k}{1 + \lambda_J(W_i - k)} = 0.$$

The Jackknife empirical log-likelihood ratio asymptotically follows a chi-square distribution with 1 degree of freedom. i.e.,

$$l_J(k) \xrightarrow{d} \chi_1^2.$$

Therefore, the Jackknife empirical-based confidence interval for k can be constructed as follows:

$$\{k : l_J(k) \leq \chi_1^2(1-\alpha)\}$$

where $\chi_1^2(1-\alpha)$ is the $(1-\alpha)-th$ quantile of χ_1^2.

CHAPTER 3

SIMULATION STUDIES

Newly introduced methods need to be examined and we need to show more evidence to illustrate the advantages of new methods over the existing ones. In this section, adequate simulation studies have been conducted to give a clear comparison between newly proposed confidence intervals and existing confidence intervals. The intervals being examined in the simulation study are Vangel's modified approximate method (Vangel), the Generalized Pivotal Quantity method (GPQ), Plug-in empirical likelihood with bootstrap-based confidence interval (BEL) and the Jackkinfe empirical likelihood confidence interval(JEL).

Considering the practical application for agriculture and engineering fields, the finite sample performance is the main issue we focus on. In the simulation studies, we carried out the simulation with sample size $n = 30, 50, 100, 200$, to examine the performance from smaller sample size to larger sample size. According to the property of the coefficient of variation, we selected the true value $k = 0.2, 0.5, 1$ to investigate the performance of all methods mentioned above. Large k may result in a non-negligible probability of obtaining a negative observation, and based on practical application, the scenario is of less interest. Thus the selected true values are narrowed down to the range from 0 to 1. The iteration times $R = 1000$ and bootstrap replication $B = 200$ were chosen for calculating BEL intervals because of the computational extensiveness with bootstrap method, and $R = 10000$ was chosen for calculating Vangel, GPQ and JEL intervals in each simulation setting, respectively.

Two underly distributions, normal distribution and Chi-square distribution, were selected to generate the random observations. For the setting with normal distribution as the pre-selected underly distribution, we consider all the methods mentioned above and make the comparison. However, as k getting larger, the vangel's method failed to work properly due to the property of the pivotal quantity. We mark N/A as comparing the rest. In the other setting, Chi-square distribution was selected as the underly distribution to examine the proposed non-parametric approaches. Since the normality assumption is invalid, Vangel and GPQ can not be applied to the setting.

We calculated the coverage probability (CP) and average length (AL) as the criteria to evaluate the performance of these intervals. At the nominal confidence level 90%, we calculated the percentages that the confidence intervals cover the true value of k. The closer the percentage is to the nominal level, the better performance of the confidence interval. Average length (AL) is the summation of all length of confidence intervals divided by total number of iteration. With similar coverage probabilities, the shorter average length is, the better performance of the confidence interval. All the simulation studies were conducted with the statistical package R.

The simulation results are displayed in table 3.1 to table 3.6 . From table 3.1, we conclude that GPQ interval has the best performance among the four intervals regardless the sample size in this scenario. Its coverage probabilities are very close to the nominal level. Vangel, BEL and JEL intervals undercover the true value with small to moderate sample sizes ($n = 30, 50, 100$), but they are acceptable as sample size n increases to 200. Vangel's method got relatively shorter average length but it is not impressive enough to top other methods.

Table 3.2 shows the similar performances for all the methods as we selected $k = 0.5$. As k getting larger, the average lengths increase. Table 3.3 shows the simulation results for $k = 1$. Due to the property of approximate pivotal, Vangel's method is not available (N/A). And the proposed methods still perform with satisfactory, especially the GPQ method. Tables 3.4 to 3.6 are the results from the setting of non-normality assumption. With the Chi-square distribution as the underly distribution, GPQ and Vangel's methods are no longer valid, but we can apply the proposed non-parametric methods. The results show that, BEL and JEL undercover the true values of k with small to moderate sample sizes ($n = 30, 50, 100$). However, as sample size increases, the coverage probabilities are approaching to the nominal level. Among different k values in this setting, BEL performs slightly better than JEL in terms of coverage probability.

In sum, under normality assumption, GPQ has the best performance with the appropriate coverage probability and stability. Comparing to the existing method, GPQ is more powerful as it expands the range of availability for the value of coefficient of variation. We recommend GPQ when the underly distribution is normal. Under non-normality assumption, the performances of the JEL and BEL intervals are acceptable when sample size is big enough. Therefore, JEL and BEL are recommended when the underlying distribution is unknown.

Table 3.1. The coverage probability and average length of 90 percent, $k = 0.2$, Underly distribution :Normal

N	*Method*	*CP*	*AL*
30	Vangel	0.851	0.081
	GPQ	0.906	0.094
	BEL	0.879	0.073
	JEL	0.838	0.090
50	Vangel	0.871	0.065
	GPQ	0.904	0.071
	BEL	0.893	0.060
	JEL	0.857	0.070
100	Vangel	0.884	0.047
	GPQ	0.902	0.049
	BEL	0.901	0.041
	JEL	0.882	0.049
200	Vangel	0.892	0.033
	GPQ	0.898	0.035
	BEL	0.900	0.021
	JEL	0.899	0.035

Table 3.2. The coverage probability and average length of 90 percent, $k = 0.5$,underly distribution: normal

N	*Method*	*CP*	*AL*
30	Vangel	0.870	0.269
	GPQ	0.909	0.290
	BEL	0.899	0.275
	JEL	0.851	0.261
50	Vangel	0.887	0.206
	GPQ	0.899	0.214
	BEL	0.894	0.221
	JEL	0.861	0.203
100	Vangel	0.894	0.145
	GPQ	0.902	0.147
	BEL	0.881	0.113
	JEL	0.887	0.144
200	Vangel	0.894	0.102
	GPQ	0.899	0.103
	BEL	0.894	0.053
	JEL	0.899	0.101

Table 3.3. The coverage probability and average length of 90 percent, $k = 1$,underly distribution: normal

N	*Method*	*CP*	*AL*
30	Vangel	N/A	N/A
	GPQ	0.900	1.005
	BEL	0.880	na
	JEL	0.820	0.778
50	Vangel	N/A	N/A
	GPQ	0.899	0.672
	BEL	0.906	na
	JEL	0.862	0.592
100	Vangel	N/A	N/A
	GPQ	0.908	0.430
	BEL	0.894	na
	JEL	0.900	0.412
200	Vangel	N/A	N/A
	GPQ	0.898	0.296
	BEL	0.896	na
	JEL	0.911	0.287

Table 3.4. The coverage probability and average length of 90 percent, $k = 0.2$,underly distribution: Chi-square

N	*Method*	*CP*	*AL*
30	BEL	0.85	0.072
	JEL	0.843	0.087
50	BEL	0.891	0.065
	JEL	0.860	0.068
100	BEL	0.878	0.045
	JEL	0.883	0.048
200	BEL	0.894	0.026
	JEL	0.895	0.034

Table 3.5. The coverage probability and average length of 90 percent, $k = 0.5$,underly distribution: Chi-square

N	*Method*	*CP*	*AL*
30	BEL	0.900	0.261
	JEL	0.836	0.227
50	BEL	0.898	0.180
	JEL	0.841	0.180
100	BEL	0.903	0.086
	JEL	0.860	0.130
200	BEL	0.876	0.030
	JEL	0.874	0.093

Table 3.6. The coverage probability and average length of 90 percent, $k = 1$,underly distribution: Chi-square

N	*Method*	*CP*	*AL*
30	BEL	0.887	0.347
	JEL	0.784	0.519
50	BEL	0.87	0.268
	JEL	0.800	0.422
100	BEL	0.873	0.152
	JEL	0.836	0.313
200	BEL	0.878	0.073
	JEL	0.844	0.227

CHAPTER 4

REAL EXAMPLE

In this chapter, a real case is studied to illustrate the methods we introduced in the previous chapters. The practice sample data set named Beef Council Check-off is obtained from The Data and Story Library (DASL) at Carnegie Mellon University. Among 7 variables in the data set, we chose 2 of them to conduct the experiment, Average size of farm (hundreds of acres) and Average value of products sold (thousands). Each of the variable contains 56 observations, which is similar amount as moderate sample size we chose to conduct the simulation study.

In order to check the normality of each variable, the normality test called Shapiro-Wilk test has been conducted. The null hypothesis of Shapiro-Wilk test is that sample data $x_1, x_2, ..., x_n$ is from normal distribution. The small p-value (smaller than the significant level one selected) indicates the rejection of null hypothesis which means the sample data is from non-normal distribution. And if one fails to reject the null hypothesis, we can treat the sample data comes from normal distribution. More detailed information can be referred to Shapiro and Wilk (1965).

For variable of average value of products sold (thousands), the sample coefficient of variation $\hat{k} = 0.4733$. And regarding to Shapiro-Wilk test, the calculated p-value is 0.5037 which indicates that sample data is from a normal distribution. Based on the test results, we applied all the previously introduced methods to obtain the 90 percent confidence intervals

for CV. Table 4.1 shows the simulation results. From Table 4.1, we can see that confidence intervals calculated from all mentioned methods are very similar. We can conclude that BEL and JEL are relatively better due to the slightly shorter Exact Length.

Similarly, the sample coefficient of variation for variable of average size of farm (hundreds of acres) is $\hat{k} = 0.6594$. Shapiro-Wilk test was conducted to test the normality. With calculated p-value 0.002981,the null hypothesis is rejected at the significant level of 0.05, which indicates the sample is from a population with non-normality distribution. Therefore, we can only apply the newly proposed non-parametric methods to calculate the confidence interval for CV with the nominal level 0.9. The simulation results can be found in Table 4.2. And we can conclude that JEL and BEL give very similar performance in this real example.

Table 4.1. The 90 percent Confidence Interval and Exact Length for CV of Average value of products sold (thousands)

Method	Confidence Interval	Exact Length
Vangel	(0.3986, 0.5755)	0.1769
GPQ	(0.3973, 0.5872)	0.1898
BEL	(0.4068, 0.5606)	0.1538
JEL	(0.4067, 0.5579)	0.1512

Table 4.2. The 90 percent Confidence Interval and Exact Length for CV of Average size of farm (hundreds of acres)

Method	Confidence Interval	Exact Length
BEL	(0.5858, 0.7726)	0.1869
JEL	(0.5707, 0.7662)	0.1955

CHAPTER 5

DISCUSSION AND CONCLUSION

The significance of the coefficient of variation may be underestimated by statisticians. Comparing to other frequently mentioned indexes for variation, CV draws less attention from statisticians. However, with the properties of unitlessness and mean-related, coefficient of variation appears more and more in applied researches such as agriculture field and engineering field. In literature, several methods of constructing confidence interval for CV have been introduced in the past decades. Nevertheless, most of existing methods are of deficiency. The dependency of underlying distribution and narrowed availability range for CV are the main motivation for this thesis study. In this thesis, we proposed a new parametric interval and two non-parametric intervals for CV. Under the normality assumption, the GPQ-based confidence interval shows a nearly perfect small sample performance, and successfully extend the availability range for CV. Therefore, it is recommended to use GPQ-based confidence interval for CV under normality assumption. Through plenty of simulation studies, the BEL and JEL confidence intervals are shown to have acceptable finite sample performances when sample size is big enough ($n \geq 200$). Since they are non-parametric intervals, they are particularly useful to obtain confidence intervals for CV from the population with non-normality assumption. Thus, these non-parametric methods are suggested to obtain confidence intervals for the coefficient of variation with unknown underlying distribution. The JEL-based and BEL-based confidence intervals give more liberal results as

shown in the simulation study when sample size is small. The future research will focus on developing better non-parametric method with better performance as sample size is small.

REFERENCES

[1] Barndorff-Nielsen, O.E. (1986), Inference on full or partial parameters based on the standardized Signed log-likelihood ratio, Biometrika 73: 307-322.

[2] Barndorff-Nielsen, O. E. (1991). Modified Signed log-likelihood ratio. Biometrika 78: 557-563.

[3] David, F. N. (1949). Note on the application of Fisher's k-statistics. Biometrika 36: 383-393.

[4] Hanning, J., Iyer, H.K., and Patterson, P.. (2006). Fiducial Generalized Confidence Intervals. Journal of the American Statistical Association, 101, 254-269.

[5] Hjort, N. L.., McKeague, I. W. and Keilegom, I. V. (2009). Extending the scope of empirical likelihood. Ann. Statistics 37, 1079-1111.

[6] Girma Taye and Peter Njuho (2008). Monitroing Field Variability Using Confidence Interval for Coefficient of Variation, Communications in Statistics-Theory and Methods, 37: 831-846.

[7] Gomez, K. A., Gomez, A. A. (1984). Statistical Procedures for Agricultural Research. 2nd ed. New York: John Wiley and Sons, Inc.

[8] Hall P. and La Scala B. (1990), Methodology and Algorithms of Empirical Likelihood, International Statistical Review, S8, 2, pp. 109-127.

[9] Jing B., Yuan J. and Zhou W. (2009), Jackknife Empirical Likelihood, Journal of the American Statistical Association, Vol.195, No. 487, pp. 1224-1232.

[10] Johnson, N. L., Welch, B. L. (1940), Application of the no- central t-distribution. Biometrika 31: 362-389.

[11] Lehmann, E. L. (1986), Testing Statistical Hypothesis. 2nd ed. New York: Wiley.

[12] Loh W. (1991) , Bootstrap calibration for confidence interval construction and selection, Statistica Sinika, Vol.1, pp. 477-491.

[13] McKay, A. T.(1932). Distribution of the coefficient of variation and the exextended t-distribution. J. Roy. Statist. Soc. B 95: 695-698.

[14] Owen A.B. (1988), Empirical likelihood ratio confidence intervals for a single functional, Biometrika, Vol.75, pp. 237-249.

[15] Owen A.B. (1990), Empirical likelihood ratio confidence regions, Biometrika, Vol.18, pp. 90-120.

[16] Peter Sprent and Nigel C. Sweeton (2007), Applied Nonparametric Statistical Methods, Chapman Hall/CRC, fourth edition.

[17] Pierce, D. A., Peters, D. (1992). Practical use of higher order asymptotic for multiparameter exponential families (with discussion). J.Roy. Statist. Soc. B 54:701-738.

[18] Qin, G.S., and Zhou, X.H. (2006), Empirical likelihood inference for the area under the ROC curve, Biometrics, Vol. 62, pp. 613-622.

[19] Qin J. and Lawless J. (1994), Empirical likelihood and general estimating equations, The Annals of Statistics, Vol. 22, No.1, pp. 300-325.

[20] Reid, N. (1996), The roles of conditioning in inference. Statist. Soc. B 54: 701-738.

[21] Shapiro, S. S.; Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika 52 (3-4): 591C611.

[22] Steel, R. G. D., Torrie, J. H. (1980). Principles and Procedures of Statistics, 2nd ed. New York: McGraw-Hill.

[23] Tukey, J. W. (1958). Bias and confidence in not-quite large samples. Ann. Statist. 29, 614.

[24] Vangel, M. G. (1996). Confidence intervals for a normal coefficient of variation. Am. Statist. 50: 21-26.

[25] Walter A. Hendricks and Kate W. Robey (1936) . The Sampling Distribution of the Coefficient of Variation. Ann. Math. Statist. Vol. 7, No. 3, pp.129-132.

[26] Weerahandi S. (1993), Generalized confidence intervals, Journal of the American Statistical Association, Vol. 88, No. 423, pp. 899-905.