A non-parametric coverage interval

2008 Metrologia 45 L1

(http://iopscience.iop.org/0026-1394/45/1/N01)

Metrologia **45** (2008) L1–L4 doi:10.1088/0026-1394/45/1/N01

SHORT COMMUNICATION

A non-parametric coverage interval

Shuo-Huei Lin[1], **Wenyaw Chan**[2] **and Lin-An Chen**[1]

[1] Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan
[2] Division of Biostatistics, School of Public Health, University of Texas–Health Science Center at Houston, The University of Texas, TX, USA

Received 23 October 2007
Published 8 January 2008
Online at stacks.iop.org/Met/45/L1

Abstract
Classically the non-parametric coverage interval is estimated by empirical quantiles. We introduce an alternative way for estimating the coverage interval by symmetric quantiles given by Chen and Chiang (1996 *J. Nonparametric Stat.* **7** 171–85). We further show that this alternative has a better precision in the sense that its asymptotic variances are smaller than the classical one.

1. Introduction

The coverage interval, called the reference interval in laboratory chemistry, refers to population-based reference values obtained from a well-defined group of reference individuals. This is an interval with two confidence limits which covers the measurement values in the population in some probabilistic sense. Laboratory test results are commonly compared with a coverage interval before caregivers make physiological assessments, medical diagnoses or management decisions. An individual who is being screened for a disorder based on a measurement is suspected to be abnormal if his/her measurement value lies outside the coverage interval.

The coverage interval can be estimated either parametrically or non-parametrically. The parametric method classically assumes that the underlying distribution of the measurement variable is normal whereas, recently, Chen *et al* (2007) have proposed a technique for constructing coverage intervals for asymmetric distributions. On the other hand, the non-parametric approach estimates the quantiles (percentile) directly; the most popular technique for estimating the unknown quantiles is through the empirical quantile. Most authorities now recommend the non-parametric method because it makes no assumptions concerning the type of reference variable. It is easy and reliable whether the reference variable follows normal or non-normal distribution (see these points in Reed *et al* (1971) and Solberg (2006)).

It is vitally important to establish a coverage interval so that users can diagnose diseases with precision. Some factors that may increase the precision have been considered. The number of 120 or more of healthy subjects required for the determination of coverage interval has been recommended by the International Federation of Clinical Chemistry. The determination of the confidence interval of the quantile, that is, the limits within which a true quantile is located with a specified confidence, is strongly recommended. However, Friedberg *et al* (2007) have observed that analytic imprecision is a very important factor for the quality of an established coverage interval. Hence, the search for an alternative technique in developing a coverage interval to increase the analytic precision of the computed coverage interval is an interesting and important topic.

We consider a non-parametric approach in constructing a coverage interval. Is there an alternative quantile such that its produced coverage interval may gain precision better than a quantile constructed empirically?' To improve the efficiency of the trimmed mean for estimating the location parameter, Kim (1992) and Chen and Chiang (1996) introduced the symmetric quantile to construct an alternative trimmed mean. Basically a symmetric quantile pair is a parameter plus and minus a classical quantile defined for an absolute value of the error variable defined as the measurement variable minus the parameter. They observed that this trimmed mean can have asymptotic variances very close to the Cramer–Rao lower bounds for several distributions, including heavy tail ones. Then, from the point of robust estimation, the symmetric quantile is efficient in the detection of outliers. We then consider the question whether it can be more accurate than the empirical quantiles to detect the central part of the underlying distribution for establishing the coverage interval. Our aim in this research is to construct an alternative coverage interval by symmetric quantiles and show that it does gain better

0026-1394/08/010001+04$30.00

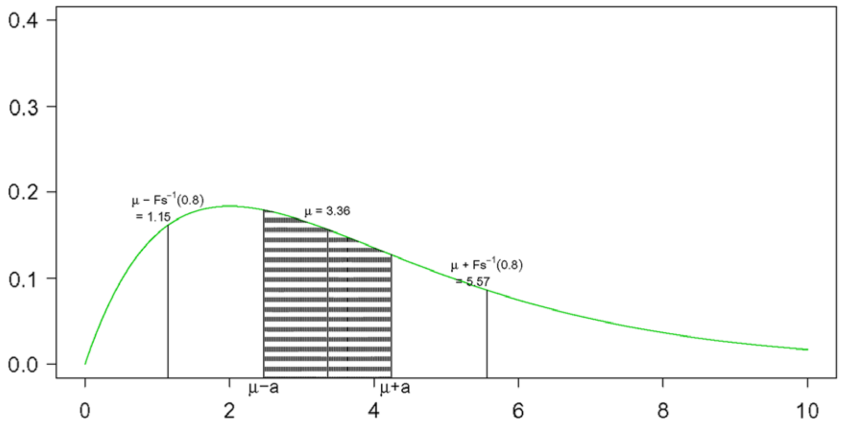

Figure 1. Folded distribution function and symmetric coverage interval for Gamma distribution $\Gamma(2, 2)$. (This figure is in colour only in the electronic version)

precision than the classical version constructed by empirical quantiles.

2. Symmetric coverage interval

For random variable y with cumulative distribution function F, the λth quantile is defined as

$$F^{-1}(\lambda) = \inf\{c\ :\ F(c) \geqslant \lambda\}.$$

The classical $1-\alpha$ coverage interval is defined as

$$C(1-\alpha) = \left(F^{-1}\left(\frac{\alpha}{2}\right), F^{-1}\left(1-\frac{\alpha}{2}\right)\right).$$

Suppose that we now have a random sample $y_1, \ldots, y_n$ from distribution F. The corresponding empirical $1-\alpha$ coverage interval is

$$C_n(1-\alpha) = \left(F_n^{-1}\left(\frac{\alpha}{2}\right), F_n^{-1}\left(1-\frac{\alpha}{2}\right)\right), \qquad (2.1)$$

where we let F_n^{-1} be the empirical quantile, a quantile function with distribution function of the sample type as $F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \leqslant y)$.

Unlike the way in which the empirical quantile is constructed based on the cumulative distribution function, the so-called symmetric quantile of Chen and Chiang (1996) is formulated based on a folded distribution function. Let us consider the folded cumulative function about μ, known or unknown, as

$$F_s(a) = P(|y-\mu| \leqslant a), \quad a \geqslant 0.$$

Extending from that given by Chen and Chiang (1996), we define the $1-\alpha$ symmetric coverage interval as

$$C_s(1-\alpha) = (\mu - F_s^{-1}(1-\alpha), \mu + F_s^{-1}(1-\alpha)),$$

where $F_s^{-1}(\lambda) = \inf\{a\ :\ F_s(a) \geqslant \lambda\}$. If F is continuous, the $1-\alpha$ symmetric coverage interval satisfies $1-\alpha = P(\mu - F_s^{-1}(1-\alpha) \leqslant y \leqslant \mu + F_s^{-1}(1-\alpha))$. If we further assume that F is symmetric at μ, it can be seen that

$$C_s(1-\alpha) = C(1-\alpha), \qquad (2.2)$$

the classical one and the symmetric one are identical in the sense of containing the same set of reference individuals.

We interpret the folded cumulative function and the symmetric coverage interval through a picture (see figure 1). Considering the Gamma distribution $\Gamma(2, 2)$ which has the probability density function (pdf) as the curve in the figure, we consider the folded distribution about the median. With this distribution, the median μ is 3.36. For a given $a > 0$, the value of this folded distribution at a represents the probability of a region as a part of a shadow. Suppose that our interest is to construct an 80% coverage interval. For this continuous distribution, we search for $F_s^{-1}(0.8) = a^*$ such that $0.8 = P(\mu - a^* \leqslant y \leqslant \mu + a^*)$ with $y \sim \Gamma(2, 2)$, which indicates that $a^* = 2.21$. Hence the 80% symmetric coverage interval is

$$C_s(0.8) = (\mu - F_s^{-1}(0.5), \mu + F_s^{-1}(0.5)) = (1.15, 5.57)$$

(see the limits $\mu - F_s^{-1}(0.5)$ and $\mu + F_s^{-1}(0.5)$ in figure 1).

Let $\hat{\mu}$ be an estimate of μ. We may define the sample type $1-\alpha$ symmetric coverage interval as

$$C_{sn}(1-\alpha) = (\hat{\mu} - F_{sn}^{-1}(1-\alpha), \hat{\mu} + F_{sn}^{-1}(1-\alpha)), \qquad (2.3)$$

where $F_{sn}(a) = \frac{1}{n}\sum_{i=1}^{n} I(|y_i - \hat{\mu}| \leqslant a)$ is the sample type folded cumulative distribution function and $F_{sn}^{-1}(1-\alpha) = \inf\{a\ :\ F_{sn}(a) \geqslant 1-\alpha\}$.

Let us give a simple example to describe the construction of the sample symmetric coverage interval. Suppose that we have a set of observations that are ordered as

$$-5, -3, -2, -1, -0.5, 0.5, 1, 3, 50, 100.$$

We want to construct 80% empirical and symmetric coverage intervals. With $F_n^{-1}(0.1) = -5$ and $F_n^{-1}(0.9) = 50$, the 80% empirical coverage interval is

$$C_n(0.8) = (-5, 50). \qquad (2.4)$$

For construction of a symmetric coverage interval, we choose the sample median as the estimate of μ. That is,

$$\hat{\mu} = F_n^{-1}(0.5) = \inf\left\{a\ :\ \frac{1}{10}\sum_{i=1}^{10} I(y_i \leqslant a) \geqslant 0.5\right\} = -0.5.$$

Let us denote residuals $e_i = y_i - \hat{\mu}, i = 1, ..., 10$. The residuals are

$$-4.5, -2.5, -1.5, -0.5, 0, 1, 1.5, 3.5, 50.5, 100.5.$$

The sample type folded cumulative distribution function is

$$F_{sn}(a) = \frac{1}{10}\sum_{i=1}^{10} I(|e_i| \leqslant a).$$

For examples, $F_{sn}(0) = \frac{1}{10}$, $F_{sn}(1) = \frac{1}{10}\big[I(|-0.5| \leqslant 1) + I(|0| \leqslant 1) + I(|1| \leqslant 1)\big] = \frac{3}{10}$. Then we have

$$F_{sn}^{-1}(0.8) = \inf\left\{a : \frac{1}{10}\sum_{i=1}^{10} I(|e_i| \leqslant a) \geqslant 0.8\right\} = 4.5.$$

This indicates that the 80% symmetric coverage interval is

$$\begin{aligned} C_{sn}(0.8) &= (\hat{\mu} - F_{sn}^{-1}(0.8), \hat{\mu} + F_{sn}^{-1}(0.8)) \\ &= (-0.5 - 4.5, -0.5 + 4.5) = (-5, 4). \end{aligned} \tag{2.5}$$

Comparing the resulting sample empirical and symmetric coverage intervals in (2.4) and (2.5), the benefit of using the latter one that is shorter than the former one is seen. This will happen very often when the observations are drawn from asymmetric distributions.

3. Precision study of symmetric coverage interval

The equality of (2.2) does not hold when the underlying distribution F is not symmetric so that there is no fair criterion to compare their corresponding sample coverage intervals. Hence, we may set the case that F is symmetric to compare the precision of these two coverage intervals through the asymptotic variances of their sample type coverage intervals.

We consider that μ is the median parameter and let $\hat{\mu}$ be the sample median as

$$\hat{\mu} = \text{arginf}_{\mu \in R} \sum_{i=1}^{n} |y_i - \mu|.$$

Suppose that we assume that F is continuous and symmetric at μ. From Ruppert and Carroll (1980), we have a Bahadur representation for this sample median as

$$n^{1/2}(\hat{\mu} - \mu) = n^{-1/2}\frac{1}{f(\mu)}\sum_{i=1}^{n}(0.5 - I(y_i \leqslant \mu)) + o_p(1). \tag{3.1}$$

On the other hand, a Bahadur representation for $F_{sn}^{-1}(1-\alpha)$ developed by Chen and Chiang (1996) is

$$\begin{aligned} & n^{1/2}\left(F_{sn}^{-1}(1-\alpha) - \left(F^{-1}\left(1 - \frac{\alpha}{2}\right) - \mu\right)\right) \\ &\quad = \frac{1}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)} n^{-1/2} \\ &\quad\quad \times \sum_{i=1}^{n}\left\{1 - \alpha - I\left(F^{-1}\left(\frac{\alpha}{2}\right) \leqslant y_i \leqslant F^{-1}\left(1-\frac{\alpha}{2}\right)\right)\right\} \\ &\quad\quad + o_p(1). \end{aligned} \tag{3.2}$$

The assumption of a symmetric distribution indicates that $\hat{\mu} - F_{sn}^{-1}(1-\alpha)$ and $\hat{\mu} + F_{sn}^{-1}(1-\alpha)$ have the same asymptotic variance and, from (3.1) and (3.2), we have a Bahadur representation for $\hat{\mu} - F_{sn}^{-1}(1-\alpha)$ as

$$\begin{aligned} & n^{1/2}\left((\hat{\mu} - F_{sn}^{-1}(1-\alpha)) - F^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &\quad = n^{-1/2}\sum_{i=1}^{n}\left\{\left[-\frac{1}{2f(\mu)} - \frac{1-\alpha}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)}\right]\right. \\ &\quad\quad \times I\left(y_i \leqslant F^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &\quad\quad + \left[-\frac{1}{2f(\mu)} + \frac{\alpha}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)}\right] \\ &\quad\quad \times I\left(F^{-1}\left(\frac{\alpha}{2}\right) \leqslant y_i \leqslant \mu\right) \\ &\quad\quad + \left[\frac{1}{2f(\mu)} + \frac{\alpha}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)}\right] \\ &\quad\quad \times I\left(\mu < y_i \leqslant F^{-1}\left(1-\frac{\alpha}{2}\right)\right) \\ &\quad\quad + \left[\frac{1}{2f(\mu)} - \frac{1-\alpha}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)}\right] \\ &\quad\quad \left. \times I\left(y_i \geqslant F^{-1}\left(1-\frac{\alpha}{2}\right)\right)\right\} + o_p(1). \end{aligned} \tag{3.3}$$

Since $y_1, ..., y_n$ is a random sample from distribution F, we may see that the asymptotic variance of $n^{1/2}(\hat{\mu} - F_{sn}^{-1}(1-\alpha) - F^{-1}(\frac{\alpha}{2}))$ is

$$\begin{aligned} \sigma_s^2 &= \frac{\alpha}{2}\left[\left(\frac{1}{2f(\mu)} + \frac{1-\alpha}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)}\right)^2\right. \\ &\quad \left. + \left(\frac{1}{2f(\mu)} - \frac{1-\alpha}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)}\right)^2\right] \\ &\quad + \left(1 - \frac{\alpha}{2}\right)\left[\left(-\frac{1}{2f(\mu)} + \frac{\alpha}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)}\right)^2\right. \\ &\quad \left. + \left(\frac{1}{2f(\mu)} + \frac{\alpha}{2f\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)}\right)^2\right]. \end{aligned} \tag{3.4}$$

On the other hand, in this situation where y has a continuous and symmetric distribution, we may see that $n^{1/2}(F_n^{-1}(\frac{\alpha}{2}) - F^{-1}(\frac{\alpha}{2}))$ and $n^{1/2}(F_n^{-1}(1-\frac{\alpha}{2}) - F^{-1}(1-\frac{\alpha}{2}))$ also have the

Table 1. The efficiencies, Eff, of the symmetric coverage interval.

	0.6	0.8	0.9	0.95	0.98
$N(0,1)$	0.87	0.87	1.02	1.21	1.48
$t(r)$					
$r = 1$	0.98	1.78	2.13	2.1	2.04
$r = 5$	0.89	1.01	1.31	1.61	1.85
$r = 10$	0.88	0.94	1.16	1.42	1.7
Cauchy(s)	0.98	1.78	2.13	2.1	2.04
$s = 1, 5, 10$					
Lap(b)	1.2	1.6	1.8	1.9	1.96
$b = 1, 5, 10$					

same asymptotic variance (see, for example, Sen and Singer (1993, p 168)) as

$$\sigma_e^2 = \frac{\alpha}{2}\left(1-\frac{\alpha}{2}\right) f^{-2}\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right). \tag{3.5}$$

Since these two sample coverage intervals estimate the same population coverage interval, it is fair that we evaluate the efficiency of the symmetric type coverage interval defined as the following:

$$\mathrm{Eff} = \frac{\sigma_e^2}{\sigma_s^2}. \tag{3.6}$$

Let us consider several distributions for the computation of asymptotic variances of (3.4) and (3.5) to compare their corresponding efficiencies of (3.6) where distributions include the standard normal distribution $N(0,1)$, the t-distribution $t(r)$, where r is the degrees of freedom, the Cauchy distribution (Cauchy(s), $s > 0$) with pdf

$$f(y) = \frac{1}{\pi}\frac{s}{y^2+s^2}, \qquad y \in R$$

and the Laplace distribution (Lap(b)) with pdf

$$f(y) = \frac{1}{2b}\mathrm{e}^{-\frac{|y|}{b}}, \qquad y \in R.$$

We display the resulting efficiencies in table 1.

It is relatively efficient to use the empirical quantile to construct the coverage interval when the quantile percentage is close to 0.5 in either direction. This means that when we want a $1-\alpha$ coverage interval with coverage probability $1-\alpha$ of value 0.6 or even smaller, the one estimated by empirical quantiles is the right choice. On the other hand, we see that it gains more precision to use a symmetric quantile to construct the coverage interval when $1-\alpha$ has a value of 0.8 or more. This alternative coverage interval is then attractive since it is common that we apply the coverage interval only for large $1-\alpha$; for example, the reference interval in medical diagnosis chooses a value of 0.95. In fact, in the case where the underlying distribution is the Laplace one the coverage interval constructed by symmetric quantiles totally dominates the one by empirical quantiles.

This interesting result is not surprising. The surprising fact is that, unlike the estimation of location and scale parameters that have received much attention in the statistical literature for proposing techniques and developing theories in gaining better precision, not much attention has been paid to developing alternative ways for constructing coverage intervals for gaining better precision than the classical one in the statistical and metrological literature.

Acknowledgment

The authors would like to express their gratitude to the editor and a referee for their valuable comments. The constructive and insightful comments from the referee greatly improved the quality of the paper. This research work was partially supported by the National Science Council of Taiwan Grant No NSC95-2118-M-009-007.

References

Chen L-A and Chiang Y C 1996 Symmetric type quantile and trimmed means for location and linear regression model *J. Nonparametric Stat.* **7** 171–85

Chen L-A, Huang J-Y and Chen H-C 2007 Parametric coverage interval *Metrologia* **44** L7–9

Friedberg R C, Souers R, Wagar E A, Stankovic A K and Valenstein P N 2007 The origin of reference intervals *Arch. Pathol. Lab. Med.* **131** 348–57

Kim S J 1992 The metrically trimmed means as a robust estimator of location *Ann. Stat.* **20** 1534–47

Reed A H, Henry R J and Mason W B 1971 Influence of statistical method used on the resulting estimate of normal range *Clin. Chem.* **17** 275

Ruppert D and Carroll R J 1980 Trimmed least squares estimation in the linear model *J. Am. Stat. Assoc.* **75** 828–38

Sen P K and Singer J M 1993 *Large Sample Methods in Statistics* (New York: Chapman and Hall)

Solberg H E 2006 Establishment and use of reference values *Textbook of Clinical Chemistry* ed N W Tietz (St Louis: Saunders)