Mathematical Modelling and Analysis
Volume 13 Number 2, 2008, pages 195–202
© 2008 Technika ISSN 1392-6292 print, ISSN 1648-3510 online

ESTIMATION OF CONFIDENCE INTERVALS FOR QUANTILES IN A FINITE POPULATION

V. CHADYŠAS

Vilnius Gediminas Technical University
Saulėtekio al. 11, LT-10223 Vilnius, Lithuania
E-mail: viktoras.chadysas@fm.vgtu.lt

Received September 28, 2007; revised November 8, 2007; published online May 1, 2008

Abstract. Confidence intervals provide a way of reporting an estimate of a population quantile along with some information about the precision of estimates. Some procedures that may be used to obtain estimates of confidence intervals for quantiles in a finite population (most of which are based on resampling) are compared in the paper. A simulation study, based on two different artificial populations, is performed and comparisons of the estimation methods proposed for confidence intervals of population quantiles are made.

Key words: quantile, confidence interval, finite population, bootstrap, jackknife.

1 Introduction

This study is concerned with the methods of estimation of the confidence intervals for quantiles when a simple random sample is drawn without replacement from a finite population. Several procedures that may be used to obtain estimates of the confidence intervals for quantiles, using a simple random sampling design [4] in the finite population, are considered.

An initial way of obtaining the confidence interval for a median when the distribution of the median estimator is asymptotically normal is described in [4]. If the population is not normally distributed, then the confidence level of such an interval may be inaccurate for small sample sizes, because the distribution of the sample quantile may be not well-approximated by a normal distribution. In this case, we construct estimates of the confidence intervals for quantiles using resampling methods [5].

2 Methodology

2.1 Definitions

Definition 1. ([3]) For a probability distribution $F(x) = P(\xi < x)$ of some random variable ξ and real number x $(-\infty < x < \infty)$, $q \in (0,1)$, the quantile K_q of level q is defined by

$$K_q = \inf\{x : F(x) \geq q\}. \tag{2.1}$$

If the function $F(x)$ is continuous, the quantile can be found as

$$K_q = F^{-1}(q). \tag{2.2}$$

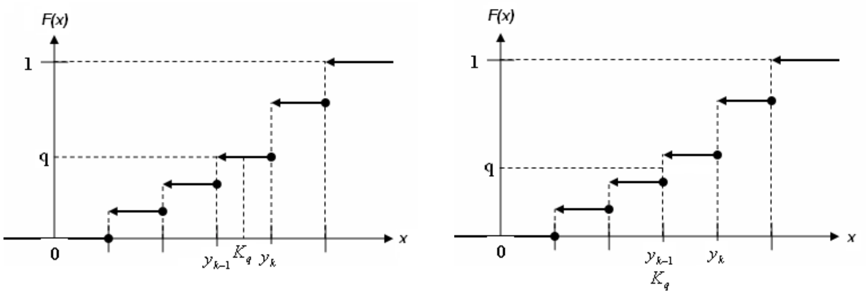
Here F^{-1} is the inverse function of F.

Suppose $\mathcal{U} = \{1, 2, \ldots, N\}$ is a finite population, y is a study variable defined for the elements of the population $\mathcal{U}$ with the values $\{y_1, y_2, \ldots, y_N\}$. For any given number x $(-\infty < x < \infty)$, the finite population distribution function $F(x)$ is defined as a proportion of elements k in the population for which $y_k < x$. The function $F(x)$ can be written as

$$F(x) = \frac{\sharp A_x}{N}. \tag{2.3}$$

Here the set $A_x = \{l \in \mathcal{U} : y_l < x\}$, and $\sharp A_x$ denotes the number of elements in the set A_x.

Let us introduce another definition of a finite population distribution quantile starting from equation (2.2). Since the distribution function of the study variable in a finite population is a step function, equation (2.2) may have an infinite number of solutions or no solutions at all. The solutions of (2.2) are shown in Fig. 1.



(a)

(b)

Figure 1. Determination of the population quantile.

In case (a), K_q corresponds to the whole interval $(y_{k-1}, y_k]$ of x values satisfying (2.2). In case (b), equation (2.2) does not have any solutions. In both cases, the q-level quantile defined by (2.1) is equal to y_{k-1}. Let us define an inverse function of a finite population distribution function and quantile in some different way. Let us arrange the population elements in increasing order of the variable y: $y_1 \le y_2 \le \ldots \le y_N$.

DEFINITION 2. The finite population q-level quantile K_q, $0 < q < 1$, is defined as

$$K_q = F^{-1}(q) = \begin{cases} 0.5(y_{k-1} + y_k), & \text{if } F(y_k) = q; \\ y_{k-1}, & \text{if } F(y_{k-1}) < q < F(y_k), \end{cases} \tag{2.4}$$

for some $k \in \mathcal{U}$.

We are interested in the estimation of the population quantile K_q in the case of simple random sampling. Simple random sampling (SRS) is a sampling design in which all the possible collections s, s $\subset \mathcal{U}$ of n different elements, have the same probability $1/C_N^n$ of selection [1]. The SRS design may be obtained when n elements from the finite population are drawn with equal selection probabilities without replacement.

The general procedure of quantile estimation:

1. First, obtain an estimate $\widehat{F}(x)$ of the distribution function $F(x)$.
2. Then estimate $K_q = F^{-1}(q)$ by $\widehat{K}_q = \widehat{F}^{-1}(q)$.

2.1.1 Estimator of the distribution function

Let us define an indicator variable $z(x)$ with the values

$$z_k(x) = \begin{cases} 1, & \text{if } \ y_k < x, \\ 0, & \text{if } \ y_k \ge x, \end{cases}$$

$k = 1, 2, \ldots, N$, and the total $t_{z(x)} = \sum_{k=1}^{N} z_k(x)$. Then the distribution function $F(x)$ in (2.3) can be expressed as the population mean:

$$F(x) = \frac{t_{z(x)}}{N} = \mu_{z(x)}.$$

The estimator of $F(x) = \mu_{z(x)}$ for SRS can be constructed as

$$\widehat{F}(x) = \widehat{\mu}_{z(x)} = \frac{\widehat{t}_{z(x)}}{N},$$

where $\widehat{t}_{z(x)} = \frac{N}{n} \sum_{k\in} \ z_k(x)$.

2.1.2 Quantile estimator

We will consider a simple random sample s drawn from $\mathcal{U}$. Let the values of the sample elements be sorted in increasing order by $y_1 \le y_2 \le \ldots \le y_n$.

We define the quantile estimator as follows:

$$\widehat{K}_q = \widehat{F}^{-1}(q) = \begin{cases} 0.5(y_{k-1} + y_k), & \text{if } \ \widehat{F}(y_k) = q; \\ y_{k-1}, & \text{if } \ \widehat{F}(y_{k-1}) < q < \widehat{F}(y_k), \end{cases}$$

for any $q \in (0, 1)$. The simulation results show that the distribution of the quantile estimator is non-symmetric. Some examples of simulation are presented in the following sections.

2.2 Confidence intervals

Definition 3. ([4]). If there exist c_1 and c_2 such that

$$P\{c_1 \le \widehat{F}(K_q) \le c_2\} \ge 1-\alpha,$$

then the interval $[\widehat{F}^{-1}(c_1), \widehat{F}^{-1}(c_2)]$ is the level α $(0 < \alpha < 1)$ confidence interval for quantile K_q.

Some methods used to estimate confidence intervals for quantiles are considered below.

2.2.1 The normal distribution-based estimator (NDB)

If $\widehat{K}_q$, is normally distributed, the $1-\alpha$ level confidence interval estimator for the population quantile can be estimated by

$$\left[\widehat{F}^{-1}\left(q - K_{\alpha/2}^{\mathcal{N}(0,1)}\sqrt{\widehat{\mathbf{D}}\widehat{F}(\widehat{K}_q)}\right),\ \widehat{F}^{-1}\left(q + K_{\alpha/2}^{\mathcal{N}(0,1)}\sqrt{\widehat{\mathbf{D}}\widehat{F}(\widehat{K}_q)}\right)\right]. \tag{2.5}$$

Here $K_{\alpha/2}^{\mathcal{N}(0,1)}$ is the $1-\alpha/2$ quantile of the standard normal distribution [2].

In the case of simple random sampling, confidence interval estimator (2.5) for the finite population quantile K_q, is obtained by setting

$$\widehat{\mathbf{D}}\widehat{F}(\widehat{K}_q) = \left(1-\frac{n}{N}\right)\frac{\widehat{s}^2}{n}, \quad \widehat{s}^2 = \frac{1}{n-1}\sum_{k\in}\left(z_k(\widehat{K}_q) - \overline{z}\right)^2,$$

$$\overline{z} = \frac{1}{n}\sum_{l\in} z_l(\widehat{K}_q), \quad z_l(\widehat{K}_q) = \begin{cases} 1, & \text{if } y_l < \widehat{K}_q, \\ 0, & \text{otherwise.} \end{cases}$$

2.2.2 The traditional bootstrap method (TB)

The procedure is:

- Draw a simple random sample of size n from a population.
- From the sample drawn, draw a simple random subsample with replacement of size n (bootstrap sample). Let $\widehat{K}_q^{(1)}$ be its quantile. Repeat this process B times, obtain quantiles $\widehat{K}_q^{(1)}$, $\widehat{K}_q^{(2)}$,…, $\widehat{K}_q^{(B)}$.
- Taking the quantiles of this set $\widehat{K}_q^{\mathrm{TB}}(\alpha/2)$ and $\widehat{K}_q^{\mathrm{TB}}(1-\alpha/2)$ of the levels $\alpha/2$ and $1-\alpha/2$, we get the estimate of the confidence interval of the level $1-\alpha$ for the quantile K_q:

$$\left[\widehat{K}_q^{\mathrm{TB}}(\alpha/2),\ \widehat{K}_q^{\mathrm{TB}}(1-\alpha/2)\right].$$

2.2.3 The rescaling bootstrap method (RB)

This method is similar to the previous method, except that the confidence interval for a quantile is constructed for the rescaled study variable:

- Draw a sample of size n from the population.
- From the sample drawn with the values $\{y_1, y_2, \ldots, y_n\}$, of the study variable y draw a bootstrap sample of any size m, with the values of y $\{y_1^*, y_2^*, \ldots, y_m^*\}$ (subsample).
- Rescale the values of the subsample

$$\widetilde{y}_i = \overline{y} + \sqrt{(1-\frac{n}{N})\frac{m}{n-1}}(y_i^* - \overline{y}), \quad i = 1, 2, \ldots, m, \quad \overline{y} = \frac{1}{n}\sum_{l\in} y_l.$$

- Taking the quantiles of this set $\widehat{K}_q^{\mathrm{RB}}(\alpha/2)$ and $\widehat{K}_q^{\mathrm{RB}}(1-\alpha/2)$ of the levels $\alpha/2$ and $1-\alpha/2$, we get the estimate of the confidence interval of the level $1-\alpha$ for the quantile K_q:

$$\Big[\widehat{K}_q^{\mathrm{RB}}(\alpha/2),\, \widehat{K}_q^{\mathrm{RB}}(1-\alpha/2)\Big].$$

2.2.4 The jackknife method (J)

The procedure is given by:

- The quantile estimate is calculated from the sample eliminating the i-th observation, by means of the same estimator, $\widehat{K}_q^{(i)}$, $i = 1, \ldots, n$.
- n estimates are obtained: $\widehat{K}_q^{(1)}, \widehat{K}_q^{(2)}, \ldots, \widehat{K}_q^{(n)}$.
- Taking the quantiles of this set $\widehat{K}_q^{\mathrm{J}}(\alpha/2)$ and $\widehat{K}_q^{\mathrm{J}}(1-\alpha/2)$ of the levels $\alpha/2$ and $1-\alpha/2$, we get the estimate of the confidence interval of the level $1-\alpha$ for the quantile K_q:

$$\Big[\widehat{K}_q^{\mathrm{J}}(\alpha/2),\, \widehat{K}_q^{\mathrm{J}}(1-\alpha/2)\Big].$$

3 Simulation Study

The data of artificial population of size $N = 1000$ was used for the simulation study. Two collections of variable values were generated.

Case 1. The values of the random variable having the standard normal distribution with the density function

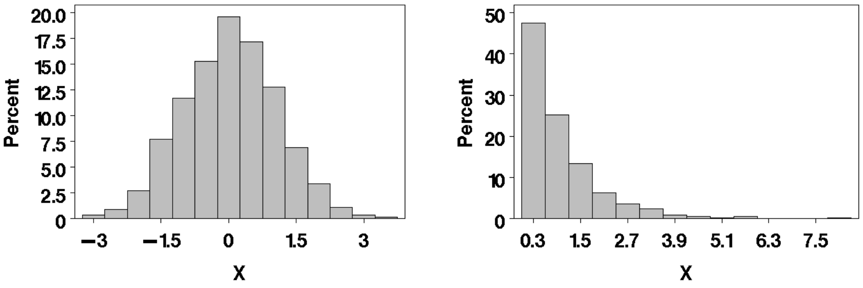
$$f(x) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{x^2}{2}\}, \quad -\infty < x < \infty,$$

where generated.

Case 2. The values of the random variable having the exponential distribution with the density function

$$f(x) = \begin{cases} e^{-x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise}, \end{cases}$$

where generated. The histograms of finite populations are shown in Fig. 2.

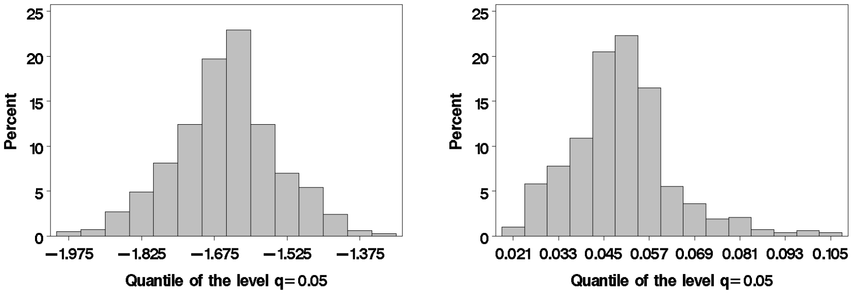


(a) Case 1

(b) Case 2

Figure 2. Histograms of finite populations.

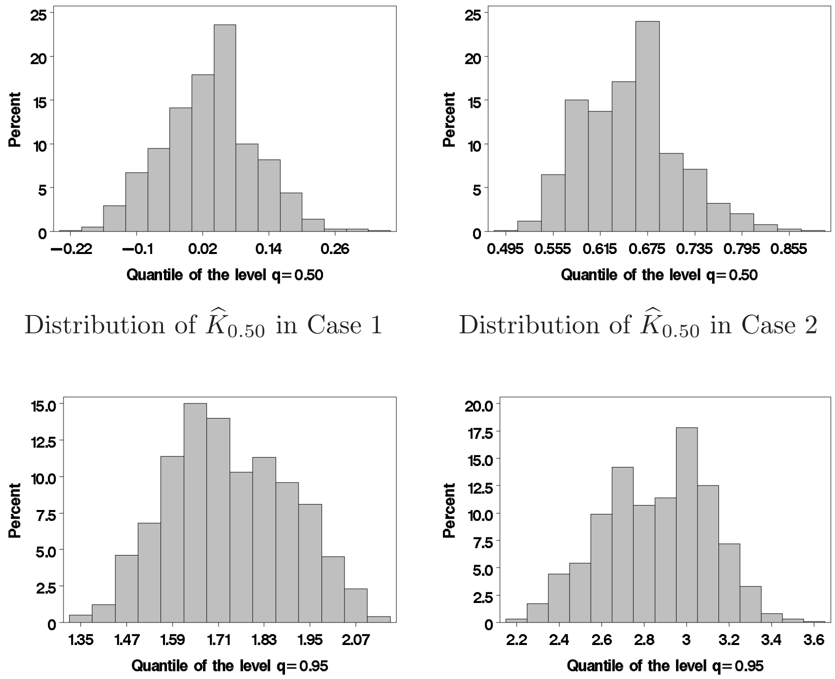
1000 simple random samples of size $n = 200$ have been drawn from a finite population. Given a sample drawn without replacement from a finite population, $B = 1000$ bootstrap samples are drawn of size $m = 200$ with replacement from the original sample in the cases TB and RB. Parameters for which we estimate confidence intervals by different methods are quantiles of the levels q equal to 0.05, 0.25, 0.50, 0.75 and 0.95. The distribution of the quantile estimators are presented in Fig. 3 and Fig. 4.



Distribution of $\widehat{K}_{0.05}$ in Case 1

Distribution of $\widehat{K}_{0.05}$ in Case 2

Figure 3. Distribution of the quantile estimator for the level 0.05. Histograms of 1000 quantile estimates in each case.



Distribution of $\widehat{K}_{0.50}$ in Case 1

Distribution of $\widehat{K}_{0.50}$ in Case 2

Distribution of $\widehat{K}_{0.95}$ in Case 1

Distribution of $\widehat{K}_{0.95}$ in Case 2

Figure 4. Distribution of the quantile estimator for the levels $0.5, 0.95$. Histograms of 1000 quantile estimates in each case.

The confidence level of the confidence interval is $\alpha = 0.1$. The number of coverage of 90% confidence intervals for 1000 samples have been calculated in each case. The averages of the estimated confidence intervals in each case are presented in Table 1 and Fig. 5.

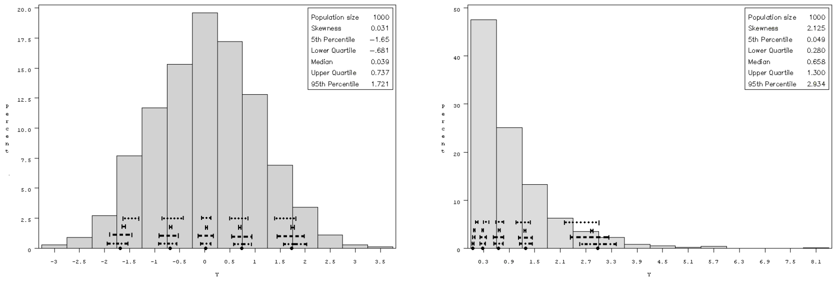
In this figure, dot dash, dashed, solid, and dotted lines represent respectively the averages of the estimated confidence intervals, by using NDB, TB, J, and RB methods (see Sections 2.2.1, 2.2.2, 2.2.3 and 2.2.4). True quantiles $K_{0.05}$, $K_{0.25}$, $K_{0.50}$, $K_{0.75}$, $K_{0.95}$ are indicated on the horizontal axis from left to right by points.

4 Conclusions

The simulation results show that in the case of simple random sampling, the quality of the normal distribution-based method and traditional bootstrap method based estimates of the confidence interval for quantiles is similar. The normal distribution-based method underestimates the confidence interval. The traditional bootstrap method overestimates the confidence interval.

Table 1. The number of coverage of 90% confidence intervals for 1000 samples.

	Case 1				Case 2			
	NDB	TB	RB	J	NDB	TB	RB	J
$K_{0.05}$	918	927	531	79	913	928	0	71
$K_{0.25}$	882	927	845	75	883	917	0	75
$K_{0.50}$	885	916	883	47	885	913	876	47
$K_{0.75}$	893	926	773	63	891	925	856	63
$K_{0.95}$	887	911	727	43	896	904	852	90



(a) Case 1

(b) Case 2

Figure 5. The averages of the estimated confidence intervals.

The estimates of the confidence interval obtained by the jackknife method are too narrow.

The estimates of the confidence intervals for a median, obtained by the rescaling bootstrap method, are more accurate than for other quantiles. The estimates of the confidence intervals obtained by this method are too narrow.

References

[1] W.G. Cochran. *Sampling Techniques*. John Wiley and Sons, New York, 1977.

[2] D. Krapavickaitė. Kvantilio vertinimas baigtinėje populiacijoje. *Lietuvos matematikos rinkinys*, **42**(Special issue):541–547, 2002.

[3] P.J̃. McCarthy. Stratified sampling and distribution-free confidence intervals for a median. *Journal of the American Statistical Association*, **60**:772–783, 1965.

[4] C.-E. Särndal, B. Swensson and J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.

[5] J. Shao and D. Tu. *The Jackknife and Bootstrap*. New-York: Springer-Verlag, 1995.