

NOTE ON A BAHADUR REPRESENTATION OF SAMPLE QUANTILES FROM A FINITE POPULATION

Motoyama, Hitoshi

Institute of Social Sciences, School of Humanities and Social Sciences, Academic Assembly,
Shinshu University

<https://doi.org/10.5109/1798145>

出版情報 : Bulletin of informatics and cybernetics. 46, pp.37-46, 2014-12. Research Association
of Statistical Sciences

バージョン :

権利関係 :



NOTE ON A BAHADUR REPRESENTATION OF SAMPLE
QUANTILES FROM A FINITE POPULATION

by

Hitoshi MOTOYAMA

*Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.46*

—◆◆◆—
FUKUOKA, JAPAN
2014

NOTE ON A BAHADUR REPRESENTATION OF SAMPLE QUANTILES FROM A FINITE POPULATION

By

Hitoshi MOTOYAMA*

Abstract

In this note, we provide the details of the proof of a Bahadur representation of sample quantiles from a finite population.

Key Words and Phrases: Bahadur representation, quantile, finite population, asymptotic normality, survey sampling, empirical distribution function.

1. Introduction

Sample quantiles such as sample medians have been attracting many survey researchers as well as traditional measures such as means, since skewed distributions are very common in socio-economic data. Sample quantiles also have information of the tails of the distribution of the quantity of interest. Therefore, sample quantiles are widely used in practices of survey sampling. For example, estimates of median and other quantiles of yearly income of Japan household are regularly reported by the Japan Statistics Bureau.

Using combinatorial arguments, Thompson (1939) gave design-based exact confidence intervals for the sample median under simple random sampling from a finite population. For other sample quantiles, Wilks (1962) and Sedransk and Smith (1988) have described design-based exact confidence intervals. Their result was extended to various complex surveys such as stratified sampling and cluster sampling (Sedransk and Meyer (1978) and Blesseos (1976)). However, the design-based approach to the construction of confidence intervals is not practical, especially for complex surveys, because of its complicated combinatorial form. So various attempts were made for evaluating the distributions of quantiles and constructing their confidence intervals. As for the results obtained till the early 1980's, we refer the readers to Sedransk and Smith (1988).

One method in evaluating the distributions is asymptotic distribution based approach. For the asymptotic normality of sample quantiles for samples from a finite population, Rosén (1964) first proved the asymptotic normality for simple random sampling without replacement. Francisco and Fuller (1992) derived the asymptotic normality via establishing the Bahadur representation for sample quantiles for stratified cluster sampling from a finite population. Shao (1994) also showed the asymptotic normality via Bahadur representation under stratified multistage sampling under more mild conditions. Motoyama and Takahashi (2008) established the asymptotic normality

* Institute of Social Sciences, School of Humanities and Social Sciences, Academic Assembly, Shinshu University 3-1-1 Asahi, Matsumoto Nagano 390-8621 Japan. E-mail: hitoshi@shinshu-u.ac.jp

by using the methods of statistical functionals. Motoyama (2012) succeeded to prove the asymptotic normality elementarily and included some Monte Carlo evaluations of the confidence intervals. For recent attempt for this problem, we refer the reader to Chatterjee (2011) and the references therein.

In this short note, we shall provide the details of the proof of a Bahadur representation of sample quantiles from a finite population. As we noted above, Francisco and Fuller (1992) proved the Bahadur representation for sample quantiles and Shao (1994) also proved the Bahadur representation under mild conditions. However, the majority of the technical details of the proof in Shao (1994) were left to the reader. This has sometimes discouraged the potential users from acceptance of this powerful tool. The aim of this note is making its proof more accessible to wide users. The proof of this paper is essentially based on the method of the proof of Ghosh (1971) for IID case but the proof includes some new ideas for treating the finite population. The rest of the paper is organized as follows. Section 2. describes definitions and notation. The Bahadur representation and asymptotic normality are established in Section 3. Section 4. presents some Monte Carlo results.

2. Definitions and notation

To fix ideas, let $\{\mathcal{P}_k, k = 1, 2, \dots\}$ be a sequence of finite populations. Throughout the paper k is used as the index of the finite population. Each \mathcal{P}_k has a characteristic x_1, \dots, x_{N_k} of the population size N_k with the population distribution function

$$F_{N_k}(x) = \frac{1}{N_k} \sum_{i=1}^{N_k} I(x_i \leq x),$$

where I is the indicator function such that

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{elsewhere} \end{cases}.$$

Simple random samples X_1, \dots, X_{n_k} of size n_k are chosen from the populations \mathcal{P}_k (Cochran (1977), p.18). More precisely, let $(\pi_1, \dots, \pi_{N_k})$ take all possible permutations of $(1, \dots, N_k)$ with common probability $(N_k!)^{-1}$, and $X_i = x_{\pi_i}, 1 \leq i \leq n_k$.

We define the empirical distribution function of X_1, \dots, X_{n_k} as follows:

$$F_{n_k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} I(X_i \leq x),$$

where I is the indicator function.

Let $\theta_k = F_{N_k}^{-1}(p) = \inf\{x : F_{N_k}(x) \geq p\}$ be the population p th quantile, and let $\hat{\theta}_k = F_{n_k}^{-1}(p) = \inf\{x : F_{n_k}(x) \geq p\}$ be the sample p th quantile. We shall give the proof of the Bahadur representation and asymptotic normality of the sample quantiles.

3. Bahadur representation and asymptotic normality

In what follows, we consider the asymptotics $n_k, N_k - n_k \rightarrow \infty$ as $k \rightarrow \infty$. To prove the Bahadur representation and asymptotic normality of the sample quantiles, we assume that the following assumptions hold.

- (A1) The sequence $\{\theta_k\}$ is bounded.
- (A2) There is a sequence of functions $\{f_k\}$ such that

$$\lim_{k \rightarrow \infty} \left[\frac{F_{N_k}(\theta_k + \delta_k) - p}{\delta_k} - f_k(\theta_k) \right] = 0,$$

for any sequence $\{\delta_k\}$ of order $\sim O(a_k^{-1})$ and

$$0 < \inf_k f_k(\theta_k) \leq \sup_k f_k(\theta_k) < \infty.$$

Here and in the sequel, $a_k = (n_k(N_k - 1)/(N_k - n_k))^{1/2}$.

The assumptions (A1) and (A2) are essentially the same as the assumption (A6) and (A7) of Shao (1994) p.961. Condition (A2) is a kind of smoothness condition so that F_{N_k} is nearly differentiable at θ_k when k is large, although F_{N_k} is not differentiable for each fixed k .

Under these assumptions, we have the Bahadur representation of the sample quantile $\hat{\theta}_k$.

THEOREM 3.1. *Under the assumptions (A1) and (A2), we have as $k \rightarrow \infty$*

$$\hat{\theta}_k = \theta_k + \frac{p - F_n(\theta_k)}{f_k(\theta_k)} + o_p(a_k^{-1}). \quad (1)$$

To prove the theorem, we need the following lemma.

LEMMA 3.2. *(Ghosh (1971) Lemma) Let $\{V_n\}$ and $\{W_n\}$ be two sequence of random variables satisfying the following conditions.*

(I) *For all $\epsilon > 0$ there exists λ depending on ϵ such that $P(|W_n| > \lambda) < \epsilon$ for large $n > n(\epsilon)$.*

(II) *For all y and all $\epsilon > 0$*

$$\begin{aligned} \lim_{n \rightarrow \infty} P(V_n \leq y, W_n \geq y + \epsilon) &= 0 \\ \lim_{n \rightarrow \infty} P(V_n \geq y + \epsilon, W_n \leq y) &= 0. \end{aligned}$$

Then $V_n - W_n \rightarrow 0$ in probability as $n \rightarrow \infty$.

The proof of the lemma is given in Ghosh (1971)(or Rao (1987) p.157, David and Nagaraja (2003) pp.286-287). Now we shall prove the theorem.

PROOF. Define $G_{n_k}(x) = 1 - F_{n_k}(x)$, $G_{N_k}(x) = 1 - F_{N_k}(x)$, and let $V_k = a_k(\hat{\theta}_k - \theta_k)$. Note that the event

$$[V_k \leq t]$$

is equivalent to the event

$$[p \leq F_{n_k}(\theta_k + ta_k^{-1})]$$

which in turn is equivalent to the event

$$[Z_{t,k} \leq t_k],$$

where

$$Z_{t,k} = a_k \{G_{n_k}(\theta_k + ta_k^{-1}) - G_{N_k}(\theta_k + ta_k^{-1})\} / f_k(\theta_k)$$

and

$$t_k = a_k (F_{N_k}(\theta_k + ta_k^{-1}) - p) [f_k(\theta_k)]^{-1}.$$

From the assumptions (A1) and (A2), we have

$$F_{N_k}(\theta_k + ta_k^{-1}) - p - f_k(\theta_k)ta_k^{-1} = o(a_k^{-1}).$$

So, $t_k = a_k [ta_k^{-1}(f_k(\theta_k) + o(1))] / f_k(\theta_k) \rightarrow t$ as $k \rightarrow \infty$. Let

$$W_k = a_k \{G_{n_k}(\theta_k) - G_{N_k}(\theta_k)\} / f_k(\theta_k).$$

Then for every $\epsilon > 0$, we have

$$P(V_k \leq t, W_k \geq t + \epsilon) = P(Z_{t,k} \leq t_k, W_k \geq t + \epsilon) \quad (2)$$

and similarly

$$P(V_k \geq t + \epsilon, W_k \leq t) = P(Z_{t,k} \geq t'_k, W_k \leq t), \quad (3)$$

where

$$\begin{aligned} t'_k &= a_k (F_{N_k}(\theta_k + (t + \epsilon)a_k^{-1}) - p) [f_k(\theta_k)]^{-1} \\ &= a_k [(t + \epsilon)a_k^{-1}(f_k(\theta_k) + o(1))] / f_k(\theta_k) \rightarrow t + \epsilon \quad \text{as } k \rightarrow \infty. \end{aligned}$$

Since

$$\begin{aligned} W_k - Z_{t,k} &= \frac{a_k}{f_k(\theta_k)} \{ (G_{n_k}(\theta_k) - G_{N_k}(\theta_k)) - (G_{n_k}(\theta_k + ta_k^{-1}) - G_{N_k}(\theta_k + ta_k^{-1})) \} \\ &= \frac{a_k}{f_k(\theta_k)} \{ F_{n_k}(\theta_k + ta_k^{-1}) - F_{n_k}(\theta_k) - (F_{N_k}(\theta_k + ta_k^{-1}) - F_{N_k}(\theta_k)) \}, \end{aligned}$$

we have (from the mean and variances of hypergeometric distribution)

$$E[W_k - Z_{t,k}] = 0$$

and

$$\begin{aligned} \text{Var}[W_k - Z_{t,k}] &= E[(W_k - Z_{t,k})^2] = a_k^2 p_k^* (1 - p_k^*) (N_k - n_k) / \{f_k(\theta_k)\}^2 n_k (N_k - 1) \\ &= p_k^* (1 - p_k^*) / \{f_k(\theta_k)\}^2, \end{aligned}$$

where $p_k^* = |F_{N_k}(\theta_k + ta_k^{-1}) - F_{N_k}(\theta_k)| \rightarrow 0$. So $\text{Var}[W_k - Z_{t,k}] \rightarrow 0$, we have

$$W_k - Z_{t,k} \rightarrow 0 \quad \text{in probability.} \quad (4)$$

From the finite population central limit theorem (Erdős and Rényi (1959), Hájek (1960). See also Eeden and Runnenburg (1960).), W_k has an asymptotic normal distribution as $n_k, N_k - n_k \rightarrow \infty$ which ensures the assumption (I) of the lemma. And, by (2), (3) and (4), W_k and V_k satisfy the conditions (II) of the lemma. Thus $V_k - W_k \rightarrow 0$ in probability and the theorem is proved.

Using this representation, we have the following asymptotic normality of sample quantiles.

THEOREM 3.3. *Under the assumptions (A1) and (A2), we have*

$$P(a_k(\hat{\theta}_k - \theta_k)/(p(1-p)/f_k^2(\theta_k))^{1/2} \leq x) \xrightarrow{d} N(0, 1) \quad \text{as } k \rightarrow \infty. \quad (5)$$

PROOF. Using the Bahadur representation,

$$a_k(\hat{\theta}_k - \theta_k)/(p(1-p)/f_k^2(\theta_k))^{1/2} = a_k(p - F_{n_k}(\theta_k))/(p(1-p))^{1/2} + o_p(1)$$

Since $n_{n_k}F_{n_k}(\theta_k)$ has the hypergeometric distribution, $F_{n_k}(\theta_k)$ has the expectation $F_{N_k}(\theta) \xrightarrow{p} p$ and $a_kF_{n_k}(\theta_k)$ has the variance $(F_{N_k}(\theta_k)(1 - F_{N_k}(\theta_k))) \xrightarrow{p} p(1-p)$. Hence, using the finite population central limit theorem (Erdős and Rényi (1959), Hájek (1960). See also Eeden and Runnenburg (1960).) and Slutsky's lemma, we have the desired result.

4. Monte Carlo Simulations

In order to evaluate the fruits of theoretical facts, we compare the empirical distributions of studentized quantiles to the cumulative distribution of the standard normal distribution. (As for the performance of the interval estimations, we refer the readers to Motoyama (2012).) Finite populations of size $N_k = 1000, 5000, 10000$ are generated from the lognormal distribution with mean 3 and standard deviation 0.4 of the distribution on the log scale, and fixed over all simulation runs to observe properties in simple random sampling without replacement from the finite populations. The simulated samples of sampling fractions 10% and 30% are chosen 1000 times repeatedly, then we compare the empirical distributions of studentized quantiles to the cumulative distribution of the standard normal distribution.

In implementing the Monte Carlo simulations, we estimate $1/f_k(\theta_k)$ by

$$\widehat{\frac{1}{f_k(\theta_k)}} = \frac{F_{n_k}^{-1}(p + h_{n_k}) - F_{n_k}^{-1}(p - h_{n_k})}{2h_{n_k}}$$

Such estimators were originally suggested by Siddiqui (1960) and are consistent for $1/f_k(\theta_k)$ under some regularity conditions (e.g. Shao (1994)). In this study, we adopt $h_{n_k} = n_k^{-1/2}$ which is identical with that of Shao (1994).

The simulation results are presented in Figure 1(1st quantiles), Figure 2(2nd quantiles), and Figure 3(3rd quantiles). The left columns of the each figure are for the samples of fraction 10% and the right columns are for the samples of fraction 30%. The 1st rows of the each figure are for the samples from the population of size 1000, the 2nd rows are for the samples from the population of size 5000, and the last rows are for the samples from the population of size 10000. So, as we see the each figure down, we can evaluate effects of the large sample size.

From these figures, we can see the following features: (i) The case when the sampling fraction is 10% is slightly better than the case of a sampling fraction 30% in small sample situations. (ii) As the sample size n_k increases, the normal approximations provide the better approximations to the distributions of the sample quantiles.

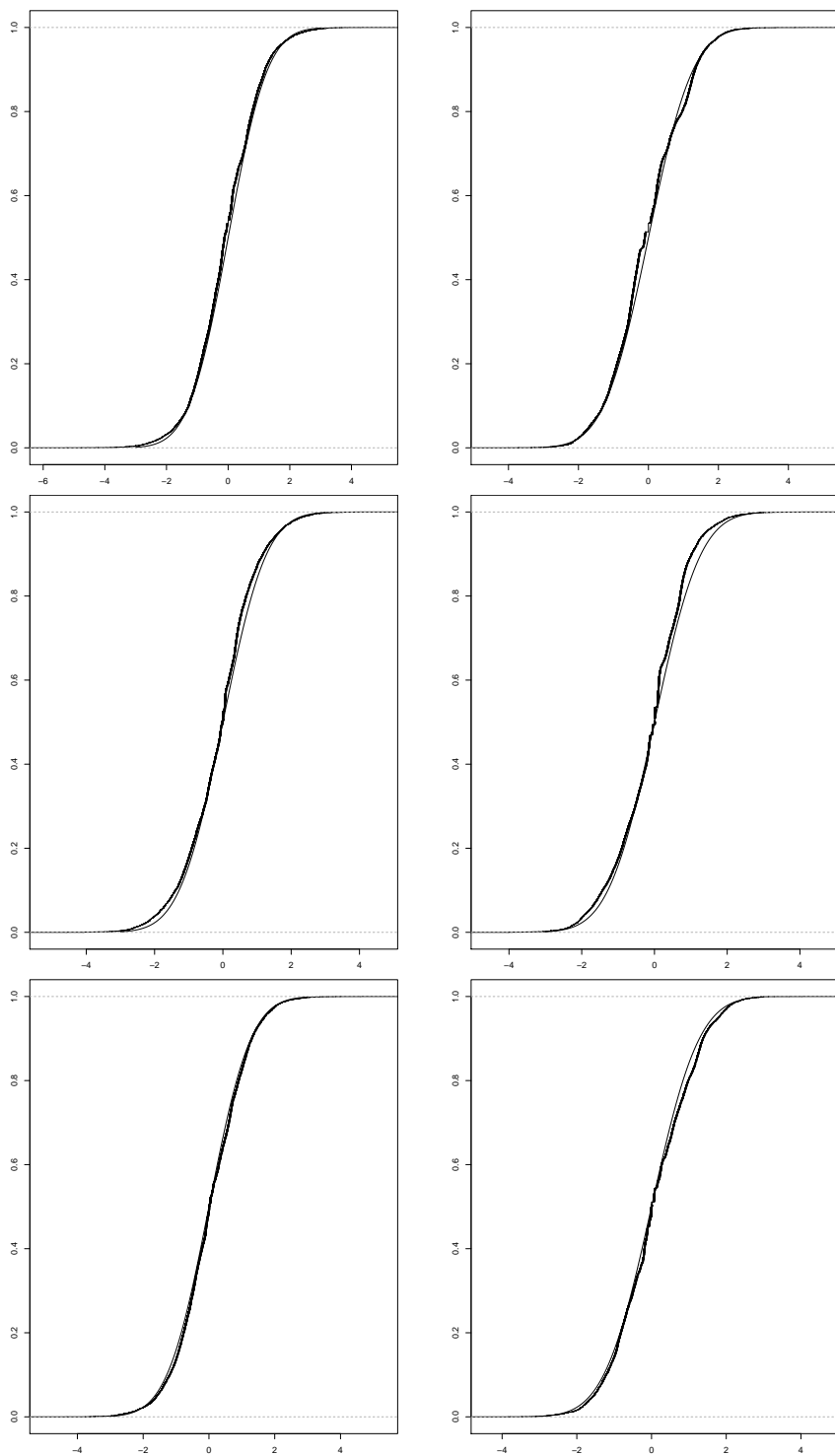


Figure 1: 1st quantiles.

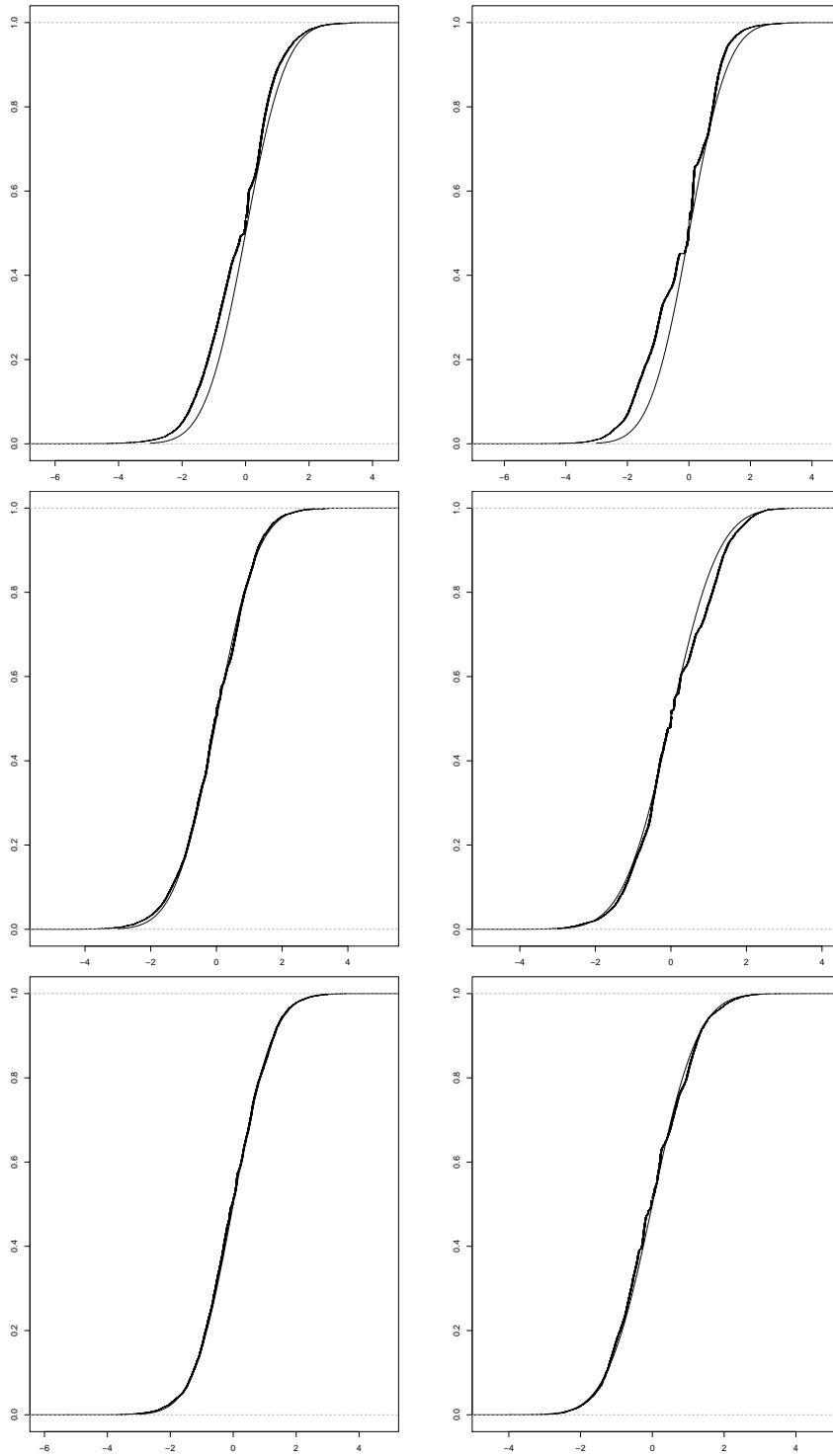


Figure 2: 2nd quantiles.

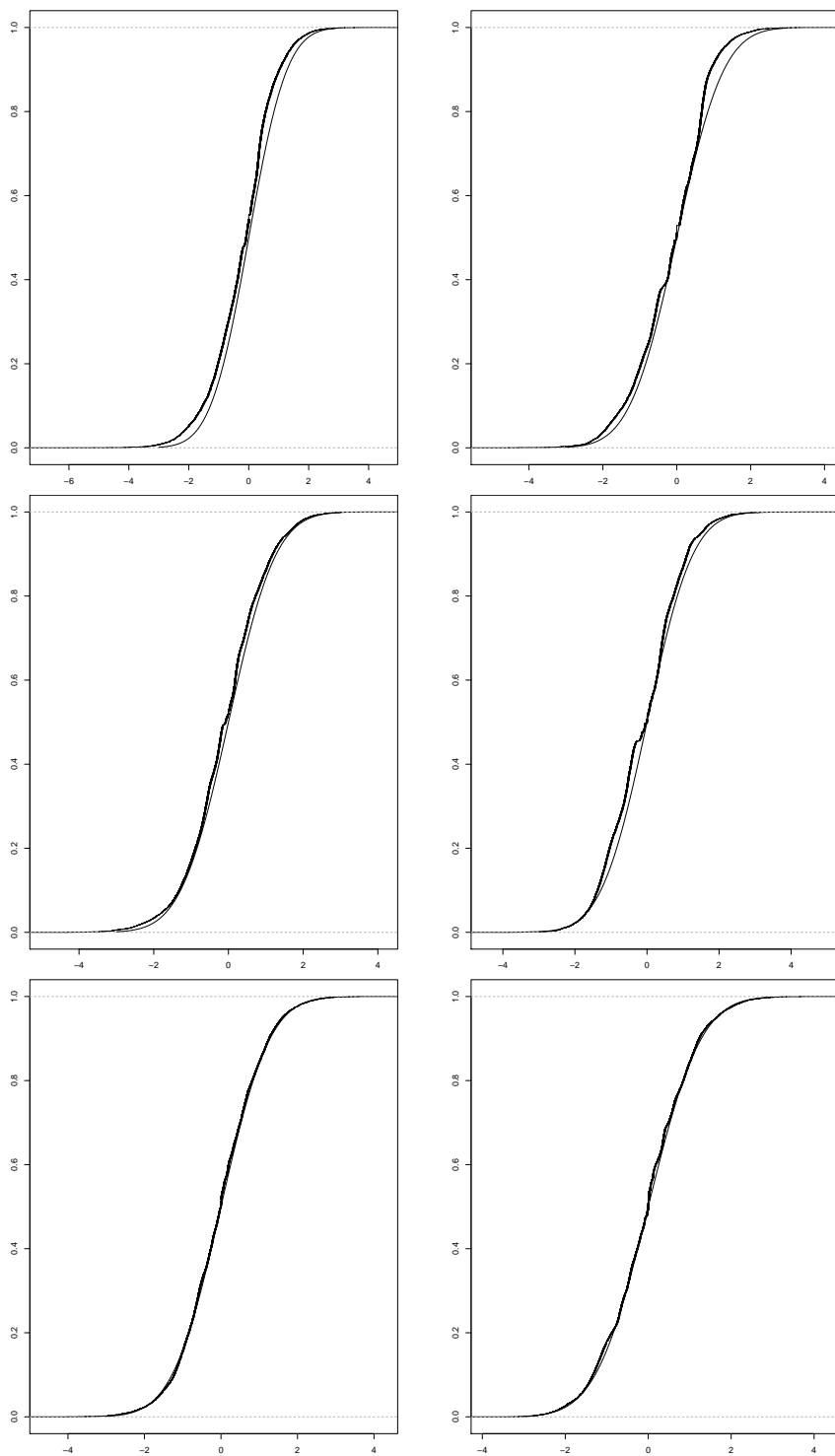


Figure 3: 3rd quantiles.

We can conclude that the normal approximations for the distributions of the sample quantiles are appropriate for large size samples, so they are very useful for application to large scale sample surveys.

Acknowledgement

We would like to thank an anonymous referee for valuable comments and constructive suggestions that helped to much improve the paper. This research is supported in part by the Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology 21730172, 21330048, in part by a research grant from Shinshu University, and in part by the ISM Corporate Research Program (2014-ISM-General Cooperative Research 2-2058).

References

- Blesseos, N.P. (1976). *Distribution-Free Confidence Intervals for Quantiles in Stratified and Cluster Sampling* Ph.D. Thesis, Illinois Institute of Technology.
- Chatterjee, A. (2011). Asymptotic Properties of Sample Quantiles from a Finite Population, *Annals of the Institute of Statistical Mathematics*. **63**, 157-179.
- Cochran, W. G. (1977). *Sampling Techniques*, 3rd. ed. John Wiley & Sons.
- David, H.A. and H.N. Nagaraja (2003). *Order Statistics*, Third ed. John Wiley & Sons.
- Erdős, P. & A. Rényi (1959). On the Central Limit Theorem for Samples from a Finite Population, *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*. **4**, 49-61.
- Eeden, van C. & J.Th. Runnenburg (1960). Conditional Limit-distributions for the Entries in a 2×2 -table, *Statistica Neerlandica*. **14**, 111-126.
- Francisco, C.A. & W.A. Fuller (1992). Quantile Estimation with a Complex Survey Design, *The Annals of Statistics*. **19**, 454-464.
- Ghosh, J.K. (1971). A New Proof of the Bahadur Representation of Quantiles and an Application, *The Annals of Mathematical Statistics*. **42**, 1957-1961.
- Hájek, J. (1960). Limiting Distributions in Simple Random Sampling from a Finite Population, *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*. **5**, 361-374.
- Motoyama, H. & H. Takahashi (2008). Smoothed Versions of Statistical Functionals from a Finite Population, *Journal of the Japan Statistical Society*. **38**, 475-504.
- Motoyama, H. (2012). Note on a Simple Derivation of the Asymptotic Normality of Sample Quantiles from a Finite Population, *Behaviormetrika*. **39**, 1-8.
- Rao, B.L.S.P. (1987). *Asymptotic Theory of Statistical Inference*. John Wiley & Sons.
- Rosén, B. (1964). Limit Theorems for Sampling from Finite Populations, *Arkiv för Matematik*. **5**, 383-424.
- Sedransk, J. & J. Meyer (1978). Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling, *Journal of the Royal Statistical Society. Series B(Methodological)*. **40**, 239-252.

- Sedransk, J. & P.J. Smith (1988). Inference for Finite Population Quantiles, in P.R. Krishnaiah and C.R. Rao, eds., *Handbook of Statistics, Vol.6*. Elsevier Science Publishers B.V., 267-289.
- Shao, J. (1994). L -Statistics in Complex Survey Problem, *The Annals of Statistics*. **22**, 946-967.
- Siddiqui, M.M. (1960). Distribution of Quantiles in Samples from a Bivariate Population. *Journal of Research of the National Bureau of Standards-B. Mathematics and Mathematical Physics*. **64B**, 145-150.
- Wilks, S.S. (1962). *Mathematical Statistics*. John Wiley & Sons.

Received March 14, 2014

Revised October 15, 2014