JIRSS (2015)
Vol. 14, No. 2, pp 53-70
DOI: 10.7508/jirss.2015.02.004

Estimation of E(Y) from a Population with Known Quantiles

Ehsan Zamanzade, Hamed Mohammad Ghasemi

Department of Statistics, University of Isfahan, Isfahan, Iran.

Abstract. In this paper, we consider the problem of estimating $E(Y)$ based on a simple random sample when at least one of the population quantiles is known. We propose a stratified estimator of $E(Y)$, and show that it is strongly consistent. We then establish the asymptotic normality of the suggested estimator, and prove that it is asymptotically more efficient than the standard mean estimator in simple random sampling. For finite sample sizes, Monte Carlo simulation is used to show that the proposed method considerably improves the standard procedure. Finally, a real data example is used to illustrate the application of the proposed method.

Keywords. Mean estimation, Relative efficiency, Monte Carlo simulation.

MSC: 62F10; 62D05.

1 Introduction

The use of auxiliary information to improve statistical inference has been widely discussed in sampling theory. The most widely used auxiliary information are mean, median and coefficient of variation of the auxiliary variable, or correlation coefficient between the auxiliary variable and the variable of interest. Srivastava and Jhajj (1981) considered a class of estimators for the population mean under assumption that mean and variance of auxiliary variable are available. Upadhyaya and Singh (1999) proposed

Ehsan Zamanzade(✉)(e.zamanzade@sci.ui.ac.ir)

two different ratio type estimators for situations in which the coefficient of variation and the coefficient of kurtosis of the auxiliary variable are known. The problem of estimation of variance using information of auxiliary variable has been considered by Das and Tripathi (1978), Isaki (1983) and Yadav and Kadilar (2014).

Assuming that the mean of the auxiliary variable X is known, the ratio estimator of the mean of the variable of interest Y is given by

$$\hat{\mu}_R = \mu^X \left(\frac{\hat{\mu}^Y_{SRS}}{\hat{\mu}^X_{SRS}} \right),$$

where μ^X is the true mean of the auxiliary variable X, and $\hat{\mu}^Y_{SRS}$, $\hat{\mu}^X_{SRS}$ are the sample means of auxiliary variable X and variable of interest Y, respectively.

The ratio estimator $\hat{\mu}_R$ is not unbiased but it generally has less mean square error than $\hat{\mu}^Y_{SRS}$, specially when the variable of interest and auxiliary variable are highly correlated. See Cochran (1977) for more details about the properties of this estimator.

The problem of estimating the population mean when the mean and the first/third quartiles of the auxiliary variable are known, has been considered by Al-Omari (2012). Let μ^X be mean of the auxiliary variable X, Al-Omari (2012) proposed the ratio estimators based on the first (q_1) and the third (q_3) quartiles of X as

$$\hat{\mu}_{SRS1} = \hat{\mu}^Y_{SRS}(\frac{\mu_X + q_1}{\hat{\mu}^X_{SRS} + q_1}), \ \hat{\mu}_{SRS3} = \hat{\mu}^Y_{SRS}(\frac{\mu_X + q_3}{\hat{\mu}^X_{SRS} + q_3}),$$

respectively, where $\hat{\mu}^X_{SRS}$ and $\hat{\mu}^Y_{SRS}$ are the sample means of auxiliary variable X and variable of interest Y, respectively.

In the case of μ^X being unknown, Al-Omari (2012) proposed to use double sampling method to estimate μ^X. In doing so, one first selects a large sample size n' to estimate μ^X, then a sub-sample of size n'' is selected from the population of interest to compute $\hat{\mu}^Y_{SRS}$.

Breidt (2004) considered control variate method for improving the efficiency of quantile estimation when the population of interest has known mean. He proposed to estimate population quantile $\left(Q_p\right)$ by

$$\tilde{Q}_p = \hat{Q}_p + \left(\mu^Y - \hat{\mu}^Y_{SRS}\right),$$

where μ^Y is the known mean of the population of interest, $\hat{Q}_p$ and $\hat{\mu}^Y_{SRS}$ are sample quantile and sample mean based on a simple random sample, respectively.

In this paper, we consider the problem of estimating the population mean when at least one of the population quantiles is known under the infinite population setting. We introduce a stratified mean estimator in which the strata are based on the quantiles information. The proposed method is different from standard post stratification (see Lohr (1999)) in that the strata are based on the quantiles information rather than covariate information. The mathematical development is similar to judgment post stratification due to MacEachern et al. (2004).

It is worth mentioning that the results presented here are also applicable in estimating of $E\left(h\left(Y\right)\right)$ instead of $E(Y)$ as long as h is monotone and the variance of $h(Y)$ exists. Examples of $h(Y)$ include $h(Y) = Y^l, l = 1, 2, \ldots$, corresponding to estimation of the population moments for random variables with non-negative supports, and $h(Y) = I_{\{Y \leq c\}}$ corresponding to estimation of distribution function.

The rest of the paper is organized as follows. In Section 2 of the paper, we propose a non-parametric mean estimator for the case that at least one of the population quantiles is known. We then prove that the proposed estimator is strongly consistent. We also establish its asymptotic normality and show that it is asymptotically more efficient than the standard mean estimator in simple random sampling. In Section 3, we examine the performance of the introduced estimator for finite sample sizes via Monte Carlo simulation. In Section 4, the application of the proposed method is illustrated using a real data set. We end in Section 5 with a summary.

2 Estimation of $E(Y)$

Let Y be the variable of interest with distribution function F. The pth order quantile of random variable of Y is defined as

$$Q_p = inf\{y : F(y) \geq p\}.$$

We adopt the following notations in the rest of the paper. Let $Y_1, \ldots, Y_n$ be a simple random sample of size n from a population with distribution function F, mean μ and variance σ^2. Let $\mathbf{Q}_0 = (Q_{p_1} \ldots, Q_{p_k})$ be the vector of known quantiles of the population of interest, and $m = k + 1$. Let $T_1, \ldots, T_n$ be auxiliary random variables corresponding to the simple random sample of $Y_1, \ldots, Y_n$, where $T_i = j$ if $Q_{p_{j-1}} < Y_i < Q_{p_j}$, and zero otherwise, for $i = 1 \ldots n$; $j = 1, \ldots m$, where $p_0 = 0$ and $p_m = 1$. Therefore the simple random sample of $Y_1, \ldots, Y_n$ may be represented as $(Y_1, T_1), \ldots, (Y_n, T_n)$. We define $\mu_j = E(Y_i|T_i = j)$ and $\sigma_j^2 = var(Y_i|T_i = j)$, for $j = 1, \ldots, m$. Since the pairs (Y_i, T_i), $i = 1, \ldots, n$, form a random sample, μ_j and σ_j^2 do not depend on i. Note that μ_j and

σ_j^2 are in fact, the mean and the variance of the random variable Y, which is truncated between $Q_{p_{j-1}}$ and Q_{p_j}, respectively.

Let I_{ij} be one if $T_i = j$ and zero otherwise. Then, the number of observations which are between $Q_{p_{j-1}}$ and Q_{p_j} is denoted by $N_j = \sum_{i=1}^{n} I_{ij}$. One can simply show that the vector $\mathbf{N} = (N_1, \ldots, N_m)$ follows a multinomial distribution with mass parameter n and probability vector $\mathbf{d} = (d_1, \ldots, d_m)$, where $d_j = p_j - p_{j-1}$, for $j = 1, \ldots, m$. Therefore, the probability that some of the N_j be zero is positive. Let $I_j = 1$ if $N_j > 0$ and zero otherwise and $S_n = \sum_{j=1}^{m} d_j I_j$. Furthermore, let

$$J_j = \begin{cases} 0 & N_j = 0 \\ \frac{1}{N_j} & N_j > 0. \end{cases}$$

The following lemma states that the population mean can be expressed in terms of the weighted average of truncated means.

Lemma 2.1. *The population mean satisfies* $\mu = \sum_{j=1}^{m} d_j \mu_j$.

Proof. We have

$$\begin{aligned} \mu &= E(E(Y_1|T_1)) \\ &= \sum_{j=1}^{m} P(T_1 = j) E(Y_1|_1 T = j) \\ &= \sum_{j=1}^{m} d_j \mu_j. \quad \square \end{aligned}$$

□

The next lemma shows that how population variance σ^2 can be expressed as weighted averages of variances of between truncated component and within truncated component.

Lemma 2.2. *The population variance satisfies*

$$\sigma^2 = \sum_{j=1}^{m} d_j \sigma_j^2 + \sum_{j=1}^{m} d_j (\mu_j - \mu)^2$$

Proof. We can write

$$\begin{aligned}\sigma^2 &= E(Var(Y_1|T_1)) + Var(E(Y_1|T_1)) \\ &= \sum_{j=1}^{m} d_j \sigma_j^2 + \sum_{j=1}^{m} d_j(\mu_j - \mu)^2. \quad \square\end{aligned}$$

□

Suppose that, we want to draw inference for the population mean μ based on (Y_i, T_i). Consider

$$L(\mu) = \sum_{j=1}^{m} \frac{d_j I_j J_j}{S_n} \sum_{i=1}^{n} (Y_i - \mu)^2 I_{ij},$$

as loss function for μ. In order to justify using this loss function, note that $J_j \sum_{i=1}^{n} (Y_i - \mu)^2 I_{ij}$ can be regarded as mean squared error loss function of the sample units which fall between $Q_{p_{j-1}}$ and Q_{p_j}, for $j = 1, \ldots, m$. Therefore, $L(\mu)$ is in fact the weighted average of those mean squared error loss functions.

We propose to estimate the parameter of μ by the minimizing $L(\mu)$ with respect to μ, which results in the estimator

$$\hat{\mu}_S = \sum_{j=1}^{m} w_j \hat{\mu}_j,$$

where $w_j = \frac{d_j I_j}{S_n}$, and $\hat{\mu}_j$ is the mean of the observations between $Q_{p_{j-1}}$ and Q_{p_j}.

In the next theorem, we show that the proposed estimator is strongly consistent.

Theorem 2.1. *$\hat{\mu}_S$ is a strongly consistent estimator of the population mean.*

Proof. Note that $I(N_j > 0) \xrightarrow{a.s.} 1$ as n tends to infinity. So it turns out from Theorem 11.1 in (Gut , 2005, p 247) that $w_j \xrightarrow{a.s.} d_j$, for $j = 1, \ldots, m$.
Since $\hat{\mu}_j$ is the sample mean of the observations which fall between $Q_{P_{j-1}}$ and Q_{P_j}, it follows from strong law of large numbers that $\hat{\mu}_j$ is a strongly consistent estimator of μ_j. Therefore, the theorem is proven by Lemma 2.1. □

The next theorem establishes the asymptotic normality of the proposed mean estimator.

Theorem 2.2. *Let $Y_1, \ldots, Y_n$ be a simple random sample of size n and $\mathbf{Q}_0 = (Q_{P_1}, \ldots, Q_{P_k})$ be the vector of known quantiles of the population. As the sample size n approaches to infinity $\sqrt{n}\,(\hat{\mu}_S - \mu)$ converges to a normal distribution with mean zero and variance $\sum_{j=1}^{m} d_i \sigma_j^2$.*

Proof. Note that

$$\sqrt{n}(\hat{\mu}_S - \mu) = \sqrt{n}\sum_{j=1}^{m} w_j(\hat{\mu}_j - \mu_j) + \sqrt{n}\sum_{j=1}^{m} \mu_j(w_j - d_j).$$

On the other hand, since by conditioning on $\mathbf{N} = (N_1, \ldots, N_m)$, $\hat{\mu}_1, \ldots, \hat{\mu}_m$ are independent random variables. Thus, for $(t_1, \ldots, t_m) \in \mathbb{R}^m$, we can write

$$\begin{aligned} P(\bigcap_{j=1}^{m} \sqrt{N_j}(\hat{\mu}_j - \mu_j) \leq t_j) &= E\{P(\bigcap_{j=1}^{m} \sqrt{N_j}(\hat{\mu}_j - \mu_j) \leq t_j|\mathbf{N})\} \\ &= E\{\prod_{j=1}^{m} P(\sqrt{N_j}(\hat{\mu}_j - \mu_j)) \leq t_j|\mathbf{N})\} \\ &\longrightarrow E\{\prod_{j=1}^{m} P(Z_j \leq t_j)\} \\ &= P(\bigcap_{j=1}^{m} Z_j \leq t_j), \end{aligned}$$

where Z_j follows a normal distribution with mean zero and variance σ_j^2. Thus, we can conclude that the vector

$$\mathbf{U}^T = (\sqrt{N_1}\,(\hat{\mu}_1 - \mu_1), \ldots, \sqrt{N_m}(\hat{\mu}_m - \mu_m))$$

converges in distribution to an m-dimensional normal distribution with mean zero and variance covariance matrix $\mathbf{\Sigma}$, where $\mathbf{\Sigma}$ is a diagonal matrix of the vector $\left(\sigma_1^2, \ldots, \sigma_m^2\right)$. Let $\mathbf{C}_n = \left(\sqrt{\frac{n}{N_1}}w_1, \ldots, \sqrt{\frac{n}{N_m}}w_m\right)$, then it is clear that $\mathbf{C}_n$ converges in probability to the vector $\mathbf{C} = \left(\sqrt{d_1}, \ldots, \sqrt{d_m}\right)$. So, by using Slutsky's theorem we have

$$\sqrt{n}\sum_{j=1}^{m} w_i\left(\hat{\mu}_j - \mu_j\right) = \mathbf{C}_n\mathbf{U} \xrightarrow{d} N(0, \sum_{j=1}^{m} d_j\sigma_j^2).$$

It remains to show that $\sqrt{n}\sum_{j=1}^{m}\mu_j(w_j - d_j) \xrightarrow{p} 0$. To see this, note that since $I\left(N_j > 0\right) \xrightarrow{a.s.} 1$, then for large enough values of n, we have $\mid w_j - d_j \mid \leq d_j o\left(\left(1-d_j\right)^n\right)$. Therefore, $\forall\epsilon > 0$ we have

$$P\left(\sqrt{n} \mid \mu_j\left(w_j - d_j\right) \mid \geq \epsilon\right) \leq \frac{\sqrt{n} \mid \mu_j \mid E\left(\mid w_j - d_j \mid\right)}{\epsilon}$$

$$\leq \frac{\sqrt{n} \mid \mu_j \mid}{\epsilon\left(1-d_j\right)^n} \longrightarrow 0 \quad as \quad n \to \infty,$$

and this completes the proof. □

Remark 1. One can conclude from above theorem and Lemma 2.2 that $\hat{\mu}_S$ is asymptotically more efficient than the standard mean estimator in simple random sampling, $\hat{\mu}_{SRS}$. To see this, note that according to the central limit theorem $\sqrt{n}\left(\hat{\mu}_{SRS} - \mu\right)$ converges to a normal distribution with mean zero and variance σ^2. On the other hand, it follows from Lemma 2.2 that $\sum_{j=1}^{m} d_j\sigma_j^2 \leq \sigma^2$, and thus $\hat{\mu}_S$ asymptotically more efficient than $\hat{\mu}_{SRS}$.

For finite sample sizes, the proposed mean estimator is not unbiased in general and the computation of its bias and variance requires tedious calculations which depends on the distribution of the vector $(w_1, \ldots, w_m)$. However, in the special case that the vector $(d_1, \ldots, d_m)$ being $\left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$, we show that $\hat{\mu}_S$ is unbiased and has less variance than standard mean estimator. The following lemma states some distributional properties of $(w_1, \ldots, w_m)$ in this case.

Lemma 2.3. *Let* $(d_1, \ldots, d_m) = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$, *then*

(I) $E\left(w_1\right) = \frac{1}{m}$,

(II) $Var(w_1) = \frac{1}{m^2}\sum_{j=1}^{m-1}\left(\frac{j}{m}\right)^{n-1}$,

(III) $Cov(w_1, w_2) = \frac{-1}{m-1}Var(w_1)$,

(IV) $\frac{nm^2}{m-1}Var(w_1) < 1$, $\forall n \geq 3$,

(V) $E\left(J_1 w_1^2\right) = \frac{1}{m^n}\left(\frac{1}{n} + \sum_{k=2}^{m}\sum_{j=1}^{k-1}\sum_{t=1}^{n-k-1}\frac{(-1)^{j-1}}{k^2 t}\binom{m-1}{k-1}\binom{k-1}{j-1}\binom{n}{t}(k-j)^{n-t}\right)$.

The proof follows from Lemma 4 in Dastbaravarde et al. (2016).

Theorem 2.3. *Let $Y_1, \ldots, Y_n$ be a simple random sample of size n and $\mathbf{Q}_0 = (Q_{P_1}, \ldots, Q_{P_k})$ be the vector of known quantiles of the population. Let $(d_1, \ldots, d_m) = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$, then $\hat{\mu}_S$ is an unbiased estimator of the population mean and its variance is given by*

$$E({J_1 w_1}^2) \sum_{j=1}^{m} \sigma_j^2 + \frac{m}{m-1} Var(w_1) (\sum_{j=1}^{m} (\mu_j - \mu)^2).$$

Proof. One can write

$$\begin{aligned} E(\hat{\mu}_S) &= E(\sum_{j=1}^{m} E(w_j \hat{\mu}_j | \mathbf{T})) \\ &= E(\sum_{j=1}^{m} w_j \mu_j) \\ &= E(w_1) \sum_{j=1}^{m} \mu_j \\ &= \frac{1}{m} \sum_{j=1}^{m} \mu_j = \mu. \end{aligned}$$

The third equality holds because in the case that

$$(d_1, \ldots, d_m) = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right),$$

w_j's are identically distributed.

The variance of $\hat{\mu}_S$ can be written as

$$\begin{aligned}
Var(\sum_{j=1}^{m} w_j\hat{\mu}_j) &= E(Var(\sum_{j=1}^{m} w_j\hat{\mu}_j|\mathbf{T})) + Var(E(\sum_{j=1}^{m} w_j\hat{\mu}_j|\mathbf{T}))\\
&= E(\sum_{j=1}^{m} w_j^2 J_j\sigma_j^2) + Var(\sum_{j=1}^{m} w_j\mu_j)\\
&= E(w_1^2 J_1)\sum_{j=1}^{m}\sigma_j^2 + \sum_{j=1}^{m}\mu_j^2 Var(w_j) + \sum_{j_1\neq j_2}^{m}\mu_{[j_1]}\mu_{[j_2]}Cov(w_{[j_1]}, w_{[j_2]})\\
&= E(w_1^2 J_1)\sum_{j=1}^{m}\sigma_j^2 + Var(w_1)\sum_{j=1}^{m}\mu_j^2 - \frac{1}{m-1}Var(w_1)(m^2\mu^2 - \sum_{j=1}^{m}\mu_j^2)\\
&= E(w_1^2 J_1)\sum_{j=1}^{m}\sigma_j^2 + \frac{m}{m-1}Var(w_1)(\sum_{j=1}^{m}(\mu_j - \mu)^2).
\end{aligned}$$

The third equality holds because in the case that $(d_1, \ldots, d_m) = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$, w_j's and $J_j w_j^2$'s are both identically distributed. □

Corollary 2.1. One can conclude from Theorem 2.3 that in the case that $(d_1, \ldots, d_m) = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$, the variance of the proposed mean estimator is less than the variance of the standard mean estimator in simple random sampling. To see this, note that

$$\begin{aligned}
&Var(\sum_{j=1}^{m} w_j\hat{\mu}_j) - Var(\frac{1}{n}\sum_{i=1}^{n} Y_i)\\
&= (mE(J_1 w_1^2) - \frac{1}{n})\frac{1}{m}\sum_{j=1}^{m}\sigma_j^2 + (\frac{m^2}{m-1}Var(w_1) - \frac{1}{n})\frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu)^2.
\end{aligned}$$

Besides, $\frac{m^2}{m-1}Var(w_1) \leq \frac{1}{n}$, as it was shown in Lemma 2.3. On the other hand, by using Cauchy Schwarz inequality and Lemma 2.3, one can write $nmE(J_1 w_1^2) = m^2 E(N_1)E(J_1 w_1^2) \geq m^2(E(w_1)^2) = 1$, and this gives the result.

3 Simulation study

In this section, we compare the performance of the proposed mean estimator with that of $\hat{\mu}_{SRS}$ via Monte Carlo simulation. For this purpose, we have generated 1,000,000 random samples of sizes $n = 5, 10, 20, 30, 40, 50$ from six distributions: standard normal distribution ($N(0,1)$), standard exponential distribution ($Exp(1)$), uniform distribution ($U(0,1)$), gamma distribution with scale parameter 1 and shape parameter 5 ($Gamma(5)$), beta distribution with parameters 0.5, 0.5 ($Beta(0.5, 0.5)$), and Student's t distribution with 3 degrees of freedom ($t3$). So, we allow the sample size to be small ($n = 5, 10$), moderate ($n = 20, 30$) and large ($n = 40, 50$). Furthermore, symmetric and skewed distributions, heavy-tailed and light-tailed distributions and distributions with bounded and unbounded supports are all considered in the simulation study. We assume that at least one of the population quartiles is known. The efficiency of $\hat{\mu}_S$ relative to $\hat{\mu}_{SRS}$ is defined as

$$\widehat{RE} = \frac{\widehat{Var}(\hat{\mu}_{SRS})}{\widehat{MSE}(\hat{\mu}_S)}.$$

Thus, $\widehat{RE} > 1$ indicates that the proposed estimator has better performance than $\hat{\mu}_{SRS}$. Furthermore, since $\hat{\mu}_S$ is not unbiased in general, we also report the estimated bias of $\hat{\mu}_S$. The simulation results are presented in Tables 1-3.

Table 1: Estimated relative efficiency ($\widehat{RE}$) and estimated bias of the proposed mean estimator when parent distributions are standard normal and standard exponential.

		Sample size (n)											
		5		10		20		30		40		50	
Parent distribution	Known quartile(s)($\mathbf{Q}_0$)	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$
$N(0,1)$	Q_1	1.31	0.09	1.65	0.02	2.02	0.00	2.10	0.00	2.13	0.00	2.13	0.00
	Q_2	1.57	0.00	2.33	0.00	2.59	0.00	2.65	0.00	2.68	0.00	2.69	0.00
	Q_3	1.31	-0.09	1.66	-0.02	2.01	0.00	2.10	0.00	2.12	0.00	2.13	0.00
	Q_1, Q_2	1.37	0.10	2.37	0.02	3.43	0.00	3.71	0.00	3.80	0.00	3.84	0.00
	Q_1, Q_3	1.67	0.00	2.32	0.00	4.07	0.00	4.69	0.00	4.87	0.00	4.96	0.00
	Q_2, Q_3	1.37	-0.10	2.36	-0.02	3.43	0.00	3.72	0.00	3.80	0.00	3.84	0.00
	Q_1, Q_2, Q_3	1.40	0.00	2.49	0.00	5.12	0.00	6.22	0.00	6.53	0.00	6.68	0.00
$Exp(1)$	Q_1	0.96	0.06	1.17	0.01	1.29	0.00	1.31	0.00	1.32	0.00	1.32	0.00
	Q_2	1.28	0.00	1.65	0.00	1.82	0.00	1.85	0.00	1.88	0.00	1.88	0.00
	Q_3	1.74	-0.10	1.90	-0.02	2.33	0.00	2.52	0.00	2.60	0.00	2.63	0.00
	Q_1, Q_2	0.98	0.09	1.46	0.02	1.83	0.00	1.90	0.00	1.92	0.00	1.93	0.00
	Q_1, Q_3	1.58	-0.03	1.90	-0.01	2.66	0.00	2.96	0.00	3.09	0.00	3.14	0.00
	Q_2, Q_3	1.58	-0.09	2.15	-0.02	2.86	0.00	3.15	0.00	3.29	0.00	3.35	0.00
	Q_1, Q_2, Q_3	1.28	0.00	1.89	0.00	2.91	0.00	3.31	0.00	3.45	0.00	3.52	0.00

Table 2: Estimated relative efficiency ($\widehat{RE}$) and estimated bias of the proposed mean estimator when parent distributions are uniform and gamma with scale parameter 1 and shape parameter 5.

		Sample size (n)											
		5		10		20		30		40		50	
Parent distribution	Known quartile(s)($\mathbf{Q}_0$)	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$
$U(0,1)$	Q_1	1.30	0.02	1.73	0.00	2.16	0.00	2.24	0.00	2.25	0.00	2.26	0.00
	Q_2	1.86	0.00	3.31	0.00	3.77	0.00	3.85	0.00	3.89	0.00	3.91	0.00
	Q_3	1.31	-0.02	1.73	0.00	2.17	0.00	2.24	0.00	2.25	0.00	2.26	0.00
	Q_1, Q_2	1.50	0.03	3.13	0.00	5.40	0.00	6.02	0.00	6.15	0.00	6.20	0.00
	Q_1, Q_3	1.62	0.00	2.52	0.00	5.13	0.00	5.97	0.00	6.14	0.00	6.21	0.00
	Q_2, Q_3	1.50	-0.03	3.12	0.00	5.42	0.00	6.01	0.00	6.14	0.00	6.21	0.00
	Q_1, Q_2, Q_3	1.49	0.00	3.12	0.00	9.75	0.00	13.59	0.00	14.63	0.00	14.95	0.00
$Gamma(5)$	Q_1	1.12	0.19	1.40	0.04	1.63	0.00	1.67	0.00	1.68	0.00	1.69	0.00
	Q_2	1.50	0.00	2.14	0.00	2.37	0.00	2.42	0.00	2.44	0.00	2.46	0.00
	Q_3	1.53	-0.23	1.86	-0.05	2.36	0.00	2.49	0.00	2.54	0.00	2.56	0.00
	Q_1, Q_2	1.18	0.23	1.95	0.06	2.65	0.00	2.80	0.00	2.83	0.00	2.86	0.00
	Q_1, Q_3	1.65	-0.03	2.21	-0.01	3.65	0.00	4.16	0.00	4.30	0.00	4.39	0.00
	Q_2, Q_3	1.52	-0.23	2.51	-0.06	3.65	0.00	4.04	0.00	4.18	0.00	4.27	0.00
	Q_1, Q_2, Q_3	1.37	0.00	2.33	0.00	4.37	0.00	5.20	0.00	5.45	0.00	5.57	0.00

Table 3: Estimated relative efficiency ($\widehat{RE}$) and estimated bias of the proposed mean estimator when parent distributions are beta distribution with parameters 0.5, 0.5 and t-distribution with 5 degrees of freedom.

		Sample size (n)											
		5		10		20		30		40		50	
Parent distribution	Known quartile(s)($\mathbf{Q}_0$)	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$	$\widehat{RE}$	$\widehat{Bias}$
$Beta(0.5, 0.5)$	Q_1	1.26	0.03	1.68	0.00	2.07	0.00	2.14	0.00	2.15	0.00	2.16	0.00
	Q_2	2.07	0.00	4.28	0.00	4.98	0.00	5.09	0.00	5.14	0.00	5.18	0.00
	Q_3	1.27	-0.03	1.67	0.00	2.07	0.00	2.15	0.00	2.16	0.00	2.16	0.00
	Q_1, Q_2	1.57	0.04	3.60	0.01	6.90	0.00	7.83	0.00	8.00	0.00	8.01	0.00
	Q_1, Q_3	1.50	0.00	2.32	0.00	4.43	0.00	5.02	0.00	5.11	0.00	5.17	0.00
	Q_2, Q_3	1.57	-0.04	3.60	-0.01	6.88	0.00	7.80	0.00	8.00	0.00	8.10	0.00
	Q_1, Q_2, Q_3	1.50	0.00	3.25	0.00	11.34	0.00	16.77	0.00	18.11	0.00	18.55	0.00
$t3$	Q_1	1.23	0.14	1.33	0.03	1.48	0.00	1.52	0.00	1.56	0.00	1.56	0.00
	Q_2	1.22	0.00	1.43	0.00	1.58	0.00	1.60	0.00	1.64	0.00	1.65	0.00
	Q_3	1.23	-0.14	1.34	-0.03	1.46	0.00	1.53	0.00	1.56	0.00	1.56	0.00
	Q_1, Q_2	1.16	0.15	1.44	0.04	1.75	0.00	1.83	0.00	1.90	0.00	1.92	0.00
	Q_1, Q_3	1.49	0.00	1.53	0.00	1.86	0.00	2.07	0.00	2.10	0.00	2.18	0.00
	Q_2, Q_3	1.12	-0.15	1.46	-0.04	1.75	0.00	1.85	0.00	1.87	0.00	1.89	0.00
	Q_1, Q_2, Q_3	1.17	0.00	1.46	0.00	1.92	0.00	2.12	0.00	2.24	0.00	2.29	0.00

Table 1 gives the results when the parent distributions are standard normal and standard exponential, respectively. It is clear from this table that, when the parent distribution is $N(0,1)$, the proposed estimator is always better than $\hat{\mu}_{SRS}$. It is worth noting that $\hat{\mu}_S$ is almost unbiased except when the sample size is too small ($n = 5$). For this distribution, the best performance of $\hat{\mu}_S$ for $n = 5$ occurs when $\mathbf{Q}_0 = (Q_1, Q_3)$ is the vector of known quartiles. The largest REs for $n \geq 10$ are achieved when $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$ is the vector of known quartiles. In the case that standard exponential to be parent distribution, $\hat{\mu}_S$ is the best estimator except for the cases that the sample size to be too small ($n = 5$) and $\mathbf{Q}_0 = (Q_1)$, $\mathbf{Q}_0 = (Q_1, Q_2)$ to be vector of known quartiles where the performance of $\hat{\mu}_S$ is slightly worse than $\hat{\mu}_{SRS}$. The proposed estimator is almost unbiased for $n \geq 10$. In this case, the highest relative efficiency for $n = 5$, $n = 10$ and $n \geq 20$ happens when $\mathbf{Q}_0 = (Q_3)$, $\mathbf{Q}_0 = (Q_2, Q_3)$ and $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$ to be vector of known quartiles, respectively.

The simulation results for uniform and gamma distributions are presented in Table 2. In the case of $U(0,1)$ being the parent distribution, the proposed estimator is almost unbiased and always superior to $\hat{\mu}_{SRS}$. For small sample sizes ($n = 5, 10$), the best performance of $\hat{\mu}_S$ is obtained for $\mathbf{Q}_0 = (Q_2)$. For moderate and large sample sizes ($n \geq 20$), the largest REs are occurred when $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$ to be the vector of known quartiles. In the case of $Gamma(5)$ being the parent distribution, $\hat{\mu}_S$ is biased for $n = 5$ and almost unbiased for $n \geq 10$. The relative efficiency is always larger than one. The best performance of $\hat{\mu}_S$ happens when $\mathbf{Q}_0 = (Q_1, Q_3)$, $\mathbf{Q}_0 = (Q_2, Q_3)$ and $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$ to be vector of known quartiles for $n = 5$, $n = 10$, and $n \geq 20$, respectively.

Table 3 presents the simulation results for $Beta(0.5, 0.5)$ and $t3$ distributions. For $Beta(0.5, 0.5)$ distribution, $\hat{\mu}_S$ is almost unbiased. We observe that for this distribution the proposed estimator is always better than $\hat{\mu}_{SRS}$, and the improvement over $\hat{\mu}_{SRS}$ is considerable for $n \geq 20$ and $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$. In this case, the best performance of $\hat{\mu}_S$ is occurred when $\mathbf{Q}_0 = (Q_2)$ to be the vector of known quartiles for $n = 5, 10$ and $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$ for $n \geq 20$. In the case of $t3$ being the parent distribution, $\hat{\mu}_S$ is biased for $n = 5$ and almost unbiased for $n \geq 10$ and always superior to $\hat{\mu}_{SRS}$. For this distribution, the best performance of $\hat{\mu}_S$ is obtained when $\mathbf{Q}_0 = (Q_1, Q_3)$, $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$ for $n \leq 10$ and $n \geq 20$, respectively.

It is worth mentioning that, for all considered cases, the relative efficiency increases as the sample size increases while the other parameters are fixed. The proposed estimator is almost unbiased when the sample size is not too small ($n \geq 10$). The highest relative efficiency for $n \geq 20$ happens when $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$ to be the vector of known quartiles for all considered distributions. Furthermore, note that $\hat{\mu}_S$ is an unbiased estimator of population mean in the case of $\mathbf{Q}_0 = (Q_2)$ and $\mathbf{Q}_0 = (Q_1, Q_2, Q_3)$ being the vector of known quartiles of the population.

4 Performance comparison using a real data set

In this section, we assess the performance of $\hat{\mu}_S$ with real data set instead of Monte Carlo simulation. The data set, we have used for this purpose, contains the percentage of body fat determined by underwater weighing and different body circumference measurements for 252 men, and is available online at http://lib.stat.cmu.edu/datasets/bodyfat. We take the percentage of body fat as the variable of interest (Y), and we assume that the population median ($M = 19.2$) is known. We compute the efficiency of the proposed estimator for samples of size $5, 10, 20, 30, 40$ and 50. The sampling is done with replacement, so the assumption of independence is preserved. Note that in this example, $m = 2$ and $\mathbf{d} = \left(\frac{1}{2}, \frac{1}{2}\right)$, so it follows from Theorem 2.3 that the proposed estimator is unbiased and its exact variance can be computed via this theorem. The exact variances of $\sqrt{n}\hat{\mu}_S$ and $\sqrt{n}\hat{\mu}_{SRS}$ with their relative efficiency are given in Table 4. The relative efficiency (RE) is defined as the ratio of variance of $\sqrt{n}\hat{\mu}_{SRS}$ to variance of $\sqrt{n}\hat{\mu}_S$. Since the REs are computed before doing any rounding, the relative efficiencies do not coincide with the ratios between the tabled variances, which have been rounded to two decimal places.

Table 4: Efficiency of $\hat{\mu}_S$ to $\hat{\mu}_{SRS}$ for estimating the mean of the body fat of 252 men.

Sample size (n)	5	10	20	30	40	50
$Var\left(\sqrt{n}\hat{\mu}_{SRS}\right)$	14.00	7.00	3.50	2.33	1.75	1.40
$Var\left(\sqrt{n}\hat{\mu}_S\right)$	6.80	2.64	1.21	0.79	0.58	0.47
RE	2.05	2.65	2.88	2.94	2.97	2.99

We observe that the usage of auxiliary median information in the proposed mean estimator, improves the precision of estimation of population mean as compared to the standard mean estimator. The efficiency gain in estimation of population mean varies from 2.05 to 2.99 as sample size goes from 5 to 50.

5 Conclusion

In this paper, we introduced a nonparametric mean estimator when at least one of the population quantiles is known. We proved that our estimator is strongly consistent. We established asymptotic normality of the proposed estimator and showed that it is asymptotically more efficient than the standard mean estimator. For finite sample sizes, the superiority of the proposed mean estimator to the standard mean estimator was shown by using Monte Carlo simulation and a real data set.

Acknowledgements

The authors are very thankful to an associate editor and two anonymous referees for their valuable comments which led to a considerable improvement of an earlier version of this paper.

References

Al-Omari, A. I. (2012), Ratio estimation of the population mean using auxiliary information in simple random sampling and median ranked set sampling. *Statistics and Probability Letters*, **82**, 1883-18890.

Breidt, F. J. (2004), Simulation estimation of quantiles from a distribution with known mean. *Journal of Computational and Graphical Statistics*, **13**(2), 487-498.

Cochran, W. G. (1977), *Sampling technique*. 3rd Edition, New York: Wiley.

Das, A. K. and Tripathi, T. P. (1978), Use of auxiliary information in estimating the finite population variance. *Sankhya*, **40**, 139-148.

Dasbaravarde, A., Arghami, N. R. and Sarmad, M. (2016), Some theoretical results concerning non-parametric estimation by using a judgment post-stratification sample. *Communications in Statistics - Theory and Methods*, **45**(8), 2181-2203.

Gut, A. (2005), *Probability: A Graduate Course*. New York: Springer.

Isaki, C. T. (1983), Variance estimation using auxiliary information. *Journal of the American Statistical Association*, **78**, 117-123.

Lohr, S. L. (1999), *Sampling: Design and Analysis*. CA, Duxbury: Pacific Grove.

MacEachern, S. N., Stasny, E. A. and Wolfe, D. A. (2004), Judgment post-stratification with imprecise rankings. *Biometrics*, **60**(1), 207-215.

Srivastava, S. K. and Jhajj, H. S. (1981), A class on estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, **68**, 341-343.

Upadhyaya, L. N. and Singh, H. P. (1999), Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal*, **41**, 627-636.

Yadav, S. K. and Kadilar, C. (2014), A two parameter variance estimator using auxiliary information. *Applied Mathematics and Computation*, **226**, 117-122.