Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2020;12(2):218-28

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2019-72403

Performances of the Distribution Function Estimators Based on Ranked Set Sampling Using Body Fat Data

Vücut Yağı Verisi Kullanılarak Sıralı Küme Örneklemesine Dayalı Dağılım Fonksiyonu Kestiricilerinin Performansları

Yusuf Can SEVİL[a], Tuğba ÖZKAL YILDIZ[a]

[a]Dokuz Eylül University Faculty of Science, Department of Statistics, İzmir, TURKEY

This study was presented as orally at 5th International Researches, Statisticians and Young Statisticians Congress, 18-20 October 2019, Aydın, Turkey.

ABSTRACT In this study, an application of empirical distribution function (EDF) estimators based on ranked set sampling (RSS) using real-life data set (body fat data) is illustrated. In this application, three variables which are percentage of body fat (Y), abdomen circumference (X_1) and age (X_2) are used. Age and abdomen circumference are separately used in ranking process as auxiliary variables which have correlation 0.813 (for perfect ranking) and 0.291 (for imperfect ranking) with the percentage of body fat, respectively. Ranked set samples are constructed by using three different sampling designs which are level-0 level-1 and level-2. The effects of perfect and imperfect ranking on the estimators of the sampling designs are investigated. Relative efficiencies of the EDF estimators are obtained by using their mean squared errors (MSE) and integrated mean squared errors (IMSE), numerically. For both perfect and imperfect ranking, these EDF estimators based on sampling designs have outperformance against EDF estimator based on SRS.

Keywords: Ranked set sampling; sampling designs; empirical distribution function; mean squared error; integrated mean squared error; body fat data

ÖZET Bu çalışmada, sıralı küme örneklemesine dayalı ampirik dağılım fonksiyonu kestiricilerinin gerçek bir veri seti (vücut yağı verisi) kullanılarak uygulaması gerçekleştirildi. Bu uygulamada, vücut yağı yüzdesi (Y), karın çevresi (X_1) ve yaş (X_2) olmak üzere 3 değişken kullanılmıştır. Vücut yağ yüzdesiyle korelasyonları sırasıyla 0.813 (kusursuz sıralama) ve 0.291 (kusurlu sıralama) olan yaş ve karın çevresi, sıralama prosedüründe ayrı ayrı yardımcı değişkenler olarak kullanılmıştır. Sıralı küme örneklemleri, düzey-0, düzey-1 ve düzey-2 olan 3 farklı örneklem tasarımı kullanılarak elde edilmiştir. Kusursuz ve kusurlu sıralamanın örneklem tasarımlarının kestiricileri üzerine etkileri araştırılmıştır. Ampirik dağılım fonksiyonu kestiricilerinin göreceli etkinlikleri, hata kareler ortalamaları ve integrallenmiş hata kareler ortalamaları kullanılarak sayısal olarak elde edilmiştir. Kusurlu ve kusursuz sıralama için örneklem tasarımlarına dayalı ampirik dağılım fonksiyonu kestiricilerinin, basit rasgele örneklemeye dayalı ampirik dağılım kestiricisi karşısında daha iyi bir performansa sahip olduğu görülmüştür.

Anahtar kelimeler: Sıralı küme örneklemesi; örneklem tasarımları; ampirik dağılım fonksiyonu; hata kareler ortalaması; integrallenmiş hata kareler ortalaması; vücut yağı verisi

Ranked set sampling (RSS) was introduced by McIntyre as an advantageous alternative to simple random sampling (SRS).[1] McIntyre studied mean estimator based on RSS and showed that this estimator is more efficient than SRS.[1] Then, mathematical theory of RSS was first suggested by Takahasi & Wakimoto.[2] Also, the effects of ranking error on RSS was investigated by Dell & Clutter.[3] Stokes proposed the estima-

Correspondence: Yusuf Can SEVİL
Dokuz Eylül University Faculty of Medicine, Department of Statistics, İzmir, TURKEY
E-mail: yusuf.sevil@ogr.deu.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 28 Nov 2019 ***Received in revised form:*** 07 Apr 2020 ***Accepted:*** 09 Apr 2020 ***Available online:*** 28 May 2020

tion of correlation coefficient of a bivariate normal population in RSS.[4] An estimator of the population variance was given by Stokes.[5] Neerchal et al. suggested an RSS estimator of the population proportion.[6] Moreover, many modified versions of RSS have been examined such as extreme ranked set sampling, ERSS, median ranked set sampling, MRSS and percentile ranked set sampling, PRSS.[7-9]

In RSS, the ranking process plays vital role to construct a ranked set sample. Many different mechanisms such as expert opinion, visual inspection or auxiliary variable are used to rank the observations. In many studies on RSS, the interested variable is ranked by using expert opinion or visual inspections. In cases which visual inspections are time consuming or costly, use of auxiliary variable is suggested by Stokes as an alternative to these two ranking process.[10] The auxiliary variable has to be highly correlated, either positively or negatively, to the variable of interest.

In literature, the estimation of cumulative distribution function (CDF) with various settings of the RSS has been studied by many authors. Stokes & Sager suggested an unbiased estimator based on RSS for the population distribution function.[11] EDF estimators using ERSS and MRSS were introduced by Samawi & Al-Sageer.[12] Al-Subh et. al. studied goodness of fit (GOF) tests based on EDF using selective order statistics in RSS.[13] Non-parametric estimation based on partially rank-ordered set for continuous distribution function was proposed by Nazari et al.[14] EDF estimator based on level-2 sampling design was considered by Sevil & Yildiz.[15] They examined power of Kolmogorov-Smirnov GOF test for standard normal and inverse Gaussian distribution. Yildiz & Sevil proposed EDF estimators using level-0, level-1 and level-2 sampling designs.[16] They made power comparison among some GOF tests based on the EDF estimators. Yildiz & Sevil compared the proposed EDF estimators with the EDF when using SRS via their integrated mean squared errors (IMSE) under finite population.[17]

In Biostatistics, RSS and its modified methods have been used efficiently. Estimation of population mean using ERSS suggested by Samawi et al.[7] They used the suggested estimator to estimate mean of the body mass index (BMI) of diabetic women. ERSS was also used for regression analysis by Samawi & Ababneh.[18] In their study, Iowa 65 Rural Health Study data was used for numerical analysis. Also, Samawi & Al-Sageer gave an example on the bilirubin level of babies in neonatal intensive care.[12] Estimation of disease prevalence such as diabetes and hypertension proportions using RSS are studied by Chen et al.[19] The mean of BMI using RSS and judgment post-stratification was studied by Gory & Ozturk.[20] Gocoglu & Demirel considered proportion estimators based on RSS and its modified methods.[21] The suggested proportion estimators were illustrated by using the abalone data set.

The main motivation behind this work is to compare the suggested estimators in Yildiz & Sevil with the EDF based on SRS by calculating their mean squared errors (MSE) and integrated mean squared errors (IMSE) using body fat data.[16] In Section 2, we gave the sampling designs. Also, we introduced the EDFs based on the sampling designs in Section 3. For body fat data, we make performance comparisons between the EDFs based on sampling designs and the EDF based on SRS in Section 4. Finally, summary and concluding remarks are given in Section 5.

SAMPLING DESIGNS

In this section, we described the procedures of RSS based on the sampling designs. These sampling designs were proposed by Deshpande et al.[22] In literature, research in RSS drew considerable attention in finite population setting as well, e.g., Al-Saleh & Samawi, Ozdemir & Gokpinar, Gokpinar & Ozdemir, Frey, Jafari-Jozani & Johnson, Ozturk & Jafari-Jozani and Ozturk.[23-31] Some notations are as follows: k : set sizes, l : number of cycles and n : sample sizes.

Let (X,Y) be any bivariate continuous random vector and $(X_{11j}, X_{12j}, \Lambda, X_{1kj})$, $(X_{21j}, X_{22j}, \Lambda, X_{2kj})$, Λ, $(X_{k1j}, X_{k2j}, \Lambda, X_{kkj})$ are random sets of size k from jth cycle, $j = 1, \Lambda, l$. Order statistics of the ith set are

denoted by $X_{i(1)j}, X_{i(2)j}, \Lambda, X_{i(k)j}$ and their corresponding units are $Y_{i[1]j}, Y_{i[2]j}, \Lambda, Y_{i[k]j}$. Here, the parenthesis (r) indicate that X is the *rth* order statistic and the bracket $[r]$ indicate that Y is the judgement order statistic. It means that Y may not be the *rth* smallest order statistic. Ranking quality depends on correlation between *X* and *Y*. The procedure of level-0 is described as follows:

Step 1. A random sample of size *k* is selected from the population, $(X_{i1j}^{(0)}, X_{i2j}^{(0)}, \Lambda, X_{ikj}^{(0)})$, $i = 1, \Lambda, k$, $j = 1, \Lambda, l$.

Step 2. The units are ranked from the smallest to the largest, $X_{i(1)j}^{(0)}, X_{i(2)j}^{(0)}, \Lambda, X_{i(k)j}^{(0)}$.

Step 3. Then, $X_{i(i)j}^{(0)}$ and its corresponding unit $Y_{i[i]j}^{(0)}$ are selected and $Y_{i[i]j}^{(0)}$ is measured.

Step 4. The all units are replaced back into the population.

Step 5. Step 1-4 are repeated k times.

Step 6. Step 1-5 are repeated l times.

These $n = kl$ sample observations, $\{Y_{1[1]j}^{(0)}, Y_{2[2]j}^{(0)}, \Lambda, Y_{k[k]j}^{(0)}\}$, represents the final sample, $j = 1, \Lambda, l$. In the level-0 sampling design, sets are constructed by sampling with replacement. So, the same units can be appeared more than one in the final sample. The other sampling design is level-1 and the algorithmic form of the level-1 sampling design process is as follows:

Step 1. A random sample of size *k* is selected from the population, $(X_{i1j}^{(1)}, X_{i2j}^{(1)}, \Lambda, X_{ikj}^{(1)})$, $i = 1, \Lambda, k$, $j = 1, \Lambda, l$.

Step 2. The units are ranked from the smallest to the largest, $X_{i(1)j}^{(1)}, X_{i(2)j}^{(1)}, \Lambda, X_{i(k)j}^{(1)}$.

Step 3. Then, $X_{i(i)j}^{(1)}$ and its corresponding unit $Y_{i[i]j}^{(1)}$ are selected and $Y_{i[i]j}^{(1)}$ is measured.

Step 4. $Y_{i[i]j}^{(1)}$ is not replaced back into the population.

Step 5. Step 1-4 are repeated k times.

Step 6. Step 1-5 are repeated l times.

The final sample is denoted by $\{Y_{1[1]j}^{(1)}, Y_{2[2]j}^{(1)}, \Lambda, Y_{k[k]j}^{(1)}\}$, $j = 1, \Lambda, l$. In the level-1 sampling design, sets are constructed without replacement of the measured unit. Thus, the same units are not appeared more than one in the final sample. The third sampling design is level-2 sampling and the process of level-2 sampling design is given below:

Step 1. A random sample of size *k* is selected from the population, $(X_{i1j}^{(2)}, X_{i2j}^{(2)}, \Lambda, X_{ikj}^{(2)})$, $i = 1, \Lambda, k$, $j = 1, \Lambda, l$.

Step 2. The units are ranked from the smallest to the largest, $X_{i(1)j}^{(2)}, X_{i(2)j}^{(2)}, \Lambda, X_{i(k)j}^{(2)}$.

Step 3. Then, $X_{i(i)j}^{(2)}$ and its corresponding unit $Y_{i[i]j}^{(2)}$ are selected and $Y_{i[i]j}^{(2)}$ is measured.

Step 4. The all units in the set are not replaced back into the population.

Step 5. Step 1-4 are repeated k times.

Step 6. Step 1-5 are repeated l times.

The final sample is represented by $\{Y_{1[1]j}^{(2)}, Y_{2[2]j}^{(2)}, \Lambda, Y_{k[k]j}^{(2)}\}$, $j = 1, \Lambda, l$.

EMPIRICAL DISTRIBUTION FUNCTIONS

In this section, we describe the EDFs based on the sampling designs. First, these EDFs were proposed by Yildiz & Sevil.[16] In this work, they investigated the properties of the suggested EDFs. Also, power of some GOF tests were examined. They showed that the performances of the GOF tests based on suggested EDFs is higher than the GOF tests based on EDF using SRS. Then, Yildiz & Sevil compared the suggested EDFs based on sampling designs with the EDF based on SRS.[17] They obtained that the EDF estimators based on sampling designs are more efficient than the EDF based on SRS.

Let $Y_1, Y_2, \Lambda\ , Y_n$ be simple random sample and $\hat{F}(y)$ denotes the EDF based on SRS.

$$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y) \tag{1}$$

where $I(.)$ is indicator function. If the sample units are obtained by using RSS based on level-0 sampling design, then the EDF based on level-0, $\hat{F}^{*}_{L-0}(y)$ is given by following equation.

$$F^{*}_{L-0}(y) = \frac{1}{lk}\sum_{j=1}^{l}\sum_{i=1}^{k} I(Y^{(0)}_{i[i]j} \le y) \tag{2}$$

When the sample is constructed by using RSS based on level-1 sampling design, then the EDF based on level-1, $\hat{F}^{*}_{L-1}(y)$ is given in below.

$$F^{*}_{L-1}(y) = \frac{1}{lk}\sum_{j=1}^{l}\sum_{i=1}^{k} I(Y^{(1)}_{i[i]j} \le y) \tag{3}$$

If the sample units are collected by using RSS based on level-2 sampling design, then the EDF based on level-2, $\hat{F}^{*}_{L-2}(y)$ is given by the following equation.

$$F^{*}_{L-2}(y) = \frac{1}{lk}\sum_{j=1}^{l}\sum_{i=1}^{k} I(Y^{(2)}_{i[i]j} \le y) \tag{4}$$

APPLICATION

In this section, we illustrated the EDF estimators based on sampling designs by using real-life data set (body fat data). We examined the bias of the EDF based on sampling designs. Also, relative efficiencies are obtained by using their MSEs and IMSEs.

The body fat data (http://lib.stat.cmu.edu/datasets/bodyfat) for 252 men were collected by Penrose et al.[32] Suppose the set of 252 men constitutes a hypothetical population. It is made up of 15 measured variables on 252 men. Variables in the data are density, percentage of body fat (PBF), age, weight, height and 10 circumferences: neck, chest, abdominal, hip, thigh, knee, ankle, biceps, forearm and wrist. The body fat percentage of a human or other living being is the total mass of fat divided by total body mass. It is determined by underwater weighing and can be estimated using Equation 5. Let

$$\begin{aligned} D &= 1/\left[(A/a)+(B/b)\right] \\ B &= (1/D)\left[ab/(a-b)\right]-\left[b/(a-b)\right] \\ PBF &= 100\times B \end{aligned} \tag{5}$$

where $D =$ body density, $W =$ body weight, $A =$ proportion of lean tissue, $B =$ proportion of fat tissue ($A+B=1$), $a =$ density of lean tissue and $b =$ density of fat tissue.

In this application, we use three variables which are percentage of body fat (Y), abdomen circumference is a measure of obesity which is easy to calculate and readily accessible (X_1) and age (X_2). Our target parameter is the distribution function of percentage of body fat. The interested variable, Y , has normal distribution with parameters $\mu = 19.15$ and $\sigma^2 = 70.03$. In Figure 1, probability density function (PDF) and CDF of Y are given.

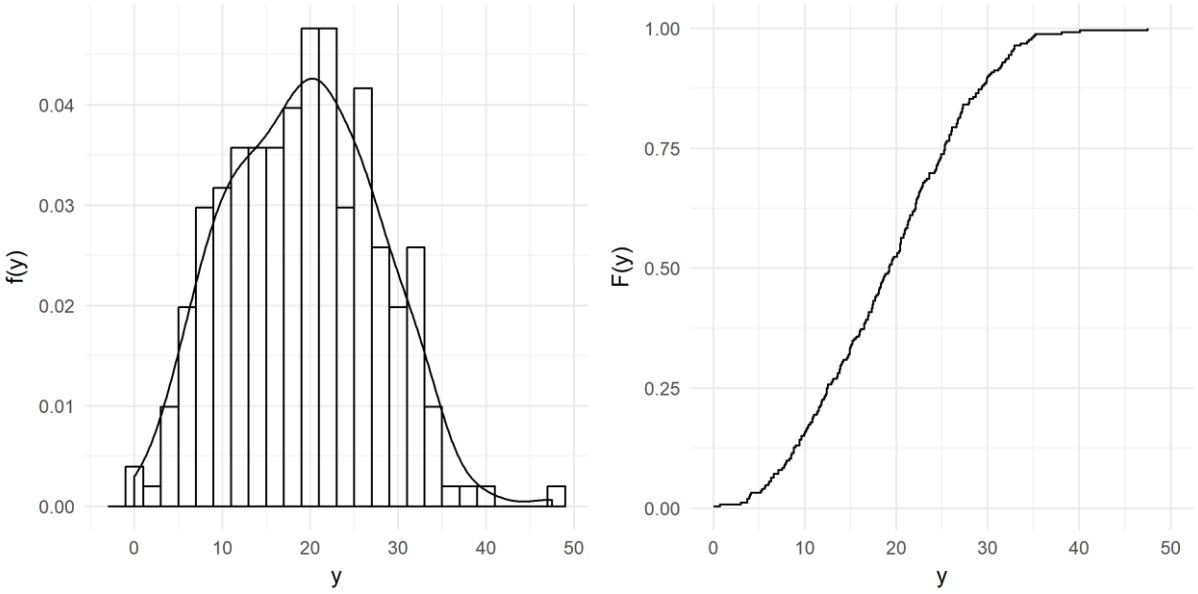

FIGURE 1: The PDF (left) and CDF (right) of the percentage of body fat.

Age and abdomen circumference are separately used in ranking process as auxiliary variables which have correlation 0.813 (for perfect ranking) and 0.291 (for imperfect ranking) with the percentage of body fat, respectively.

5,000 samples were drawn body fat data based on SRS, RSS (level-0, level-1 and level-2 sampling designs) for set sizes $k = 2,3,4,5$ and number of cycles $l = 2,5,10$. The bias values, the MSEs and IMSEs were computed from each sample, and the resulting values were used to determine relative efficiencies. To obtain the bias values, we considered $F(y) = 0.1, \Lambda\ ,0.9$. Bias values are calculated using the following equation,

$$bias(F^*_{L-t}(y)) = E(F^*_{L-t}(y)) - F(y) \tag{6}$$

where $t = 0,1,2.$ When X_1 is used as auxiliary variable in ranking process, the bias values and relative efficiencies based on MSEs are given in Table 1 and 2, respectively. In Table 1, the bias values of EDFs based on level-0, level-1 and level-2 sampling designs are presented. As seen in Table 1, bias values almost equal to zero. Then, relative efficiencies based on MSEs are given in Table 2. These efficiencies are obtained by using following equation. If $Eff\,(F^*_{L-t}(y), \hat{F}(y)) > 1$, $F^*_{L-t}(x)$ is more efficient than $\hat{F}(y)$.

$$Eff\,(F^*_{L-t}(y), \hat{F}(y)) = \frac{MSE(\hat{F}(y))}{MSE(F^*_{L-t}(y))} \tag{7}$$

where

$$MSE(\hat{F}(y)) = E(\hat{F}(y) - F(y))^2 \tag{8}$$

and

$$MSE(F^{*}_{L-t}(y)) = E(F^{*}_{L-t}(y) - F(y))^2 \quad (9)$$

According to these relative efficiencies, the suggested EDFs based on sampling designs in Yildiz & Sevil are more efficient than EDF based on SRS.[16,17] Also, it can be said that the level-2 sampling design has higher performance than level-0 and level-1. While the set size is increased, the relative efficiencies increase as well. When X_2 is used as auxiliary variable in ranking process, the bias values and relative efficiencies based on MSEs are given in Table 3 and 4, respectively. According to Table 3, it can be said that the bias values are almost equal to zero. Also, relative efficiencies in Table 4 are lower than relative efficiencies in Table 2. It is shown that the efficiencies of the EDFs based on sampling designs reduce when the correlation is decreased. Moreover, the lowest efficiencies are obtained for EDF based on level-0 in Table 4.

TABLE 1: The bias values of EDFs based on the sampling designs for perfect ranking.

		Level-0			Level-1			Level-2		
$F(x)$	k	$l=2$	$l=5$	$l=10$	$l=2$	$l=5$	$l=10$	$l=2$	$l=5$	$l=10$
0.1	2	0.001	0.001	0.001	-0.002	0.001	0.001	-0.001	-0.001	-0.001
	3	0.001	-0.001	0.000	-0.003	-0.001	0.000	0.001	-0.001	0.000
	4	0.001	0.000	-0.001	-0.001	0.001	0.000	0.001	0.000	0.000
	5	-0.001	0.001	0.001	-0.002	0.000	-0.001	-0.001	0.000	-0.001
0.2	2	-0.003	-0.001	-0.001	-0.005	0.000	0.003	0.000	-0.001	-0.001
	3	0.000	-0.001	-0.001	0.001	0.000	0.000	-0.001	-0.001	0.002
	4	0.001	0.000	0.000	0.001	-0.002	0.000	-0.001	0.000	0.000
	5	0.000	0.000	0.000	0.000	-0.001	-0.001	0.003	-0.001	-0.001
0.3	2	-0.004	-0.002	0.001	0.000	-0.001	0.001	0.000	0.002	0.000
	3	-0.002	-0.002	0.000	-0.004	-0.001	-0.001	0.003	0.000	0.002
	4	0.000	0.001	0.001	0.000	0.000	-0.001	0.000	0.000	-0.001
	5	0.002	0.001	0.000	-0.003	-0.001	-0.003	0.001	0.000	0.000
0.4	2	0.000	0.000	0.001	-0.001	0.002	-0.002	0.000	0.002	0.000
	3	0.002	0.001	0.001	-0.003	-0.001	0.000	0.001	-0.001	-0.001
	4	0.000	0.002	0.001	-0.001	-0.001	-0.001	-0.004	-0.003	-0.001
	5	0.001	-0.002	-0.001	-0.001	-0.002	-0.001	0.001	0.000	0.000
0.5	2	0.001	0.000	0.000	0.000	0.003	-0.002	0.000	0.001	0.000
	3	0.005	0.004	0.000	0.001	-0.002	0.000	0.003	0.000	-0.002
	4	0.001	0.002	-0.001	-0.006	-0.003	0.000	0.001	-0.001	0.001
	5	-0.002	0.002	-0.001	-0.002	-0.003	-0.002	0.001	0.000	0.000
0.6	2	0.000	0.001	0.000	0.002	0.001	0.002	0.000	0.003	-0.003
	3	0.003	0.000	-0.001	0.005	-0.002	0.000	0.005	0.001	0.001
	4	-0.004	0.001	0.001	0.001	-0.003	-0.001	-0.003	0.000	0.000
	5	0.001	-0.001	0.002	-0.002	-0.002	-0.002	-0.003	0.000	-0.001
0.7	2	0.003	0.000	0.001	0.002	0.002	0.000	0.005	0.001	0.000
	3	0.002	-0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	4	0.004	-0.001	-0.001	-0.002	-0.002	-0.001	0.001	-0.001	0.001
	5	0.002	-0.002	0.000	-0.005	-0.001	-0.002	0.000	-0.001	-0.001
0.8	2	0.004	0.002	0.000	-0.002	-0.001	-0.002	0.000	-0.001	-0.002
	3	-0.001	0.000	0.000	0.001	-0.003	-0.001	0.002	0.001	0.001
	4	-0.002	0.000	0.001	-0.002	-0.002	-0.001	-0.002	0.000	-0.001
	5	-0.002	0.001	-0.001	0.002	-0.003	-0.001	0.005	-0.001	0.000
0.9	2	0.001	0.003	0.001	0.003	0.001	-0.002	0.002	0.000	0.001
	3	0.001	-0.001	0.000	-0.003	0.001	-0.001	-0.004	0.000	0.001
	4	0.000	0.001	0.000	-0.001	-0.001	-0.001	0.000	0.001	0.000
	5	0.000	0.001	0.001	0.000	-0.002	-0.001	0.003	0.001	-0.001

TABLE 2: The relative efficiencies by using MSEs of EDF based on the sampling designs for perfect ranking.

		Level-0			Level-1			Level-2		
$F(x)$	k	$l=2$	$l=5$	$l=10$	$l=2$	$l=5$	$l=10$	$l=2$	$l=5$	$l=10$
0.1	2	1.050	0.989	0.999	1.078	1.040	1.077	1.088	1.081	1.103
	3	1.073	1.052	1.023	1.142	1.096	1.142	1.107	1.103	1.100
	4	1.118	1.072	0.945	1.148	1.157	1.127	1.131	1.159	1.125
	5	1.172	1.067	0.942	1.187	1.232	1.213	1.176	1.202	1.247
0.2	2	1.085	1.098	1.047	1.125	1.110	1.052	1.080	1.186	1.117
	3	1.187	1.151	1.021	1.249	1.168	1.172	1.220	1.228	1.206
	4	1.205	1.110	1.082	1.308	1.196	1.317	1.275	1.233	1.376
	5	1.226	1.118	1.038	1.295	1.266	1.281	1.283	1.349	1.362
0.3	2	1.199	1.167	1.176	1.159	1.166	1.218	1.243	1.191	1.244
	3	1.230	1.216	1.098	1.277	1.260	1.300	1.312	1.293	1.298
	4	1.351	1.314	1.166	1.414	1.415	1.378	1.383	1.504	1.455
	5	1.421	1.282	1.174	1.473	1.451	1.400	1.511	1.448	1.661
0.4	2	1.200	1.253	1.076	1.189	1.237	1.232	1.173	1.266	1.203
	3	1.284	1.250	1.191	1.294	1.343	1.369	1.312	1.326	1.406
	4	1.380	1.299	1.197	1.419	1.360	1.416	1.449	1.482	1.535
	5	1.515	1.313	1.225	1.570	1.543	1.470	1.575	1.640	1.726
0.5	2	1.174	1.192	1.096	1.205	1.210	1.254	1.217	1.219	1.237
	3	1.293	1.217	1.185	1.324	1.296	1.307	1.323	1.367	1.337
	4	1.354	1.366	1.195	1.476	1.438	1.423	1.471	1.449	1.464
	5	1.426	1.435	1.163	1.509	1.510	1.467	1.503	1.602	1.699
0.6	2	1.173	1.087	1.066	1.136	1.156	1.158	1.153	1.130	1.129
	3	1.274	1.221	1.090	1.286	1.249	1.227	1.295	1.249	1.262
	4	1.288	1.231	1.099	1.321	1.317	1.297	1.279	1.343	1.344
	5	1.299	1.270	1.062	1.323	1.330	1.325	1.369	1.323	1.418
0.7	2	1.137	1.169	1.068	1.168	1.240	1.186	1.187	1.217	1.197
	3	1.242	1.199	1.101	1.299	1.295	1.254	1.286	1.280	1.348
	4	1.359	1.313	1.122	1.382	1.372	1.314	1.421	1.399	1.433
	5	1.473	1.273	1.175	1.466	1.424	1.434	1.527	1.501	1.540
0.8	2	1.115	1.040	1.040	1.103	1.095	1.135	1.091	1.083	1.090
	3	1.257	1.054	1.054	1.237	1.144	1.245	1.232	1.211	1.223
	4	1.244	1.160	1.081	1.243	1.315	1.306	1.299	1.307	1.386
	5	1.259	1.212	1.059	1.324	1.368	1.292	1.293	1.347	1.462
0.9	2	1.052	1.041	0.993	1.084	1.103	1.031	1.051	1.058	1.059
	3	1.185	1.059	1.005	1.155	1.093	1.107	1.119	1.096	1.197
	4	1.197	1.112	1.059	1.220	1.171	1.290	1.225	1.251	1.287
	5	1.195	1.147	1.084	1.197	1.235	1.245	1.270	1.299	1.279

For the relative efficiencies based on IMSEs, 2,000 samples are selected from the body fat data. These efficiencies are computed by using the following equation. As a performance comparisons criteria, Wang et al. described IMSE of EDF using SRS as follows:[33]

$$\begin{aligned} IMSE(\hat{F}(y)) &= \int_{-\infty}^{\infty} \left\{\hat{F}(y) - F(y)\right\}^2 dx \\ &= \int_{0}^{1} \left\{\hat{F}(F^{-1}(p)) - p\right\}^2 dF^{-1}(p) \\ &= \int_{0}^{1} \left\{\hat{F}(F^{-1}(p)) - p\right\}^2 \frac{1}{f(F^{-1}(p))} dp \end{aligned} \tag{10}$$

where $y = F^{-1}(p)$ and $p \in [0,1]$. This $IMSE(\hat{F}(y))$ can be calculated approximately by using composite trapezoidal rule

$$IMSE(\hat{F}(y)) = \frac{b-a}{2L}\left\{\sum_{i=1}^{L}\left|F(q_i)-\hat{F}(q_i)\right| + \sum_{i=2}^{L-1}\left|F(q_i)-\hat{F}(q_i)\right|\right\} \quad (11)$$

where b and a are upper and lower limits of integral, respectively. L is the number of cut points q_i, $L=(b-a)/w$ where w is the width of intervals. In the interval $[a,b]$, cut points q_i are obtained by using $q_i = a + i(b-a)/L$, $i = 1,\Lambda\ ,L$. IMSEs based on sampling designs are given by

$$IMSE(\hat{F}(y)) = \frac{b-a}{2L}\left\{\sum_{i=1}^{L}\left|F(q_i)-\hat{F}^{*}_{L-t}(q_i)\right| + \sum_{i=2}^{L-1}\left|F(q_i)-\hat{F}^{*}_{L-t}(q_i)\right|\right\} \quad (12)$$

where $t = 0,1,2$. $F^{*}_{L-t}(q_i)$ are calculated by using (2) – (4).

TABLE 3: The bias values of EDFs based on the sampling designs for imperfect ranking.

		Level-0			Level-1			Level-2		
$F(x)$	k	$l=2$	$l=5$	$l=10$	$l=2$	$l=5$	$l=10$	$l=2$	$l=5$	$l=10$
0.1	2	-0.001	-0.002	0.000	0.001	-0.002	-0.001	0.000	0.000	0.001
	3	0.000	0.000	-0.001	0.003	0.000	0.001	-0.005	0.003	0.000
	4	-0.002	-0.001	-0.001	-0.001	-0.001	0.000	-0.001	0.000	0.000
	5	0.001	0.000	0.000	0.001	0.000	-0.001	0.002	-0.001	-0.001
0.2	2	0.002	0.000	0.000	0.001	0.001	0.000	0.001	0.002	-0.002
	3	-0.001	0.000	0.001	-0.002	0.001	-0.001	-0.001	-0.001	-0.001
	4	0.002	0.000	0.000	-0.001	-0.002	0.000	-0.001	0.001	-0.002
	5	0.001	-0.001	-0.001	-0.003	0.001	-0.001	0.000	0.001	0.000
0.3	2	-0.002	-0.001	-0.002	-0.004	0.001	-0.001	-0.001	0.000	0.000
	3	0.000	0.001	-0.001	-0.001	0.001	0.000	-0.002	0.002	0.001
	4	0.004	-0.001	-0.002	-0.001	0.002	0.002	0.004	0.000	0.001
	5	0.000	0.000	0.002	0.001	-0.002	0.000	0.001	0.000	-0.001
0.4	2	0.003	0.000	-0.003	-0.003	0.002	0.000	-0.002	-0.002	-0.001
	3	0.000	0.002	0.000	0.000	0.001	0.000	0.000	-0.002	0.000
	4	0.001	-0.001	-0.001	0.000	0.000	0.000	0.004	0.000	0.000
	5	-0.002	0.000	0.001	0.001	0.002	0.000	0.002	0.000	0.000
0.5	2	0.005	-0.002	-0.002	0.003	0.003	0.000	-0.002	-0.002	-0.001
	3	0.001	0.001	-0.001	-0.004	0.000	-0.002	0.002	-0.003	0.000
	4	0.000	0.002	0.001	-0.002	0.000	-0.002	-0.001	0.001	0.000
	5	0.001	0.001	0.001	0.000	-0.001	-0.001	0.001	0.000	0.000
0.6	2	-0.001	0.001	0.001	0.001	-0.002	0.000	-0.004	-0.005	0.000
	3	0.008	0.002	0.001	-0.003	0.002	-0.001	0.001	0.000	-0.002
	4	0.002	-0.001	0.000	-0.002	-0.002	-0.001	0.003	0.002	0.001
	5	0.001	-0.001	0.000	-0.001	-0.001	0.000	0.002	0.001	0.000
0.7	2	0.003	0.001	-0.001	-0.004	-0.001	0.001	-0.001	-0.002	-0.001
	3	-0.002	-0.002	-0.001	-0.002	0.000	-0.001	-0.006	0.000	0.000
	4	0.000	0.003	0.000	0.003	-0.002	0.001	0.002	-0.003	0.000
	5	0.002	0.000	-0.002	0.001	-0.002	-0.002	0.001	0.001	-0.001
0.8	2	-0.004	0.002	-0.001	0.005	-0.001	0.001	-0.001	0.000	-0.002
	3	-0.002	0.000	-0.001	0.005	0.001	0.000	0.001	0.001	-0.001
	4	-0.001	0.001	0.000	-0.004	0.001	0.000	0.002	0.000	0.001
	5	0.001	0.000	0.001	0.002	-0.002	-0.001	0.000	0.001	0.001
0.9	2	-0.002	-0.001	0.000	0.001	-0.001	0.000	-0.004	-0.001	0.000
	3	0.001	-0.001	0.000	0.001	-0.003	-0.001	-0.001	-0.002	0.000
	4	0.000	0.000	0.001	0.001	-0.001	0.000	0.001	0.000	0.001
	5	-0.002	0.001	0.001	-0.002	0.000	-0.001	0.001	0.001	0.000

TABLE 4: The relative efficiencies by using MSEs of EDF based on the sampling designs for imperfect ranking.

		Level-0			Level-1			Level-2		
$F(x)$	k	$l=2$	$l=5$	$l=10$	$l=2$	$l=5$	$l=10$	$l=2$	$l=5$	$l=10$
0.1	2	1.018	0.979	0.969	1.036	1.033	1.027	1.021	0.999	1.036
	3	1.003	0.943	0.920	1.015	0.989	1.013	1.067	1.011	1.016
	4	1.031	0.932	0.844	1.046	1.059	1.039	1.025	1.041	0.989
	5	0.986	0.927	0.847	1.014	1.032	1.094	0.973	1.064	1.059
0.2	2	0.967	0.964	0.935	1.019	1.000	0.947	1.011	1.036	1.012
	3	1.059	0.984	0.875	1.055	1.029	0.998	1.070	1.034	0.974
	4	1.033	0.918	0.869	1.031	1.021	1.098	1.059	1.011	1.037
	5	0.966	0.910	0.788	1.062	1.030	1.023	1.047	0.999	1.021
0.3	2	1.011	1.017	0.980	1.030	1.039	1.059	1.054	1.035	1.073
	3	0.967	0.962	0.913	1.025	0.984	1.013	1.008	1.027	0.990
	4	1.008	0.975	0.870	1.008	1.036	0.983	0.985	1.079	1.005
	5	1.014	0.921	0.817	1.049	0.994	1.035	1.048	1.031	1.002
0.4	2	1.004	1.019	0.942	1.032	1.061	0.996	0.996	1.025	0.985
	3	0.979	0.952	0.880	1.005	1.049	1.028	1.004	1.041	1.029
	4	0.949	0.935	0.859	0.994	1.034	1.044	1.001	1.027	1.023
	5	1.026	0.951	0.844	1.096	1.006	1.018	1.060	1.038	1.050
0.5	2	1.024	0.983	0.938	1.019	1.027	1.068	1.018	1.010	1.011
	3	0.998	0.930	0.872	0.962	1.010	1.024	0.973	1.001	0.991
	4	1.002	0.943	0.881	0.974	1.008	1.080	1.002	1.017	0.986
	5	1.002	0.937	0.829	1.067	1.051	1.044	1.053	1.017	1.043
0.6	2	0.990	1.007	0.933	0.986	1.043	1.037	0.990	0.996	1.012
	3	1.073	0.968	0.898	1.078	1.046	1.031	1.093	1.029	0.996
	4	1.007	0.964	0.872	1.043	1.029	1.028	1.058	0.982	1.055
	5	1.041	0.935	0.819	1.052	1.009	1.008	1.043	1.067	0.996
0.7	2	1.034	1.044	0.939	1.002	1.051	1.019	1.031	1.046	1.067
	3	0.997	0.943	0.914	1.030	1.049	1.008	0.960	1.039	1.019
	4	1.041	1.019	0.877	1.079	1.069	1.038	1.057	1.051	1.054
	5	1.042	0.978	0.810	1.083	1.032	1.027	1.118	1.048	1.058
0.8	2	0.984	0.927	0.972	1.063	0.985	0.974	0.994	1.060	0.991
	3	1.020	0.939	0.895	1.110	1.001	1.020	1.062	0.964	1.019
	4	0.997	0.954	0.891	1.039	1.013	1.061	1.042	1.006	1.059
	5	0.943	0.926	0.838	1.027	1.006	1.015	1.008	1.034	1.075
0.9	2	0.973	0.961	0.925	0.976	1.002	0.999	0.948	0.977	0.995
	3	0.999	0.912	0.885	1.027	0.949	0.961	1.020	0.954	0.987
	4	1.020	0.931	0.899	1.055	0.982	1.051	1.042	1.006	1.079
	5	0.921	0.911	0.830	0.964	1.050	1.043	0.972	0.982	1.009

Relative efficiencies based on IMSEs are computed as follows:

$$Eff(F^*_{L-t}(y),\hat{F}(y)) = \frac{IMSE(\hat{F}(y))}{IMSE(F^*_{L-t}(y))} \tag{13}$$

If $Eff(F^*_{L-t}(y),\hat{F}(y)) > 1$, $F^*_{L-t}(y)$ is more efficient than $\hat{F}(y)$. The results are given in Table 5. It can be seen that the EDFs based on the sampling designs have out performances against EDF based on SRS in Table 5. For X_1, the relative efficiencies based on IMSEs increase while the set size is increased. For X_2, this is not true. Because, relative efficiencies decrease due to ranking error while the set size increases.

TABLE 5: The relative efficiencies by using IMSEs of EDF based on the sampling designs.

		X_1			X_2		
l	k	**Level-0**	**Level-1**	**Level-2**	**Level-0**	**Level-1**	**Level-2**
2	2	1.173	1.187	1.178	1.128	1.134	1.120
	3	1.203	1.208	1.212	1.107	1.117	1.146
	4	1.220	1.220	1.240	1.141	1.147	1.166
	5	1.239	1.255	1.285	1.111	1.150	1.158
5	2	1.145	1.181	1.174	1.120	1.127	1.126
	3	1.174	1.210	1.211	1.113	1.126	1.119
	4	1.182	1.237	1.225	1.090	1.137	1.132
	5	1.189	1.250	1.264	1.082	1.141	1.151
10	2	1.137	1.178	1.178	1.085	1.114	1.131
	3	1.139	1.208	1.219	1.067	1.149	1.132
	4	1.131	1.218	1.257	1.060	1.121	1.138
	5	1.116	1.224	1.263	1.014	1.153	1.123

CONCLUSION

In this study, we investigated the performances of the EDFs based on sampling designs that are proposed in Yildiz & Sevil.[16] To examine the performances of these EDFs, the real data is used. In this application, samples are selected from body fat data for 252 men collected by Penrose et al. (http://lib.stat.cmu.edu/datasets/bodyfat).[32] In this data, three variables which are percentage of body fat (Y), abdomen circumference (X_1) and age (X_2) are used. The target parameter is the distribution function of percentage of body fat (Y). Abdomen circumference (X_1) and age (X_2) are separately used in ranking process as auxiliary variables which have correlation 0.813 (for perfect ranking) and 0.291 (for imperfect ranking) with the percentage of body fat, respectively. According to Table 1 and 3, it can be said that the EDFs based on sampling designs are unbiased estimators for the CDF of percentage of body fat (Y). When abdomen circumference (X_1) is used as auxiliary variable, the EDFs based on sampling designs have higher performances than the EDF based on SRS. Also, the relative efficiencies based on MSEs of the EDF using level-2 are the highest among the EDFs based on sampling designs when auxiliary variable is abdomen circumference (X_1). In general, according to the relative efficiencies based on MSEs, the EDFs based on sampling designs have outperformances against EDF based on SRS when the ranking process is performed by using age (X_2). For relative efficiencies based on IMSEs, the EDFs using sampling designs are more efficient than the EDF using SRS.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

All authors contributed equally while this study preparing.

REFERENCES

1. McIntyre GA. A method for unbiased selective sampling, using ranked sets. Aust J Agr Res. 1952;3(4):385-90. [Crossref]
2. Takahasi K, Wakimoto K. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Annals of the Institute of Statistical Mathematics. 1968; 20(1):1-31. [Crossref]
3. Dell T, Clutter J. Ranked set sampling theory with order statistics background. Biometrics. 1972;28(2):545-55. [Crossref]
4. Stokes SL. Inferences on the correlation coefficient in bivariate normal populations from ranked set samples. J Am Stat Assoc. 1980;75(372):989-95. [Crossref]
5. Stokes SL. Estimation of variance using judgement ordered ranked set samples. Biometrics. 1980;36(1):35-42. [Crossref]
6. Neerchal NK, Sinha BK, Lacayo H. Ranked set sampling from a dichotomous population. Journal of Applied Statistical Science. 1998;11(1):83-90.
7. Samawi HM, Ahmed MS, Abu-Dayyeh W. Estimating the population mean using extreme ranked set sampling. Biometrical J. 1996;38(5):577-86. [Crossref]
8. Muttlak HA. Median ranked set sampling. Journal of Applied Statistical Science. 1997;6(4):245-55.
9. Muttlak HA. Modified ranked set sampling methods. Pak J Statist. 2003;19(3):315-23.
10. Stokes SL. Ranked set sampling with concomitant variables. Comm Stat Theor Meth. 1977;6(12):1207-11. [Crossref]
11. Stokes SL, Sager TW. Characterization of a ranked-set-sample with application to estimating distribution functions. J Am Stat Assoc. 1988;83(402):374-81. [Crossref]
12. Samawi HM, Al-Sagheer OA. On the estimation of the distribution function using extreme and median ranked set sampling. Biometrical J. 2001;43(3):357-73. [Crossref]
13. Al-Subh S, Alodat M, Ibrahim K, Jemain A. EDF goodness of fit tests of logistic distribution under selective order statistics. Pak J Statist. 2009;25(3):265-74.
14. Nazari S, Jafari Jozani M, Kharrati-Kopaei M. On distribution function estimation with partially rank-ordered set samples: estimating mercury level in fish using length frequency data. Statistics. 2016;50(6):1387-410. [Crossref]
15. Sevil YC, Ozkal Yildiz T. Power comparison of the Kolmogorov-Smirnov test under ranked set sampling and simple random sampling. J Stat Comput Sim. 2017;87(11):2175-85. [Crossref]
16. Ozkal Yildiz T, Sevil YC. Performances of some goodness-of-fit tests for sampling designs in ranked set sampling. J Stat Comput Sim. 2018;88(9):1702-16. [Crossref]
17. Ozkal Yildiz T, Sevil YC. Empirical distribution function estimators based on sampling designs in a finite population using single auxiliary variable. J Appl Stat. 2019;46(16):2962-74. [Crossref]
18. Samawi HM, Ababneh FM. On regression analysis using ranked set sample. Journal of Statistical Research. 2001;35(2):93-105.
19. Chen H, Stasny EA, Wolfe DA. Improved procedures for estimation of disease prevalence using ranked set sampling. Biometrical J. 2007;49(4):530-8. [Crossref] [PubMed]
20. Gory J, Ozturk O. Analysis of the NHANES III data set using ranked set and judgement post-stratified samples. Adv Appl Stat. 2015;47(1):65-89. [Crossref]
21. Göçoğlu A, Demirel N. Estimating the population proportion in modified ranked set sampling methods. J Stat Comput Sim. 2019;89(14):2694-710. [Crossref]
22. Deshpande JV, Frey J, Ozturk O. Nonparametric ranked-set sampling confidence intervals for quantiles of a finite population. Environ Ecol Stat. 2006;13(1):25-40. [Crossref]
23. Al-Saleh MF, Samawi HM. A note on inclusion probability in ranked set sampling and some of its variations. Test. 2007;16(1):198-209. [Crossref]
24. Ozdemir YA, Gokpinar F. A generalized formula for inclusion probabilities in ranked set sampling. Hacettepe Journal of Mathematics and Statistics. 2007;36(1):89-99.
25. Ozdemir YA, Gokpinar F. A new formula for inclusion probabilities in median ranked set sampling. Comm Stat Theor Meth. 2008;37(13):2022-33. [Crossref]
26. Gokpinar F, Ozdemir YA. Generalization of inclusion probabilities in ranked set sampling. Hacettepe Journal of Mathematics and Statistics. 2010;39(1):89-95.
27. Frey J. Recursive computation of inclusion probabilities in ranked set sampling. J Stat Plan Infer. 2011;141(11):3632-9. [Crossref]
28. Jafari Jozani M, Johnson BC. Design based estimation for ranked set sampling in finite population. Environ Ecol Stat. 2011;18(4):663-85. [Crossref]
29. Jafari Jozani M, Johnson BC. Randomized nomination sampling in finite populations. J Stat Plan Infer. 2012;142(7):2103-15. [Crossref]
30. Ozturk O, Jozani MJ. Inclusion probabilities in partially rank ordered set sampling. Computational Statistics & Data Analysis. 2014;69:122-32. [Crossref]
31. Ozturk O. Estimating of population mean and total in a finite population setting using multiple auxiliary variables. J Agr Biol Envir St. 2014;19(2):161-84. [Crossref]
32. Penrose KW, Nelson AG, Fisher AG. Generalized body composition prediction equation for men using simple measurement techniques. Medicine & Science in Sports & Exercise. 1985;17(2):189. [Crossref]
33. Wang X, Wang K, Lim J. Isotonized CDF estimation from judgement poststratification data with empty strata. Biometrics. 2012;68(1):194-202. [Crossref] [PubMed]