FORTHCOMING 62D05-05-21-03

ON THE NONPARAMETRIC ESTIMATION OF THE ODDS AND DISTRIBUTION FUNCTION USING MOVING EXTREME SET SAMPLING

Mohamed S. Abdallah[1*] and Amer I. Al-Omari[2]
[1*] Department of Quantitative Techniques, Faculty of Commerce, Aswan University, Egypt. E-mail: mohamed_abdallah@com.aswu.edu.eg; statisticsms.2010@gmail.com
[2]Department of Statistics, Faculty of Science, Al al-Bayt University, Mafraq, Jordan, E-mail: alomari_amer@yahoo.com

ABSTRACT
In this article, we will consider the problem of estimating the cumulative distribution function (CDF) and the odds measure under moving extreme ranked set sampling. Using maximum likelihood method and local polynomial regression approach, new intuitive and easy-to-implement nonparametric estimators of the CDF and the odds are derived. It has been proved that the proposed estimators are consistent estimators. Simulated and empirical data are subsequently used to evaluate the performances of the newly proposed estimators. The numerical results provide that the proposed estimators are much more efficient than their alternatives at the center and the upper tail of the parent distribution even when the rankings are not perfect.

KEYWORDS: Moving Extreme Ranked Set Sampling, Distribution Function, Local Polynomial Regression, Odds Function, Ranking Errors.

MSC: 62D05, 62F03.

RESUMEN
En este artículo, consideraremos el problema de estimar la función de distribución acumulativa (CDF) y la medida de los odds bajo muestreo por conjuntos ordenados extremales. Usando el método de Máxima Verosimilitud y el enfoque de la regresión local polinomial, nuevos estimadores intuitivos y fáciles de implementar de la CDF y de los odds son derivados. Ha sido proveído que los propuestos estimadores son consistentes. Son usados simulados y empíricos subsecuentemente para evaluar el comportamiento de los nuevos estimadores propuestos. Los resultados numéricos soportan que los propuestos estimadores son mucho más eficientes que sus alternativas en el centro y en la cola superior de la distribución de origen, incluso si el rankeo es no perfecto.

PALABRAS CLAVE: Muestreo Móvil por Conjuntos Ordenados Extremal, Función de Distribución, Regresión Local Polinomial, Función Odds, Errores en en Rankeo.

1. INTRODUCTION

Moving extreme ranked set sampling (MERSS) is a sampling strategy proposed by Al-Odat and Al-Saleh (2000) for estimating the population mean as an alternative sampling technique to the conventional ranked set sampling (RSS) scheme. To attain MERSS, one can act the following steps:

1. Randomly draw k sets of sizes $1, 2, \ldots, k$ from the interested population.
2. Exactly quantify the maximum ordered sampling unit from each set sample.
3. Repeat the preceding steps, if needed, m times (cycles) to get a sample of size $n = km$ for actual quantification.

Let $Y_{i[1:k]j}, Y_{i[2:k]j} \ldots Y_{i[k:k]j}$ be the judgment order statistics of the i^{th} sample $(i = 1,2, \ldots k,)$ in the j^{th} cycle $(j = 1,2, \ldots, m)$. Then $\{Y_{i[i:i]j}: \ i = 1,2, \ldots k \ ; \ j = 1,2, \ldots, m\}$ are denoted to the MERSS. The term judgment (subjective) order statistic and square brackets $[\cdot]$ are used to emphasis that the ranking process may not be completely accurate, i.e. the sampling units can be ranked with errors. This situation is known as imperfect ranking. Perfect ranking, however, is a situation in which the judgment rank of each unit matches with its actual rank, and hence the square brackets $[\cdot]$ can be replaced with the round ones $(\cdot)$. It is interesting to note here that for each i, the measured units $\{Y_{i(i:i)1}, Y_{i(i:i)2} \ldots Y_{i(i:i)m}\}$ are independent and identically distributed (iid) random variables. While for each j, $\{Y_{1(1:1)j}, Y_{2(2:2)j}, \ldots, Y_{k(k:k)j}\}$ are only independent random variables.

Despite the superiority of RSS over MERSS in estimating several population parameters, there are many considerations which make MERSS a desirable choice for many practitioners. One can justify this preference as MERSS depends only on $\frac{k(k-1)}{2} \times m$ sample units to get a sample of size n, yet RSS needs $k^2 \times m$ sample units to obtain the same sample size. In addition, MERSS can practically be preferable as identifying the maximum/minimum rank is much easier than determining the intermediate ranks. Consequently, research literature in MERSS has expanded rapidly in the last two decades, among which we can briefly cite Al-Saleh and Al-Hadhrami (2003) and Al-Omari (2021) regarding to parametric inference about the population parameters, Al-Saleh and Al-Ananbeh (2005), relating to estimation of correlation coefficient under the bivariate normal distribution, Al-Omari (2015, 2016), Zamanzade et al. (2020) and Al-Omari and Abdallah (2021), respecting to estimation of cumulative distribution function (CDF), Rahmani and Razmkhah (2017), for testing the quality ranking, and Zamanzade and Mahdizadeh (2020), concerning to nonparametric estimation about the population proportion. For other interesting studies of MERSS can be found in the monograph of Bouza and Al-Omari (2019) and the references cited therein.

One of the most commonly used technique in the area of nonparametric analysis is Local polynomial regression (LPR). Stone (1977) was the first to suggest the use of the LPR for discovering the association between dependent and independent variables. The idea behind LPR is that any function can be well approximated in some neighborhood points by a low-order polynomial and that simple model can be fitted to data easily leads to a smooth function over the support of the data. Further information can be found in Fan and Gijbels (1996).

According to our best knowledge, despite the superiority and the popularity of the LPR in estimating several statistical measures, for interesting examples see Jayasinghe and Zeephongsekul (2012) and Cattaneo, et al. (2019), no works in the literature have so far carried out on LPR under MERSS. To fill this gap, this work aims to incorporate LPR in estimating the CDF and the odds measure under MERSS.

The layout of this study is organized as follows: Section 2 provides the CDF estimator introduced by Al-Saleh and Ahmad (2019) as well as our proposed one. The odds estimator introduced by Al-Saleh and Samawi (2010) as well as our novel ones are explained in Section 3. A comparison study between the proposed procedures with their competitors is studied in Section 4. In Section 5, the findings are illustrated using real data example. Eventually, some concluding remarks and future research points appear in Section 6.

2. ESTIMATION OF CDF USING MERSS

The problem of the CDF is a comprehensively studied work in a nonparametric analysis because through CDF, one can identify a lot about the population properties, such as odds, hazard, entropy…etc. In this part, we will address the CDF estimator under MERSS published by Al-Saleh and Ahmed (2019), then we will modify this estimator using LPR model.

2.1 Al-Saleh and Ahmad (2019) 's CDF estimator

Recall again that $\{Y_{i(i:i)j}: \ i = 1,2, \dots k \ ; \ j = 12, \dots, m\}$ are MERSS drawn from a population with the probability density function (pdf) $f(\cdot)$ and CDF $F(\cdot)$. Al-Saleh and Ahmad (2019) used the maximum likelihood estimation (MLE) method for estimating $F(y)$ based on MERSS. Their idea is based on the fact that $\{I(Y_{i(i:i)1} \le y), I(Y_{i(i:i)2} \le y) \dots I(Y_{i(i:i)m} \le y)\}$ are iid each has a Bernoulli distribution with success probability $B_{i,1}(F(y))$, where $I(\cdot)$ is the indicator function and $B_{a,b}(y)$ is the CDF of the Beta distribution with parameters a and b at point y. Let

$$Y_i = \sum_{j=1}^{m} I(Y_{i(i:i)j} \le y), \qquad i = 1, \dots, k.$$

Then Y_is are independent random variables each with binomial distribution with mass parameter m, and success probability $B_{i,1}(F(y))$. Hence the corresponding likelihood function of Y_is is:

$$L(F|Y) = \prod_{i=1}^{k} \binom{m}{Y_i} \left(B_{i,1}(F(y))\right)^{Y_i} \left(1 - B_{i,1}(F(y))\right)^{m-Y_i}.$$

The CDF estimator intuitively can be obtained by maximizing $L(F|Y)$ or equivalently maximizing $\log L(F|Y)$ as shown below:

$$\hat{F}(y) = \underset{F \in [0,1]}{\operatorname{argmax}} \log L(F|Y).$$

Based on simulation studies, Al-Saleh and Ahmad (2019) concluded, and also confirmed later by Zamanzade et al. (2020), that $\hat{F}(y)$ is the best one compared to its competitors at the upper tail of the population of the distribution provided that the perfectness assumption is assumed.

Remark 1: Zamanzade et al. (2020) stated, under the perfectness, that:

$$\frac{\hat{F}(y)-F(y)}{Var\left(\hat{F}(y)\right)} \xrightarrow{p} N(0,1),$$

where $Var\left(\hat{F}(y)\right) = -\left(E\left(\frac{d^2 \log L(F|Y)}{dF^2}\right)\right)^{-1} = \left(\sum_{i=1}^{k} \frac{m\left(b_{i,1}(F(y))\right)^2}{B_{i,1}(F(y))\left(1-B_{i,1}(F(y))\right)}\right)^{-1}$, $b_{a,b}(y)$ is pdf of the Beta distribution with parameters a and b at the point y and $\xrightarrow{p}$ indicates convergence in probability.

2.2 Proposed CDF estimator

Since the CDF enjoys with an important property, as it is sufficiently smooth function leads its derivatives are typically exist. Using Taylor Series expansion, $F(y)$ can be linearized as:

$$F(y) \approx F(y^*) + F^{(1)}(y^*)(y-y^*) + \frac{F^{(2)}(y^*)}{2!}(y-y^*)^2 + \cdots + \frac{F^{(p)}(y^*)}{p!}(y-y^*)^p, \quad (1)$$

where $\approx$ indicates approximate equality, $F^{(j)}(y^*) = \frac{dF(y)}{dy}\Big|_{y=y^*}$ $j = 1,2,\dots,p$, and y^* is an observation from the data neighborhood around y. Since $F(y)$ is unknown, then (1) can be rewritten as:

$$F(y^*) \approx F(y) + F^{(1)}(y^*)(y^*-y) + \frac{F^{(2)}(y^*)}{2!}(y^*-y)^2 + \cdots + \frac{F^{(p)}(y^*)}{p!}(y^*-y)^p. \quad (2)$$

since we need all the sampling units share in estimating $F(y)$, then (2) becomes:

$$F\left(Y_{i[i:i]j}\right) \approx F(y) + F^{(1)}\left(Y_{i[i:i]j}\right)\left(Y_{i[i:i]j}-y\right) + \frac{F^{(2)}\left(Y_{i[i:i]j}\right)}{2!}\left(Y_{i[i:i]j}-y\right)^2 + \cdots + \frac{F^{(p)}\left(Y_{i[i:i]j}\right)}{p!}\left(Y_{i[i:i]j}-y\right)^p. \quad i = 1,2,\dots k\,;\ j = 1,2,\dots,m \quad (3)$$

which is equivalent to:

$$F\left(Y_{i(i:i)j}\right) \approx \beta_o + \beta_1\left(Y_{i(i:i)j}-y\right) + \beta_2\left(Y_{i(i:i)j}-y\right)^2 + \cdots + \beta_p\left(Y_{i(i:i)j}-y\right)^p, \quad (4)$$

where $\beta_j = \frac{F^{(j)}\left(Y_{i(i:i)j}\right)}{j!}$ $j = 0,1,\dots,p$.

By estimating $F\left(Y_{i(i:i)j}\right)$ with $\hat{F}\left(Y_{i(i:i)j}\right)$, (4) will become:

$$\hat{F}\left(Y_{i(i:i)j}\right) \approx \beta_o + \beta_1\left(Y_{i(i:i)j}-y\right) + \beta_2\left(Y_{i(i:i)j}-y\right)^2 + \cdots + \beta_p\left(Y_{i(i:i)j}-y\right)^p, \quad (5)$$

It is obvious that (5) is similar to the standard linear regression equation and hence the unknown coefficients, β's, can be estimated by ordinary least square (OLS). According to Fan and Gijbels (1996), we need only points within a neighborhood of y are given higher weights than the remaining. Alternatively, weighted least square (WLS) method is adopted and the unknown coefficients in (4) can be estimated as:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = (Y^T W Y)^{-1} Y^T W \hat{F}, \quad (6)$$

where $Y = \begin{pmatrix} 1 & \left(Y_{1(1:1)1}-y\right) & \cdots & \left(Y_{1(1:1)1}-y\right)^p \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \left(Y_{k(k:k)1}-y\right) & \cdots & \left(Y_{k(k:k)1}-y\right)^p \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \left(Y_{k(k:k)m}-y\right) & \cdots & \left(Y_{k(k:k)m}-y\right)^p \end{pmatrix}$, $\hat{F} = \begin{pmatrix} \hat{F}\left(Y_{1(1:1)1}\right) \\ \vdots \\ \hat{F}\left(Y_{k(k:k)1}\right) \\ \vdots \\ \hat{F}\left(Y_{k(k:k)m}\right) \end{pmatrix}$,

$W = \text{diag}\left(\frac{1}{h}K\left(\frac{Y_{i(i:i)j}-y}{h}\right)\right)$ $i = 1 \dots k\,;\ j = 1 \dots m$, $K(\cdot)$ denotes a kernel function assigns the weights and h is the bandwidth controls the size of the neighborhood points around y. Comparing (5) with (3), one can conclude that $F(y)$ can be estimated by $\hat{\beta}_o$, and hence the proposed CDF estimator can be formulated as:

$$\tilde{F}(y) = e_1\hat{\beta} = \hat{\beta}_o = e_1(Y^T W Y)^{-1} Y^T W \hat{F},$$

where $e_1 = (1 \quad 0 \quad \cdots \quad 0)$ is the first $(p+1)$-dimensional unit row vector.

Remark 2: It is worth noting that Y^TWY is invertible if rank$(Y^TWY) = p + 1$. This needs that rank$(WY) = p + 1$. Since Y is a full column rank matrix as its columns are everywhere linearly independent. Hence, we can almost sure that X^TWX is an invertible matrix if W is also a full column rank matrix.

It is observed that the idea behind our proposed estimator, $\tilde{F}(y)$, is to take the estimator, $\hat{F}(y)$, introduced by Al-Saleh and Ahmad (2019) as a starting point, then obtain a smooth local approximation estimator using a polynomial expansion. As expectedly, the performance of $\tilde{F}(y)$ will strongly depend on $\hat{F}(y)$. The following proposition shows that $\tilde{F}(y)$ is a consistent estimator to $F(y)$.

Proposition 1. Let $\{Y_{i(i:i)j}: i = 1,2, \dots, k; j = 1,2, \dots, m\}$ be a perfect MERSS and Y^TWY be a full column rank, then $\sup_y |\tilde{F}(y) - F(y)| \to 0$ as n $\to \infty$.

Proof.

$$\sup_y |\tilde{F}(y) - F(y)| = \sup_y |e_1\hat{\beta} - e_1\beta| = e_1 \sup_y |(Y^TWY)^{-1}Y^TW(\hat{F} - Y\beta)|$$
$$= e_1(Y^TWY)^{-1}Y^TW \sup_y |\hat{F} - Y\beta|.$$

In the light of the Remark 1 and from (4), the proof completes.

It is clear that $\tilde{F}(y)$ has a practical problem for which it depends on unknown quantities to be estimated. First, the degree of the local polynomial, p, has a strong influence on the quality of the fitted regression. As, large (small) values provide less (greater) bias with large (less) variability in the estimates. According to Cattaneo, et al. (2019), being $p = 2$ gives a good approximation to the underlying function with reasonable precise fitting. Therefore, hereafter, we will set $p = 2$. For choosing the kernel function, it has been shown that the kind of the kernel function has much less effect on the performance of the estimator. However, those with bounded support are most commonly used. Toward this end, Epanechnikov kernel function is adopted.

Due to its role in controlling the smoothness of the fit, the most important factor which has a critical effect on the estimates shown in (6) is the bandwidth, h, selection. There, therefore, exist several automatic schemes for choosing h. An easy popular choice is the nearest-neighbor bandwidth method. The main advantages of this method are it is intuitive, not required heavy computations and avoiding plug-in estimates. For further discussion, reader can refer to Loader (1999). The steps of the nearest neighbor bandwidth method can be explained as follows:

1. Compute the distances between the point y and all the data points, i.e. $d(Y_{i(i:i)j} - y) = |Y_{i(i:i)j} - y| \quad i = 1,2, \dots, k \; ; \; j = 1,2, \dots, m.$
2. Choose the bandwidth, denoted by $\hat{h}_1$, to be the q^{th} smallest distance obtained in step 1, where $q = \lfloor \alpha n \rfloor$, $\lfloor x \rfloor$ is integer part of x, α is the selected percentage such that minimizes the cross-validation (CV) index:

   $CV(\alpha) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(\hat{F}(Y_{i(i:i)j}) - \sum_{t=0}^{p}\hat{\beta}_t^{-Y_{i[i:i]j}}(Y_{i(i:i)j} - y)^t\right)^2$ and $\hat{\beta}_t^{-Y_{i[i:i]j}}$ is the estimate of regression coefficient in (6) without observation $Y_{i[i:i]j}$.

Remark 3: It is interesting to highlight that the proposed estimator shown in (5), $\hat{\beta}$, is not only a benefit for estimating the CDF, $\hat{\beta}_0$, but also for estimating the pdf. As, $\hat{\beta}_1$ can be considered as a density estimator under MERSS. More discussion about density estimation can be found in Cattaneo, et al. (2019).

3. ESTIMATION OF ODDS BASED ON MERSS

The odds is sometimes a better measure than the CDF to represent chance, it has a crucial important role in several statistical disciplines such as linear models, especially logistic regression models, and survival analysis. For a random variable Y, the odds measure, denoted by $O(y_o)$, is defined as:

$$O(y) = \frac{F(y)}{1 - F(y)},$$

Obviously, $O(y) \geq 0$ and it is a strictly monotone increasing function of y. Additionally, $O(y)$ is finite until $F(y) < 1$. Therefore, we will assume, throughout this study, that $F(y) < 1$. In this part, we will address the odds estimator under MERSS published by Al-Saleh and Samawi (2010), then our proposed estimators using MLE and LPR model are presented.

3.1. Al-Saleh and Samawi's odds estimator

Al-Saleh and Samawi (2010) constructed their odds estimator based on the fact that the sum of a geometric series, for any $g(y) \in (0,1)$, could be expressed as:

$$\sum_{i=1}^{\infty} (g(y))^i = \frac{g(y)}{1-g(y)}. \tag{7}$$

Recall that the CDF of $Y_{i(i:i)j}$ at point y is given by:

$$F_{(i:i)}(y) = \int_0^y f_{(i:i)}(u)\, du = \int_0^y iF(u)^{i-1} f(u)\, du = F(y)^i, \tag{8}$$

where $f_{(i:i)}(u)$ is the pdf of i^{th} order statistics from the sample of size i. By replacing $g(\cdot)$ in (7) with $F_{(i:i)}(\cdot)$, we can obtain:

$$\sum_{i=1}^{\infty} F_{(i:i)}(y) = \sum_{i=1}^{\infty} (F(y))^i = \frac{F(y)}{1-F(y)} = O(y). \tag{9}$$

Consequently, the suggested estimator of Al-Saleh and Samawi (2010), denoted by $\hat{O}(y)$, can be obtained as first estimating $F_{(i:i)}(y)$ with:

$$\hat{F}_{(i:i)}(y) = \frac{1}{m}\sum_{j=1}^{m} I(Y_{i(i:i)j} \le y). \tag{10}$$

then plugging (10) in (9) with restricting the summation of i to k, as shown below:

$$\hat{O}(y) = \sum_{i=1}^{k} \hat{F}_{(i:i)}(y) = \frac{1}{m}\sum_{i=1}^{k}\sum_{j=1}^{m} I(Y_{i(i:i)j} \le y).$$

The following proposition shows that $\hat{O}(y)$ is a consistent estimator to $O(y)$.

Proposition 2. Let $\{Y_{i(i:i)j}: i = 1,2,\dots,k; j = 1,2,\dots,m\}$ be a perfect MERSS and $F(y) < 1$, then $\sup_y \left|\hat{O}(y) - O(y)\right| \to 0$ as $k \to \infty$ and $m \to \infty$.

Proof.

It is proven by Al-Saleh and Samawi (2010) that:

$\text{Bias}\left(\hat{O}(y)\right) = O(y)(F(y))^k$ and $Var\left(\hat{O}(y)\right) = \frac{1}{m}\left(O(y)\frac{(1-(F(y))^k)(1-(F(y))^{k+1})}{1+F(y)}\right)$.

Since $\text{MSE}\left(\hat{O}(y)\right) = \text{Var}\left(\hat{O}(y)\right) + \left(\text{Bias}\left(\hat{O}(y)\right)\right)^2$, where MSE refers to the mean square error. Hence $\lim_{k,m\to\infty} \text{MSE}\left(\hat{O}(y)\right) = 0$, for which the proof completes.

3.1. Modified Al-Saleh and Samawi's odds estimator

In this part, we plan to use the CDF estimators described in the preceding section to construct new odds estimators. In the light of (8), one can write:

$$F_{(i:i)}(y) = \int_0^y iF(u)^{i-1} f(u)\, du = B_{i,1}(F(y)). \tag{11}$$

and hence $O(y)$ can be rewritten as:

$$O(y) = \sum_{i=1}^{\infty} F_{(i:i)}(y) = \sum_{i=1}^{\infty} B_{i,1}(F(y)).$$

Consequently, our first proposed odds estimator, denoted by $\hat{O}_1(y)$, can be constructed as estimating $F_{(i:i)}(y)$ with, see (11),:

$$\tilde{F}_{(i:i)}(y) = B_{i,1}\left(\hat{F}(y)\right). \tag{12}$$

then plugging (12) in (9) with restricting the summation of i to k, as shown below:

$$\hat{O}_1(y) = \sum_{i=1}^{k} \tilde{F}_{(i:i)}(y) = \sum_{i=1}^{k} B_{i,1}\left(\hat{F}(y)\right).$$

Of course, $F_{(i:i)}(y)$ can also be estimated in terms with $\tilde{F}(y)$ rather than $\hat{F}(y)$. This leads to our second proposed odds estimator takes the form:

$$\hat{O}_2(y) = \sum_{i=1}^{k} B_{i,1}\left(\tilde{F}(y)\right).$$

The consistency of $\hat{O}_1(y)$ as well as $\tilde{O}_1(y)$ is shown by the following proposition.

Proposition 3. Let $\{Y_{i(i:i)j}: i = 1,2, \dots, k; j = 1,2, \dots, m\}$ be a perfect MERSS and $F(y) < 1$, then as $k \to \infty$, we have:

(a) $\sup_y |\hat{O}_1(y) - O(y)| \to 0$.

(b) $\sup_y |\hat{O}_2(y) - O(y)| \to 0$.

Proof. (a) In the light of Remark 1, $\hat{O}_1(y) \xrightarrow{p} \sum_{i=1}^{k} B_{i,1}(F(y))$. By comparing (11) with (8), we can get:

$$\hat{O}_1(y) \xrightarrow{p} \sum_{i=1}^{k} F(y)^i = O(y) \qquad \text{as } k \to \infty.$$

which completes the proof.

(b) From Proposition 1, it is shown that $\tilde{F}(y)$ is a consistent estimator to $F(y)$. Hence the proof can be done in the same way presented in (a).

4. MONTE CARLO COMPARISONS

In this part, the performance of the proposed procedures in different designs is conducted in terms of the relative efficiency (RE) criterion. The RE of $\tilde{F}(y)$ to $\hat{F}(y)$ can be defined as:

$$RE_1(\tilde{F}(y)) = \frac{MSE(\tilde{F}(y))}{MSE(\hat{F}(y))}. \tag{13}$$

Analogously, The RE of the proposed odds estimators to $\hat{O}(y)$ can be computed as:

$$RE_2\left(\hat{O}_L(y)\right) = \frac{MSE\left(\hat{O}_L(y)\right)}{MSE\left(\hat{O}(y)\right)}. \quad L \in [1,2]. \tag{14}$$

With the definition of $RE_1(\tilde{F}(y))$ in (13), if $RE_1(\tilde{F}(y))$ is greater than one indicates $\tilde{F}(y)$ outperforms $\hat{F}(y)$ at the point y. This argument can also be extended to $RE_2\left(\hat{O}_L(y)\right)$. To generate MERSS, we assume that the ranking process is done using imperfect ranking model suggested by Dell and Clutter (1972) which assumes that (Y, X) has a bivariate normal distribution with correlation coefficient ρ. The selected levels of ρ are: $\rho = 1$ for perfect ranking, $\rho = 0.9$ for nearly perfect ranking, and $\rho = 0.5$ for nearly imperfect ranking. Also, two configurations of the set sizes $k = 2,5$ with three levels of set sizes $m = 10,15,20$ are considered. The values of both $RE_1(\tilde{F}(y))$ and $RE_2\left(\hat{O}_L(y)\right)$ are computed for $y = Q_p$, $p \in \{0.1,0.2, \dots,0.9\}$. where Q_p is the p^{th} quantile of the standard normal distribution. For each combination of k, m and ρ, $10,000$ samples are generated based on MERSS. Here, we only consider the experimental results for normal distribution, as we have noticed that the pattern of the results is not much affected by changing the parent distribution. Based on the results presented in Fig. (1- 3), we can observe the following:

- It is seen that the all REs related to all the proposed procedures depend heavily on both the location of the point y and the quality of ranking ρ, yet the effect of the values of ρ is more pronounced.
- $\tilde{F}(y)$ ($\hat{O}_2(y)$) tends to be considerably better than $\hat{F}(y)$ $\left(\hat{O}(y)\right)$ as y moves around the center of the parent distribution provided that the quality of ranking is not weak.
- $\hat{O}_1(y)$ is slightly more efficient than $\hat{O}(y)$ at the lower tail of the distribution in the case that the quality of ranking is reasonable.
- It is clear that when the ranking is perfect, increasing the sample size has a positive effect on the $RE_1(\tilde{F}(y))$ and $RE_2\left(\hat{O}_2(y)\right)$ as well. However, the sample size has a week effect on the behavior of $\hat{O}_1(y)$.
- For a fixed sample, increasing the set size improves all REs related to all the proposed procedures rather increasing than the cycle assuming the perfectness. This effect becomes weaker corresponding to $\hat{O}_1(y)$.
- It is apparent that using LPR considerably improves the behavior of $\hat{O}_1(y)$ at the upper tail of the distribution regardless the quality of ranking.
- In summary, it is evident that all the proposed procedures based on LPR can be the best choice in the case that the point y is far from the boundaries of the parent distribution as well

as the quality of ranking is fairly good. Otherwise, $\hat{F}(y)$ and $\hat{O}(y)$ can be suggested for estimating the CDF and the odds function respectively.

5. REAL DATA APPLICATION

In this part, we investigate the performance of the proposed procedures using an empirical dataset. We consider the Australian sport dataset, available at http://www.stats ci.org/data/oz/ais.html, as a hypothetical population in which "lean body mass (LBM)", whose summary statistics shown by Fig. 4, is considered as the variable of interest. The Pearson correlation between "LBM" and "weight in kg (WT) " equals 93%, while for "LBM" and body mass index (BMI) it is 71%. Thus we consider "WT" and "BMI" as concomitant variables for ranking purpose in MERSS scheme. Note that the "LBM" itself is also used for the ranking purpose. Consequently, our study includes the perfect ranking and two different levels of imperfect ranking.

For the same values k and m mentioned in Section 4, 10,000 samples with replacement are drawn using MERSS schemes. Again, for each the selected samples, all the aforementioned estimators are computed and $RE_1\left(\tilde{F}(y)\right)$ and $RE_2\left(\hat{O}_L(y)\right)$ are obtained at each p as displayed by Table (2). One can observe from the results shown in Table (2) that when the quality ranking is not weak, $\tilde{F}(y)$ and $\hat{O}_2(y)$ are the best estimators for all the considered situations except for the cases that the point y is near the boundaries. It is also evident that the REs for LPR-based estimators improves as increasing k rather than m as long as the quality of ranking is strong. Consequently, all of these results are consistent with what we mentioned in the preceding section. Eventually, it may be important to mention that all the tabulated results and presented figures are coded using R package and it is available upon request from the second author.

6. CONCLUSION

This article is concerned with the problem of estimating CDF and odds function based on MERSS. New CDF estimator based on LPR is derived. The resulting proposed estimator is used to introduce two different odds estimators. It turned out that, the estimators based on LPR can have some advantages over their competitors for the points at the center of the parent distribution and also the quality of rankings is reasonable. Under the perfectness setup, a considerable efficiency gain is observed as increasing the set size rather than cycle size for a fixed sample size. Since MERSS is less prone to ranking errors, hence we recommend to use our LPR-based estimators.

Concerning to possible future topics, a plenty of work has to be done. Further investigation to determine the theoretical properties of LPR-based estimators may be needed, as Cattaneo, et al. (2019) studied in depth the different properties of the CDF estimator based on LPR under simple random sample design. It may be important to employ our proposed odds estimators to introduce new odds ratio estimators using the same methodologies explained in Samawi and Al-Saleh (2013) and Huang et al. (2018). It is also interesting to note that our proposed strategy explained in Section 2 can be useful for estimating other important statistical functions, such as the hazard function and reversed hazard function. This can be done by just replacing $\hat{F}\left(Y_{i(i:i)j}\right)$ in (5) with $\log\left(1-\hat{F}\left(Y_{i(i:i)j}\right)\right)$ for estimating the hazard function and with $\log\left(\hat{F}\left(Y_{i(i:i)j}\right)\right)$ for estimating the reversed hazard function, then selecting the second element of (6). The authors plan to take these topics in their subsequent works.

Acknowledgments: The authors are grateful to anonymous referees and an editor for their valuable comments and suggestions which lead to this improved version.

RECEIVED: NOVEMBER, 2020.
REVISED: APRIL, 2021.

REFERENCES

[1]. AL-OMARI, A.I. (2015): The efficiency of L ranked set sampling in estimating the distribution function, **Afrika Matematika**. 26, 1457–1466.

[2]. AL-OMARI, A.I. (2016): Quartile ranked set sampling for estimating the distribution function. **Journal of the Egyptian Mathematical Society**, 24, 303-308.

[3]. AL-OMARI, A. I. (2021): Maximum likelihood estimation in location-scale families using varied L ranked set sampling. **RAIRO-Operations Research**, 55, S2759-S2771.

[4]. AL-OMARI, A. I. and ABDALLAH, M. S. (2021): Estimation of the distribution function using moving extreme and MiniMax ranked set sampling. **Communications in Statistics-Simulation and Computation. Doi:10.1080/03610918.2021.1891433.**

[5]. AL-ODAT, M. and AL-SALEH, M.F. (2000): A variation of ranked set sampling, **Journal of Applied Statistical Science.** 10, 137-146.
[6]. AL-SALEH M.F. and AHMAD D.M. (2019): Estimation of the Distribution Function Using Moving Extreme Ranked Set Sampling (MERSS): In **Ranked Set Sampling: 65 Years Improving the Accuracy in Data Gathering**, Academic Press, Amsterdam, 43-58.
[7]. AL-SALEH, M.F. and SAMAWI, H. (2010): On estimating the odds using Moving Extreme Ranked Set Sampling. **Statistical Methodology**, 7, 133–140.
[8]. AL-SALEH, M.F. and AL-ANANBEH, A.M. (2007): Estimation of the means of the bivariate normal using moving extreme ranked set sampling with concomitant variable. **Statistical Papers** 48, 179–195.
[9]. AL-SALEH, M.F. and AL-HADHRAMI, S.A. (2003): Estimation of the mean of the exponential distribution using moving extremes ranked set sampling. **Statistical Papers** 44, 367–382.
[10]. BOUZA, C.N. and AL-OMARI, A.I. (2019**): Ranked Set Sampling, 65 Years Improving the Accuracy in Data Gathering.** Elsevier, ISBN: 978-0-12-815044-3.
[11]. CATTANEO, M. JANSSON, M. and MA, X. (2019): Simple Local Polynomial Density Estimators. **Journal of the American Statistical Association**. https://doi.org/10.1080/01621459.2019.1635480.
[12]. DELL, T. R. and CLUTTER, J. L. (1972): Ranked set sampling theory with order statistics background. **Biometrics** 28, 545-555.
[13]. FAN, J. and GIJBELS, I. (1996): **Local polynomial modeling and its applications. In: Monographs on Statistics and Applied Probability**. Chapman and Hall, London.
[14]. HUANG, Y. YIN, J. and SAMAWI, H. (2018): Methods improving the estimate of diagnostic odds ratio. **Communications in Statistics-Simulation and Computation** 47, 353-366.
[15]. JAYASINGHE,C. and ZEEPHONGSEKUL, P. (2012): On the nonparametric smooth estimation of the reversed Hazard Rate function. **Statistical Methodology** 9, 364–380.
[16]. LOADER, C. (1999): Local regression and likelihood. Springer Science & Business Media.
[17]. Rahmani, H. and Razmkhah, M. (2017): Perfect ranking test in moving extreme ranked set sampling. **Statistical Papers** 58, 855–875.
[18]. SAMAWI, H. and AL-SALEH. M.F. (2013): Valid estimation of odds ratio using two types of moving extreme ranked set sampling. **Journal of the Korean Statistical Society** 42, 17–24.
[19]. STONE, C. (1977): Consistent nonparametric regression. **Annals of Statistics**. 5, 595–645.
[20]. ZAMANZADE, E. and MAHDIZADEH, M. (2020): Using ranked set sampling with extreme ranks in estimating the population proportion. **Statistical Methods in Medical Research** 29, 165–177.
[21]. ZAMANZADE, E. MAHDIZADEH, M. and SAMAWI, H. (2020): Efficient estimation of cumulative distribution function using moving extreme ranked set sampling with application to reliability. **Statistical Papers.** https://doi.org/10.1007/s10182-020-00368-3.

Appendix

Table 1. The efficiency values of the CDF and odds estimators using real data set

$k=2$		$m=10$			$m=15$			$m=20$		
	p	$\tilde{F}(y)$	$\hat{O}_1(y)$	$\tilde{O}_1(y)$	$\tilde{F}(y)$	$\hat{O}_1(y)$	$\tilde{O}_1(y)$	$\tilde{F}(y)$	$\hat{O}_1(y)$	$\tilde{O}_1(y)$
LBM	.10	0.798	1.087	0.944	0.704	1.673	0.875	0.654	1.045	0.712
	.25	1.234	1.001	1.131	1.320	1.032	1.333	1.309	1.056	1.342
	.50	1.287	0.998	1.232	1.298	0.998	1.473	1.345	1.002	1.523
	.75	1.110	0.996	1.087	1.132	0.954	1.077	1.135	0.998	1.063
	.90	1.027	0.999	1.003	1.022	0.978	1.001	1.045	0.999	1.013
BMI	.10	0.623	1.043	0.611	0.654	1.077	0.688	0.609	1.039	0.601
	.25	1.226	1.004	1.286	1.165	1.047	1.254	1.165	1.030	1.287
	.50	1.165	1.029	1.322	1.264	1.017	1.432	1.321	1.020	1.276
	.75	0.908	0.998	1.079	1.123	1.001	1.367	1.132	0.990	1.065
	.90	1.026	0.998	0.987	1.034	1.000	0.998	1.017	1.000	0.990
WT	.10	0.607	0.999	0.576	0.612	0.976	0.576	0.556	0.925	0.532
	.25	1.110	0.987	1.098	1.114	0.998	1.109	1.047	0.911	1.035
	.50	1.056	0.967	1.250	1.078	1.023	1.140	1.064	1.017	1.143
	.75	0.976	1.002	1.094	0.945	1.001	1.076	0.998	1.000	1.067
	.90	0.998	1.000	1.000	1.015	1.000	0.999	0.967	1.000	0.997
$k=5$		$m=4$			$m=6$			$m=8$		
	p	$\tilde{F}(y)$	$\hat{O}_1(y)$	$\tilde{O}_1(y)$	$\tilde{F}(y)$	$\hat{O}_1(y)$	$\tilde{O}_1(y)$	$\tilde{F}(y)$	$\hat{O}_1(y)$	$\tilde{O}_1(y)$
LBM	.10	0.736	1.286	0.897	0.699	1.223	0.765	0.675	1.176	0.699
	.25	1.023	1.086	1.043	1.132	1.132	1.187	1.234	1.154	1.276
	.50	1.654	1.045	1.277	1.704	1.043	1.434	1.765	1.046	1.654
	.75	1.365	0.985	1.143	1.264	0.976	1.220	1.165	0.987	1.398
	.90	1.107	1.002	1.000	1.122	1.001	0.977	1.187	1.002	0.997
BMI	.10	0.684	1.143	0.678	0.603	1.187	0.570	0.593	1.087	0.598
	.25	0.998	1.010	0.889	1.650	0.987	1.010	1.125	1.023	1.014
	.50	1.176	0.998	1.087	1.234	0.967	1.254	1.287	0.998	1.2576
	.75	1.065	0.979	1.062	1.055	0.998	1.198	1.025	1.001	1.033
	.90	1.010	0.977	1.010	1.043	0.997	0.967	1.002	0.998	0.932
WT	.10	0.501	0.944	0.532	0.554	0.908	0.504	0.501	0.992	0.512
	.25	1.016	0.765	0.857	1.030	0.872	0.882	1.017	0.965	0.871
	.50	0.932	0.889	0.965	0.993	0.879	0.998	1.009	0.889	0.976
	.75	0.997	0.901	1.001	0.886	1.012	1.054	0.943	1.018	1.039
	.90	0.897	1.002	0.997	0.911	1.000	1.010	0.912	0.865	0.954

$m=10$ $m=15$ $m=20$

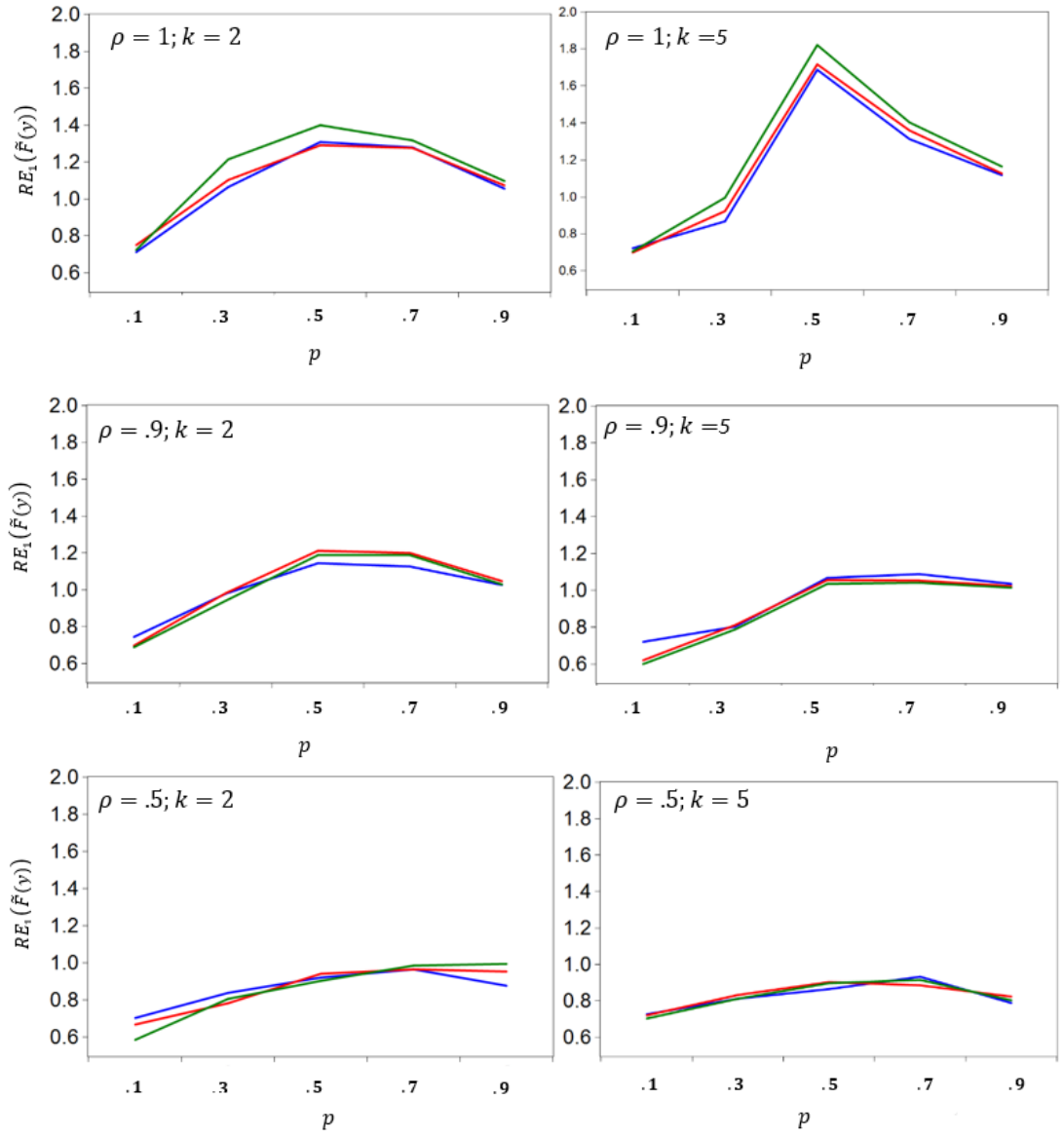


Fig. 1: The RE of $\tilde{F}(y)$ with respect to $\hat{F}(y)$

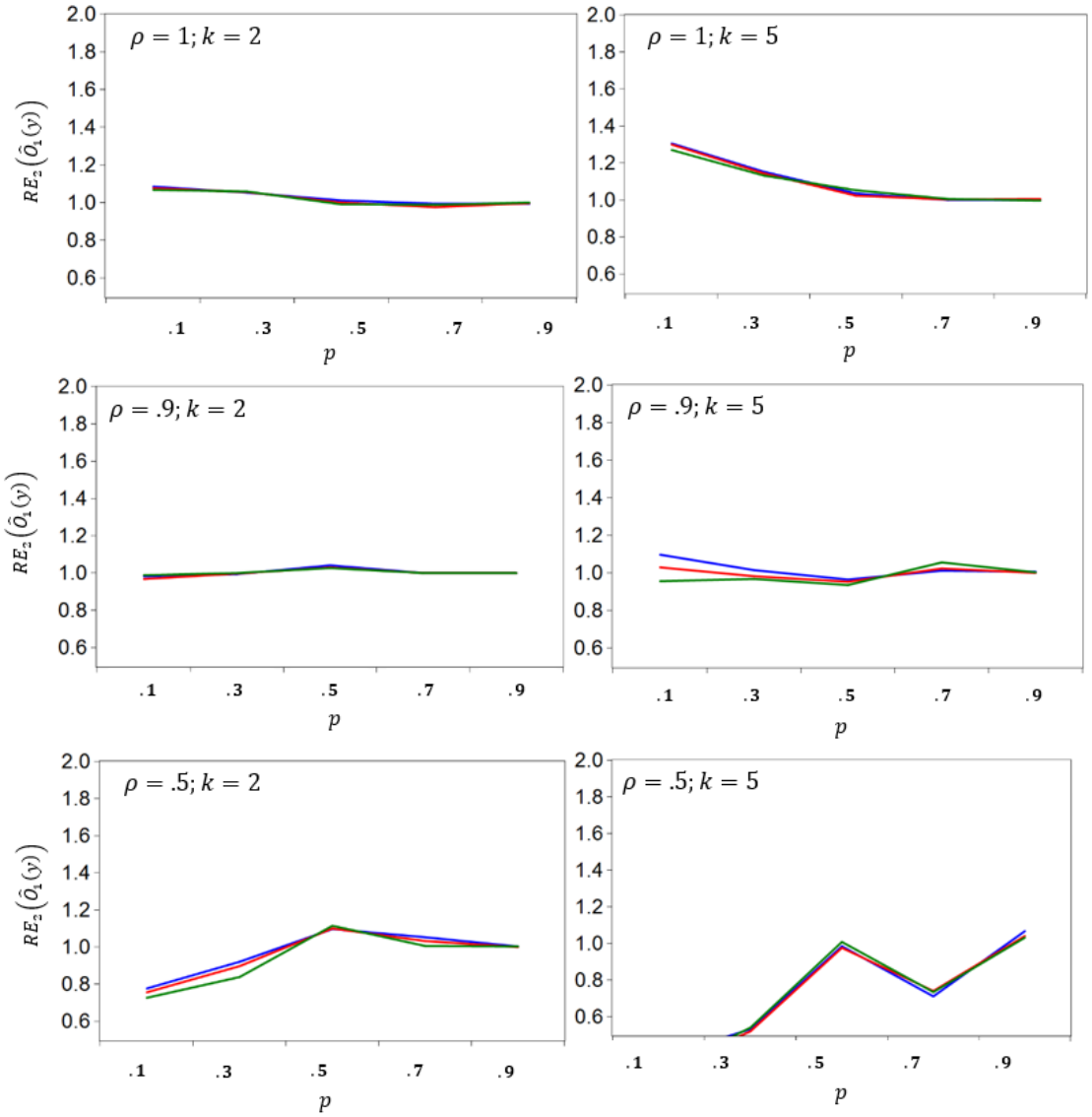
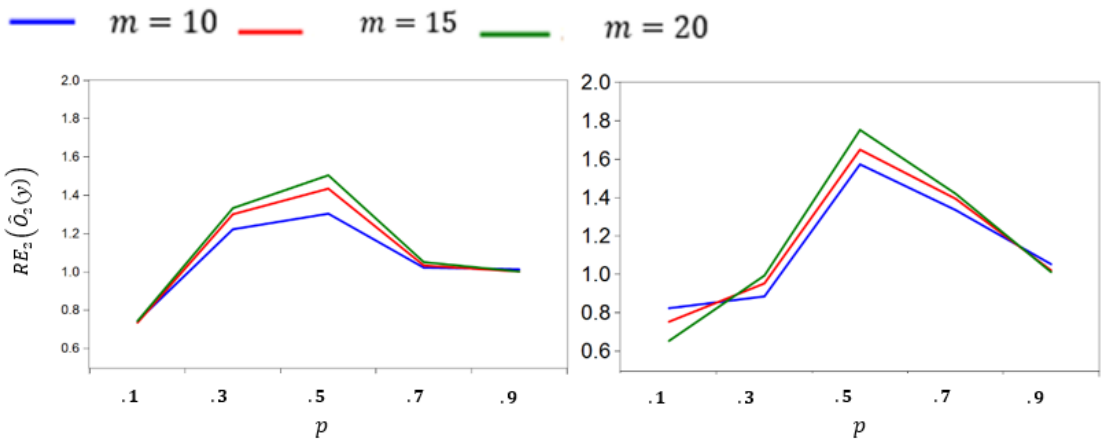


Fig. 2: The RE of $\hat{O}_1(y)$ with respect to $\hat{O}(y)$



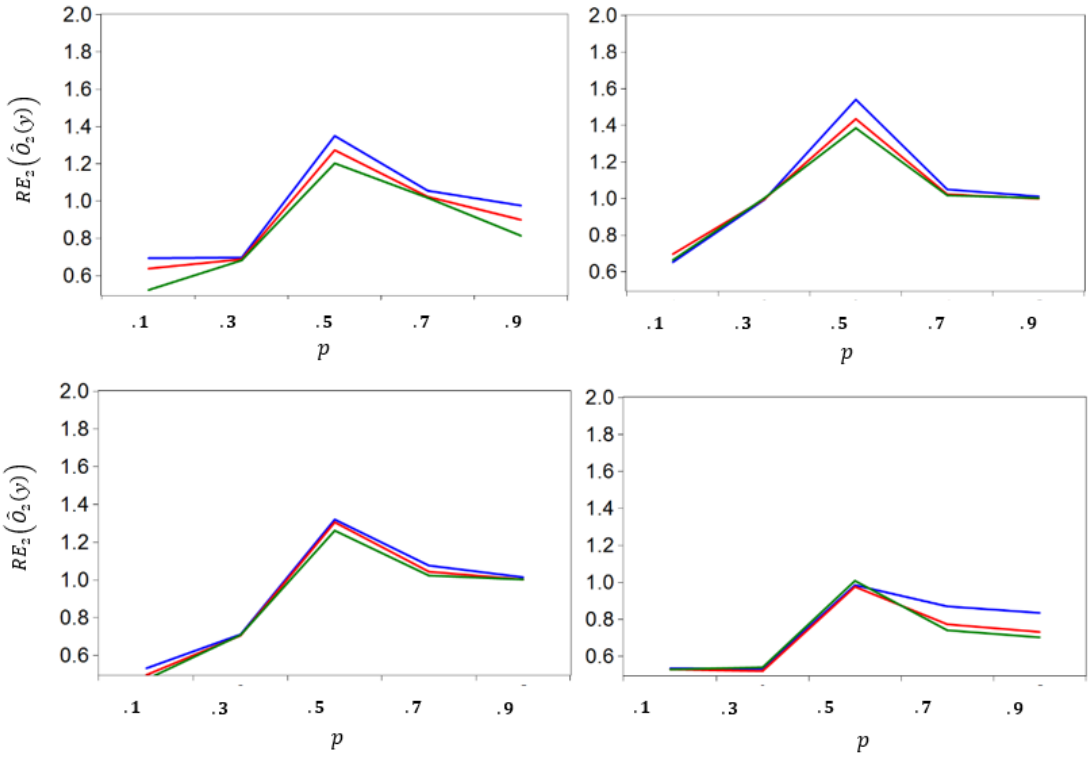


Fig. 3: The RE of $\hat{O}_2(y)$ with respect to $\hat{O}(y)$

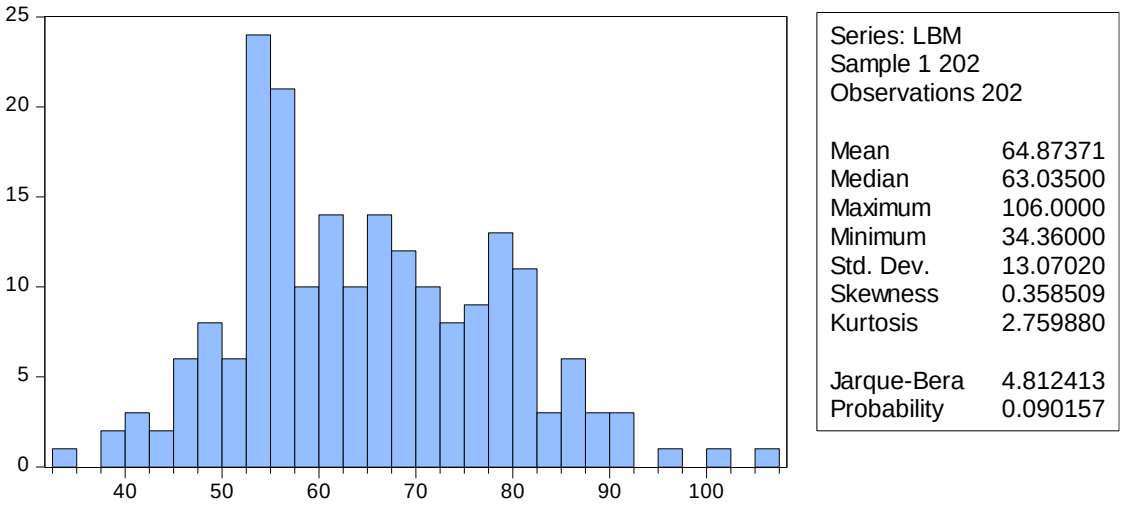


Fig. 4: The summary statistics of the lean body mass (LBM)