Pakistan Journal of Statistics and Operation Research

A closed-form estimator of $R = P(X < Y)$ based on ranked set sampling for a family of statistical distributions with application in agriculture

Hossein Pasha-Zanoosi[1*]

*Corresponding author

1. Department of Basic Sciences and General Courses, Faculty of Economics and Management, Khorramshahr University of Marine Science and Technology, Khorramshahr, Iran, pashazanoosi@yahoo.com

Abstract

In this article, we will derive a closed-form estimator for the probability $R = P(X < Y)$ based on a ranked set sampling (RSS) scheme when the he random variables X and Y are assumed to follow the Lehmann Type-II (L-II) family of distributions. Estimating R through the maximum likelihood (ML) method within the RSS framework does not yield an analytical solution because of the non-linear components present in the likelihood equations. In this context, we employ a modified maximum likelihood (MML) estimation approach to derive a closed-form estimator for R. Estimates of R under both ML and MML techniques along with their corresponding asymptotic confidence intervals are determined and compared in a simulation study under one of the distributions of the L-II family called the inverse Topp-Leone distribution. At the end, the simulation results are strengthened using a real example in the field of agriculture.

Key Words: Lehmann Type-II family of distributions; Ranked set sampling; $R = P(X < Y)$; Modified maximum likelihood estimation; Inverse Topp-Leone distribution.

Mathematical Subject Classification: 62F12, 62F10, 62N05.

1. Introduction

Ranked Set Sampling (RSS) is an innovative statistical method aimed at enhancing data collection efficiency, particularly in scenarios where obtaining measurements is challenging, costly, or labor-intensive. This technique involves the strategic selection of samples that leverage ranking information, thereby increasing the accuracy of statistical estimates. Originally developed by McIntyre (1952) for use in agricultural research, RSS has since been recognized for its ability to optimize sampling processes in a variety of fields. For example, in reliability engineering, consider a scenario where you are evaluating the durability of critical mechanical parts, such as bearings, that play a vital role in machinery. The goal is to determine the average time before these components fail. Performing actual failure tests can be prohibitively expensive. Instead, an alternative approach is to evaluate the components of each assembly using rapid, non-invasive testing methods or expert assessments that bypass the need for direct failure testing. For instance, you might rate them based on visual inspection for surface defects or initial performance in a short operational test. As another example, in medicine, suppose the purpose of a study is to estimate the prevalence of diabetes in a community. It would be time-consuming and expensive to perform detailed tests; instead, a simple blood glucose test (which is relatively inexpensive) can be used to rank people. A subset of people from each rank can then undergo more extensive testing, such as measuring HbA1c, to confirm diabetes.

The index $R = P(X < Y)$ is used as a statistical tool to evaluate two groups by measuring the probability that one random variable Y exceeds another random variable X. This metric is valuable in many disciplines, such as economics, psychology, and medical research, where it is critical to make comparisons between different groups or populations. For example, in clinical trials evaluating the effectiveness of two different treatments, researchers can use $R = P(X < Y)$ to determine whether one treatment is more effective than the other. Similarly, economists might use this index to analyze income disparities among different demographic segments. Since the introduction of $R = P(X < Y)$ parameter by Birnbaum (1956), numerous studies have explored its inference under various conditions. These investigations include scenarios where the variables X and Y are either independent or dependent, different methods of data collection (whether complete or incomplete), and different underlying distributions for X and Y. For example see Kundu and Gupta (2006), Kundu and Raqab (2009), Saraçoğlu et al. (2012), Nasiri (2016), Jovanović (2017), Pak et al. (2022), TANIŞ et al. (2023), and Aljohani (2024). The increasing use of the RSS design strategy in various fields has generated considerable interest among researchers in estimating the parameter $R = P(X < Y)$ associated with this design. Recent studies, including those by Akgül and Şenoğlu (2017), Akgül and Şenoğlu (2022), Chacko and Mathew (2019), and Hassan (2022) have highlighted the importance of this parameter in improving the effectiveness of RSS applications.

The Lehmann Type II (L-II) family of statistical distributions, developed by Cordeiro and De Castro (2011), is of paramount importance for the examination of reliability and survival data. This distribution family is held in high esteem for its efficacy in these domains, providing substantial insights and methodologies that enhance data analysis and inform decision-making processes.

The cumulative distribution function (CDF) of the L-II family of distributions with the domain $x\,(x > 0)$ and the shape parameter $\alpha\,(\alpha > 0)$ is defined as follows

$$F(x; \alpha) = 1 - [1 - G(x)]^{\alpha}, \tag{1}$$

where $G(x)$ is a baseline CDF. The corresponding probability density function (PDF) for the L-II family of distributions can be derived by differentiating the CDF with respect to x as follows

$$f(x; \alpha) = \alpha g(x) [1 - G(x)]^{\alpha - 1}, \tag{2}$$

where $g(x)$ is a baseline PDF corresponding to $G(x)$.

The L-II family includes many well-known distributions, such as the exponential, Kumaraswamy, Burr Type XII, Lomax, Chen, and Weibull distributions. In recent years, the extensive range of applications of this family of distributions has led to the introduction of new distributions.

To simplify notation, a random variable X following the L-II distribution with parameter α will be denoted by $X \sim L\text{-II}(\alpha)$. Now, suppose $X \sim L\text{-II}(\alpha_1)$ and $Y \sim L\text{-II}(\alpha_2)$ are independent. Then

$$\begin{aligned} R = P(X < Y) &= \int_0^{\infty} F_X(y)\, f_Y(y)\, dy \\ &= \int_0^{\infty} \Big[1 - [1 - G(y)]^{\alpha_1}\Big] \cdot \alpha_2\, g(y)\, [1 - G(y)]^{\alpha_2 - 1}\, dy. \end{aligned}$$

Applying the change of variable $t = 1 - G(y)$, $dt = -g(y)\, dy$ yields

$$R = \alpha_2 \int_0^1 \left[1 - t^{\alpha_1}\right] t^{\alpha_2 - 1}\, dt = \frac{\alpha_1}{\alpha_1 + \alpha_2}. \tag{3}$$

The findings from this study are applicable to all members of the L-II distribution family. For a particular distribution within this family, one simply needs to modify the CDF of its baseline, $G(x)$.

The article introduces two significant contributions to the field. Firstly, it provides the first estimation of R for the L-II family of distributions using the RSS scheme, which has not been previously explored. Secondly, it presents a closed-form estimator for R derived from an MML approach. It is important to highlight that closed-form ML estimates for R under the RSS scheme have not been available due to the nonlinear nature of the likelihood equations, which typically lack explicit solutions.

The subsequent sections will detail the continuation of this paper. In Sect. 2, we discuss the estimation of R using the ML method, which includes an iterative algorithm for optimization. We also provide a detailed explanation of how to construct the corresponding asymptotic confidence interval (ACI). In Sect. 3, we derive a closed-form estimator for R using an MML approach, and similarly, we elaborate on the process of developing the associated ACI. In Sect. 4, we explore the computations performed in this research under the inverse Topp-Leone distribution as an example of the L-II family of distributions. In Sect. 5, a simulation study is conducted to evaluate the effectiveness of the MML estimation approach in comparison to the traditional ML estimation method. This evaluation includes an analysis of their respective asymptotic intervals, which provides insight into the performance differences between the two approaches. In Sect. 6, the article examines real-world agricultural data using the methodologies discussed in previous sections. Sect. 7 then provides a summary and concluding thoughts on the findings presented throughout the paper.

2. Estimation of R using ML Method

This section begins by outlining the data collection method employed under the RSS framework. Following this, it discusses the estimation of R using the ML method, which includes the calculation of its asymptotic confidence interval.

To implement an RSS scheme, start by randomly selecting m (called set size) units from the population using the SRS scheme. Next, rank these m units based on some attribute that can be assessed using techniques such as visual comparison or expert judgement. From this ranked list, measure the unit identified as the smallest and include it in your sample. Continue this process for the second smallest unit and so on until all m ranks have been sampled. To increase the representativeness of your sample, repeat this entire process for multiple cycles, ultimately aiming to collect a total of $n = mr$ observations, where r is the number of cycles performed.

Now, let $X_{(i)ik}\,(k = 1, ..., r_x,\ i = 1, ..., m_x)$ be a collection of RSS scheme with a sample size of m, where r_x is the number of cycles and m_x is the set size taken from $X \sim L - II\,(\alpha_1)$. Similarly, let $Y_{(j)jl}(l = 1, ..., r_y,\ j = 1, \ldots, m_y)$ be a collection of RSS schemes with sample size n, also defined by set size m_y and number of cycles r_y, drawn from $Y \sim L - II\,(\alpha_2)$. The PDF of $X_{(i)ik}$, which is corresponds to its ith ordered statistic, is defined in the following manner.

$$f_i\,(x_{ik}) = \frac{m_x!}{(i-1)!\,(m_x - i)!}[F\,(x_{ik})]^{i-1}[1 - F\,(x_{ik})]^{m_x - i} f\,(x_{ik})\,, \tag{4}$$

where $X_{(i)ik}$ is represented by X_{ik} for convenience. The PDF of $Y_{(j)jl}$ is defined similarly. We will proceed to apply the ML approach to determine the unknown parameters of the model, (α_1 and α_2). Following this, the ML estimate for R can be derived from Equation (3) by utilizing the invariance property inherent in ML estimation. By drawing samples from the distributions $X \sim L - II\,(\alpha_1)$ and $Y \sim L - II\,(\alpha_2)$ using the RSS scheme, with sample sizes m and n respectively, the likelihood function based on Equations (1), (2), and (4) with constant C is formulated as follows:

$$\begin{aligned}
L &= \prod_{k=1}^{r_x}\prod_{i=1}^{m_x} f\,(x_{ik}) \prod_{l=1}^{r_y}\prod_{j=1}^{m_y} f\,(y_{jk}) \\
&= C{\alpha_1}^m{\alpha_2}^n \prod_{k=1}^{r_x}\prod_{i=1}^{m_x} g\,(x_{ik}) \prod_{l=1}^{r_y}\prod_{j=1}^{m_y} g\,(y_{jk}) \\
&\quad\times \prod_{k=1}^{r_x}\prod_{i=1}^{m_x} [1 - [1 - G\,(x_{ik})]^{\alpha_1}]^{i-1} \prod_{l=1}^{r_y}\prod_{j=1}^{m_y} [1 - [1 - G\,(y_{jk})]^{\alpha_2}]^{j-1} \\
&\quad\times \prod_{k=1}^{r_x}\prod_{i=1}^{m_x} [1 - G\,(x_{ik})]^{\alpha_1(m_x - i + 1) - 1} \prod_{l=1}^{r_y}\prod_{j=1}^{m_y} [1 - G\,(y_{jk})]^{\alpha_2(m_y - j + 1) - 1}.
\end{aligned}$$

To determine the log-likelihood function, we begin by applying the logarithm to Equation 1, resulting in the following calculation.

$$l = \ln C + m\ln\alpha_1 + n\ln\alpha_2 + \sum_{k=1}^{r_x}\sum_{i=1}^{m_x} \ln\left[g\,(x_{ik})\right] + \sum_{l=1}^{r_y}\sum_{j=1}^{m_y} \ln\left[g\,(y_{jl})\right]$$

$$+\sum_{k=1}^{r_x}\sum_{i=1}^{m_x}(i-1)\ln\left[1-\left[1-G\left(x_{ik}\right)\right]^{\alpha_1}\right]$$
$$+\sum_{l=1}^{r_y}\sum_{j=1}^{m_y}(j-1)\ln\left[1-\left[1-G\left(y_{jk}\right)\right]^{\alpha_2}\right]$$
$$+\sum_{k=1}^{r_x}\sum_{i=1}^{m_x}\left[\alpha_1\left(m_x-i+1\right)-1\right]\ln\left[1-G\left(x_{ik}\right)\right]$$
$$+\sum_{l=1}^{r_y}\sum_{j=1}^{m_y}\left[\alpha_2\left(m_y-j+1\right)-1\right]\ln\left[1-G\left(y_{jl}\right)\right],$$

and the corresponding likelihood equations are

$$\frac{m}{\alpha_1}-\sum_{k=1}^{r_x}\sum_{i=1}^{m_x}(i-1)\frac{\left[1-G\left(x_{ik}\right)\right]^{\alpha_1}\ln\left[1-G\left(x_{ik}\right)\right]}{1-\left[1-G\left(x_{ik}\right)\right]^{\alpha_1}}$$
$$+\sum_{k=1}^{r_x}\sum_{i=1}^{m_x}\left(m_x-i+1\right)\ln\left[1-G\left(x_{ik}\right)\right]=0, \tag{5}$$
$$\frac{n}{\alpha_2}-\sum_{l=1}^{r_y}\sum_{j=1}^{m_y}(j-1)\frac{\left[1-G\left(y_{jl}\right)\right]^{\alpha_2}\ln\left[1-G\left(y_{jl}\right)\right]}{1-\left[1-G\left(y_{jl}\right)\right]^{\alpha_2}}$$
$$+\sum_{l=1}^{r_y}\sum_{j=1}^{m_y}\left(m_y-j+1\right)\ln\left[1-G\left(y_{jl}\right)\right]=0. \tag{6}$$

The presence of non-linear terms in Equations (5) and (6) prevents us from deriving closed-form solutions for the ML estimates of α_1 and α_2. Consequently, we estimate the unknown parameters through the *optim* function in R software. Following this process, we can compute the ML estimate of R, denoted as $\hat{R}_{ML}$, using the formula $\hat{R}_{ML}=\hat{\alpha}_{1ML}/(\hat{\alpha}_{1ML}+\hat{\alpha}_{2ML})$.

To evaluate the precision of estimators, it is essential to derive the asymptotic variance-covariance matrix associated with them. This matrix can be computed by taking the inverse of the observed Fisher information matrix. The observed information matrix for ML estimator of $\delta=(\alpha_1,\alpha_2)$ is

$$Q\left(\delta\right)=-\begin{bmatrix}\frac{\partial^2 l}{\partial{\alpha_1}^2} & \frac{\partial^2 l}{\partial\alpha_1\partial\alpha_2}\\ \frac{\partial^2 l}{\partial\alpha_2\partial\alpha_1} & \frac{\partial^2 l}{\partial{\alpha_2}^2}\end{bmatrix}=\begin{bmatrix}q_{11} & q_{12}\\ q_{21} & q_{22}\end{bmatrix}.$$

The elements of the information matrix that has been analyzed are derived as follows:

$$q_{11}=\frac{m}{{\alpha_1}^2}+\sum_{k=1}^{r_x}\sum_{i=1}^{m_x}(i-1)\frac{\left[1-G\left(x_{ik}\right)\right]^{\alpha_1}\ln^2\left[1-G\left(x_{ik}\right)\right]}{\left(1-\left[1-G\left(x_{ik}\right)\right]^{\alpha_1}\right)^2},\qquad q_{12}=0, \tag{7}$$
$$q_{22}=\frac{n}{{\alpha_2}^2}+\sum_{l=1}^{r_y}\sum_{j=1}^{m_y}(j-1)\frac{\left[1-G\left(y_{jl}\right)\right]^{\alpha_2}\ln^2\left[1-G\left(y_{jl}\right)\right]}{\left(1-\left[1-G\left(y_{jl}\right)\right]^{\alpha_2}\right)^2},\qquad q_{21}=0. \tag{8}$$

The expectation of certain elements, which rely on the baseline CDF and PDF, is not always derived analytically. In such instances, the Fisher information matrix for the parameter δ, denoted as $I\left(\delta\right)=E\left[Q\left(\delta\right)\right]$, can be estimated using numerical methods. Alternatively, the observed information matrix can serve as a reliable estimator for $I\left(\delta\right)$. Consequently, when the analytical computation of the Fisher information matrix is not feasible, the observed information matrix $Q\left(\delta\right)$ plays a crucial role in establishing the asymptotic normality of the ML estimator.

As the sample size increases towards infinity, the ML estimator for the R parameter tends to follow a normal distribution. This distribution has a mean equal to R and an asymptotic variance denoted as $V(R)$, which can be obtained from the formula proposed by Rao as below

$$V(R) = \sum_{j=1}^{2}\sum_{i=1}^{2} \frac{\partial R}{\partial \delta_i}\frac{\partial R}{\partial \delta_j} p_{ij},$$

where the elements of matrix P, denoted as p_{ij}, represent the elements of the matrix that serves as the inverse of matrix Q. Thus we have

$$V(R) = \left(\frac{\partial R}{\partial \alpha_1}\right)^2 q_{11} + \left(\frac{\partial R}{\partial \alpha_2}\right)^2 q_{22},$$

where

$$\frac{\partial R}{\partial \alpha_1} = \frac{\alpha_2}{(\alpha_1+\alpha_2)^2}, \qquad \frac{\partial R}{\partial \alpha_2} = -\frac{\alpha_1}{(\alpha_1+\alpha_2)^2}.$$

As a result, the ACI for R based on the ML method at a level of $100(1-\omega)$ can be expressed as follows

$$\hat{R}_{ML} \pm Z_{\frac{\omega}{2}} \hat{V}_{ML},$$

where $Z_{\frac{\omega}{2}}$ signifies the upper $\frac{\omega}{2}$ quantile from the standard normal distribution, whereas V_{ML} refers to the value of V at the point of ML estimation for the parameters.

3. Estimation of R using MML Approach

In the previous discussions, we noted that equations 1 and 2 do not yield explicit solutions because they involve nontractable elements. In this section, we will use an MML approach, originally introduced by Mehrotra and Nanda (1974), to estimate the model parameters that facilitates the derivation of explicit solutions for these parameters. The MML approach offers a significant advantage over the ML method by providing an explicit solution for parameter estimation. Estimates from the MML approach can serve as initial values in the iterative process used to solve the likelihood equations, thereby facilitating the determination of the ML estimates. In this context, Mehrotra and Nanda (1974) suggest a technique that involves replacing nonlinear elements with their expected values in the likelihood equations.

Let us denote the nonlinear expression in Equation (5) as $S_i(ik)$ and its expected value as S. In this case we will have

$$\begin{aligned} S = E\left[s_i(x_{ik})\right] &= E\left[(i-1)\frac{[1-G(x_{ik})]^{\alpha_1}\ln[1-G(x_{ik})]}{1-[1-G(x_{ik})]^{\alpha_1}}\right] \\ &= (i-1)\int_0^{+\infty}\frac{[1-G(x_{ik})]^{\alpha_1}\ln[1-G(x_{ik})]}{1-[1-G(x_{ik})]^{\alpha_1}} f_i(x_{ik})\,dx_{ik}. \end{aligned} \tag{9}$$

By replacing Equation (4) in Equation (9) and putting $A = \frac{(i-1)m_x!}{(i-1)!(m_x-i)!}$, we get the following integral

$$\begin{aligned} S = A\int_0^{+\infty} & [1-G(x_{ik})]^{\alpha_1(m_x-i+1)}\ln[1-G(x_{ik})] \\ & \times (1-[1-G(x_{ik})])^{i-2} f(x_{ik})\,dx_{ik}. \end{aligned}$$

The integral S is calculated by using the formula 4.253 (1) in Gradshteyn and Ryzhik (1994) and by changing the variable $Z = [1-G(x_{ik})]^{\alpha_1}$ as follows

$$\begin{aligned} S &= \frac{A}{\alpha_1}\int_0^{+\infty} z^{m_x-i+1}\ln z(1-z)^{i-2}dz \\ &= \frac{1}{\alpha_1}(m_x-i+1)\left[\psi(m_x-i+2)-\psi(m_x+1)\right], \end{aligned} \tag{10}$$

where the symbol $\psi(.)$ represents the digamma function. By replacing Equation (10) in Equation (5), we get the following likelihood equation

$$\frac{m}{\alpha_1} - \frac{r_x}{\alpha_1}\sum_{i=1}^{m_x}(m_x - i + 1)\left[\psi(m_x - i + 2) - \psi(m_x + 1)\right] + \sum_{k=1}^{r_x}\sum_{i=1}^{m_x}(m_x - i + 1)\ln\left[1 - G(x_{ik})\right] = 0.$$

Therefore, the estimation of parameter α_1 is based on the MML approach, say $\hat{\alpha}_{1MML}$, is calculated from the following equation

$$\hat{\alpha}_{1MML} = \frac{r_x\sum\limits_{i=1}^{m_x}(m_x - i + 1)\left[\psi(m_x - i + 2) - \psi(m_x + 1)\right] - m}{\sum\limits_{k=1}^{r_x}\sum\limits_{i=1}^{m_x}(m_x - i + 1)\ln\left[1 - G(x_{ik})\right]}. \tag{11}$$

The estimation of parameter α_2, say $\hat{\alpha}_{2MML}$, is determined as follows in a similar process according to the MML approach

$$\hat{\alpha}_{2MML} = \frac{r_y\sum\limits_{j=1}^{m_y}(m_y - j + 1)\left[\psi(m_y - j + 2) - \psi(m_y + 1)\right] - n}{\sum\limits_{l=1}^{r_y}\sum\limits_{j=1}^{m_y}(m_y - j + 1)\ln\left[1 - G(y_{jl})\right]}. \tag{12}$$

Following this process, we can compute the MML estimate of R, denoted as $\hat{R}_{MML}$, using the formula $\hat{R}_{MML} = \hat{\alpha}_{1MML}/(\hat{\alpha}_{1MML} + \hat{\alpha}_{2MML})$. In this case, the elements of the observed Fisher information matrix, say $Q^*(\delta)$, are calculated from the following equations to create a confidence interval.

$$Q^*(\delta) = \begin{bmatrix} q_{11}^* & q_{12}^* \\ q_{21}^* & q_{22}^* \end{bmatrix},$$

where

$$q_{11}^* = \frac{m}{{\alpha_1}^2} - \frac{r_x}{{\alpha_1}^2}\sum_{i=1}^{m_x}(m_x - i + 1)\left[\psi(m_x - i + 2) - \psi(m_x + 1)\right], \quad q_{12}^* = 0, \tag{13}$$

$$q_{22}^* = \frac{n}{{\alpha_2}^2} - \frac{r_y}{{\alpha_2}^2}\sum_{j=1}^{m_y}(m_y - j + 1)\left[\psi(m_y - j + 2) - \psi(m_y + 1)\right], \quad q_{21}^* = 0, \tag{14}$$

The subsequent steps are completely similar to those proposed to construct the ACI for R based on the estimates obtained by the ML method.

4. Example

This section examines the inverse Topp-Leone distribution, introduced by Hassan et al. (2020) as a member of the L-II family of distributions. The CDF and PDF for the inverse Topp-Leone distribution are defined as follows

$$F(x;\alpha) = 1 - \left[1 - \left(\frac{x}{1+x}\right)^2\right]^\alpha, \quad x \geq 0, \quad \alpha > 0,$$

and

$$f(x;\alpha) = \frac{2\alpha x}{(1+x)^3}\left[1 - \left(\frac{x}{1+x}\right)^2\right]^{\alpha-1}, \quad x \geq 0, \quad \alpha > 0.$$

where $G(x) = (\frac{x}{1+x})^2$ and $g(x) = \frac{2\alpha x}{(1+x)^3}$. The distribution in question offers numerous benefits that justify its emphasis. Firstly, it features a straightforward form for its CDF. Secondly, both its PDF and hazard rate function

(HRF) exhibit a unimodal and right-skewed characteristic. Additionally, the distribution provides explicit formulas for the quantiles and mode, which facilitates the exploration of various statistical properties. Lastly, with only one shape parameter, it presents a valuable tool for statisticians conducting inference through diverse methodologies (Elgarhy et al., 2023).

The outcomes presented in Sect. 2 yield the following findings.

(i) The estimates of α_1 and α_2 based on Equations (5) and (6) are obtained by solving the following nonlinear equations using the ML method

$$\frac{m}{\alpha_1} - \sum_{k=1}^{r_x}\sum_{i=1}^{m_x}(i-1)\frac{\left[1-\left(\frac{x_{ik}}{1+x_{ik}}\right)^2\right]^{\alpha_1}\ln\left[1-\left(\frac{x_{ik}}{1+x_{ik}}\right)^2\right]}{1-\left[1-\left(\frac{x_{ik}}{1+x_{ik}}\right)^2\right]^{\alpha_1}}$$
$$+\sum_{k=1}^{r_x}\sum_{i=1}^{m_x}(m_x-i+1)\ln\left[1-\left(\frac{x_{ik}}{1+x_{ik}}\right)^2\right]=0,$$
$$\frac{n}{\alpha_2} + \sum_{l=1}^{r_y}\sum_{j=1}^{m_y}(j-1)\frac{\left[1-\left(\frac{y_{jl}}{1+y_{jl}}\right)\right]^{\alpha_2}\ln\left[1-\left(\frac{y_{jl}}{1+y_{jl}}\right)\right]}{1-\left[1-\left(\frac{y_{jl}}{1+y_{jl}}\right)\right]^{\alpha_2}}$$
$$+\sum_{l=1}^{r_y}\sum_{j=1}^{m_y}(m_y-j+1)\ln\left[1-\left(\frac{y_{jl}}{1+y_{jl}}\right)\right]=0.$$

To solve the above equations, we use the *optim* function in R software.

(ii) The elements of the observed information matrix based on Equations (7) and (8) are obtained as

$$q_{11}=\frac{m}{{\alpha_1}^2}+\sum_{k=1}^{r_x}\sum_{i=1}^{m_x}(i-1)\frac{\left[1-\left(\frac{x_{ik}}{1+x_{ik}}\right)^2\right]^{\alpha_1}\ln^2\left[1-\left(\frac{x_{ik}}{1+x_{ik}}\right)^2\right]}{\left(1-\left[1-\left(\frac{x_{ik}}{1+x_{ik}}\right)^2\right]^{\alpha_1}\right)^2},\qquad q_{12}=0,$$
$$q_{22}=\frac{n}{{\alpha_2}^2}+\sum_{l=1}^{r_y}\sum_{j=1}^{m_y}(j-1)\frac{\left[1-\left(\frac{y_{jl}}{1+y_{jl}}\right)^2\right]^{\alpha_2}\ln^2\left[1-\left(\frac{y_{jl}}{1+y_{jl}}\right)^2\right]}{\left(1-\left[1-\left(\frac{y_{jl}}{1+y_{jl}}\right)^2\right]^{\alpha_2}\right)^2},\qquad q_{21}=0.$$

(iii) The estimates of α_1 and α_2 based on Equations (11) and (12) and using the MML approach are

$$\hat{\alpha}_{1MML}=\frac{r_x\sum\limits_{i=1}^{m_x}(m_x-i+1)\left[\psi(m_x-i+2)-\psi(m_x+1)\right]-m}{\sum\limits_{k=1}^{r_x}\sum\limits_{i=1}^{m_x}(m_x-i+1)\ln\left[1-\left(\frac{x_{ik}}{1+x_{ik}}\right)^2\right]},$$
$$\hat{\alpha}_{2MML}=\frac{r_y\sum\limits_{j=1}^{m_y}(m_y-j+1)\left[\psi(m_y-j+2)-\psi(m_y+1)\right]-n}{\sum\limits_{l=1}^{r_y}\sum\limits_{j=1}^{m_y}(m_y-j+1)\ln\left[1-\left(\frac{y_{jl}}{1+y_{jl}}\right)^2\right]}.$$

The elements of the observed information matrix in this case for all members of this family are determined from Equations (13) and (14) and do not depend on the baseline distribution.

5. Simulation

In this section, we conduct a simulation study to evaluate the effectiveness of the $R = P(X < Y)$ estimator derived from the MML approach in comparison to the ML method. We focus on evaluating both point and interval estimates, where X and Y are drawn from the inverse Topp-Leone distribution with parameters of α_1 and α_2, respectively. The data used for this analysis are collected according to the RSS scheme. We use several key criteria for comparison, including average bias (AB), mean square error (MSE), and average length (AL) of intervals, along with their corresponding coverage probability (CP). Also, we present the relative efficiency (RE) calculated as $RE = MSE\left(\hat{R}_{ML}\right)/MSE\left(\hat{R}_{MML}\right)$. When RE equals 1, it signifies that $\hat{R}_{ML}$ and $\hat{R}_{MML}$ exhibit identical performance levels.

In the configuration of the simulation, we define the set sizes as $m_x = m_y =$ 1, 2, 3, 4, and 5, while the number of cycles is established as $r_x = r_y =$10. Consequently, the resulting sample sizes are determined to be $m = n =$10, 20, 30, 40, and 50. It is important to highlight that when $m_x = m_y = 1$, the samples derived from the RSS design are essentially identical to those obtained from the simple random sampling (SRS) design. We consider the (α_1, α_2) pairs as (2,18), (4,16), (6,14), (8,12), (10,10), (12,8), (14,6), (16,4), and (18,2) for parameter settings corresponding to true $R = P(X < Y)$ values of 0.1, 0.2, . . . , 0.9, respectively. R software with 5000 samples was used for all calculations.

The simulation results are shown in Tables 1 and 2. The values of AB in estimating R have been negligible. In this context, the MML approach has demonstrated a marginally superior performance compared to the ML method. When R is less than 0.5, the estimates of R tend to be overestimated, while values above 0.5 show a tendency for underestimation. Notably, at $R = 0.50$, the AB values are significantly low. The results indicate that the performance of the two methods examined is nearly equivalent, as evidenced by the RE values, which are all approximately equal to 1. At $R = 0.5$ the simulation shows the highest values for both MSEs and ALs. In contrast, the lowest MSEs and ALs are recorded at $R = 0.1$ and $R = 0.9$. The values of MSEs and ALs decreased significantly with the increase in the set sizes. The effectiveness of the asymptotic intervals in the two methods studied, evaluated by both the AL and CP criteria, proved to be highly satisfactory. In particular, the ML method outperformed under the AL criterion, while the MML approach outperformed under the CP criterion.

6. Real Data Analysis

In a study conducted by Murray et al. (2000), apple trees were treated with a fluorescent tracer, Tinopal CBS-X, at a concentration of 2% diluted in water. The experiment involved two plots of ninety trees each, where one was subjected to high-volume spraying using coarse nozzles, resulting in larger droplets, while the other experienced low-volume spraying with fine nozzles, producing smaller droplets. To assess the effectiveness of the spray coverage, a total of fifty sets, each consisting of five leaves, were selected from the central five trees within each plot. These sets were then utilized to create an RSS scheme with a set size of 5 and a cycle of 10. The primary focus was on determining the percentage of leaf surface area covered by the spray. Although chemical analysis of the leaf surfaces provided precise measurements, it was noted to be labor-intensive and costly. Alternatively, visual assessments under ultraviolet light were employed to evaluate spray deposits, offering a more economical and reasonably accurate method when performed by trained observers. The data collected from both spraying methods were compared to ascertain whether variations in nozzle settings influenced the extent of spray coverage on the leaves. The information is presented in Table 3, which includes measurements from the plot treated with high volume spray, serving as the control group, and those treated with low volume spray, representing the experimental group. The primary focus is to determine if the settings of the sprayer influence the percentage area coverage. In this context, if X and Y represent the response variables for the treatment and control groups respectively, then $R = P(X < Y)$ can be utilized as an indicator of the treatment effect. Therefore, the objective of this analysis is to estimate $R = P(X < Y)$.

Before delving into the analysis of this data set, we first verify that the inverse Topp-Leone distribution adequately models the percentage of sprayed leaves on apple trees, considering both high and low volume scenarios (i.e., X and Y). To evaluate the fit of the inverse Topp-Leone distribution to the data sets, we employ the Kolmogorov-Smirnov (K-S) goodness of fit test. The calculated K-S test statistics for the data sets X and Y are 0.1151 and 0.1199, respectively. Additionally, their associated P-values are 0.4871 and 0.4341. These results indicate that the Topp-Leon inverse distribution adequately models both data sets.

We now present the estimates of R derived from ML and MML techniques within the framework of the RSS scheme. Additionally, we construct 95% ACI for R based on these two methodologies. The findings are summarized in Table 4. As indicated in this table, both ML and MML approaches have demonstrated comparable effectiveness in terms of point and interval estimation. Moreover, the estimated value of R is approximately 0.65, indicating that the probability of the coverage area for apple tree leaves sprayed at a high volume is lower than that for those sprayed at a low volume.

Table 1: Estimates of R based on the ML and MML techniques, along with the corresponding ABs and MSE values

		ML			MML			
(m_x, m_y)	R	$\hat{R}_{ML}$	AB	MSE	$\hat{R}_{MML}$	AB	MSE	RE
(1,1)	0.1	0.10846	0.00846	0.00210	0.10683	0.00683	0.00201	1.04
(2,2)		0.10279	0.00279	0.00065	0.10285	0.00285	0.00064	1.02
(3,3)		0.10127	0.00127	0.00031	0.10120	0.00120	0.00033	0.94
(4,4)		0.10107	0.00107	0.00018	0.10072	0.00072	0.00020	0.90
(5,5)		0.10086	0.00086	0.00012	0.10068	0.00068	0.00012	1.00
(1,1)	0.2	0.20883	0.00883	0.00556	0.20948	0.00948	0.00563	0.99
(2,2)		0.20365	0.00365	0.00193	0.20328	0.00328	0.00195	0.99
(3,3)		0.20077	0.00077	0.00096	0.20248	0.00248	0.00098	0.98
(4,4)		0.20166	0.00166	0.00058	0.20170	0.00170	0.00063	0.92
(5,5)		0.20060	0.00060	0.00038	0.20039	0.00039	0.00039	0.97
(1,1)	0.3	0.30868	0.00868	0.00905	0.30790	0.00790	0.00896	1.01
(2,2)		0.30231	0.00231	0.00314	0.30264	0.00264	0.00315	1.00
(3,3)		0.30155	0.00155	0.00166	0.30205	0.00205	0.00164	1.01
(4,4)		0.30039	0.00039	0.00099	0.30067	0.00067	0.00100	0.99
(5,5)		0.30100	0.00100	0.00068	0.30046	0.00046	0.00068	1.00
(1,1)	0.4	0.40530	0.00530	0.01060	0.40420	0.00420	0.01117	0.95
(2,2)		0.40149	0.00149	0.00408	0.40219	0.00219	0.00406	1.00
(3,3)		0.40060	0.00060	0.00210	0.40117	0.00117	0.00220	0.95
(4,4)		0.40053	0.00053	0.00126	0.40021	0.00021	0.00034	0.94
(5,5)		0.40133	0.00133	0.00090	0.40028	0.00028	0.00089	1.01
(1,1)	0.5	0.50276	0.00276	0.01218	0.50043	0.00043	0.01222	1.00
(2,2)		0.50018	0.00018	0.00427	0.49960	-0.00040	0.00450	0.95
(3,3)		0.50025	0.00025	0.00226	0.50076	0.00076	0.00235	0.96
(4,4)		0.50016	0.00016	0.00141	0.49967	-0.00033	0.00140	1.01
(5,5)		0.50091	0.00091	0.00094	0.49906	-0.00094	0.00097	0.97
(1,1)	0.6	0.59780	-0.00220	0.01114	0.59305	-0.00695	0.01132	0.98
(2,2)		0.59928	-0.00072	0.00405	0.59804	-0.00196	0.00411	0.99
(3,3)		0.59862	-0.00138	0.00217	0.59963	-0.00037	0.00214	1.01
(4,4)		0.59961	-0.00039	0.00132	0.59973	-0.00027	0.00136	0.97
(5,5)		0.59982	-0.00018	0.00088	0.59963	-0.00037	0.00090	0.98
(1,1)	0.7	0.69268	-0.00732	0.00881	0.69232	-0.00768	0.00875	1.01
(2,2)		0.69784	-0.00216	0.00317	0.69663	-0.00337	0.00332	0.95
(3,3)		0.69882	-0.00118	0.00165	0.69925	-0.00075	0.00162	1.02
(4,4)		0.69891	-0.00109	0.00100	0.69912	-0.00088	0.00103	0.97
(5,5)		0.69976	-0.00024	0.00066	0.69909	-0.00091	0.00069	0.96
(1,1)	0.8	0.79194	-0.00806	0.00546	0.79050	-0.00950	0.00565	0.97
(2,2)		0.79477	-0.00523	0.00197	0.79688	-0.00312	0.00191	1.03
(3,3)		0.79862	-0.00138	0.00094	0.79808	-0.00192	0.00100	0.94
(4,4)		0.79901	-0.00099	0.00059	0.79897	-0.00103	0.00060	0.98
(5,5)		0.79869	-0.00131	0.00040	0.79944	-0.00056	0.00041	0.98
(1,1)	0.9	0.89258	-0.00742	0.00197	0.89258	-0.00742	0.00209	0.94
(2,2)		0.89745	-0.00255	0.00062	0.89777	-0.00223	0.00062	1.00
(3,3)		0.89875	-0.00125	0.00030	0.89839	-0.00161	0.00032	0.94
(4,4)		0.89885	-0.00115	0.00019	0.89963	-0.00037	0.00019	1.00
(5,5)		0.89951	-0.00049	0.00012	0.89942	-0.00058	0.00013	0.92

7. Conclusions

In this study, a modified maximum likelihood (MML) approach was employed as an alternative to the traditional maximum likelihood (ML) method to derive an explicit solution for estimating the probability of $R = P(X < Y)$. The variables X and Y are drawn from a distribution belonging to the Lehmann Type-II (L-II) family, with samples collected using the ranked set sampling (RSS) scheme.

The simulation results indicated that both the ML and MML estimation techniques produced highly satisfactory and nearly identical results. In particular, the MML method showed slightly better performance in terms of average bias (AB) and coverage probability (CP) compared to the ML approach. Conversely, the ML method outperformed MML

Table 2: ALs for estimates of R based on the ML and MML techniques, along with the corresponding CPs values

		ML		MML	
(m_x, m_y)	R	AL	CP	AL	CP
(1,1)	0.1	0.16596	0.9224	0.16382	0.9174
(2,2)		0.09581	0.9352	0.10160	0.9464
(3,3)		0.06827	0.9404	0.07490	0.9564
(4,4)		0.05344	0.9530	0.05988	0.9670
(5,5)		0.04390	0.9470	0.05013	0.9728
(1,1)	0.2	0.28003	0.9168	0.28058	0.9196
(2,2)		0.16765	0.9404	0.17742	0.9488
(3,3)		0.12008	0.9378	0.13264	0.9634
(4,4)		0.09454	0.9458	0.10627	0.9604
(5,5)		0.07754	0.9582	0.08861	0.9728
(1,1)	0.3	0.35836	0.9158	0.35798	0.9154
(2,2)		0.21739	0.9368	0.23051	0.9514
(3,3)		0.15729	0.9458	0.17286	0.9634
(4,4)		0.12326	0.9438	0.13866	0.9692
(5,5)		0.10165	0.9494	0.11614	0.9700
(1,1)	0.4	0.40400	0.9258	0.40262	0.9162
(2,2)		0.24711	0.9392	0.26207	0.9528
(3,3)		0.17915	0.9430	0.19671	0.9602
(4,4)		0.14073	0.9498	0.15816	0.9694
(5,5)		0.11600	0.9492	0.13258	0.9712
(1,1)	0.5	0.41690	0.9074	0.41684	0.9130
(2,2)		0.25706	0.9444	0.27219	0.9454
(3,3)		0.18650	0.9446	0.20466	0.9606
(4,4)		0.14649	0.9508	0.16472	0.9680
(5,5)		0.12073	0.9524	0.13805	0.9710
(1,1)	0.6	0.40198	0.9160	0.40333	0.9134
(2,2)		0.24695	0.9394	0.26197	0.9494
(3,3)		0.17922	0.9416	0.19663	0.9600
(4,4)		0.14070	0.9440	0.15815	0.9650
(5,5)		0.11592	0.9464	0.13259	0.9712
(1,1)	0.7	0.35782	0.9166	0.35819	0.9208
(2,2)		0.21728	0.9372	0.23064	0.9458
(3,3)		0.15717	0.9424	0.17246	0.9608
(4,4)		0.12343	0.9458	0.13870	0.9646
(5,5)		0.10153	0.9484	0.11624	0.9716
(1,1)	0.8	0.27939	0.9200	0.28057	0.9206
(2,2)		0.16861	0.9384	0.17735	0.9492
(3,3)		0.12035	0.9450	0.13235	0.9588
(4,4)		0.09430	0.9466	0.10602	0.9702
(5,5)		0.07774	0.9470	0.08866	0.9698
(1,1)	0.9	0.16474	0.9164	0.16453	0.9150
(2,2)		0.09565	0.9406	0.10107	0.9474
(3,3)		0.06826	0.9474	0.07517	0.9612
(4,4)		0.05347	0.9464	0.05970	0.9650
(5,5)		0.04375	0.9498	0.05008	0.9712

in terms of mean squared error (MSE) and average length (AL) of the asymptotic confidence intervals.

Table 3: The RSS scheme for the real data

Group	Cycle	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
X	1	0.003	0.028	0.244	0.057	0.143
	2	0.039	0.119	0.126	0.105	0.565
	3	0.034	0.118	0.130	0.218	0.296
	4	0.051	0.104	0.193	0.210	0.150
	5	0.032	0.141	0.130	0.250	0.229
	6	0.069	0.070	0.260	0.225	0.285
	7	0.100	0.091	0.244	0.130	0.347
	8	0.012	0.096	0.069	0.373	0.133
	9	0.046	0.117	0.126	0.223	0.273
	10	0.028	0.083	0.108	0.212	0.261
Y	1	0.036	0.137	0.183	0.270	0.487
	2	0.250	0.181	0.290	0.328	0.715
	3	0.089	0.032	0.269	0.419	0.315
	4	0.180	0.111	0.130	0.194	0.742
	5	0.100	0.009	0.184	0.277	0.122
	6	0.042	0.089	0.199	0.269	0.395
	7	0.044	0.083	0.227	0.177	0.742
	8	0.044	0.171	0.067	0.192	0.336
	9	0.009	0.017	0.217	0.438	0.544
	10	0.071	0.132	0.310	0.343	0.379

Table 4: The point estimate of the unknown parameters and the point and interval estimates of R for real data under ML and MML techniques

	Point estimates			Interval estimates	
Method	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{R}$	Interval	Length
ML	46.1000	24.6563	0.6515	(0.5960,0.7071)	0.1111
MML	45.9251	25.0163	0.6474	(0.5841,0.7106)	0.1266

References

Akgül, F. G. and Şenoğlu, B. (2017). Estimation of $P(X < Y)$ using ranked set sampling for the weibull distribution. *Quality Technology & Quantitative Management*, 14(3):296–309.

Akgül, F. G. and Şenoğlu, B. (2022). Inferences for stress–strength reliability of burr type x distributions based on ranked set sampling. *Communications in Statistics-Simulation and Computation*, 51(6):3324–3340.

Aljohani, H. M. (2024). Estimation for the $P(X > Y)$ of lomax distribution under accelerated life tests. *Heliyon*, 10(3).

Birnbaum, Z. (1956). On a use of the mann-whitney statistic. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 13–18. University of California Press.

Chacko, M. and Mathew, S. (2019). Inference on $P(X < Y)$ for bivariate normal distribution based on ranked set sample. *Metron*, 77(3):239–252.

Cordeiro, G. M. and De Castro, M. (2011). A new family of generalized distributions. *Journal of statistical computation and simulation*, 81(7):883–898.

Elgarhy, M., Alsadat, N., Hassan, A. S., Chesneau, C., and Abdel-Hamid, A. H. (2023). A new asymmetric modified topp–leone distribution: Classical and bayesian estimations under progressive type-ii censored data with applications. *Symmetry*, 15(7):1396.

Gradshteyn, I. and Ryzhik, I. (1994). *Table of integrals, series and products*. Academic Press, Boston.

Hassan, A. S., Elgarhy, M., and Ragab, R. (2020). Statistical properties and estimation of inverted topp-leone distribution. *J. Stat. Appl. Probab*, 9(2):319–331.

Hassan, M. K. (2022). Ranked set sampling on estimation of $P[Y < X]$ for inverse weibull distribution and its applications. *International Journal of Quality & Reliability Management*, 39(7):1535–1550.

Jovanović, M. (2017). Estimation of $P(X < Y)$ for geometric—exponential model based on complete and censored samples. *Communications in Statistics-Simulation and Computation*, 46(4):3050–3066.

Kundu, D. and Gupta, R. D. (2006). Estimation of $P[Y < X]$ for weibull distributions. *IEEE Trans. Reliab.*, 55(2):270–280.

Kundu, D. and Raqab, M. Z. (2009). Estimation of $R = P(Y < X)$ for three-parameter weibull distribution. *Statistics & Probability Letters*, 79(17):1839–1846.

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian journal of agricultural research*, 3(4):385–390.

Mehrotra, K. and Nanda, P. (1974). Unbiased estimation of parameters by order statistics in the case of censored samples. *Biometrika*, pages 601–606.

Murray, R., Ridout, M., Cross, J., et al. (2000). The use of ranked set sampling in spray deposit assessment. *Aspects of Applied Biology*, 57:141–146.

Nasiri, P. (2016). Estimation parameter of $R = P(Y < X)$ for lomax distribution with presence of outliers. In *International Mathematical Forum*, volume 11, pages 239–248.

Pak, A., Raqab, M., Mahmoudi, M. R., Band, S. S., and Mosavi, A. (2022). Estimation of stress-strength reliability $R = P(X > Y)$ based on weibull record data in the presence of inter-record times. *Alexandria Engineering Journal*, 61(3):2130–2144.

Saraçoğlu, B., Kinaci, I., and Kundu, D. (2012). On estimation of $R = P(Y < X)$ for exponential distribution under progressive type-ii censoring. *Journal of Statistical Computation and Simulation*, 82(5):729–744.

TANIŞ, C., Saraçoğlu, B., Asgharzadeh, A., and Mousa, A. (2023). Estimation of pr $(X < Y)$ for exponential power records. *Hacettepe Journal of Mathematics and Statistics*, pages 1–13.