A ranked-based estimator of the mean past lifetime with an application

Elham Zamanzade · Majid Asadi · Afshin Parvardeh · Ehsan Zamanzade

Received: 26 May 2021 / Revised: 12 March 2022 / Accepted: 10 April 2022 / Published online: 30 April 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
The mean past lifetime (MPL) is an important tool in reliability and survival analysis for measuring the average time elapsed since the occurrence of an event, under the condition that the event has occurred before a specific time $t > 0$. This article develops a nonparametric estimator for MPL based on observations collected according to ranked set sampling (RSS) design. It is shown that the proposed estimator is a strongly uniform consistent estimator of MPL. It is also proved that the introduced estimator tends to a Gaussian process under some mild conditions. A Monte Carlo simulation study is employed to evaluate the performance of the proposed estimator with its competitor in simple random sampling (SRS). Our findings show the introduced estimator is more efficient than its counterpart estimator in SRS as long as the quality of ranking is better than random. Finally, an illustrative example is provided to describe the potential application of the developed estimator in assessing the average time between the infection and diagnosis in HIV patients.

Keywords Asymptotic Gaussian · Mean past lifetime · Ranked set sampling · Judgement ranking

Mathematics Subject Classification 62D05 · 62N02

1 Introduction

The mean past lifetime (MPL) of a nonnegative random variable $X$ with the cumulative distribution function (CDF) $F(t) = \mathbb{P}(X \leq t)$ is defined as

$$K(t) = \mathbb{E}(t - X | X \leq t).$$ (1)
The functional form of the MPL function \( K(t) \) in terms of the CDF \( F \) is given by

\[
K(t) = \int_0^t F(x) \, dx \quad \frac{F(t)}{t}, \quad t > 0,
\]

provided that \( F(t) > 0 \). Asadi and Berred (2012) investigated some of MPL properties in connection with other reliability measures and discussed its estimation based on simple random sampling (SRS). Parvardeh (2015) studied the asymptotic behaviour of the empirical estimator of MPL function based on SRS.

Ranked set sampling (RSS), introduced by McIntyre (1952), is a useful sampling technique when the exact values of the sample units are relatively difficult (expensive, destructive, or time-consuming) to quantify but one can use prior information to rank sample units in a small set, without their actual quantifications. The prior information can be taken as judgement ranking mechanism and be result of a visual comparison, an expert opinion or a concomitant variable. To obtain a ranked set sample, one first determines the parameters \( k \) and \( m \), referred as the set size and cycle size, respectively. One then draws \( k \) simple random samples (sets) of size \( k \) from the population of interest. In the first set, the unit that judged to be smallest is maintained and measured \((X_{[1]}(1))\). In the second set, the unit that judged to be second smallest among \( k \) units is maintained and measured \((X_{[2]}(1))\). This process continues until the unit that judged to be largest is maintained and measured \((X_{[k]}(1))\). These measured values \( X_{[r]}(1), \ r = 1, \ldots, k \), construct the first cycle of the sampling. For increasing the size of sampling, this process is replicated \( m \) times (cycles) to reach a ranked set sample of size \( n = m \times k \), represented by \( \{X_{[r]}(j), \ r = 1, \ldots, k, \ j = 1, \ldots, m\} \). We call a ranking procedure consistent if

\[
F(t) = \frac{1}{k} \sum_{r=1}^{k} F_{[r]}(t),
\]

where \( F_{[r]} \) is the CDF of a sample unit in a set of size \( k \) which judged to have rank \( r \). Presnell and Bohn (1999) showed that the consistent assumption in ranking process holds under some mild conditions. Note that if the ranking processes leads to no error, then the distribution of \( X_{[r]}(j) \) is the similar to the distribution of the \( r \)th order statistic of a sample of size \( k \) (for \( j = 1, \ldots, m \)) and the equality (3) holds owing to binomial expansion.

Although, at the beginning, McIntyre (1952) considers the problem of mean estimation in agricultural context, through the years, many standard statistical topics have been discussed in RSS and substantial works have been done in different fields. For example, the problem of estimation of the population mean is considered by Takahasi and Wakimoto (1968). Stokes and Sager (1988), and Huang (1997) discussed the CDF estimation in RSS. The problem of the variance and the proportion estimation in RSS are addressed by Stokes (1980), MacEachern (2002) and Chen et al. (2007), Zamanzade and Mahdizadeh (2017), respectively. Mahdizadeh and Zamanzade (2018a) and Mahdizadeh and Zamanzade (2018b) discussed reliability estimation in RSS and Zamanzade et al. (2019) considered estimation of the mean residual life-
time. Al-Omari and Haq (2011), Haq and Al-Omari (2014), Haq et al. (2013, 2014) studied constructing statistical control charts using RSS. Samawi et al. (2017), and Samawi et al. (2018) applied RSS on logistic regression analysis. Chen et al. (2019), Qian et al. (2021), and He et al. (2020, 2021) studied parametric estimation in RSS, and Wang et al. (2016, 2017) described how RSS can be used in multi-stage sampling designs. We refer the interested reader to recent survey paper of Wolfe (2012).

To justify the use of RSS for lifetime data, note that in the biomedical sciences, the outcome could be time to death, recurrence or progression of a certain type of a disease. In a situation where patients can be recruited easily, in a set of small size, one could rank them according to the severity of the disease at baseline. For instance, suppose that a medical researcher is interested in estimating the survival function of the patients who suffer a certain type of cancer. Note that although exact measurement of the survival time of people who suffer from the cancer is a time-consuming job, the medical researcher can simply rank them in a set of small size based on his/her personal judgment about their health status. This sampling design can be also applied in clinical trials, as a more efficient way of sampling participants from the population. Apart from time and cost considerations, the use of RSS in clinical trials also has an advantage from ethical point of view as fewer patients would be exposed to harmful therapies. See Zamanzade et al. (2019) for more information on this topic.

In Sect. 2, we propose an RSS-based estimator for MPL function and investigate some of its asymptotic properties. In Sect. 3, we compare the estimator in RSS with its competitor in SRS. The comparison results show the preference of the introduced methodology. In Sect. 4, a potential application of the proposed method in practice is illustrated using a medical example. Section 5 is devoted to some concluding remarks.

### 2 The proposed estimator

Let $X_1, \ldots, X_n$ be a simple random sample of size $n$ with CDF function $F$. The empirical estimator of $K(t)$ based on a simple random sample of size $n$ is given by

$$K_{SRS}(t) = \frac{\sum_{r=1}^{n}(t - X_r) \mathbb{I}(X_r \leq t)}{\sum_{r=1}^{n} \mathbb{I}(X_r \leq t)}, \quad (4)$$

where $\mathbb{I}(.)$ is the usual indicator function.

The asymptotic behaviour of $K_{SRS}(t)$ is studied by Parvardeh (2015). He proved that as $n$ tends to infinity,

$$\left\{ \sqrt{n} (K_{SRS}(t) - K(t)) , \quad t \geq 0 \right\} \rightarrow Z \text{ in } D([\tau, T]) \text{ for every } 0 < \tau < T < \infty,$$

where $D([\tau, T])$ is the usual $D$ space on $[\tau, T]$ with the Skorokhod topology (Billingsley 1999) and $Z = \{Z(t), t \geq 0\}$ is a mean zero Gaussian process with variance function $\sigma^2_{SRS}(t) = \sigma^2(t)/F(t)$ where

$$\sigma^2(t) = \mathbb{V}(t - X|X < t) = \frac{2\int_0^t (t - x) F(x)dx}{F(t)} - K^2(t).$$
Let \( \{X_{[r]j}; r = 1, \ldots, k; j = 1, \ldots, m\} \) be a ranked set sample of size \( n = mk \) from the population of interest. To estimate the parameter \( K(t) \), one can replace the CDF \( F \) in Eq. (2) with its empirical counterpart, \( \text{FRSS}(t) = \frac{1}{mk} \sum_{r=1}^{k} \sum_{j=1}^{m} \mathbb{I}(X_{[r]j} \leq t) \), in RSS. This leads to the following estimator

\[
K_{\text{RSS}}(t) = W \sum_{r=1}^{k} U_r, \tag{5}
\]

in which

\[
W = \frac{\mathbb{I}(V_1 + \cdots + V_k > 0)}{V_1 + \cdots + V_k}, \quad V_r = \sum_{j=1}^{m} \mathbb{I}(X_{[r]j} \leq t), \quad \text{and} \quad U_r = \sum_{j=1}^{m} (t - X_{[r]j}) \mathbb{I}(X_{[r]j} \leq t) \text{ for } r = 1, \ldots, k.
\]

The entire expression in equation (5) is taken to be 0 if \( \mathbb{I}(V_1 + \cdots + V_k > 0) = 0 \), i.e., if \( V_1 + \cdots + V_k = 0 \). Note that conditional on \( V = (V_1, \ldots, V_k) \), the sums \( U_r, r = 1, \ldots, k \), are independent random variables. Also, for a fixed value of \( r \), \( U_r \) depends on \( V \) only through \( V_r \).

Thus, it follows that

\[
\mathbb{E}[K_{\text{RSS}}(t)|V] = W \sum_{r=1}^{k} \mathbb{E}[U_r|V_r],
\]

\[
\mathbb{V}[K_{\text{RSS}}(t)|V] = W^2 \sum_{r=1}^{k} \mathbb{V}[U_r|V_r].
\]

Let \( K_{[r]}(t) \triangleq \mathbb{E}(t - X_{[r]}|X_{[r]} \leq t) \) and \( \sigma^2_{[r]}(t) \) denotes its variance. One can write

\[
\mathbb{E}[K_{\text{RSS}}(t)|V] = W \sum_{r=1}^{k} V_r K_{[r]}(t),
\]

\[
\mathbb{V}[K_{\text{RSS}}(t)|V] = W^2 \sum_{r=1}^{k} V_r \sigma^2_{[r]}(t).
\]

Since \( V_1, \ldots, V_k \) are independent random variables with \( V_r \sim \text{Bin}(m, F_{[r]}(t)) \), the mean and the variance of \( K_{\text{RSS}}(t) \) can be obtained as

\[
\mathbb{E}[K_{\text{RSS}}(t)] = \mathbb{E}[\mathbb{E}[K_{\text{RSS}}(t)|V]] = \sum_{v_1 + \cdots + v_k > 0} \frac{p(v)}{v_1 + \cdots + v_k} \sum_{r=1}^{k} v_r K_{[r]}(t), \tag{6}
\]

where

\[
p(v) = \mathbb{P}(V = v) = \prod_{r=1}^{k} \binom{m}{v_r} F_{[r]}(t)^{v_r} [1 - F_{[r]}(t)]^{m - v_r},
\]

and

\[
\mathbb{V}[K_{\text{RSS}}(t)] = \mathbb{E}(\mathbb{V}[K_{\text{RSS}}(t)|V]) + \mathbb{V}(\mathbb{E}[K_{\text{RSS}}(t)|V])
\]
A ranked-based estimator of the mean past lifetime with an application

\[ = \sum_{v_1 + \cdots + v_k > 0} \frac{p(v)}{(v_1 + \cdots + v_k)^2} \sum_{r=1}^{k} v_r \sigma_{[r]}^2(t) \]

\[ + \sum_{v_1 + \cdots + v_k > 0} \frac{p(v)}{(v_1 + \cdots + v_k)^2} \left\{ \sum_{r=1}^{k} v_r K_{[r]}(t) \right\}^2 \]

\[ - \left\{ \sum_{v_1 + \cdots + v_k > 0} \frac{p(v)}{(v_1 + \cdots + v_k)^2} \sum_{r=1}^{k} v_r K_{[r]}(t) \right\}^2. \]

The first theorem in this section establishes the consistency of \( K_{RSS}(t) \).

**Theorem 1** Let \( \{X_{[r]}; r = 1, \ldots, k; j = 1, \ldots, m\} \) be a ranked set sample of size \( n = mk \). Under consistent ranking assumption, if the set size \( k \) is fixed and the number of cycles (\( m \)) goes to infinity, then

\[ \sup_{\tau < t < T} |K_{RSS}(t) - K(t)| \longrightarrow 0, \quad \text{for every } 0 < \tau < T < \infty. \]

**Proof** We have

\[ K_{RSS} - K(t) = \frac{1}{k} \sum_{r=1}^{k} \hat{K}_{[r]}(t) \hat{F}_{[r]}(t) - \frac{1}{k} \sum_{r=1}^{k} K_{[r]}(t) F_{[r]}(t) \]

\[ = \frac{1}{k} \sum_{r=1}^{k} \frac{\hat{F}_{[r]}(t)}{F_{RSS}(t)} \left( \hat{K}_{[r]}(t) - K_{[r]}(t) \right) \]

\[ + \frac{1}{k} \sum_{r=1}^{k} \frac{\hat{F}_{[r]}(t)}{F_{RSS}(t)} K_{[r]}(t) - \frac{1}{k} \sum_{r=1}^{k} \frac{F_{[r]}(t)}{F_{RSS}(t)} K_{[r]}(t) \]

\[ + \frac{1}{k} \sum_{r=1}^{k} \frac{F_{[r]}(t)}{F_{RSS}(t)} K_{[r]}(t) - \frac{1}{k} \sum_{r=1}^{k} \frac{F_{[r]}(t)}{F(t)} K_{[r]}(t) \]

\[ = \frac{1}{k} \sum_{r=1}^{k} \frac{\hat{F}_{[r]}(t)}{F_{RSS}(t)} \left( \hat{K}_{[r]}(t) - K_{[r]}(t) \right) \]

\[ + \frac{1}{kF_{RSS}(t)} \sum_{r=1}^{k} K_{[r]}(t) \left( \hat{F}_{[r]}(t) - F_{[r]}(t) \right) \]

\[ + \left( \frac{1}{F_{RSS}(t)} - \frac{1}{F(t)} \right) \frac{1}{k} \sum_{r=1}^{k} K_{[r]}(t) F_{[r]}(t). \]
where $\hat{K}_{[r]}(t)$ and $\hat{F}_{[r]}(t)$ are empirical estimates of $K_{[r]}(t) = \mathbb{E}(t - X_{[r]}|X_{[r]} \leq t)$ and $F_{[r]}(t)$, respectively. Thus, we get

$$|K_{RSS}(t) - K(t)| \leq \max_{1 \leq r \leq k} \sup_{\tau \leq t \leq T} \left| \hat{K}_{[r]}(t) - K_{[r]}(t) \right|$$

$$+ \max_{1 \leq r \leq k} \sup_{\tau \leq t} \left| \hat{F}_{[r]}(t) - F_{[r]}(t) \right| \frac{1}{k F_{RSS}(t)} \sum_{r=1}^{k} K_{[r]}(t)$$

$$+ \sup_{\tau \leq t} \left| \frac{1}{F_{RSS}(t)} - \frac{1}{F(t)} \right| \frac{1}{k} \sum_{r=1}^{k} K_{[r]}(t) F_{[r]}(t).$$

(7)

Since $\frac{1}{k F_{RSS}(t)} \sum_{r=1}^{k} K_{[r]}(t)$ and $\frac{1}{k} \sum_{r=1}^{k} K_{[r]}(t) F_{[r]}(t)$ are both bounded on a finite interval, it follows from Parvardeh (2015) that the first and last term of summation (7) tends to zero as $m$ goes to infinity. Moreover, one can conclude from Glivenko-Cantelli theorem that the second term of summation (7) tends to zero as $m \to \infty$. Hence

$$\lim_{m \to \infty} \sup_{\tau \leq t \leq T} |K_{RSS}(t) - K(t)| = 0,$$

which completes the proof.

Now, to obtain asymptotic distribution of $K_{RSS}(t)$, we should first note that

$$\{ \sqrt{m} (F_{RSS}(t) - F(t)), \ t \geq 0 \} \to \frac{1}{k} \sum_{r=1}^{k} \mathbb{B}_{[r]} \text{ in } D([\tau, T])$$

for every $0 < \tau < T < \infty$, where $\mathbb{B}_{[r]}$ is a Gaussian process with mean zero and covariance function (Shorack and Wellner 2009)

$$\mathbb{E}(\mathbb{B}_{[r]}(t_1) \mathbb{B}_{[r]}(t_2)) = F_{[r]}(t_1 \wedge t_2) - F_{[r]}(t_1) F_{[r]}(t_2), \ t_1, t_2 \geq 0.$$

The following theorem provides the asymptotic distribution of $K_{RSS}$.

**Theorem 2** Let $\{X_{[r]}; r = 1, \ldots, k; j = 1, \ldots, m\}$ be a ranked set sample of size $n = mk$. Under consistent ranking process assumption, if the set size $k$ is fixed and the number of cycles $(m)$ goes to infinity, then

$$\{ \sqrt{n} (K_{RSS} - K(t)), \ t \geq 0 \} \to Z \text{ in } D([\tau, T])$$

for every $0 < \tau < T < \infty$, where $Z = \{Z(t), t \geq 0\}$ a mean zero Gaussian process of the form

$$Z_{RSS}(t) = \frac{1}{\sqrt{k F(t)}} \left[ \sum_{r=1}^{k} \int_{0}^{t} \mathbb{B}_{[r]}(x)dx - K(t) \mathbb{B}_{[r]}(t) \right] \ t \geq 0,$$
with variance function

\[ \forall(Z^{RSS}(t)) = \frac{1}{k F^2(t)} \sum_{r=1}^{k} \left( \sigma_r^2(t) F_r(t) + F_r(t) (1 - F_r(t)) \left( K_r(t) - K(t) \right) \right)^2, \quad t \geq 0. \]

**Proof** Note that

\[ K_{RSS}(t) - K(t) = \frac{1}{F_{RSS}(t)} \int_0^t \frac{1}{k} \sum_{r=1}^{k} \hat{F}_r(x) dx - \frac{1}{F(t)} \int_0^t \frac{1}{k} \sum_{r=1}^{k} F_r(x) dx \]

\[ = \frac{1}{k} \frac{1}{F_{RSS}(t)} \sum_{r=1}^{k} \int_0^t \left( \hat{F}_r(x) - F_r(x) \right) dx \]

\[ + \frac{1}{k} \frac{1}{F_{RSS}(t)} \sum_{r=1}^{k} \int_0^t F_r(x) dx - \frac{1}{k} \frac{1}{F(t)} \sum_{r=1}^{k} \int_0^t F_r(x) dx \]

\[ = \frac{1}{k} \frac{1}{F_{RSS}(t)} \sum_{r=1}^{k} \int_0^t \left( \hat{F}_r(x) - F_r(x) \right) dx \]

\[ + \frac{1}{F_{RSS}(t)} \sum_{r=1}^{k} \int_0^t F_r(x) dx \]

\[ = \frac{1}{k} \frac{1}{F_{RSS}(t)} \sum_{r=1}^{k} \int_0^t \left( \hat{F}_r(x) - F_r(x) \right) dx \]

\[ + K(t) F(t) \left( \frac{1}{F_{RSS}(t)} - \frac{1}{F(t)} \right). \]

Therefore, \( \{Z_{n}^{RSS}(t) = \sqrt{n} \left( K_{RSS}(t) - K(t) \right), \quad t \geq 0 \} \) can be written as

\[ Z_{n}^{RSS}(t) = \frac{1}{k} \frac{1}{F_{RSS}(t)} \sum_{r=1}^{k} \int_0^t \sqrt{n} \left( \hat{F}_r(x) - F_r(x) \right) dx \]

\[ + \frac{K(t)}{F_{RSS}(t)} \sqrt{n} (F(t) - F_{RSS}(t)). \]

Hence, the corresponding limiting process \( Z^{RSS} \) is obtained as

\[ Z^{RSS}(t) = \frac{1}{\sqrt{k}} \frac{1}{F(t)} \sum_{r=1}^{k} \int_0^t B_r(x) dx - \frac{K(t)}{\sqrt{k} F(t)} \sum_{r=1}^{k} B_r(t). \]
\[
= \frac{1}{\sqrt{k}} \frac{1}{F(t)} \left[ \sum_{r=1}^{k} \int_{0}^{t} B_{[r]}(x)dx - K(t)B_{[r]}(t) \right], \quad t \geq 0.
\]

The variance function of \( Z^{RSS} \) is given by

\[
\mathbb{V}(Z^{RSS}(t)) = \frac{1}{k} \frac{1}{F^2(t)} \text{Cov} \left( \sum_{r=1}^{k} \int_{0}^{t} B_{[r]}(x)dx - K(t)B_{[r]}(t), \sum_{r=1}^{k} \int_{0}^{t} B_{[r]}(y)dy - K(t)B_{[r]}(t) \right)
\]

\[
= \sum_{r=1}^{k} \int_{0}^{t} \int_{0}^{t} F_{[r]}(x \land y) - F_{[r]}(x)F_{[r]}(y)dx dy - 2K(t)
\]

\[
\times \sum_{r=1}^{k} \int_{0}^{t} F_{[r]}(x \land t) - F_{[r]}(x)F_{[r]}(t)dx
\]

\[
+ K^2(t) \sum_{r=1}^{k} F_{[r]}(t)(1 - F_{[r]}(t))
\]

\[
= \sum_{r=1}^{k} \int_{0}^{t} \int_{0}^{t} F_{[r]}(x \land y) - F_{[r]}(x)F_{[r]}(y)dx dy - 2K(t)
\]

\[
\times \sum_{r=1}^{k} (1 - F_{[r]}(t))K_{[r]}(t)F_{[r]}(t)
\]

\[
+ K^2(t) \sum_{r=1}^{k} F_{[r]}(t)(1 - F_{[r]}(t))
\]

\[
= \frac{1}{kF^2(t)} \left( \sum_{r=1}^{k} (\sigma^2_{[r]}(t) + K^2_0(t)) F_{[r]}(t) - K^2_0(t)F^2_{[r]}(t) \right)
\]

\[
- 2K(t) \sum_{r=1}^{k} (1 - F_{[r]}(t))K_{[r]}(t)F_{[r]}(t) + K^2(t) \sum_{r=1}^{k} F_{[r]}(t)(1 - F_{[r]}(t))
\]

\[
= \frac{1}{kF^2(t)} \sum_{r=1}^{k} (\sigma^2_{[r]}(t)F_{[r]}(t) + F_{[r]}(t)(1 - F_{[r]}(t))[K_{[r]}(t) - K(t)]^2).
\]

\[\square\]

3 Comparisons

We now compare the performance of the introduced estimator with its competitor in SRS. The comparison is done based on the following imperfect ranking models:

- **Fraction of random ranking model** (Frey et al. 2007): In this model, we assume that with probability \( p \) the rank of sample unit, with judgement rank \( r \), is identified correctly and with probability of \( (1 - p) \) is identified randomly. Hence, the CDF

\[\]
of \( X_{[r]j} \) in this model is a mixture given by \( F_{[r]} = pF_{(r)} + (1 - p)F \), where \( p \in [0, 1] \).

- **Fraction of neighbour ranking model** (Vock and Balakrishnan 2011): Here we assume that with probability \( p \) the rank of sample unit, with judgement rank \( r \), is identified correctly and with probability of \( \frac{1-p}{2} \) is confused with one of its adjacents. Thus, the CDF of \( X_{[r]j} \) in this model is a mixture given by \( F_{[r]} = \frac{1-p}{2}F_{(r-1)} + pF_{(r)} + \frac{1-p}{2}F_{(r+1)} \), where \( p \in [0, 1] \), \( F_{(0)} = F_{(1)} \) and \( F_{(k+1)} = F_{(k)} \).

Note that the fraction of neighbour ranking model often leads to less severe ranking error than the random ranking model, but it is more conceivable to happen in practice if the ranking is done using personal judgement of an expert. The comparison is done for three distributions with different MPL curve shapes: Standard exponential distribution (Exp(1)), Weibull distribution with shape parameter 4 and scale parameter 3 (Weibull(4,3)) and Rescaled beta distribution with probability density function \( \frac{3}{2}(1-x)^2I_{(0,2)}(x) \) (Rbeta(1,3)). The MPL curves of these three distributions are depicted in Fig. 1.

### 3.1 Finite sample size comparisons

In this subsection, the performance of MPL estimators in RSS and SRS are compared for finite sample sizes using Monte Carlo simulation. To this end, we first set \( n \in \{15, 30, 90\} \), \( k \in \{3, 5\} \), and for each combination of \((n, k)\), we have generated 1,000,000 random samples from both RSS and SRS designs. In order to control ranking quality, the values of \( p \) in the imperfect ranking models are selected from the set \( p \in \{0.2, 0.5, 0.8, 1\} \). Since MPL estimators in SRS and RSS are not unbiased, the relative efficiency (RE) of \( K_{RSS}(t) \) to \( K_{SRS}(t) \) is defined as the ratio of their mean square errors, i.e., \( RE(t) = \frac{MSE(K_{SRS}(t))}{MSE(K_{RSS}(t))} \) (see Wackerly et al. 2008, p. 445), and is estimated based on 1,000,000 repetitions for \( t \in \{Q_{0.05}, \ldots, Q_{0.95}\} \), where \( Q_q \) is the
Fig. 2 Simulated $RE(t)$ for $k \in \{3, 5\}$, $n = 15$ and $p \in \{1, 0.8, 0.5, 0.2\}$ under fraction of random ranking model for Weibull(4,3), Rbeta(1,3) and Exp(1) distributions.

$q$th quantile of the parent distribution. Note that $RE(t) > 1$ indicates the preference of $K_{RSS}(t)$ over $K_{SRS}(t)$.

Here, we only report the simulation results for settings with $n = 15$ and $k \in \{3, 5\}$ in Figs. 2 and 3 for fraction of random and neighbor ranking models, respectively. This is so because our simulation results show that the $RE$ values are not much affected by sample size $n$ when the other parameters are kept fixed (see the simulation results for $n \in \{15, 30, 90\}$ and $k \in \{3, 5\}$ in Figs. S1–S3 for fraction of random ranking model and in Figs. S4–S6 for fraction of neighbor ranking model in the Supplementary Materials). This observation is consistent with some results in RSS literature for the empirical estimation of $E(h(X))$, where $h(.)$, as a function of $X$, has finite second moment (see Chen et al. 2004, p. 16).

Figure 2 presents the simulation results under fraction of random ranking model. We observe from this figure that although the $RE$ has different patterns for different...
A ranked-based estimator of the mean past lifetime with an application

1.5 2.0 2.5 3.0 3.5 4.0
1.0 1.2 1.4 1.6 1.8 2.0 2.2

Weibull(4,3) Weibull(4,3) Weibull(4,3)
k=5 k=5 k=5
p=1 p=0.8 p=0.5

Weibull(4,3) Weibull(4,3) Weibull(4,3)
k=3 k=3 k=3
p=0.2

Weibull(4,3) Weibull(4,3) Weibull(4,3)
k=5 k=5 k=5
p=1 p=0.8 p=0.5

Rbeta(1,3) Rbeta(1,3) Rbeta(1,3)
k=5 k=5 k=5
p=0.2

Rbeta(1,3) Rbeta(1,3) Rbeta(1,3)
k=3 k=3 k=3
p=1 p=0.8 p=0.5

Exp(1) Exp(1) Exp(1)
k=5 k=5 k=5
p=0.2

Exp(1) Exp(1) Exp(1)
k=3 k=3 k=3
p=1 p=0.8 p=0.5

Simulated $RE(t)$ for $k \in \{3, 5\}$, $n = 15$ and $p \in \{1, 0.8, 0.5, 0.2\}$ under fraction of neighbor ranking model for Weibull(4,3), Rbeta(1,3) and Exp(1) distributions

distributions, it never falls below one which indicates superiority of $K_{RSS}(t)$ over $K_{SRS}(t)$. Furthermore, the relative efficiency increases as the value of set size ($k$) increases while the other parameters are kept fixed, and the efficiency gain is considerable if the quality of ranking is fairly good ($p \geq 0.8$). As one expects, the relative efficiency decreases as the value of $p$ decreases and $RE$ reaches to almost one when $p = 0.2$. This can be justified by the fact that statistical properties of $K_{RSS}(t)$ under fraction of random ranking model with $p = 0$ coincides to $K_{SRS}(t)$.

Simulation results under fraction of neighbor ranking model are given in Fig. 3. The pattern of the performance of the estimators in Fig. 3 is similar to that of Fig. 2, with an obvious difference that the $RE$ values in Fig. 3 are higher than what we observed in Fig. 2, specially for small values of $p$. This can be justified by the fact that fraction of neighbor ranking model leads to less serve ranking error than the random ranking model.
3.2 Asymptotic comparisons

In the SRS design, Parvardeh (2015) showed that the \( \sqrt{n}(K_{SRS}(t) - K(t)) \) goes to a Gaussian process with mean zero and variance function \( \sigma^2_{SRS}(t) = \sigma^2(t)/F(t) \) as \( n \) goes to infinity where

\[
\sigma^2(t) = \mathbb{V}(t - X \mid X < t) = \frac{2 \int_0^t (t - x) F(x)dx}{F(t)} - K^2(t).
\]

We define the asymptotic relative efficiency (ARE) as the ratio of asymptotic variance of \( K_{SRS}(t) \) to asymptotic variance of \( K_{RSS}(t) \), i.e.,

\[
ARE(t) = \frac{\sigma^2_{SRS}(t)}{\sigma^2_{RSS}(t)}.
\]

In the next theorem, we show that \( ARE(t) \) never falls below one.

**Theorem 3** Let \( \{X_{[r]}; r = 1, \ldots, k; j = 1, \ldots, m\} \) be ranked set sample of size \( n = mk \). If the ranking process is consistent, then \( ARE(t) \geq 1 \).

**Proof** Using the fundamental identity \( F(t) = \sum_{r=1}^{k} F_{[r]}(t)/k \), we have

\[
k\sigma^2(t) = k \left( \frac{\int_0^t 2(t-x) F(x)dx}{F(t)} - K^2(t) \right)
\]

\[
= k \left( \frac{1}{k} \sum_{r=1}^{k} \frac{F_{[r]}(t)}{F(t)} \int_0^t 2(t-x) \frac{F_{[r]}(x)}{F_{[r]}(t)} dx - K^2(t) \right)
\]

\[
= \sum_{r=1}^{k} \left( \frac{F_{[r]}(t)}{F(t)} \left( \int_0^t 2(t-x) \frac{F_{[r]}(x)}{F_{[r]}(t)} dx - K^2_{[r]}(t) \right) + \frac{F_{[r]}(t)}{F(t)} K^2_{[r]}(t) \right) - kK^2(t)
\]

\[
= \sum_{r=1}^{k} \left( \frac{K_{[r]}(t)}{F(t)} \sigma^2_{[r]}(t) + \frac{K_{[r]}(t)}{K(t)} K^2_{[r]}(t) \right) - kK^2(t)
\]

\[
= \sum_{r=1}^{k} \left( \frac{F_{[r]}(t)}{F(t)} \sigma^2_{[r]}(t) + \sum_{r=1}^{k} \frac{K_{[r]}(t)}{K(t)} K^2_{[r]}(t) - K(t) \sum_{r=1}^{k} \frac{F_{[r]}(t)}{F(t)} M_{[r]}(t) \right)
\]

\[
= \sum_{r=1}^{k} \left( \frac{F_{[r]}(t)}{F(t)} \sigma^2_{[r]}(t) + \sum_{r=1}^{k} \frac{F_{[r]}(t)}{F(t)} K_{[r]}(t) \left[ K_{[r]}(t) - K(t) \right] \right)
\]

\[
= \sum_{r=1}^{k} \frac{F_{[r]}(t)}{F(t)} \sigma^2_{[r]}(t) + \sum_{r=1}^{k} \frac{F_{[r]}(t)}{F(t)} \left[ K_{[r]}(t) - K(t) \right]^2.
\]
A ranked-based estimator of the mean past lifetime with an application

Therefore, it follows that

\[ k\sigma_{SRS}^2(t)F^2(t) = k\sigma^2(t)F(t) \]
\[ = \sum_{r=1}^{k} F_{[r]}(t)\sigma_{[r]}^2(t) + \sum_{r=1}^{k} F_{[r]}(t)[K_{[r]}(t) - K(t)]^2 \]
\[ \geq \sum_{r=1}^{k} F_{[r]}(t)\sigma_{[r]}^2(t) + \sum_{r=1}^{k} F_{[r]}(t)(1 - F_{[r]}(t))[K_{[r]}(t) - K(t)]^2 \]
\[ = k\sqrt{V(F(t)Z_{RSS}(t))} \]
\[ = k\sigma_{RSS}^2(t)F^2(t), \]

and this completes the proof. \(\square\)

To see the amount of asymptotic efficiency gain obtained using \(K_{RSS}\) instead of \(K_{SRS}\), we have computed the \(ARE(t)\) for \(k \in \{3, 5\}\) and three different distributions, under two above imperfect ranking models with \(p \in \{0.2, 0.5, 0.8, 1\}\). Here, for brevity, we only report results for the perfect ranking case \((p = 1)\) in Fig. 4 and refer the interested reader to see complete results in Figs. S7–S8 in the Supplementary Material for fraction of random ranking and neighbour ranking models, respectively.

Figure 4 presents the \(ARE(t)\) for three different distributions under assumption of perfect ranking. We observe from this figure that the pattern of the performance of the

\[ \text{Fig. 4 Exact values of ARE(t) for } k \in \{3, 5\} \text{ under perfect ranking case for Weibull(4,3), Exp(1) and Rbeta(1,3) distributions} \]
estimator is very close to what we observed for finite sample sizes when the ranking is perfect ($p = 1$). This is also true for imperfect ranking case (see Figs. S7–S8 in the Supplementary Material). Thus, the $RE(t)$ values can be well approximated by $ARE(t)$ ones.

4 Estimation of the gap between HIV transmission and diagnosis

The human immunodeficiency virus (HIV) is a dangerous virus that weakens the body natural defence system against illness (immune system) by destroying a type of white blood cells in the body called CD4 cells. As more CD4 cells are destroyed by the HIV viruses, the patient’s immune system becomes weaker so he/she finds it harder to fight off infections. The late-stage of HIV is called acquired immune deficiency syndrome (AIDS) in which the body’s immune system goes to severely weaker condition such that it cannot defend itself at all. Usually HIV leads to AIDS in 15 years if it does not treated appropriately. According to the World Health Organization (WHO), more than 95% of HIV infections happen in developing countries. This disease implies a catastrophe not only for the individuals and households affected, but also for the entire nation, as it is likely to lead to an intensification of poverty, push some non-poor into poverty and some of the very poor into destitution.

Many people in developing countries who suffer from HIV are not aware of their infection. This is so because valid tests for HIV detection may not be available for public in these countries, or they could be very expensive. On the other hand, the unaware people of their HIV contribute much larger in HIV transmission. For example, Hall et al. (2012) estimated that around half of HIV transmissions are due to people who are unaware of their infections. Therefore, it is crucial for both government and health organizations to use an efficient method for estimating the gap between HIV infection and diagnosis. Suppose that at time $t$, a person gets a medical test to check out about HIV and the test result is positive. Let the random variable $X$ be the infection time, then we know that $X \leq t$. Thus $K(t) = \mathbb{E}(t - X | X \leq t)$ is the parameter of interest.

Note that although obtaining time of HIV infection is difficult because most patients have an established infection of unknown duration at diagnosis, a medical expert can simply rank the sample units according to the infection time using his personal judgement, interviewing with test subjects or their health status. Thus, RSS scheme can be regarded as an alternative for SRS for estimating MPL of people living with HIV at the diagnosis time $t$. Assume that one is interested in estimating MPL function of people living with HIV at the diagnosis time $t$ based on an RSS sample. However, since an RSS sample of infection time of patients living with HIV is not available, we simulate an RSS sample with set size 5 and cycle size 6 from Gamma distribution with scale parameter 40 and shape parameter 8 under fraction of random ranking model with $p = 0.8$ and round it to its nearest integer to represent the infection time of HIV in weeks. The data set is presented in Table 1 and the estimated MPL function along with its 95% normal approximation (NA) confidence interval (CI) is shown in Fig. 5, where the variance of $K_{RSS}(t)$ is estimated by replacing each of its components with its empirical counterpart.
Table 1 A simulated RSS data set for the infection time of HIV in weeks

| Cycle | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|-------|--------|--------|--------|--------|--------|
| Cycle 1 | 192    | 277    | 293    | 329    | 376    |
| Cycle 2 | 158    | 256    | 236    | 411    | 508    |
| Cycle 3 | 137    | 280    | 451    | 2447   | 478    |
| Cycle 4 | 171    | 229    | 143    | 268    | 248    |
| Cycle 5 | 203    | 232    | 238    | 367    | 363    |
| Cycle 6 | 287    | 274    | 415    | 413    | 323    |

Fig. 5 Estimation of MPL function using data set in Table 1 (represented by solid line) along with its 95% normal approximation (NA) confidence interval (represented by dash lines)

5 Conclusion

The ranked set sampling (RSS) is a sampling plan which is designed to employ auxiliary ranking information for improving the estimation of the population parameters. The auxiliary ranking information can be obtained through eye inspection, subjective judgement, or a cheap concomitant variable. The ranking process is done before any actual measurements on the variable of interest, and leads to select more informative units to include in our sample for measurement. This approach often is used when it is easy and cheap to rank units in a set without measuring their accurate values.

In this paper, we studied an empirical estimate of mean past lifetime (MPL) based on RSS. We showed that the estimator is a strongly uniformly consistent estimator and we proved that it converges to a Gaussian process under some mild conditions. We then compared the introduced estimator with the empirical estimator in simple random sampling (SRS). To this end, we considered two imperfect ranking models, which lead
to severe and mild ranking errors, four different degrees for ranking quality which move from perfect ranking to almost random ranking, eight different combination of set size and sample size, and three distributions with different MPL shapes. According to our comparison results, RSS estimator of MPL is more efficient than the SRS one, and the efficiency gain of using RSS estimator can be sizeable in some certain circumstances. Finally, we illustrated a potential application of the developed procedure in estimating the gap between HIV infection to the diagnosis.

Supplementary Information  The online version contains supplementary material available at https://doi.org/10.1007/s00362-022-01314-y.

Acknowledgements  The authors are grateful to two anonymous referees for helpful comments that have resulted in an improved paper. Ehsan Zamanzade’s research was supported in part by Iran National Science Foundation (INSF).

References

Al-Omari AI, Haq A (2011) Improved quality control charts for monitoring the process mean, using double-ranked set sampling methods. J Appl Stat 39(4):745–763
Asadi M, Berred A (2012) Properties and estimation of the mean past lifetime. Statistics 46:405–417
Billingsley P (1999) Convergence of probability measures, 2nd edn. Wiley, New York
Chen Z, Bai ZD, Sinha BK (2004) Ranked set sampling: theory & applications. Springer, New York
Chen H, Stasny EA, Wolfe DA (2007) Improved procedures for estimation of disease prevalence using ranked set sampling. Biom J 49:530–538
Chen W, Yang R, Yao D, Long C (2019) Pareto parameters estimation using moving extremes ranked set sampling. Stat Papers 62:1195–1211
Frey J, Ozturk O, Deshpande JV (2007) Nonparametric tests for perfect judgment rankings. J Am Stat Assoc 102(478):708–717
Hall HI, Holtgrave DR, Maulsby C (2012) HIV transmission rates from persons living with HIV who are aware and unaware of their infection. AIDS 26(7):893–896
Haq A, Al-Omari AI (2014) A new Shewhart control chart for monitoring process mean based on partially ordered judgment subset sampling. Qual Quant 49(3):1185–1202
Haq A, Brown J, Moltchanova E, Al-Omari AI (2013) Partial ranked set sampling design. Environmetrics 24(3):201–207
Haq A, Brown J, Moltchanova E, Al-Omari A (2014) Effect of measurement error on exponentially weighted moving average control charts under ranked set sampling schemes. J Stat Comput Simul 85(6):1224–1246
He X, Chen W, Qian W (2020) Maximum likelihood estimators of the parameters of the log–logistic distribution. Stat Pap 61:1875–1892
He X, Chen W, Rui Y (2021) Modified best linear unbiased estimator of the shape parameter of log–logistic distribution. J Stat Comput Simul 91(2):383–395
Huang J (1997) Asymptotic properties of the npmle of a distribution function based on ranked set samples. Ann Stat 25:1036–1049
MacEachern SN, Ozturk O, Wolfe DA, Stark GV (2002) A new ranked set sample estimator of variance. J R Stat Soc Ser B 62:177–188
Mahdizadeh M, Zamanzade E (2018) A new reliability measure in ranked set sampling. Stat Pap 59(3):861–891
Mahdizadeh M, Zamanzade E (2018) Smooth estimation of a reliability function in ranked set sampling. Statistics 52(4):750–768
McIntyre GA (1952) A method for unbiased selective sampling using ranked set sampling. Aust J Agric Res 3:385–390
Parvardeh A (2015) A note on the asymptotic distribution of the estimation of the mean past lifetime. Stat Pap 56(1):205–215
Presnell B, Bohn LL (1999) U-statistics and imperfect ranking in ranked set sampling. J Nonparametr Stat 10:111–126
Qian W, Chen W, He X (2021) Parameter estimation for the Pareto distribution based on ranked set sampling. Stat Pap 62:395–417
Samawi HM, Rochani H, Linder D, Chatterjee A (2017) More efficient logistic analysis using moving extreme ranked set sampling. J Appl Stat 44(4):753–76
Samawi HM, Helu A, Rochani H, Yin J, Yu L, Vogel R (2018) Reducing sample size needed for accelerated failure time model using more efficient sampling methods. J Stat Theory Pract 12(3):530–541
Shorack GR, Wellner JA (2009) Empirical processes with applications to statistics, vol 59. SIAM, Philadelphia
Stokes SL (1980) Estimation of variance using judgement ordered ranked set samples. Biometrics 36:35–42
Stokes SL, Sager TW (1988) Characterization of a ranked-set sample with application to estimating distribution functions. J Am Stat Assoc 38:374–381
Takahasi K, Wakimoto K (1968) On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann Inst Stat Math 20(1):1–31
Vock M, Balakrishnan N (2011) A Jonckheere–Terpstra-type test for perfect ranking in balanced ranked set sampling. J Stat Plan Inference 141(2):624–630
Wackerly DD, Mendenhall W, Scheaffer RL (2008) Mathematical statistics with applications, 7th edn. Thomson Brooks/Cole, Belmont
Wang X, Lim J, Stokes SL (2016) Using ranked set sampling with cluster randomized designs for improved inference on treatment effects. J Am Stat Assoc 111(516):1576–1590
Wang X, Ahn S, Lim J (2017) Unbalanced ranked set sampling in cluster randomized studies. J Stat Plan Inference 187:1–16
Wolfe DA (2012) Ranked set sampling: its relevance and impact on statistical inference. ISRN Probab Stat 568385:1–32
Zamanzade E, Mahdizadeh M (2017) A more efficient proportion estimator in ranked set sampling. Stat Probab Lett 129:28–33
Zamanzade E, Parvardeh A, Asadi M (2019) Estimation of mean residual life based on ranked set sampling. Comput Stat Data Anal 135:35–55

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.