Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

# Probability-proportional-to-size ranked-set sampling from stratified populations


by Omer Ozturk

Release date: December 15, 2020

Statistics Canada Statistique Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

**Email at** STATCAN.infostats-infostats.STATCAN@canada.ca

**Telephone,** from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

**Depository Services Program**

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada



**An HTML version is also available.**

*Cette publication est aussi disponible en français*.

Survey Methodology, December 2020
Vol. 46, No. 2, pp. 243-264
Statistics Canada, Catalogue No. 12-001-X


# Probability-proportional-to-size ranked-set sampling from stratified populations

**Omer Ozturk**[1]

Abstract

This paper constructs a probability-proportional-to-size (PPS) ranked-set sample from a stratified population. A PPS-ranked-set sample partitions the units in a PPS sample into groups of similar observations. The construction of similar groups relies on relative positions (ranks) of units in small comparison sets. Hence, the ranks induce more structure (stratification) in the sample in addition to the data structure created by unequal selection probabilities in a PPS sample. This added data structure makes the PPS-ranked-set sample more informative then a PPS-sample. The stratified PPS-ranked-set sample is constructed by selecting a PPS-ranked-set sample from each stratum population. The paper constructs unbiased estimators for the population mean, total and their variances. The new sampling design is applied to apple production data to estimate the total apple production in Turkey.

**Key Words:** PPS sampling; Stratified PPS; Ranked-set sample; Sample allocation.

## 1 Introduction

In survey sampling studies, selection of a sampling design depends on the structure of the population. In this paper, we consider a population structure having two main features. It must contain a size variable $X$ and a variable of interest $Y$. The values of the size variable should be approximately proportional to the values of the $Y$-variable, and the values of the $X$-variable should be available for all population units prior to sampling. The second feature of the population structure is that a small percentage of population units should produce extreme values in both the $Y$- and $X$-variables with different proportionality constants. These units usually produce larger means and variances than the rest of the units in the population in both the $Y$- and $X$-variables. This population structure is very common in practice. In agricultural sampling, a farm population in a state or country may contain two variables, the crop production $Y$ and the farm size $X$ in acres. Farms can be divided into two groups, the farms that have small/normal sizes and the mega-farms that have extremely large values in the $X$- and $Y$-variables. The percentage of the mega-farms would be small, but they may have larger means and variances in the $Y$- and $X$-variables and the proportionality constant between the $Y$- and $X$-values may be larger.

The population structure in the Monthly Retail Trade Survey performed by the United States Census Bureau would be another example. In this case, the population is defined by the business establishments that have Employer Identification Numbers (EINs). The Census Bureau uses a very complex design in which the previous years' annual revenues are used to construct a size variable. The structure of the population fits the setting we consider in this paper. Revenues from the previous years would be approximately proportional to the current revenues. The revenues for most of the businesses would take typical values, while revenues of a certain percentage of businesses would be extremely large, producing a

1. Omer Ozturk, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH, 43210, U.S.A. E-mail: omer@stat.osu.edu.

larger mean and variance and a different proportionality constant. Different proportionality constant may happen because large businesses may be more/less productive than the rest of the businesses.

We provide a third example using apple production data in Turkey in 2002. The data set was collected by the Turkish Statistical Institute and reported in Kadilar and Cingi (2003) and Ozturk and Bayramoglu-Kavlak (2018). It contains two variables, apple production $(Y)$ (in 1,000kg) and the numbers of apple trees $(X)$ in townships (or localities). The sampling units are the 851 localities in the data set. The $X$-values in all townships are available in the sampling frame prior to sampling. Figure 1.1 provides the scatter plot of the $Y$- and $X$-values, where we see that the value of the $Y$-variable is an increasing function of the $X$-value. We also observe that red-colored points marked with "$X$" in the plot have large values for both $X$- and $Y$-variables and their proportionality constant is different from the other points. Hence, this population fits into our population structure.

The apple production data have additional structures. The entire population is stratified into seven different geographical regions: Marmara, Aegean, Mediterranean, Central Anatolia, Black Sea, Eastern Anatolia and Southeastern Anatolia. These regions have different climate patterns and apple production changes significantly from one region to another. The extreme observations marked with "$X$" in Figure 1.1 come from the Marmara, Aegean, Mediterranean and Central Anatolia regions. This is a natural setting to construct a PPS-ranked-set sample from each sub-population.

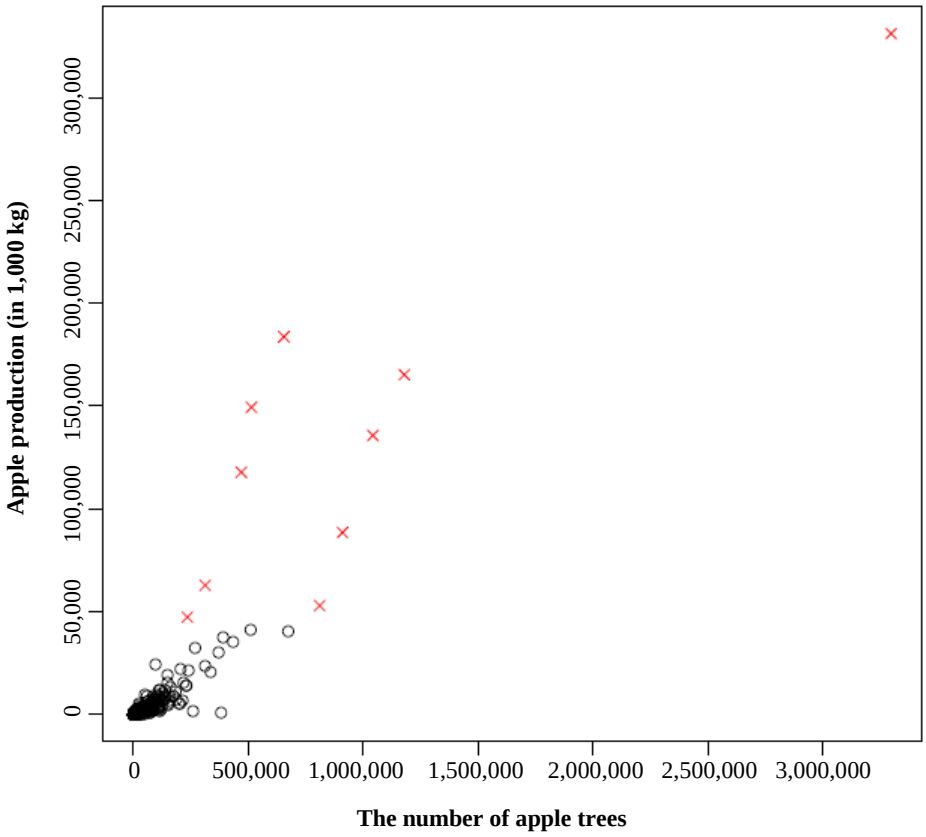


**Figure 1.1 Scatter plot of apple production.**

The variable $X$ in our population setting provides information about the relative size of the units in the entire population. Since the variable $Y$ is approximately proportional to the variable $X$, the size of the unit indicates the importance of its contribution to the variable of interest $Y$. Hence, important units should have higher probability of being included in the sample. Probability-proportional-to-size (PPS) sampling deliberately imposes higher selection probabilities for important units to produce unbiased and highly efficient estimators for the population mean and/or total. The main contribution of this paper is to introduce a new sampling design which combines the ranking information in a ranked-set sample (RSS) with the advantage of unequal selection probabilities as used in a PPS sample.

A typical PPS sample contains the triplets $(Y_i, \pi_i, u_i)$, $i = 1, \ldots, n$, where $Y_i$ and $\pi_i$ are the value of $Y$ and the selection probability of the unit $u_i$ for each draw under sampling with replacement selection. Though PPS sampling can be done without replacement, here all references to PPS sampling refers to sampling with replacement. Readers may refer to Thompson (2002, page 53) for the details of PPS sampling. In a PPS sample, the size variable is not necessarily used directly in the construction of the estimators. On the other hand, the values of the $X$-variable are available for all population units even for the units not included in the PPS sample. Hence the $X$-variable could help us to borrow additional information from a comparison set of $H$ unmeasured units.

For the construction of a typical data point $(Y_i, \pi_i, u_i)$ in a PPS-ranked-set sample, we select $H$ units from the population using PPS sampling with replacement to form a comparison set $\{u_1, \ldots, u_H\}$. We rank these units without measurement based on the $X$-variable with no additional cost. We then measure the value of the $Y$-variable for only one unit, the unit having rank $R_i$, obtaining $(Y_{[R_i]}, \pi_{[R_i]}, u_{[R_i]})$, where $Y_{[R_i]}$ and $\pi_{[R_i]}$ are the value of the $Y$-variable and the selection probability of the unit $u_{[R_i]}$ at each draw. A data point in the comparison set, $(Y_{[R_i]}, \pi_{[R_i]}, u_{[i]})$, provides more information than a data point in a PPS sample, $(Y_i, \pi_i, u_i)$, since the rank $R_i$ borrows additional information from the other $H-1$ unmeasured units in the comparison set. In this paper, we use this idea to construct a PPS-ranked-set sample that is more informative than a PPS sample. The details of this sampling procedure will be provided in Section 2.

The position information is used in a slightly different context in ranked-set sample (RSS) and judgment-post-stratified (JPS) sampling designs to borrow information from the unmeasured population units. Construction of a ranked-set sample of size $n$ requires one to determine two integers $d$ and $H$, $n = dH$, where $H$ and $d$ are the set and cycle sizes, respectively. The set size $H$ controls the amount of information that can be borrowed from the units in comparison sets. The cycle size $d$ is used to increase the total sample size in a RSS. Once $H$ and $d$ are chosen, one then selects $nH$ units from the population and partitions them into $n$ disjoint comparison sets, each having $H$ units. Units in each comparison set are ranked without measurement using the $X$-variable and the value of the $Y$-variable $(Y_{[h]j};\ j = 1, \ldots, d)$ associated with the $h^{\text{th}}$ ranked $X$ is measured in $d$ different comparison sets, $h = 1, \ldots, H$. The measured values $Y_{[h]j}$, $h = 1, \ldots, H$; $j = 1, \ldots, d$, are called a ranked-set sample.

The construction of a JPS sample of size $n$ starts with a simple random sample of size $n$ and measures all of them, $Y_i$, $i = 1, \ldots, n$. For each measured unit in this sample, one then selects an additional $H-1$

units from the population to form a comparison set of size $H$. The rank $R_i$ of the measured unit $Y_i$ in each of these comparison sets is determined. The pairs of $(Y_i, R_i)$, $i = 1, \ldots, n$, constitute a JPS sample.

The RSS and JPS samples create induced order statistics for the $Y$-variable through the ranks of the $X$-variable in comparison sets. Hence, the random variable $Y_{[h]i}$ ($Y_i$ given that $R_i = h$ for the JPS sample) is stochastically smaller than the random variable $Y_{[h']j}$ ($Y_i$ given that $R_j = h'$) for $h \leq h'$. This stochastic ordering property induces an implicit stratification among the measured sample units. Efficiency improvements of RSS and JPS samples over a simple random sample can be anticipated from the partition of the total variation into between- and within-strata variation. For further details on these sampling designs, readers may refer to the review paper in Wolfe (2012) and references therein. Both RSS and JPS samples use the position information of the units in the comparison sets, but they do not completely use the information provided by the selection probabilities in a PPS sample. All units in comparison sets for RSS and JPS are selected with equal probabilities. Hence, they may not be appropriate for the population structure that we consider in this paper.

MacEachern, Stasny and Wolfe (2004) introduced the JPS design in an infinite population setting. In a finite population setting, constructions of the JPS and RSS samples depend on whether the comparison sets are selected with or without replacement. Patil, Sinha and Taillie (1995) considered an RSS design in a finite population, where none of the units in a comparison set is returned to the population prior to selection of the next comparison set. Deshpande, Frey and Ozturk (2006) expanded the RSS sampling design with three different without replacement selection policies and constructed nonparametric confidence intervals for population quantiles.

Probability sampling has also generated extensive research interest in RSS and JPS sampling. Al-Saleh and Samawi (2007), Ozdemir and Gokpinar (2007 and 2008), Gokpinar and Ozdemir (2010), Ozturk and Jafari Jozani (2013), Frey (2011) and Ozturk (2014) computed inclusion probabilities and constructed Horwitz-Thompson type estimators for the population mean and total based on a ranked-set sample. These research papers show that an RSS design yields a substantial amount of improvement in efficiency over the usual simple random sampling design. Ozturk (2016) developed estimators for the population mean based on a JPS sample, where he showed that the estimator needs a finite population correction factor similar to the one used in a simple random sample.

A few researchers have applied the RSS methodology to existing survey sampling designs. Muttlak and McDonald (1992) incorporated the RSS sampling design with a line intersect method. Sroka (2008) used it in stratified sampling by constructing an RSS sample from each stratum. Wang, Lim and Stokes (2016) considered the RSS design in a cluster randomized design with a mixed effect model, where the cluster effect is treated as random. They showed that use of RSS at the cluster level has much bigger impact on efficiency than using the RSS at the within-cluster level. Nematollahi, Salehi and Aliakbari Saba (2008) used the RSS design in a finite population setting only in the second stage of a two-stage sampling with replacement selection scheme. Since they use the RSS design only in the second stage with replacement, the efficiency improvement of their estimator with respect to a two-stage SRS sample

estimator was minimal. Sud and Mishra (2006) also used a two-stage cluster sample with ranked set sampling design in a finite population setting under the assumption that the cluster population sizes are all equal. Ozturk (2019a) developed design based statistical inference for a two-stage clustered ranked-set sample in a finite population setting.

In this paper, we develop statistical inference for the PPS-ranked-set sampling design in a population setting where the values of the size variable are roughly proportional to the values of the variable of interest and a small percentage of population units produces large $X$- and $Y$-values with a different proportionality constant. We motivate the new sampling design using apple production data. Section 2 introduces the PPS-ranked-set sample in a finite population setting. It constructs unbiased estimators for the population mean, total and their variances. We show that the PPS-ranked-set sample estimator has smaller variance than a PPS sample estimator. Section 3 extends the PPS-ranked-set sample to a stratified population and constructs unbiased estimators for the population mean, total and their variances. Section 4 considers four different sample size allocation procedures to minimize the variance of the estimator under a cost model and different stratum population structures. Section 5 provides an efficiency comparison for the PPS-ranked-set sample estimator of the population mean with respect to other competing estimators. Section 6 illustrates the use of PPS-ranked-set sample data to estimate the apple production in Turkey. Section 7 provides some concluding remarks. All proofs are given in the Appendix.

# 2 Probability-proportional-to-size ranked-set sample

We consider a finite population of size $N$, $\mathcal{P}^N = \{u_1, \ldots, u_N\}$. Assume that each unit in the population possesses two characteristics $Y$ and $X$. The values of $X$ are available for all units and the values of $Y$ are roughly proportional to the values of $X$. A small percentage of the population units may produce extreme values for both $Y$- and $X$-variables with a different proportionality constant. Let $\pi_i$ be the probability that the unit $u_i$ is selected from $\mathcal{P}^N$, for each draw.

$\pi_i = P$ (unit $u_i$ is selected from $\mathcal{P}^N$ under sampling with replacement selection for each draw),

where $\pi_i$ is proportional to the size of $X$ for unit $u_i$. The values of the variable $Y$ in the population $\mathcal{P}^N$ are denoted by $y_1, \ldots, y_N$. The mean and variance of this population are defined as

$$\mu_N = \frac{1}{N}\sum_{k=1}^{N} y_k \text{ and } \sigma_N^2 = \frac{1}{N}\sum_{k=1}^{N}(y_k - \mu_N)^2.$$

We first briefly introduce the notation for a PPS sample. Let $n$ be the sample size. We consider a PPS sample, $(Y_i, \pi_i, u_i)$, $i = 1, \ldots, n$, constructed from population $\mathcal{P}^N$ under a sampling with replacement selection scheme with selection probability $\pi_i$. The probability mass function (PMF) and the cumulative distribution function (CDF) of the random variable $Y_i$ are given by

$$P(Y_i = y) = f(y) = \sum_{j=1}^{N} \pi_j I(y_j = y), \quad F(y) = \sum_{z=y_1}^{y}\sum_{j=1}^{N} \pi_j I(y_j = z).$$

We note that, since the sample units $u_i$, $i = 1, \ldots, n$, are selected with replacement, the $Y_i$, $i = 1, \ldots, n$, are independent and identically distributed. We now consider the order statistics $Y_{(h:n)}$ in a sample of size $n$. The CDF of the $h^{\text{th}}$ order statistics in a sample of size $n$ is given by

$$F_{(h:n)}(y) = \sum_{r=h}^{n} \binom{n}{r} F^r(y) \{1 - F(y)\}^{n-r}$$

$$f_{(h:n)}(y) = F_{(h:n)}(y) - F_{(h:n)}(y^-), \tag{2.1}$$

where $F_{(h:n)}(y^-)$ is the left-limit at $y$.

We now construct a PPS-ranked-set sample that combines the ranking information in comparison sets with the information provided by the selection probabilities in a PPS sample. Let $H$ be the set size. Using a PPS sampling design, we select $H$ units from the population with replacement to form a comparison set $S = \{u_1, \ldots, u_H\}$ with selection probabilities, $\{\pi_{u_1}, \ldots, \pi_{u_H}\}$. Units in this set are ranked from the smallest to the largest based on the values of the size variable $X$, yielding $S^* = \{(Y_{[1]}, \pi_{[1]}, u_{[1]}, X_{(1)}), \ldots, (Y_{[H]}, \pi_{[H]}, u_{[H]}, X_{(H)})\}$, where $Y_{[h]}$ is the value of the $Y$-variable and $\pi_{[h]}$ is the selection probability of the unit $u_{[h]}$ that correspond to the $h^{\text{th}}$ smallest $X$ $(X_{(h)})$ in the set. The smallest ranked unit in set $S^*$ is selected and measured for the variable $Y$, $Y_{[1]}$, and its selection probability, $\pi_{[1]}$, is recorded. The remaining $H - 1$ units are not measured. They are used only to obtain the rank of $u_1$ based on ranking of the $X$ - measurements. We construct another comparison set using a PPS sample and rank the units based on the $X$-variable. This time, we measure the $Y$-variable on the unit that corresponds to the second smallest $X$-value and record its selection probability, $(Y_{[2]}, \pi_{[2]})$. We continue constructing comparison sets and measuring the $Y$-variables until we have the measurement from the unit that corresponds to the largest $X$-value, $(Y_{[H]}, \pi_{[H]})$. The measured values $(Y_{[h]}, \pi_{[h]})$, $h = 1, \ldots, H$, are called a cycle. To increase the sample size to $n = Hd$, the entire process is repeated for $d$ cycles. The measured values $Y_{[h]j}, \pi_{[h]j}$, $h = 1, \ldots, H$, $j = 1, \ldots, d$ are called a PPS-ranked-set sample, where $Y_{[h]j}, \pi_{[h]j}$ are the $Y$ - measurement and selection probability of the unit $u_{[h]}$ that corresponds to the $h^{\text{th}}$ smallest value of the $X$ - variable in the set $h$ and cycle $j$. We refer to $Y_{[h]i}$ as the induced $h^{\text{th}}$ order statistic since its position is induced by the $X$-values in the set. We note that the induced order statistic $Y_{[h]i}$ and induced ordered unit $u_{[h]i}$ are defined in comparison sets with set size $H$. To simplify the notation we omit the set size $H$ and write $Y_{[h]i} = Y_{[h;H]i}$, $u_{[h]i} = u_{[h:H]i}$.

The PPS-ranked-set sample is illustrated in Table 2.1 for the set size $H = 3$ and the cycle size $d = 2$. For each cycle, the table contains three comparison sets (rows). Each set has three units. The units in each set are ranked based on the $X$-variable. The units on the diagonal (bold faced) are measured for the values of the $Y$-variables, and their selection probabilities are recorded. The last column contains the measured values of the units in the PPS-ranked-set sample. Each measured data point in Table 2.1 provides three pieces of information: (1) the value of $Y$, (2) the selection probability of the unit under sampling with replacement, and (3) the relative position of the unit in its comparison set. One can make an

intuitive comparison between the PPS-ranked-set sample and other sampling designs in the literature. For example, a simple random sample provides the information in item (1), a ranked-set sample provides the information in items (1) and (3), and a PPS sample provides the information in items (1) and (2). We anticipate (and show in Theorem 1) that the PPS-ranked-set sample is more informative than all three of these sampling designs, and therefore has smaller variance, since it provides the information in items (1), (2) and (3).

**Table 2.1**
**Illustration of the PPS-ranked-set sample for the set size $H = 3$ and cycle size $d = 2$**

| Cycle | Set | Ranked units in comparison sets | Measurements |
|---|---|---|---|
| 1 | 1 | $\{\mathbf{Y}_{[1]1}, \boldsymbol{\pi}_{[1]1}, \mathbf{X}_{(1)1}\}, \{Y_{[2]1}, \pi_{[2]1}, X_{(2)1}\}, \{Y_{[3]1}, \pi_{[3]1}, X_{(3)1}\}$ | $(Y_{[1]1}, \pi_{[1]1})$ |
| | 2 | $\{Y_{[1]1}, \pi_{[1]1}, X_{(1)1}\}, \{\mathbf{Y}_{[2]1}, \boldsymbol{\pi}_{[2]1}, \mathbf{X}_{(2)1}\}, \{Y_{[3]1}, \pi_{[3]1}, X_{(3)1}\}$ | $(Y_{[2]1}, \pi_{[2]1})$ |
| | 3 | $\{Y_{[1]1}, \pi_{[1]1}, X_{(1)1}\}, \{Y_{[2]1}, \pi_{[2]1}, X_{(2)1}\}, \{\mathbf{Y}_{[3]1}, \boldsymbol{\pi}_{[3]1}, \mathbf{X}_{(3)1}\}$ | $(Y_{[3]1}, \pi_{[3]1})$ |
| 2 | 1 | $\{\mathbf{Y}_{[1]2}, \boldsymbol{\pi}_{[1]2}, \mathbf{X}_{(1)2}\}, \{Y_{[2]2}, \pi_{[2]2}, X_{(2)2}\}, \{Y_{[3]2}, \pi_{[3]2}, X_{(3)2}\}$ | $(Y_{[1]2}, \pi_{[1]2})$ |
| | 2 | $\{Y_{[1]2}, \pi_{[1]2}, X_{(1)2}\}, \{\mathbf{Y}_{[2]2}, \boldsymbol{\pi}_{[2]2}, \mathbf{X}_{(2)2}\}, \{Y_{[3]2}, \pi_{[3]2}, X_{(3)2}\}$ | $(Y_{[2]2}, \pi_{[2]2})$ |
| | 3 | $\{Y_{[1]2}, \pi_{[1]2}, X_{(1)2}\}, \{Y_{[2]2}, \pi_{[2]2}, X_{(2)2}\}, \{\mathbf{Y}_{[3]2}, \boldsymbol{\pi}_{[3]2}, \mathbf{X}_{(3)2}\}$ | $(Y_{[3]2}, \pi_{[3]2})$ |

We note that $Y_{[h]1}$ is not necessarily the same as the $Y$-value of the unit having the $h^{\text{th}}$ smallest $Y$-value $(Y_{(h)1})$ since its rank is induced based on the size variable $X$. The square brackets are used to denote the possibility of within-set ranking error. If there is no ranking error, the square brackets are replaced with round parentheses. In this case $Y_{(h)1}$ becomes the $h^{\text{th}}$ order statistic in a set of size $H$.

In a recent study, Ozturk (2019b) used the induced ranks post-experimentally in a PPS-judgment-post-stratified sample. The key difference between a PPS-ranked-set sample and a PPS-judgment-post-stratified sample is in the implementation of the ranking process. The ranks in a PPS-ranked-set sample are obtained prior to measurement of the $Y$-variable, but the ranks in a PPS-judgment-post-stratified sample are obtained post-experimentally after measuring the $Y$-variables in a PPS sample.

Throughout the paper, our ranking procedure satisfies the consistency requirement

$$\mathrm{HF}(y) = \sum_{h=1}^{H} F_{[h:H]}(y), \tag{2.2}$$

where $F_{[h:H]}(y)$ is the CDF of $Y_{[h]i}$. The proof of equation (2.1) is provided in Presnell and Bohn (1999). The consistency in the ranking process indicates that the same ranking procedure, however imperfect it might be, is applied in all comparison sets. Hence, the equality in equation (2.1) holds for ranking methods that use the size variable $X$.

We construct an estimator for the population mean from the PPS-ranked-set sample:

$$\bar{Y}_{\mathrm{PR},N} = \frac{1}{HdN} \sum_{h=1}^{H} \sum_{i=1}^{d} \frac{Y_{[h]i}}{\pi_{[h]i}}.$$

An estimator for the population total is given by $T_{\mathrm{PR},N} = N\bar{Y}_{\mathrm{PR},N}$. The standard PPS estimator, often referred to as the Hansen-Hurwitz estimator, with sample size $n = dH$ has the same form as the estimator $\bar{Y}_{\mathrm{PR}}$, but it does not use the ranking information:

$$\bar{Y}_{P,N} = \frac{1}{nN}\sum_{i=1}^{n}\frac{Y_i}{\pi_i}.$$

The variance of $\bar{Y}_{P,N}$ is given in standard text books to be (see, for example, Thompson, 2002, page 52)

$$\sigma^2_{\bar{Y}_{P,N}} = \mathrm{Var}\left(\bar{Y}_{P,N}\right) = \frac{1}{nN^2}\,\mathrm{Var}\left(\frac{Y_1}{\pi_1}\right) = \frac{1}{nN^2}\sum_{k=1}^{N}\pi_k\left(\frac{y_k}{\pi_k} - N\mu_N\right)^2.$$

**Theorem 1.** *Let* $(Y_{[h]i}, \pi_{[h]i})$, $h = 1, \ldots, H$, $i = 1, \ldots, d$ *be a PPS-ranked-set sample from population* $\mathcal{P}^N$. *Under any consistent ranking scheme satisfying equation (2.1), the estimator* $\bar{Y}_{\mathrm{PR},N}$ $(T_{\mathrm{PR},N})$ *is unbiased for the population mean (total). Their variances are equal to* $\sigma^2_{\bar{Y}_{\mathrm{PR},N}}$ *and* $\sigma^2_{T_{\mathrm{PR},N}}$

$$\begin{aligned}
\sigma^2_{\bar{Y}_{\mathrm{PR},N}} &= \frac{1}{H^2dN^2}\sum_{h=1}^{H}\left\{\sum_{k=1}^{N}\frac{y_k^2}{\pi_k^2}\,f_{[h:H]}(y_k) - \left[\sum_{k=1}^{N}\frac{y_k}{\pi_k}\,f_{[h:H]}(y_k)\right]^2\right\} \\
&= \frac{1}{H^2dN^2}\sum_{h=1}^{H}\sigma^2_{[h:H]} = \sigma^2_{\bar{Y}_{p,N}} - \frac{1}{nHN^2}\sum_{h=1}^{H}\left(\mu_{[h:H]} - N\mu_N\right)^2 \le \sigma^2_{\bar{Y}_{p,N}}, \\
\sigma^2_{T_{\mathrm{PR},N}} &= N^2\sigma^2_{\bar{Y}_{\mathrm{PR},N}},
\end{aligned}$$

*where* $\mu_{[h:H]} = E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)$ *and* $\sigma^2_{[h:H]} = E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \mu^2_{[h:H]}$ *are the mean and variance of* $Y_{[h]i}/\pi_{[h]i}$.

We note that the last two expected values in Theorem 1 are computed using the randomization distribution. Theorem 1 shows that the estimator $\bar{Y}_{\mathrm{PR},N}$ has always smaller variance than the variance of the mean of a PPS estimator as long as there is meaningful information to rank the sample units in a comparison set. For settings where the PPS sample is appropriate, ranking information would be available since the size variable $X$ is approximately proportional to the variable $Y$. Hence, it provides reasonably accurate ranking for the units in the comparison sets.

The probability mass function in Theorem 1, $f_{[h:H]}(y_k)$, is given for perfect ranking as in equation (2.1). Under imperfect ranking, $f_{[h:H]}(y_k)$ is the PMF of the induced order statistic $Y_{[h]1}$, and its form is not known. In the next theorem, we provide an unbiased estimator for $\sigma^2_{\bar{Y}_{\mathrm{PR},N}}$ $\left(\sigma^2_{T_{\mathrm{PR},N}}\right)$ regardless of the quality of the ranking information.

**Theorem 2.** *Let* $(Y_{[h]i}, \pi_{[h]i})$, $h = 1, \ldots, H$; $i = 1, \ldots, d$, *be a PPS-ranked-set sample from population* $\mathcal{P}^N$. *Under any consistent ranking scheme satisfying equation (2.1), unbiased estimators of* $\sigma^2_{\bar{Y}_{\mathrm{PR},N}}$ *and* $\sigma^2_{T_{\mathrm{PR},N}}$ *are given by*

$$\begin{aligned}
\hat{\sigma}^2_{\bar{Y}_{\mathrm{PR},N}} &= \frac{1}{2d^2(d-1)H^2N^2}\sum_{h=1}^{H}\sum_{i=1}^{d}\sum_{j\ne i}^{d}\left\{\frac{Y_{[h]i}}{\pi_{[h]i}} - \frac{Y_{[h]j}}{\pi_{[h]j}}\right\}^2, \quad d > 1 \\
\hat{\sigma}^2_{T_{\mathrm{PR},N}} &= N^2\hat{\sigma}^2_{\bar{Y}_{\mathrm{PR},N}}.
\end{aligned}$$

For moderately large sample sizes, we can use the normal approximation to provide approximate $(1 - \alpha)$ 100% confidence intervals for the population mean and total, namely,

$$\bar{Y}_{\mathrm{PR},\,N} \pm t_{n-H,\,\alpha/2}\hat{\sigma}^2_{\bar{Y}_{\mathrm{PR},\,N}},\quad \bar{T}_{\mathrm{PR},\,N} \pm t_{n-H,\,\alpha/2}\hat{\sigma}^2_{\bar{T}_{\mathrm{PR},\,N}},$$

where $t_{n-H,\,a}$ is the $a^{\text{th}}$ upper quantile of a t-distribution having degrees of freedom $df = n - H$. The $df = n - H$ is proposed to take into account the heterogeneity between judgment ranking classes. For smaller sample sizes, one can approximate the degrees of freedom using the Satterthwaite approximation.

We now investigate the efficiency of the PPS-ranked-set sample estimator using several populations that fit the structure presented in Section 1. The finite populations are generated using the model below.

I. For a fixed population size $N$, generate the size variable $X$ from an exponential distribution with mean 100 and order these $N$ random numbers from the smallest to the largest, $x_{(1)} < \ldots < x_{(N)}$, where $x_{(i)}$ is the $i^{\text{th}}$ smallest value of the $X$ -values.

II. Let $N^*$ be he largest integer such that $N^* \leq N(1 - \omega)$. Generate the $Y$ -values from either

$$y_{[i]} = \begin{cases} x_{(i)} + \tau\varepsilon_i & i = 1, \ldots, N^* \\ \beta x_{(i)} + \tau\varepsilon_i & i = N^* + 1, \ldots, N, \end{cases} \tag{2.3}$$

or

$$y_{[i]} = \begin{cases} x_{(i)} + \tau\sqrt{x_{(i)}}\varepsilon_i & i = 1, \ldots, N^* \\ \beta x_{(i)} + \tau\sqrt{x_{(i)}}\varepsilon_i & i = N^* + 1, \ldots, N, \end{cases} \tag{2.4}$$

where $\varepsilon_i$ is generated from a normal distribution with mean zero and variance 1 and $y_{[i]}$ is the value of the $Y$ -variable that corresponds to the value of $x_{(i)}$.

For a given integer $N$, this model generates $N$ pairs of $(Y, X)$ -measurements, for which the values of the $Y$ -variable are proportional to the values of the $X$ -variable. For the population units producing the largest $100(1 - \omega)$ % of the $Y$ -values, the slope of the regression line between the $Y$ - and $X$ -variables is $\beta$ times larger than the slope of the regression line for other units. The variance of the $Y$ -variable is constant in model (2.3) and increases with the $X$ -values in model (2.4).

We performed a simulation study to investigate the efficiency of the PPS-ranked-set sample estimator. Finite populations of size $N = 2{,}000$ are generated from models (2.3) and (2.4). The slope parameter $\beta$ is selected to be 2 or 3. The parameter $\tau$ controls the correlation between the $Y$ - and $X$ -variables, $\rho = \mathrm{cor}(X, Y)$, and is selected to be $\tau = 3, 8, 20$. The parameter $\omega$ controls the percentage of population units having a larger proportionality constant for the units with the extreme $Y$ -values. We consider $\omega$ values of 0.05, 0.10 and 0.20. For this population setting, we compare the efficiency of the PPS-ranked-set sample estimator with the PPS and ratio estimators of the population mean. The PPS with replacement samples were generated using the Lahiri (1951) method, which does not select any of the units with probability one from the population, and gives every unit in the population a positive probability of being selected in the sample. For each sample, the sample size is fixed at $n = dH$, with

$d = 5$ and $H = 5, 10$. Relatively smaller sample sizes $(n = 25, 50)$ are used to asses the small sample behaviors of the coverage probabilities of the confidence intervals of population mean. Simulation size is taken to be 20,000. Since the ratio estimator is not a design unbiased estimator, we use the mean squared error (MSE) of the ratio estimator to compare its efficiency with the PPS-ranked-set sample estimator. The MSE of the ratio estimator is computed as follows

$$\text{MSE}_R = \frac{1}{20{,}000}\sum_{i=1}^{20{,}000}\left(\bar{Y}_{R,i} - \mu_N\right)^2,\ \bar{Y}_{R,i} = \frac{\bar{Y}_i}{\bar{X}_i}\mu_X,$$

where $\bar{Y}_i$ and $\bar{X}_i$ are the sample means of the $Y$ - and $X$ -variables, respectively, in the $i^{\text{th}}$ iteration of the simulation, and $\mu_X$ is the population mean of the $X$ -variable.

**Table 2.2**
**Relative efficiency of the PPS-ranked-set sample estimator and coverage (COV) probability of the associated confidence interval for the population mean**

| Constant variance model eq. 2.3 | | | | | | | Increasing variance model, eq. 2.4 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\beta$ | $\omega$ | $\rho$ | $H$ | $\frac{\text{MSE}_R}{\sigma^2_{\bar{Y}_{\text{PR},N}}}$ | $\frac{\sigma^2_{\bar{Y}_{P,N}}}{\sigma^2_{\bar{Y}_{\text{PR},N}}}$ | COV | $\beta$ | $\omega$ | $\rho$ | $H$ | $\frac{\text{MSE}_R}{\sigma^2_{\bar{Y}_{\text{PR},N}}}$ | $\frac{\sigma^2_{\bar{Y}_{P,N}}}{\sigma^2_{\bar{Y}_{\text{PR},N}}}$ | COV |
| 2 | 0.05 | 0.933 | 5 | 5.385 | 1.657 | 0.939 | 2 | 0.05 | 0.918 | 5 | 3.335 | 1.348 | 0.949 |
| 2 | 0.05 | 0.933 | 10 | 7.281 | 2.231 | 0.936 | 2 | 0.05 | 0.918 | 10 | 4.012 | 1.587 | 0.950 |
| 2 | 0.05 | 0.932 | 5 | 4.338 | 1.523 | 0.945 | 2 | 0.05 | 0.844 | 5 | 1.607 | 1.090 | 0.950 |
| 2 | 0.05 | 0.932 | 10 | 5.389 | 1.907 | 0.944 | 2 | 0.05 | 0.844 | 10 | 1.731 | 1.144 | 0.950 |
| 2 | 0.05 | 0.926 | 5 | 2.093 | 1.235 | 0.952 | 2 | 0.05 | 0.611 | 5 | 1.128 | 1.018 | 0.947 |
| 2 | 0.05 | 0.926 | 10 | 2.158 | 1.358 | 0.951 | 2 | 0.05 | 0.611 | 10 | 1.167 | 1.035 | 0.949 |
| 2 | 0.10 | 0.951 | 5 | 5.136 | 1.900 | 0.930 | 2 | 0.10 | 0.939 | 5 | 3.404 | 1.533 | 0.952 |
| 2 | 0.10 | 0.951 | 10 | 6.982 | 2.567 | 0.938 | 2 | 0.10 | 0.939 | 10 | 4.017 | 1.778 | 0.949 |
| 2 | 0.10 | 0.951 | 5 | 4.283 | 1.748 | 0.941 | 2 | 0.10 | 0.875 | 5 | 1.660 | 1.157 | 0.952 |
| 2 | 0.10 | 0.951 | 10 | 5.339 | 2.188 | 0.945 | 2 | 0.10 | 0.875 | 10 | 1.718 | 1.181 | 0.949 |
| 2 | 0.10 | 0.946 | 5 | 2.227 | 1.372 | 0.953 | 2 | 0.10 | 0.648 | 5 | 1.134 | 1.040 | 0.950 |
| 2 | 0.10 | 0.946 | 10 | 2.273 | 1.490 | 0.950 | 2 | 0.10 | 0.648 | 10 | 1.129 | 1.029 | 0.948 |
| 2 | 0.20 | 0.975 | 5 | 4.005 | 1.989 | 0.939 | 2 | 0.20 | 0.965 | 5 | 2.888 | 1.631 | 0.948 |
| 2 | 0.20 | 0.975 | 10 | 5.253 | 2.764 | 0.941 | 2 | 0.20 | 0.965 | 10 | 3.316 | 1.941 | 0.952 |
| 2 | 0.20 | 0.974 | 5 | 3.419 | 1.843 | 0.942 | 2 | 0.20 | 0.911 | 5 | 1.581 | 1.210 | 0.950 |
| 2 | 0.20 | 0.974 | 10 | 4.100 | 2.370 | 0.947 | 2 | 0.20 | 0.911 | 10 | 1.598 | 1.238 | 0.949 |
| 2 | 0.20 | 0.970 | 5 | 1.873 | 1.442 | 0.950 | 2 | 0.20 | 0.711 | 5 | 1.151 | 1.067 | 0.951 |
| 2 | 0.20 | 0.970 | 10 | 1.819 | 1.585 | 0.954 | 2 | 0.20 | 0.711 | 10 | 1.123 | 1.047 | 0.949 |
| 3 | 0.05 | 0.873 | 5 | 5.560 | 1.679 | 0.936 | 3 | 0.05 | 0.867 | 5 | 4.700 | 1.551 | 0.945 |
| 3 | 0.05 | 0.873 | 10 | 7.596 | 2.286 | 0.935 | 3 | 0.05 | 0.867 | 10 | 6.130 | 1.998 | 0.945 |
| 3 | 0.05 | 0.873 | 5 | 5.220 | 1.636 | 0.940 | 3 | 0.05 | 0.832 | 5 | 2.697 | 1.253 | 0.950 |
| 3 | 0.05 | 0.873 | 10 | 6.973 | 2.178 | 0.938 | 3 | 0.05 | 0.832 | 10 | 3.121 | 1.414 | 0.950 |
| 3 | 0.05 | 0.870 | 5 | 3.862 | 1.462 | 0.947 | 3 | 0.05 | 0.687 | 5 | 1.416 | 1.062 | 0.949 |
| 3 | 0.05 | 0.870 | 10 | 4.610 | 1.774 | 0.946 | 3 | 0.05 | 0.687 | 10 | 1.504 | 1.100 | 0.950 |
| 3 | 0.10 | 0.914 | 5 | 5.274 | 1.924 | 0.926 | 3 | 0.10 | 0.909 | 5 | 4.601 | 1.786 | 0.944 |
| 3 | 0.10 | 0.914 | 10 | 7.257 | 2.632 | 0.937 | 3 | 0.10 | 0.909 | 10 | 5.969 | 2.289 | 0.945 |
| 3 | 0.10 | 0.914 | 5 | 5.005 | 1.877 | 0.934 | 3 | 0.10 | 0.880 | 5 | 2.791 | 1.402 | 0.953 |
| 3 | 0.10 | 0.914 | 10 | 6.717 | 2.505 | 0.940 | 3 | 0.10 | 0.880 | 10 | 3.144 | 1.551 | 0.950 |
| 3 | 0.10 | 0.912 | 5 | 3.875 | 1.674 | 0.945 | 3 | 0.10 | 0.747 | 5 | 1.451 | 1.111 | 0.951 |
| 3 | 0.10 | 0.912 | 10 | 4.635 | 2.027 | 0.946 | 3 | 0.10 | 0.747 | 10 | 1.479 | 1.119 | 0.949 |
| 3 | 0.20 | 0.957 | 5 | 4.090 | 2.009 | 0.936 | 3 | 0.20 | 0.953 | 5 | 3.694 | 1.886 | 0.944 |
| 3 | 0.20 | 0.957 | 10 | 5.434 | 2.825 | 0.940 | 3 | 0.20 | 0.953 | 10 | 4.664 | 2.497 | 0.948 |
| 3 | 0.20 | 0.957 | 5 | 3.919 | 1.968 | 0.940 | 3 | 0.20 | 0.929 | 5 | 2.446 | 1.490 | 0.950 |
| 3 | 0.20 | 0.957 | 10 | 5.073 | 2.703 | 0.942 | 3 | 0.20 | 0.929 | 10 | 2.680 | 1.680 | 0.952 |
| 3 | 0.20 | 0.956 | 5 | 3.125 | 1.768 | 0.946 | 3 | 0.20 | 0.815 | 5 | 1.414 | 1.155 | 0.950 |
| 3 | 0.20 | 0.956 | 10 | 3.588 | 2.194 | 0.949 | 3 | 0.20 | 0.815 | 10 | 1.409 | 1.162 | 0.949 |

Table 2.2 presents the efficiency results and the coverage (COV) probabilities of the approximate 95%-confidence intervals for the population mean based on the PPS-ranked-set sample mean. The efficiency results show that the PPS-ranked-set sample estimator has higher efficiencies than the PPS and ratio estimators for all simulation parameters in Table 2.2. The efficiency increases with each of the simulation parameters $\rho$, $H$, $\beta$ for fixed values of all the other parameters. For example, in the constant variance model, for fixed values of $\beta = 2$, $\omega = 0.05$ and $H = 5$, the efficiency values with respect to the ratio and PPS estimators are 2.093 and 1.235 for $\rho = 0.926$ and 5.385 and 1.657 for $\rho = 0.933$, respectively. The same efficiency values in the increasing variance model are 1.128 and 1.018 for $\rho = 0.611$ and 3.335 and 1.348 for $\rho = 0.918$, respectively. Similar observations can be made for other combination of simulation parameters.

The coverage probabilities of the confidence intervals for the population mean are relatively close to the nominal coverage probability 0.95 for both the constant and increasing variance models.

# 3 The PPS-ranked-set sample from stratified populations

In this section, we construct a PPS-ranked-set sample from a stratified population. The entire population is divided into $L$ stratum populations, $\mathcal{P}^{N_l} = \{u_{1,l}, \ldots, u_{N_l,l}\}$, where $N_l$ is the population size for the $l^{\text{th}}$ stratum population, $l = 1, \ldots, L$. The stratum population means, variances and totals are given by

$$\mu_{N_l} = \frac{1}{N_l}\sum_{i=1}^{N_l} y_{i,l}, \quad \sigma^2_{N_l} = \frac{1}{N_l}\sum_{i=1}^{N_l}\left(y_{i,l} - \mu_{N_l}\right)^2, \quad t_l = N_l\mu_{N_l}, \quad l = 1, \ldots, L,$$

where $y_{i,l}$ is the value of $Y$ on unit $u_{i,l}$ in population $\mathcal{P}^{N_l}$. The mean, total and variance of the overall population are defined as follows

$$\mu_N = \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l} y_{i,l}, \quad t_N = N\mu_N, \qquad \sigma^2_N = \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l}\left(y_{i,l} - \mu_N\right)^2,$$

where $N = \sum_{l=1}^{L} N_l$. The population total can be written as $t = N\mu_N$. From this stratified population, we construct the stratified-PPS-ranked-set (SPR) sample

$$\left\{Y_{[h]i,l}, \pi_{[h]i,l}\right\}, \quad i = 1, \ldots, d_l; \quad h = 1, \ldots, H_l; \quad l = 1, \ldots, L,$$

where $d_l$ and $H_l$ are the cycle and set sizes, respectively, in the stratum sample from population $l$. Let $n_l = d_l H_l$ be the sample size for stratum $l$, $l = 1, \ldots, L$. The estimators of the population mean and total are then given by

$$\bar{Y}_{\text{SPR},N} = \sum_{l=1}^{L}\frac{N_l}{N}\bar{Y}_{\text{PR},N_l} = \sum_{l=1}^{L}\frac{N_l}{N}\frac{1}{N_l H_l d_l}\sum_{h=1}^{H_l}\sum_{i=1}^{d_l}\frac{Y_{[h]i,l}}{\pi_{[h]i,l}}$$

and

$$T_{\text{SPR},N} = N\bar{Y}_{\text{SPR},N}.$$

If one ignores ranking information and uses PPS sampling, the stratified-PPS sample can be written as

$$\left\{Y_{i,l},\ \pi_{i,l}\right\},\quad i = 1, \ldots, n_l,\quad l = 1, \ldots, L.$$

The estimator of the population mean based on the stratified-PPS (SP) sample is given by

$$\bar{Y}_{\mathrm{SP},N} = \sum_{l=1}^{L} \frac{N_l}{N}\, \bar{Y}_{P,N_l} = \sum_{l=1}^{L} \frac{N_l}{N}\, \frac{1}{N_l n_l} \sum_{i=1}^{n_l} \frac{Y_{i,l}}{\pi_{i,l}}.$$

The variance of $\bar{Y}_{\mathrm{SP},N}$ can be found in standard text books

$$\sigma^2_{\bar{Y}_{\mathrm{SP},N}} = \sum_{l=1}^{L} \frac{N_l^2}{N^2}\, \frac{1}{n_l N_l^2} \sum_{k=1}^{N_l} \pi_k \left\{\frac{y_k}{\pi_k} - N_l \mu_{N_l}\right\}^2 = \sum_{l=1}^{L} \frac{N_l^2}{N^2}\, \sigma^2_{P,N_l}. \tag{3.1}$$

The next theorem shows that $\bar{Y}_{\mathrm{SPR},N}$ and $T_{\mathrm{SPR},N}$ are unbiased estimators for the population mean and total, respectively.

**Theorem 3.** *Let* $\left\{Y_{[h]i,l},\ \pi_{[h]i,l}\right\}$, $i = 1, \ldots, d_l$; $h = 1, \ldots, H_l$; $l = 1, \ldots, L$, *be a stratified-PPS-ranked-set sample from a stratified population. Under any consistent ranking model satisfying equation (2.1), the estimators* $\bar{Y}_{\mathrm{SPR},N}$ *and* $T_{\mathrm{SPR},N}$ *are unbiased for the population mean* $(\mu_N)$ *and total* $(t_N)$, *respectively, and their variances are given by* $\sigma^2_{\bar{Y}_{\mathrm{SPR},N}}$ *and* $\sigma^2_{T_{\mathrm{SPR},N}}$,

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},N}} = \sum_{l=1}^{L} \frac{N_l^2}{N^2}\, \sigma^2_{\bar{Y}_{\mathrm{PR},N_l}} = \sum_{l=1}^{L} \frac{N_l^2}{N^2} \left\{\sigma^2_{P,N_l} - \frac{1}{N_l^2 n_l H_l} \sum_{h=1}^{H_l} \left(\mu_{[h:H_l]} - N_l \mu_{N_l}\right)^2\right\} \le \sigma^2_{\bar{Y}_{\mathrm{SP},N}}$$

$$\sigma^2_{T_{\mathrm{SPR},N}} = N^2 \sigma^2_{\bar{Y}_{\mathrm{SPR},N}}.$$

The proof of Theorem 3 follows from Theorem 1. Theorem 3 indicates that the variance of the sample mean $\bar{Y}_{\mathrm{SPR},N}$ based on a stratified-PPS-ranked-set sample is always less than or equal to the variance of the sample mean $\bar{Y}_{\mathrm{SP},N}$ based on the stratified-PPS-sample for settings where PPS sampling is appropriate.

From Theorem 2, an unbiased estimator for the variance $\sigma^2_{\bar{Y}_{\mathrm{SPR},N}}$ is given by

$$\hat{\sigma}^2_{\bar{Y}_{\mathrm{SPR},N}} = \sum_{l=1}^{L} \frac{N_l^2}{N^2}\, \frac{1}{2d_l^2 (d_l - 1) H_l^2 N_l^2} \sum_{h=1}^{H_l} \sum_{i=1}^{d_l} \sum_{j \ne i}^{d_l} \left\{\frac{Y_{[h]i,l}}{\pi_{[h]i,l}} - \frac{Y_{[h]j,l}}{\pi_{[h]j,l}}\right\}^2,\quad d_l > 1;\quad l = 1, \ldots, L.$$

This unbiased variance estimator provides a way to construct an approximate confidence interval for the population mean $\mu_N$, namely,

$$\bar{Y}_{\mathrm{SPR},N} \pm t_{df,\alpha/2} \hat{\sigma}_{\bar{Y}_{\mathrm{SPR},N}},$$

where $df = \sum_{l=1}^{L} n_l - \sum_{l=1}^{L} H_l$. For smaller sample sizes, the degrees of freedom can also be approximated using the Satterthwaite approximation to adjust for the effect of unequal stratum population variances. A similar expression can be written for a confidence interval for the population total $t_N$.

# 4 Sample size determination

One of the objectives of a stratified sampling design is to maximize the information content of the sample. Since our sampling design involves selecting samples from each one of the stratum populations, sample size allocation to strata populations becomes an important issue and has a big impact on the information content of the sample. We consider four different allocation methods: equal, proportional, Neyman and optimal allocation for a given cost model. The sample size allocation (as it relates to the efficiency of the estimator) depends very much on the cost structure of the sampling procedure and the magnitudes of the stratum-level variances. Hence, these four allocation procedures yield different efficiency results, since they try to minimize either the cost of sampling or the contribution of the stratum-level variances.

Note that the number of strata $L$ is fixed and the sample size for stratum population $l$ is $n_l$. For equal allocation, all stratum sample sizes are equal to $n_l \equiv n/L$, $l = 1, \ldots, L$, where $n$ is the total sample size in the stratified-PPS-ranked-set sample. For proportional allocation, the sample size $n_l$ is selected to be proportional to the stratum population size $N_l$ $l = 1, \ldots, L$, namely,

$$n_l = \frac{N_l}{N} n, \quad l = 1, \ldots, L.$$

Once $n_l$ is determined in this way, one can set $d_l = n_l / H_l$ for given set size $H_l$. For the setting where PPS sampling is appropriate, ranking in comparison sets is performed based on the size variable $X$. Since the variable $X$ is proportional to the variable $Y$, we expect that the $X$- and $Y$-variables are highly correlated. Hence, we select a moderately large value for $H_l$ for given $n_l$, such as $H_l = 5$, 6 or 10. Under proportional allocation, the variance of $\bar{Y}_{\mathrm{SPR},N}$ is given by

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},N}}(P) = \sum_{l=1}^{L} \frac{1}{NN_l n H_l} \sum_{h=1}^{H_l} \sigma^2_{[h:H_l]} = \sum_{l=1}^{L} \frac{\bar{\sigma}_l^2}{NN_l n}, \tag{4.1}$$

where $\bar{\sigma}_l^2 = \frac{1}{H_l} \sum_{h=1}^{H_l} \sigma^2_{[h:H_l]}$.

The Neyman allocation minimizes the variance of the estimator with respect to the sample size $n_l$ subject to the constraint that the sum of the stratum sample sizes equals $n$. Using Lagrange multipliers one can show that

$$n_l = \frac{n\bar{\sigma}_l}{\sum_{l=1}^{L} \bar{\sigma}_l}$$

minimizes the variance of $\bar{Y}_{\mathrm{SPR},N}$. Under a Neyman allocation, the variance of $\bar{Y}_{\mathrm{SPR},N}$ reduces to the simple form

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},N}}(N) = \sum_{l=1}^{L} \frac{\bar{\sigma}_l \sum_{l=1}^{L} \bar{\sigma}_l}{N^2 n}. \tag{4.2}$$

If the survey study has a budget constraint, the sample size allocation can be optimized by minimizing the variance of the estimator with a budgetary constraint in a cost function. A simple cost function for a stratified-PPS-ranked-set sample is given by

$$C_T = C_0 + \sum_{l=1}^{L} (c_l + r_l)\, n_l, \tag{4.3}$$

where $C_T$ is the total cost, $C_0$ is the overhead cost, $c_l$ is the cost of measuring a single observation from the stratum $l$ and $r_l$ is the cost of ranking $H_l$ observations in the comparison set in a stratum $l$. For settings where PPS-ranked-set sampling is appropriate, we expect that $r_l$ is either zero or very small. Under this cost function, the optimal allocation of the sample sizes is given by

$$n_l = n\, \frac{\bar{\sigma}_l \big/ \sqrt{c_l + r_l}}{\sum_{l=1}^{L} \bar{\sigma}_l \big/ \sqrt{c_l + r_l}}, \quad l = 1, \ldots, L.$$

Under the cost model (4.3), the variance of $\bar{Y}_{\mathrm{SPR},\,N}$ is given by

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(C) = \sum_{l=1}^{L} \frac{\bar{\sigma}_l \big/ \sqrt{c_l + r_l} \sum_{l=1}^{L} \bar{\sigma}_l \big/ \sqrt{c_l + r_l}}{N^2 n}.$$

We now compare the stratified-PPS-ranked-set sample estimator under the equal, proportional and Neyman allocation procedures. Under equal allocation, each stratum sample has the same sample size $n_l = n/L$, $l = 1, \ldots, L$. The variance of $\bar{Y}_{\mathrm{SPR},\,N}$ under the equal allocation is given by

$$\sigma^2_{\bar{X}_{\mathrm{SPR},\,N}}(E) = \sum_{l=1}^{L} \frac{L\bar{\sigma}_l^2}{N^2 n}. \tag{4.4}$$

The difference between the variances of $\bar{Y}_{\mathrm{SPR},\,N}$ under the equal and proportional allocations can be written as

$$\sigma^2_{\bar{Y}_{SPR,N}}(E) - \sigma^2_{\bar{Y}_{SPR,N}}(P) = \sum_{l=1}^{L} \frac{L\bar{\sigma}_l^2}{Nn} \left\{ \frac{N_l - \bar{N}}{NN_l} \right\}. \tag{4.5}$$

We expect that this difference would be positive for settings where large stratum populations have large variances. In that case, a proportional allocation increases the sample size for a large stratum population to reduce the contribution from this stratum sample to the variance of the estimator.

For Neyman allocation, we have

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(E) - \sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(N) = \sum_{l=1}^{L} \frac{1}{N^2 n} \{\bar{\sigma}_l - \bar{\sigma}_.\}^2 \geq 0$$

and

$$\begin{aligned}
\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(P) - \sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(N) &= \sum_{l=1}^{L} \frac{\bar{\sigma}_l^2}{NN_l n} - \sum_{l=1}^{L} \frac{\bar{\sigma}_l \sum_{l=1}^{L} \bar{\sigma}_l}{N^2 n} \\
&= \sum_{l=1}^{L} \frac{L\bar{\sigma}_l^2}{\bar{N}N_l n} - \sum_{l=1}^{L} \frac{\bar{\sigma}_l \sum_{l=1}^{L} \bar{\sigma}_l}{N^2 n} \\
&\geq \sum_{l=1}^{L} \frac{L\bar{\sigma}_l^2}{N^2 n} - \sum_{l=1}^{L} \frac{\bar{\sigma}_l \sum_{l=1}^{L} \bar{\sigma}_l}{N^2 n} = \sum_{l=1}^{L} \frac{L}{N^2 n} \{\bar{\sigma}_l - \bar{\sigma}_.\}^2 \geq 0,
\end{aligned}$$

where $\bar{\sigma}_{\cdot}^2 = \sum_{l=1}^{L} \bar{\sigma}_l^2 / L$. As expected, Neyman allocation always yields a smaller variance than both equal and proportional allocations, but it requires that the variance of the induced order statistics are known prior to construction of the sample. For the setting where the set sizes $H_l \equiv H$ for all stratum samples and the stratum population variances are known (or may be estimated) from previous studies, the Neyman allocation can be approximated as follows

$$n_l = \frac{n\bar{\sigma}_l}{\sum_{l=1}^{L} \bar{\sigma}_l} \approx \frac{n\hat{\sigma}_{N_l}}{\sum_{l=1}^{L} \hat{\sigma}_{N_l}}, \quad l = 1, \ldots, L,$$

where $\hat{\sigma}_{N_l}^2$ is the estimate of the variance, $\sigma_{N_l}^2$, of stratum population $l$.

# 5 Efficiency comparison of the new sampling design and estimator

In this section, we investigate the efficiency of the stratified-PPS-ranked-set sample estimators. We consider a stratified population with three strata $(L = 3)$. To see the effects of the stratum population sizes and variances on the allocation procedures, we generated stratified populations with different population sizes and variances. For clarity of notation, we define the proportions of population sizes and variances as follows:

$$p_{N_l} = \frac{N_l}{\sum_{l=1}^{L} N_l}, \quad p_{\sigma_{N_l}} = \frac{\sigma_{N_l}}{\sum_{l=1}^{L} \sigma_{N_l}}, \quad l = 1, \ldots, L.$$

We note that proportional and Neyman allocations select stratum sample sizes proportional to $p_{N_l}$ and $p_{\sigma_{N_l}}$, respectively.

In this part of the simulation, the population values of the $Y$- and $X$-variables are generated with a model different from the models in equations (2.3) and (2.4). For stratum population $l$, we generate $\mathbf{X}_l^* = \left(X_{1,l}^*, \ldots, X_{N_l,l}^*\right)$ from

$$X_{i,l}^* = F^{-1}(i/(N+1); 0.1); \quad i = 1, \ldots, N_l,$$

where $F(;\lambda)$ is the cumulative distribution function of the exponential distribution with mean 10 (rate $\lambda = 0.1$). To simplify construction of the values of the $X$-variable from the stratum population, we re-scaled $X_{i,l}^*$ by

$$X_{i,l} = \sqrt{\frac{X_{i,l}^*}{\min(\mathbf{X}_l^*)}}, \quad i = 1, \ldots, N_l.$$

The values of the variable $Y$ in the stratum population $l$ are generated from the quantiles of a normal distribution using the variable $X_{i,l}$, $i = 1, \ldots, N_l$. We first compute

$$\varepsilon_{i,l} = G^{-1}(i/(N_l+1); \theta_l, \tau_l); \quad i = 1, \ldots, N_l,$$

where $G(;a,b)$ is the CDF of a normal distribution with mean $a$ and standard deviation $b$. The values of the $Y$-variable are then constructed from

$$Y_{i,l} = X_{i,l}\varepsilon_{i,l}, \quad i = 1, \ldots, N_l.$$

In this construction, it is clear that the values of the $Y$-variable are proportional to the values of the $X$-variable. Hence, use of the stratified-PPS-ranked-set sample would be appropriate.

For the simulation study, the total population size $N$, the sample size $n$ and the location parameter $\theta_l$ $(l = 1, 2, 3)$ are selected to be $N = 700$, $n = 90$, and $\theta_l = 5$, respectively. The values $N_l$ and $\tau_l$ are varied to establish the values of $p_{N_l}$ and $p_{\sigma_{N_l}}$ in Tables 5.1 and 5.2. For the first four rows, the population sizes and standard deviations are selected to be $N_1 = 100$, $N_2 = 200$, $N_3 = 400$ and $\sigma_{N_1} = 35$, $\sigma_{N_2} = 10$, $\sigma_{N_3} = 5$, respectively. For the last four rows, the population sizes and standard deviations are selected as $N_1 = 400$, $N_2 = 200$, $N_3 = 100$ and $\sigma_{N_1} = 35$, $\sigma_{N_2} = 10$, $\sigma_{N_3} = 5$, respectively. To make the comparison easier, the same set size $H$ $(H = 2, 3, 5, 6)$ is used in all stratum populations for any combination of particular choices of $p_{N_l}$ and $p_{\sigma_{N_l}}$, $l = 1, \ldots, L$.

An unbiased estimator of the variance of the sample mean requires that $d_l \geq 2$, for $l = 1, \ldots, L$. In Neyman and proportional allocations, this assumption may not hold in certain stratum samples when $p_{N_l}$ or $p_{\sigma_{N_l}}$ is too small. In this case, we modified the Neyman and proportional allocations to make sure that $d_l \geq 2$ by reducing the maximum $d_l$ and increasing any $d_l$ smaller than 2. These allocation procedures may not be optimal under this modification.

**Table 5.1**
**Relative efficiencies of the stratified-PPS-sample (SP) with respect to the stratified-PPS-ranked-set (SPR) sample; E: Equal allocation; P: Proportional allocation; N: Neyman allocation**

| | Proportion of $N_l$ | | | Proportion of $\sigma^2_{N_l}$ | | | Efficiencies | | |
|---|---|---|---|---|---|---|---|---|---|
| $H$ | $p_{N_1}$ | $p_{N_2}$ | $p_{N_3}$ | $p_{\sigma^2_{N_1}}$ | $p_{\sigma^2_{N_2}}$ | $p_{\sigma^2_{N_3}}$ | $\frac{\sigma^2_{\bar{Y}_{SP}}(E)}{\sigma^2_{\bar{Y}_{SPR}}(E)}$ | $\frac{\sigma^2_{\bar{Y}_{SP}}(P)}{\sigma^2_{\bar{Y}_{SPR}}(P)}$ | $\frac{\sigma^2_{\bar{Y}_{SP}}(N)}{\sigma^2_{\bar{Y}_{SPR}}(N)}$ |
| 2 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.472 | 1.408 | 2.007 |
| 3 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.927 | 1.850 | 2.627 |
| 5 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 2.803 | 3.001 | 3.823 |
| 6 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 3.229 | 3.059 | 4.402 |
| 2 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 1.468 | 1.496 | 1.506 |
| 3 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 1.915 | 1.917 | 1.965 |
| 5 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 2.769 | 2.715 | 2.689 |
| 6 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 3.180 | 3.358 | 2.440 |

Table 5.1 presents the relative efficiencies of the stratified-PPS-ranked-set sample mean with respect to the stratified-PPS sample mean for the equal, proportional and Neyman allocation procedures. The efficiencies are computed using equations (3.1), (4.1), (4.2) and (4.4). It is clear that the stratified-PPS-ranked-set sample mean has higher efficiency than the stratified-PPS sample mean for all allocation procedures. The efficiency improvement increases with the set size $H$.

**Table 5.2**
**Relative efficiencies of the stratified-PPS-ranked-set sample estimator with respect to Neyman allocation and the coverage probabilities of confidence intervals; E: Equal allocation; P: Proportional allocation; N: Neyman allocation**

| | Proportion of $N_I$ | | | Proportion of $\sigma^2_{N_I}$ | | | Efficiencies | | Coverage Prob | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $H$ | $p_{N_1}$ | $p_{N_2}$ | $p_{N_3}$ | $p_{\sigma^2_{N_1}}$ | $p_{\sigma^2_{N_2}}$ | $p_{\sigma^2_{N_3}}$ | $\frac{\sigma^2_{\bar{Y}_{SPR}}(E)}{\sigma^2_{\bar{Y}_{SPR}}(N)}$ | $\frac{\sigma^2_{\bar{Y}_{SPR}}(P)}{\sigma^2_{\bar{Y}_{SPR}}(N)}$ | Eq | Prop | Neyman |
| 2 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.021 | 1.358 | 0.951 | 0.947 | 0.946 |
| 3 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.021 | 1.354 | 0.950 | 0.945 | 0.948 |
| 5 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.021 | 1.214 | 0.949 | 0.950 | 0.953 |
| 6 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.021 | 1.372 | 0.950 | 0.933 | 0.948 |
| 2 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 2.327 | 1.357 | 0.941 | 0.947 | 0.945 |
| 3 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 2.325 | 1.381 | 0.944 | 0.949 | 0.951 |
| 5 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 2.201 | 1.334 | 0.939 | 0.946 | 0.949 |
| 6 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 1.739 | 0.979 | 0.941 | 0.943 | 0.944 |

Table 5.2 presents the efficiencies of the allocation procedures and the coverage probabilities of the approximate confidence interval for the population mean constructed from the stratified-PPS-ranked-set samples. Again the efficiencies are computed from the analytic expressions in equations (3.1), (4.1), (4.2) and (4.4), but the coverage probabilities are computed from a simulation study by generating 5,000 stratified-PPS-rankek-set samples. The PPS samples are generated using the function 'lahiri.design' in the R-package SDaA, Verbeke (2014). Efficiencies of the equal and proportional allocations are compared with respect to the Neyman allocation. Since the Neyman allocation is optimal, we see that all entries, except 0.979 in the last row of column 9, are greater than 1, as expected. The reason that the proportional allocation is better than the Neyman allocation in the last row is that the Neyman allocation is modified. The Neyman allocation yields $d_1 = 14$, $d_2 = 1$, $d_3 = 0$. This allocation is modified to $d_1 = 11$, $d_2 = 2$, $d_3 = 2$ so that the cycle size in each stratum sample is greater than 1. The proportional allocation in the last row did not need any modification. Since the Neyman allocation is no longer optimal in this case, it is not as efficient as the proportional allocation.

Neyman allocation is always better than equal allocation even when we modify it for the cycle sizes. The efficiency of proportional allocation with respect to equal allocation can be obtained by dividing column 8 by column 9 in Table 5.2. If the ratio of the entries in column 8 and column 9 is greater than 1, proportional allocation is more efficient than equal allocation.

It is clear that in the first 4 rows of Table 5.2, equal allocation is better than proportional allocation. In these populations, smaller stratum populations have larger variances. Hence, proportional allocation selects less data from the stratum having large variance and more data from the stratum having small variance. In the last four rows of Table 5.2, where large populations have large variances, proportional allocation has higher efficiency than equal allocation since it allocates larger sample sizes to strata with larger variances. These are consistent with the finding in equation (4.5), which indicates that proportional allocation is more efficient when large stratum populations have large variances.

The last three columns of Table 5.2 provide the coverage probabilities of the confidence intervals for the population mean for equal, proportional and Neyman allocation procedures. It is clear that all coverage probabilities are very close to the nominal coverage probability 0.95.

# 6 Example

In this section, we apply the stratified PPS-ranked-set sample design to apple production data. We considered that the apple production data provided by Turkish Statistical Institute is a finite population. The apple farms in this population are divided into seven different $(L = 7)$ geographical regions. The farms in each region are considered a stratum population. Table 6.1 indicates that the population sizes $(N_l)$, means $(\mu_{N_l})$ and standard deviations $(\sigma_{N_l})$ vary significantly. Hence, it is natural to use stratified sampling to reduce the sampling variation. Since the number of apple trees $(X)$ in each township is available and the correlation coefficients $(\rho_l)$ between the $X$ - and $Y$ -variables are relatively high in the stratum populations, the PPS-ranked-set sample in each stratum population would be a reasonable choice.

**Table 6.1**
**Population characteristics of the apple production data (in tons, 1 ton = 1,000kg)**

| Strata $(l)$ | $\mu_l$ | $\sigma_l$ | $N_l$ | $\rho_l$ |
|---|---|---|---|---|
| Marmara $(l = 1)$ | 1,536.8 | 6,425 | 106 | 0.816 |
| Aegean $(l = 2)$ | 2,233.7 | 11,604.9 | 105 | 0.856 |
| Mediterranean $(l = 3)$ | 9,384.31 | 29,907.5 | 94 | 0.901 |
| Black Sea $(l = 4)$ | 967 | 2,389.7 | 204 | 0.713 |
| Central Anatolia $(l = 5)$ | 5,588 | 28,643.4 | 171 | 0.986 |
| Eastern Anatolia $(l = 6)$ | 631.4 | 1,171.1 | 103 | 0.885 |
| Southeastern Anatolia $(l = 7)$ | 72.4 | 111.3 | 68 | 0.917 |

We treated the apple production data as a stratified population and simulated stratified-PPS-ranked-set (SPR) samples for sample size $n = 210$ and set sizes $H = 2, 3, 5, 6$. In order to compare the SPR with the competitor sampling designs, we also generated samples using stratified simple random sample (SSRS) and stratified-PPS (SP) sample. Samples are selected with the equal (E), proportional (P) and Neyman (N) allocation procedures in all three sampling designs.

For Neyman and proportional allocations, whenever the stratum cycle size $d_l$ is less than 2, the Neyman and proportional allocations are modified by changing all $d_l < 2$ to 2 and reducing the maximum $d_l$ so that the total sample size still equals $n$. This modification allows us to obtain an unbiased estimator for the variance of the stratified-PPS-ranked-set sample mean.

**Table 6.2**
**The relative efficiencies of the stratified SRS, PPS and PPS-ranked-set sample designs and the coverage probabilities of the approximate confidence interval for the population mean of the apple production data; E: Equal allocation; P: Proportional allocation; N: Neyman allocation**

| | Stratified SRS | | | Stratified PPS | | | Stratified PPS-RSS | | Coverage Prob | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $H$ | $\frac{\sigma^2_{\bar{Y}_{SSRS}}(E)}{\sigma^2_{\bar{Y}_{SPR}}(E)}$ | $\frac{\sigma^2_{\bar{Y}_{SSRS}}(P)}{\sigma^2_{\bar{Y}_{SPR}}(P)}$ | $\frac{\sigma^2_{\bar{Y}_{SSRS}}(N)}{\sigma^2_{\bar{Y}_{SPR}}(N)}$ | $\frac{\sigma^2_{\bar{Y}_{SP}}(E)}{\sigma^2_{\bar{Y}_{SPR}}(E)}$ | $\frac{\sigma^2_{\bar{Y}_{SP}}(P)}{\sigma^2_{\bar{Y}_{SPR}}(P)}$ | $\frac{\sigma^2_{\bar{Y}_{SP}}(N)}{\sigma^2_{\bar{Y}_{SPR}}(N)}$ | $\frac{\sigma^2_{\bar{Y}_{SPR}}(E)}{\sigma^2_{\bar{Y}_{SPR}}(N)}$ | $\frac{\sigma^2_{\bar{Y}_{SPR}}(P)}{\sigma^2_{\bar{Y}_{SPR}}(N)}$ | Equ. | Prop. | Neyman |
| 2 | 1.139 | 1.154 | 1.321 | 1.143 | 1.158 | 1.259 | 1.777 | 1.901 | 0.943 | 0.939 | 0.948 |
| 3 | 1.254 | 1.251 | 1.447 | 1.243 | 1.243 | 1.351 | 1.759 | 1.872 | 0.945 | 0.943 | 0.952 |
| 5 | 1.399 | 1.371 | 1.464 | 1.399 | 1.470 | 1.326 | 1.566 | 1.668 | 0.947 | 0.942 | 0.949 |
| 6 | 1.475 | 1.454 | 1.434 | 1.437 | 1.435 | 1.294 | 1.434 | 1.580 | 0.944 | 0.938 | 0.947 |

Table 6.2 provides the relative efficiencies of the PPS-ranked-set sample mean with respect to other competing estimators. These entries are obtained from a simulation of 10,000 replications. It is clear that the SPR sample mean is more efficient than both SSRS and SP sample means for all allocation procedures. Efficiencies in general increase with the set size $H$ when the stratum cycle sizes $(d_l = n_l / H > 1)$ are not too small ($H$ is not large). When $H$ is large, the Neyman allocation is modified to make sure that $d_l \geq 2$. In this case, the modified Neyman allocation loses its optimality properties, but it is still better than the other allocation procedures.

In Section 4, we observed that if the large populations have large variances, proportional allocation is better than equal allocation. Table 6.1 indicates that some of the smaller stratum populations have very large variances. For example, the population in the Mediterranean region (the second smallest population) has 94 farms, but its standard deviation is the largest among the 7 stratum populations. Hence, the efficiencies of the proportional and equal allocations in the SSRS and SP samples appear to be the same with respect to the SPR samples (columns 2, 3 for SSRS and columns 5, 6 for SP).

For the stratified-PPS-ranked-set sample, Neyman allocation provides a substantial amount of improvement over the equal and proportional allocations. The efficiency improvement is a deceasing function of the set size $H$, but this reduction is due to the use of the modified Neyman allocation for large set sizes (i.e, small $d_l$).

In the stratified-PPS-ranked-set sample, equal allocation is more efficient than proportional allocation. This can be seen from the ratio of columns 8 and 9. If we divide column 8 by column 9, all entries would be less than 1, which indicates that equal allocation has smaller variance than proportional allocation. This is again consistent with equation (4.5), which shows that equal allocation is better if the large stratum populations have small variances.

Table 6.2 also provides the coverage probabilities of approximate 95% confidence intervals for the population mean. It is clear that the coverage probabilities for all allocation procedures are very close to the nominal coverage probability of 0.95.

# 7 Concluding remarks

A probability-proportional-to-size sampling provides highly efficient estimators for the population mean and total when there exists a size measure for every unit in the population. The size measure

contains significant information about the importance of each unit being included in the sample. It also provides important information about the relative position (rank) of the population units. Combining these two pieces of information in a meaningful way leads to a new sampling design, the stratified-PPS-ranked-set sampling. The stratified-PPS-ranked-set sampling combines the efficiency gains of the probability sampling and the position (rank) information of the sample unit in a comparison set.

We constructed unbiased estimators for the population mean, total and their variances. The sample size allocation to each stratum plays a significance role in the efficiency of the estimators. The choice of the sample size allocation depends on sampling cost, stratum population sizes and variances. If the larger populations have larger variances, proportional allocation works reasonably well. The new sampling design is applied to an apple production data in a stratified population.

# Appendix

**Proof of Theorem 1:** We first note that $\left(Y_{[h]i},\, \pi_{[h]i}\right)$, $i = 1, \ldots, d$, are iid random variables. We then write

$$E\left(\bar{Y}_{\mathrm{PR},\,N}\right) = \frac{1}{dHN}\sum_{h=1}^{H} dE\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right) = \frac{1}{dHN}\sum_{h=1}^{H}\sum_{k=1}^{N} dy_k\,\frac{P\left(Y_{[h]1} = y_k\right)}{\pi_k}$$

$$= \frac{1}{HN}\sum_{h=1}^{H}\sum_{k=1}^{N}\frac{y_k}{\pi_k}\,f_{[h:H]}(y_k) = \frac{1}{HN}\sum_{k=1}^{N}\frac{y_k}{\pi_k}\sum_{h=1}^{H} f_{[h:H]}(y_k).$$

Using the consistency of the ranking procedure in equation (2.1), we write

$$E\left(\bar{Y}_{\mathrm{PR},\,N}\right) = \frac{1}{HN}\sum_{k=1}^{N}\left(\frac{y_k}{\pi_k}\right)Hf\,(y_k) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{y_k}{\pi_k}\right)\sum_{j=1}^{N}\pi_j I\left(y_k = y_j\right) = \frac{1}{N}\sum_{k=1}^{N} y_k = \mu_N.$$

For the proof of the variance, consider

$$\mathrm{Var}\left(\bar{Y}_{\mathrm{PR},\,N}\right) = \frac{d}{H^2 d^2 N^2}\sum_{h=1}^{H}\mathrm{Var}\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)$$

$$= \frac{1}{H^2 dN^2}\sum_{h=1}^{H}\left\{E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \left[E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)\right]^2\right\}$$

$$= \frac{1}{nHN^2}\sum_{h=1}^{H} E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \frac{1}{nHN^2}\sum_{h=1}^{H}\mu^2_{[h:H]}. \qquad \text{(A.1)}$$

Again using the consistency of the within-set ranking procedure in equation (2.1), we write

$$\frac{1}{nN^2H}\sum_{h=1}^{H} E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 = \frac{1}{nN^2}E\left(\frac{Y_1}{\pi_1}\right)^2 = \frac{1}{nN^2}\mathrm{Var}\left(\frac{Y_1}{\pi_1}\right) + \frac{1}{nN^2}\left\{E\left(\frac{Y_1}{\pi_1}\right)\right\}^2$$

$$= \mathrm{Var}\left(\bar{Y}_{\mathrm{PPS}}\right) + \mu_N^2/n.$$

We now insert this result in equation (A.1) and write

$$\begin{aligned}\mathrm{Var}\left(\bar{Y}_{\mathrm{PR},\,N}\right) &= \mathrm{Var}\left(\bar{Y}_{\mathrm{PPS}}\right) + \mu_N^2/n - \frac{1}{nHN^2}\sum_{h=1}^{H}\mu_{[h:H]}^2 \\ &= \mathrm{Var}\left(\bar{Y}_{\mathrm{PPS}}\right) - \frac{1}{nHN^2}\sum_{h=1}^{H}\left(\mu_{[h:H]} - N\mu_N\right)^2 \\ &\leq \mathrm{Var}\left(\bar{Y}_{\mathrm{PPS}}\right).\end{aligned}$$

This completes the proof.

**Proof of Theorem 2:** It is easy to see that

$$\begin{aligned}E\left(\hat{\sigma}^2_{\bar{Y}_{\mathrm{PR},\,N}}\right) &= \frac{2d\,(d-1)}{2d^2\,(d-1)\,N^2H^2}\sum_{h=1}^{H}E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \frac{2d\,(d-1)}{2d^2\,(d-1)\,N^2H^2}\sum_{h=1}^{H}E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)E\left(\frac{Y_{[h]2}}{\pi_{[h]2}}\right) \\ &= \frac{1}{dH^2N^2}\sum_{h=1}^{H}\left\{E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \left[E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)\right]^2\right\} = \frac{1}{dH^2N^2}\sum_{h=1}^{H}\sigma^2_{[h:H]}.\end{aligned}$$

# References


Al-Saleh, M.F., and Samawi, H.M. (2007). A note on inclusion probability in ranked set sampling and some of its variations. *Test*, 16, 198-209.

Deshpande, J.V., Frey, J. and Ozturk, O. (2006). Nonparametric ranked-set sampling confidence intervals for a finite population. *Environmental and Ecological Statistics*, 13, 25-40.

Frey, J. (2011). Recursive computation of inclusion probabilities in ranked set sampling. *Journal of Statistical Planning and Inference*, 141, 3632-3639.

Gokpinar, F., and Ozdemir, Y.A. (2010). Generalization of inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 39, 89-95.

Kadilar, C., and Cingi, H. (2003). Ratio estimators in stratified random sampling. *Biometrical Journal*, 45, 218-225.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute*, 33, 133-140.

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

Muttlak, H.A., and McDonald, L.L. (1992). Ranked set sampling and the line intercept method: A more efficient procedure. *Biometrical Journal*, 34, 329-346.

Nematollahi, N., Salehi, M.M. and Aliakbari Saba, R. (2008). Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. *Communications in Statistics - Theory and Methods*, 37, 2402-2415.

Ozdemir, Y.A., and Gokpinar, F. (2007). A generalized formula for inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 36, 89-99.

Ozdemir, Y.A., and Gokpinar, F. (2008). A new formula for inclusion probabilities in median ranked set sampling. *Communications in Statistics - Theory and Methods*, 37, 2022-2033.

Ozturk, O. (2014). Estimation of population mean and total in a finite population setting using multiple auxiliary variables. *Journal of Agricultural, Biological, and Environmental Statistics*, 19, 161-184.

Ozturk, O. (2016). Statistical inference based on judgment post-stratified samples in finite population. *Survey Methodology*, 42, 2, 239-262. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016002/article/14664-eng.pdf.

Ozturk, O. (2019a). Two-stage cluster samples with ranked set sampling designs. *Annals of the Institute of Statistical Mathematics*, 71, 63-91.

Ozturk, O. (2019b). Post-stratified probability-proportional-to-size sampling from stratified populations. *Journal of Agricultural, Biological, and Environmental Statistics*, https://doi.org/10.1007/s13253-019-00370-6.

Ozturk, O., and Bayramoglu-Kavlak, K. (2018). Statistical inference using stratified ranked set samples from finite populations. *Ranked Set Sampling: 65 Years Improving the Accuracy in Data Gathering*, (Eds., C. Bouza and A.I. Al-Omari), 157-170, Elsevier, New York.

Ozturk, O., and Jafari Jozani, M. (2013). Inclusion probabilities in partially rank ordered set sampling. *Computational Statistics and Data Analysis*, 69, 122-132.

Patil, G.P., Sinha, A.K. and Taillie, C. (1995). Finite population corrections for ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 47, 621-36.

Presnell, B., and Bohn, L.L. (1999). U-Statistics and imperfect ranking in ranked set sampling. *Journal of Nonparametric Statistics*, 10, 111-126.

Sroka, C.J. (2008). Extending ranked set sampling to survey methodology. PhD thesis, Department of Statistics, Ohio State University.

Sud, V., and Mishra, D.C. (2006). Estimation of finite population mean using ranked set two stage sampling designs. *Journal of the Indian Society of Agricultural Statistics*, 60, 108-117.

Thompson, S.K. (2002). *Sampling*, New York: John Wiley & Sons, Inc.

Verbeke, T. (2014). SDaA: Sampling: Design and Analysis. R package version 0.1-3, https://CRAN.R-project.org/package=SDaA.

Wang, X., Lim, J. and Stokes, L. (2016). Using ranked set sampling with cluster randomized designs for improved inference on treatment effects. *Journal of the American Statistical Association*, 516, 1576-1590.

Wolfe, D. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *International Scholarly Research Network ISRN Probability and Statistics*, ID 568385, doi:10.5402/2012/568385.