PLOS ONE

RESEARCH ARTICLE

Sensitive proportion in ranked set sampling

Azhar Mehmood Abbasi*, Muhammad Yousaf Shad

Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

* abbasiqau2007@yahoo.com

Abstract

This paper considers the concomitant-based rank set sampling (CRSS) for estimation of the sensitive proportion. It is shown that CRSS procedure provides an unbiased estimator of the population sensitive proportion, and it is always more precise than corresponding sample sensitive proportion (Warner SL (1965)) that based on simple random sampling (SRS) without increasing sampling cost. Additionally, a new estimator based on ratio method is introduced using CRSS protocol, preserving the respondent's confidentiality through a randomizing device. The numerical results of these estimators are obtained by using numerical integration technique. An application to real data is also given to support the methods.

Citation: Abbasi AM, Shad MY (2021) Sensitive proportion in ranked set sampling. PLoS ONE 16(8): e0256699. https://doi.org/10.1371/journal.pone.0256699

Editor: Alan D Hutson, Roswell Park Cancer Institute, UNITED STATES

Received: January 6, 2021

Accepted: July 22, 2021

Published: August 31, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

In some social surveys, we may encounter the problem of estimating the proportion of the population having sensitive attribute, such as drug addicts, users of heron and non-taxpayers, for which people are not inclined to respond truthfully. In such situations the techniques for collecting direct information may result in elusive, ambiguous, and even no response. To overcome these problems, [1] advised randomized response (RR) technique under simple random sampling (SRS) plan with the objective to collect truthful answers while fully preserving the respondent's privacy. This method involves a randomizing device, such as a spinning arrow or a deck of cards, to procure truthful information on the sensitive attribute. The respondent answers 'yes' or 'no' according to the outcome produced by the randomizing device. As the interviewer is kept unaware of the result produced by the said device, the use of this technique ensures that the respondent cannot be recognized on the basis of his/her response. After development of the first randomized response model [1], numerous variants have been suggested by different researchers to obtain more reliable estimates of the sensitive attribute by increasing respondent's degree of privacy. A comprehensive literature will not be demonstrated. However, some worth-mentioning work developed under SRS plan can be found in the Reference [2–5] and the references cited therein.

Ranked set sampling (RSS) was introduced by [6], as an efficient alternative to simple random sampling (SRS), for estimation of pasture and forage yields. The RSS employs ranking of the small sets of units by visually or via a concomitant information before selecting final sample for actual quantification. [7] developed the theory of RSS procedure. More detail and application of RSS (CRSS) can be explored in the References [8–13].

The estimation of non-sensitive population proportion under CRSS has been investigated by [14]. Thereafter, [15] introduced a new proportion estimator in CRSS framework and showed that it works better than that of [14] without using extra resources. Recently, [16] has highlighted some drawbacks associated with the estimator given in [15] and proposed new improved estimators. Moreover, [17] showed that how RSS can be applied to ordered categorical variables for estimating the probabilities of all categories. They used ordinal logistic regression to aid in the ranking of the ordinal variable of interest.

The idea of using RSS in the estimation of sensitive attribute is similar to its application in above discussed inference problems. The ranking can be carried out by visually or by using a concomitant variable which should be non sensitive but statistically correlated with study attribute. The following two examples develop better understanding about how to use concomitant information for ranking the units:

Example 1. Let us suppose that we want to estimate the proportion of drug addicts in a social survey through RR technique. We can easily rank (order) two or more units by a glance with respect to either their facial expressions or ages.

Example 2. Let us suppose that under study parameter is the proportion of non-taxpayers. We can order two or more households by a glance with respect to either their living styles or house-sizes.

Recently, [18] has adopted model-based ranking approach, introduced by [17], for studying sensitive proportion using concomitant based-rank set sampling. This ranking method requires estimated success probabilities by fitting the logistic regression. The main concern with this ranking method is that concomitant information is not directly used for ranking of the study variable, instead it requires fitting logistic model to the data on previous studies and ranking process is done on the basis of obtained probabilities. In this paper, a new efficient estimator is proposed using CRSS which overcomes the drawbacks of [18] procedure and also beats [1] estimator. Furthermore, a new estimator based on ratio method is also introduced under CRSS plan.

2 Background

Let (Y, X) denotes a bivariate random variable where sensitive study attribute Y follows a Bernoulli distribution and X is a continuous nonsensitive concomitant with cumulative distribution function (cdf) $F_X(x)$. Suppose that the conditional distribution of Y given $X = x$ is also Bernoulli and is denoted by $B(1, g(x))$, where $g(x) \in (0, 1)$ could be inverse logit (probit) link function defined as

$$g(x) = \begin{cases} \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} & \text{logit function} \\ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} e^{-\frac{1}{2}x^2}\, dx & \text{probit function,} \end{cases}$$

where $w = 0.551(\beta_0 + \beta_1 x)$, $\alpha, \beta \in \Re$.

It follows that the marginal distribution of Y is $B(1, \pi)$ with mean $\pi = E[g(X)]$ and variance $\sigma_y^2 = \pi(1 - \pi)$. For more detailed discussion, an interested reader can pursue [14]. The CRSS plan for selection of $m(\geq 2)$ units can be elucidated as:

Step 1 Identify m^2 units from (Y, X) and divide them into m sets of size m.

Step 2 From each set of size m, obtain the exact measurement on X, then rank the sets according to the values of X.

Step 3 Obtain the corresponding Y values of the ith ($i = 1, 2, \ldots, m$) ordered unit of X in the ith set.

Step 4 The above Steps 1–3 can be repeated for n cycles, if required, to obtain a sample of size $k = mn$.

Let $\{(Y_{[1]j}, X_{(1)j}), (Y_{[2]j}, X_{(2)j}), \ldots, (Y_{[i]j}, X_{(i)j}), \ldots, (Y_{[m]j}, X_{(m)j})\}$ be a bivariate ranked set sample of size m in jth cycle, $j = 1, 2, \ldots, n$, where $Y_{[i]j}$ denotes ith *imperfect* ranked unit in the jth cycle and $X_{(i)j}$ denotes ith *perfect* ranked unit in jth cycle. Note that the square bracket $[\cdot]$ denotes imperfect ranking and $(\cdot)$ serves for perfect ranking. Again, from [14], $Y_{[i]}$ is $B(1, \pi_{[i]})$ with mean (probability) $\pi_{[i]} = \mathrm{E}[g(X_{(i)})]$ and variance $\sigma^2_{y_{[i]}} = \pi_{[i]}(1 - \pi_{[i]})$. Let $f_{(i)}(x)$ be the probability density function (pdf) and $F_{(i)}(x)$ cumulative distribution function (cdf) of an order statistics (OS) $X_{(i)}$ then we have

$$f_{(i)}(x) = \binom{m}{i} (F(x))^{i-1} (1 - F(x))^{m-i} f(x), -\infty < x < \infty. \tag{2.1}$$

and

$$F_{(i)}(x) = \sum_{r=i}^{m} \binom{n}{r} (F(x))^{r} (1 - F(x))^{m-r}$$

The mean and variance of $X_{(i)}$ are given by

$$\bar{X}_{(i)} = \int_{-\infty}^{\infty} x f_{(i)}(x) \mathrm{dx} \quad \text{and} \quad \sigma^2_{x(i)} = \int_{-\infty}^{\infty} (x - \mu_{x(i)})^2 f_{(i)}(x) \mathrm{dx}$$

respectively; see the Reference [19].

The success ('yes') probability of $Y_{[i]}$, as given in [14], is numerically computed as

$$\mathrm{E}(Y_{[i]}) = \pi_{[i]} = \int_{-\infty}^{\infty} f_{(i)}(x) g(x) \mathrm{dx}$$

The covariance between $X_{(i)}$ and $Y_{[i]}$ is defined as

$$\mathrm{Cov}(X_{(i)}, Y_{[i]}) = \mathrm{E}(X_{(i)} Y_{[i]}) - \mathrm{E}(X_{(i)}) \mathrm{E}(Y_{[i]}),$$

By virtue of partitioning, as defined in [20], we have

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_{(i)}(x) \qquad f(y) = \frac{1}{m} \sum_{i=1}^{m} f_{[i]}(y) \tag{2.2}$$

The following well-known notations will be used in this study.

$$\begin{aligned}
\pi &= \mathrm{E}(Y) & \sigma_y^2 &= \mathrm{E}(Y-\pi)^2 & d_{x(i)} &= \bar{X}_{(i)} - \bar{X} \\
\pi_{[i]} &= \mathrm{E}(Y_{[i]}) & \sigma_{y_{[i]}}^2 &= \mathrm{E}(Y_{[i]} - \pi_{[i]})^2 & d_{y[i]} &= \pi_{[i]} - \pi \\
\bar{X} &= \mathrm{E}(X) & \sigma_x^2 &= \mathrm{E}(X-\bar{X})^2 & d_{xy_{[i]}} &= (\bar{X}_{(i)} - \bar{X})(\pi_{[i]} - \pi) \\
\bar{X}_{(i)} &= \mathrm{E}(X_{(i)}) & \sigma_{x_{(i)}}^2 &= \mathrm{E}(X_{(i)} - \bar{X}_{(i)})^2 & \sigma_{xy_{[i]}} &= \mathrm{E}(X_{(i)} - \bar{X}_{(i)})(Y_{[i]} - \pi_{[i]})
\end{aligned}$$

Also the following relationships will be used in this paper.

$$\begin{aligned}
\sum_{i=1}^{m}\pi_{[i]} &= m\pi & \sum_{i=1}^{m} d_{x(i)} &= 0 & \sum_{i=1}^{m}\sigma_{y_{[i]}}^2 &= m\sigma_y^2 - \sum_{i=1}^{m} d_{y[i]}^2 \\
\sum_{i=1}^{m}\bar{X}_{(i)} &= m\bar{X} & \sum_{i=1}^{m} d_{y_{[i]}} &= 0 & \sum_{i=1}^{m}\sigma_{x_{(i)}}^2 &= m\sigma_x^2 - \sum_{i=1}^{m} d_{x(i)}^2 \\
\sum_{i=1}^{m}\sigma_{xy_{[i]}} &= m\sigma_{xy} - \sum_{i=1}^{m} d_{xy_{[i]}}
\end{aligned}$$

For more detail, see the References [20, 21].

3 Warner's model under CRSS

As this study involves randomized response (RR) procedure, it is important to give an overview of the basic RR procedure given in the Reference [1]. Let $Y_{1j}, Y_{2j}, \ldots, Y_{mj}$ be a simple random sample with replacement (SRSWR) of size m in jth cycle, for ($j = 1, 2, \ldots, n$). Each respondent is provided with a suitable randomizing device, say a spinner, for selection of one of the two statements: (a) *I have the sensitive attribute* A (b) *I do not have the sensitive attribute* A with pre-assigned selection probabilities $p \neq 0.5$ and $1 - p$ respectively. Each respondent spins the spinner and report 'yes' ('no') if his/her status matches (does not match) with the statement pointed out by the randomization device. As the interviewer is kept unaware of the outcome of the randomization device, and this makes the respondent comfortable to truthfully report his/her actual status. Then 'yes' response of ith ($i = 1, 2, \ldots, m$) respondent at jth cycle is given by

$$\lambda = P(yes) = p\pi + (1-p)(1-\pi)$$

Let m_1 denotes number of 'yes' responses out of the sample of size $k = mn$, then [1] derived the maximum likelihood estimate of π as given by $\hat{\pi}_{(srs)} = \{\hat{\lambda} - (1-p)\}/(2p-1)$ where $\hat{\lambda} = m_1/k$ is estimate of λ. The estimator $\hat{\pi}_{(srs)}$ is unbiased and its variance is given by

$$\mathrm{Var}(\hat{\pi}_{(srs)}) = \frac{1}{n}\left\{\frac{\sigma_y^2}{m} + \frac{p(1-p)}{m(2p-1)^2}\right\} = \frac{\sigma_y^2}{k} + \frac{p(1-p)}{k(2p-1)^2} \tag{3.1}$$

Now, suppose that the respondents are selected using CRSS design and are instructed to choose one of the two above-mentioned statements (a) and (b) by using the given randomizing device. The respondent reports 'yes' ('no') according to the outcomes of the randomizing device and his/her actual status. A complete layout of ith response under CRSS is given in the

S1 Fig. Let $Y_{[i]j} = 1$ if ith ranked unit reports 'yes', otherwise $Y_{[i]j} = 0$. Then

$$\begin{aligned} P(Y_{[i]j} = 1) &= p\pi_{[i]} + (1-p)(1-\pi_{[i]}) \\ P(Y_{[i]j} = 0) &= p(1-\pi_{[i]}) + (1-p)\pi_{[i]}. \end{aligned}$$

Let $Y_{[i]1}, Y_{[i]2}, \ldots, Y_{[i]n}$ are independent and identically distributed (i.i.d) Bernoulli randomized responses under CRSS plan with parameter $p\pi_{[i]} + (1-p)(1-\pi_{[i]})$, then likelihood function of $\pi_{[i]}$ for the given data $Y_{[i]j}, j = 1, 2, \ldots n$ is

$$\begin{aligned} L(\pi_{[i]}|y_{[i]}) &= \prod_{j=1}^{n}[p\pi_{[i]} + (1-p)(1-\pi_{[i]})]^{y_{[i]j}}[p(1-\pi_{[i]}) + (1-p)\pi_{[i]}]^{1-y_{[i]j}} \\ &= [p\pi_{[i]} + (1-p)(1-\pi_{[i]})]^{z_i}[p(1-\pi_{[i]}) + (1-p)\pi_{[i]}]^{n-z_i}, i = 1, 2, \ldots, m. \end{aligned} \tag{3.2}$$

where $z_i = \sum_{j=1}^{n} y_{[i]j}$ is the total number of successes observed under ith ranking unit. It is obvious that Z_i is binomial variate with parameters n and $p\pi_{[i]} + (1-p)(1-\pi_{[i]})$. The joint likelihood function of $\pi_{[i]}$, $i = 1, 2, \ldots, m$ given CRSS data $y_{\text{crss}} = \{Y_{[i]j}, i = 1, 2, \ldots, m; j = 1, 2, \ldots, n\}$ is

$$\mathrm{L}(\pi|y_{[\text{crss}]}) = \prod_{i=1}^{m} L(\pi_{[i]}|Y_{[i]}) = \prod_{i=1}^{m}[p\pi_{[i]} + (1-p)(1-\pi_{[i]})]^{z_i}[p(1-\pi_{[i]}) + (1-p)\pi_{[i]}]^{n-z_i} \tag{3.3}$$

Note that the form of maximum likelihood (ML) function given in (3.3) is too complicated to obtain ML estimate of π. Moreover, the variance of the estimator from (3.3) will not in closed form. To avoid this situation, we separately estimate each $\pi_{[i]}$ using likelihood function given in (3.2) and then these individual proportions are combined by using the relation $\pi = \frac{1}{m}\sum_{i=1}^{m}\pi_{[i]}$ for overall estimate of π. The log of the likelihood function (3.2) is

$$\begin{aligned} \log\mathrm{L}(\pi_{[i]}|y_{[i]}) &= z_i\log[p\pi_{[i]} + (1-p)(1-\pi_{[i]})] + (n-z_i)\log[p(1-\pi_{[i]}) + (1-p)\pi_{[i]}]. \\ &= z_i\log[(2p-1)\pi_{[i]} + (1-p)] + (n-z_i)\log[-(2p-1)\pi_{[i]} + p)]. \end{aligned}$$

and necessary conditions on $\pi_{[i]}$ for a maximum give

$$\frac{z_i(2p-1)}{(2p-1)\pi_{[i]} + (1-p)} = \frac{(n-z_i)(2p-1)}{-(2p-1)\pi_{[i]} + p}$$

After simplification, we obtain

$$\hat{\pi}_{[i]} = \frac{\hat{\lambda}_{[i]} - (1-p)}{2p-1}, \quad i = 1, 2, \ldots, m.$$

where $\hat{\lambda}_{[i]} = z_i/n$. Hence, the propose measure of π under CRSS plan is given by

$$\hat{\pi}_{(crss)} = \frac{1}{m}\sum_{i=1}^{m}\hat{\pi}_{[i]} = \frac{\hat{\lambda}_{[crss]} - (1-p)}{2p-1}, \tag{3.4}$$

$\hat{\lambda}_{[crss]} = \frac{1}{m}\sum_{i=1}^{m}\hat{\lambda}_{[i]}$.

Theorem: Let $\{Y_{[i]j}, i = 1, 2, \ldots, m; j = 1, 2, \ldots, n\}$ be a ranked set sample of size k. Then

(i) $\hat{\pi}_{(crss)}$ is an unbiased estimator of the population proportion π i.e., $\mathrm{E}(\hat{\pi}_{(crss)}) = \pi$

(ii) $\hat{\pi}_{(crss)}$ is more precise than $\hat{\pi}_{(srs)}$ i.e., $\mathrm{Var}(\hat{\pi}_{(crss)}) = \mathrm{Var}(\hat{\pi}_{(srs)}) - \frac{1}{km}\sum_{i=1}^{m} d^2_{y_{[i]}} \leq \mathrm{Var}(\hat{\pi}_{(srs)})$

Proof:

(i) From (3.4), we have

$$\mathrm{E}(\hat{\pi}_{(crss)}) = \frac{\mathrm{E}(\hat{\lambda}_{[crss]}) - (1-p)}{2p-1}$$

But

$$\begin{aligned}\mathrm{E}(\hat{\lambda}_{[crss]}) &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}(\hat{\lambda}_{[i]}) \\ &= \frac{1}{m}\sum_{i=1}^{m}[p\pi_{[i]} + (1-p)(1-\pi_{[i]})] \\ &= p\pi + (1-p)(1-\pi) \\ &= (2p-1)\pi + (1-p) \\ &= \lambda\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{E}(\hat{\pi}_{(crss)}) &= \frac{\lambda - (1-p)}{2p-1} \\ &= \frac{\pi(2p-1) + (1-p) - (1-p)}{2p-1} \\ &= \pi\end{aligned}$$

This completes proof (i).

(ii) From (3.4), the variance of $\hat{\pi}_{(crss)}$ is

$$\mathrm{Var}(\hat{\pi}_{(crss)}) = \mathrm{Var}\left(\frac{\hat{\lambda}_{[crss]} - (1-p)}{2p-1}\right) \tag{3.5}$$

Since variance of a constant term is zero, (3.5) reduces to

$$\begin{aligned}\mathrm{Var}(\hat{\pi}_{(crss)}) &= \frac{1}{(2p-1)^2}\mathrm{Var}(\hat{\lambda}_{[crss]}) \\ &= \frac{1}{(2p-1)^2}\frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}(\hat{\lambda}_{[i]}) \\ &= \frac{1}{(2p-1)^2}\frac{1}{km}\sum_{i=1}^{m}\lambda_{[i]}(1-\lambda_{[i]}),\end{aligned} \tag{3.6}$$

where $\lambda_{[i]} = p\pi_{[i]} + (1-p)(1-\pi_{[i]}) = (2p-1)\pi_{[i]} + (1-p)$.

Now, substituting the value of $\lambda_{[i]}$ in (3.6) and then simplification gives

$$\begin{aligned}\mathrm{Var}(\hat{\pi}_{(crss)}) &= \frac{1}{km}\sum_{i=1}^{m}\sigma^2_{y_{[i]}} + \frac{p(1-p)}{k(2p-1)^2}.\\ &= \frac{1}{km}(m\sigma_y^2 - \sum_{i=1}^{m} d^2_{y[i]}) + \frac{p(1-p)}{k(2p-1)^2}\\ &= \frac{1}{k}\sigma_y^2 + \frac{p(1-p)}{k(2p-1)^2} - \frac{1}{km}\sum_{i=1}^{m} d^2_{y[i]}\\ &= \mathrm{Var}(\hat{\pi}_{(srs)}) - \frac{1}{km}\sum_{i=1}^{m} d^2_{y_{[i]}}\end{aligned}$$

We have used the fact $\sum_{i=1}^{m}\sigma^2_{y_{[i]}} = m\sigma_y^2 - \sum_{i=1}^{m} d^2_{y[i]}$. Note that $\frac{1}{km}\sum_{i=1}^{m} d^2_{y_{[i]}} \geq 0$, hence $\mathrm{Var}(\hat{\pi}_{(crss)}) \leq \mathrm{Var}(\hat{\pi}_{(srs)})$. This completes the proof (ii).

The relative efficiency (RE) of $\hat{\pi}_{(crss)}$ with respect to $\hat{\pi}_{(srs)}$ can be examined by the ratio

$$\mathrm{RE}[\hat{\pi}_{(crss)}, \hat{\pi}_{(srs)}] = \frac{\mathrm{Var}(\hat{\pi}_{(srs)})}{\mathrm{Var}(\hat{\pi}_{(crss)})} = \left\{1 - \frac{\sum_{i=1}^{m} d^2_{y_{[i]}}}{m^2\mathrm{Var}(\hat{\pi}_{(srs)})}\right\}^{-1} \tag{3.7}$$

The expression (3.7) is always greater than unity irrespective of the choice of $g(\cdot)$, subject to the condition that $p \neq 0.5$. In other words, $\hat{\pi}_{(crss)}$ is a superior alternative to $\hat{\pi}_{(srs)}$. It may be noted that when $p = 0.5$, $\mathrm{Var}(\hat{\pi}_{srs}) \to \infty$ and consequently $\mathrm{RE}[\hat{\pi}_{(crss)}, \hat{\pi}_{(srs)}] = 1$. It is also obvious from (3.7) that RE is independent of number of cycles n, i.e., the results can not be improved by increasing n. The following result also holds from the Reference [14] when m is fixed and $n \to \infty$.

$$\sqrt{nm}(\hat{\pi}_{(crss)} - \pi) \to \mathrm{Normal}\left(0, \frac{1}{m}\sum_{i=1}^{m}\sigma^2_{y_{[i]}} + \frac{p(1-p)}{(2p-1)^2}\right). \tag{3.8}$$

We can see that when $m = 1$, (3.8) simplifies to Warner's result [1] under SRS. Furthermore, the choice $p = 1$ i.e., selection of sensitive attribute by randomization device is sure and respondent's privacy is zero, (3.8) reduces to [14] procedure of directly asking the respondent about the attribute of interest under CRSS. Whereas the choice $p = 0$ i.e., no chance of selecting sensitive question by randomizing device, also brings (3.8) to [14] procedure. Moreover, if both p and m are equal to 1, (3.8) becomes conventional method of direct interaction with the respondent under SRS. However, for precise and reliable estimate the conditions $m \geq 2$, $0 < p < 0.5$ (or$0.5 < p < 1$) are required. Finally, a consistent estimator of variance in (3.8) can be obtained by replacing $\pi_{[i]}$ with $\hat{\pi}_{[i]} = \sum_{j=1}^{n} Y_{[i]j}/n$. In this way the variance estimate becomes free of $g(\cdot)$. Hence, asymptotic inference can easily be derived from CRSS plan.

3.1 Numerical illustration

We investigate the RE of $\hat{\pi}_{(crss)}$ with respect to $\hat{\pi}_{(srs)}$ by using the expression (3.7) for different choices of β_0, β_1, $0.1 \leq p \leq 0.9$ and assuming X follows (i) normal distribution with parameters mean = 2 and variance = 1 (ii) uniform over the range 0 and 1. It is important to recall that the RE formula as given in (3.7) is independent of n, hence we take different m(= 2, 3, 4, 5) instead of n to evaluate the performance of $\hat{\pi}_{(crss)}$. Furthermore, the magnitude of correlation

coefficient between X and Y is also computed under inverse logit (probit) link function. All results are obtained by numerical integration technique, as demonstrated in the Section 2, using Mathematica Software and are displayed in S1-S4 Tables in S1 File.

As expected, the RE is an increasing function of m and/(or) ρ. It is also symmetric about $p = 0.5$. In other words, for given m and ρ, it does not matter one assigns the design parameter p or $1 - p$ to the aforesaid sensitive statement (a). However, respondents cooperation can be increased by choosing p or $1 - p$, whichever is assigned to the statement (a), from the interval [0.10.5) and at the same time one can also achieve reasonable precision for some suitable choice of m and/(or) ρ. Generally, the results under both link functions are almost same.

4 Sensitive proportion using ratio method

In survey sampling, a concomitant information is commonly used for improving precision of the estimator pertaining to non-sensitive quantity. Such information is utilized at the designing phase for selection of appropriate sample or directly at the estimation phase by ratio (product) or regression methods or incorporated at both phases. As regards sensitive proportion, a few attempts have been made to consider concomitant information at designing or estimation phase under SRS plan. For example, [22] has constructed a ratio estimator for sensitive proportion under SRS plan. The randomizing device for this method consisting of a deck of cards showing two aforesaid statements (a) and (b). In addition, each individual is required to disclose his/her true value of nonsensitive concomitant X. Then sensitive proportion estimate under this scenario is estimated as

$$\hat{\pi}_{Y(srs)} = \frac{\hat{\lambda}_r - (1-p)}{2p-1},$$

where $\hat{\lambda}_r = \hat{\lambda}\bar{X}/\hat{\bar{X}}$ is ratio estimator of λ. The expressions of bias and MSE of $\hat{\pi}_{Y(srs)}$ are, respectively, given by

$$\text{Bias}(\hat{\pi}_{Y(srs)}) = \frac{1}{k}\left\{\pi(C_x^2 - C_{xy}) - \frac{1-p}{2p-1}C_x^2\right\}, \tag{4.1}$$

and

$$\text{MSE}(\hat{\pi}_{Y(srs)}) = \text{Var}(\hat{\pi}_{srs}) - \frac{\lambda}{k(2p-1)^2}\{2(2p-1)\pi C_{xy} - \lambda C_x^2\}, \tag{4.2}$$

where $C_{xy} = \sigma_{xy}/(\bar{X}\pi)$ and $C_x = \sigma_x/\bar{X}$. [22] showed that $\hat{\pi}_Y$ is more precise than Warner's estimator [1] under some suitable conditions. Here, it is important to point out that the concomitant variable used in [22] method is binary. However, in case of continuous concomitant variable its functional form will remain the same except estimation process shifted to numerical integration. Thereafter, [23] extended this work and presented a general form of the estimator under SRS plan. To the best of our information, no single attempt has been made so far to consider concomitant information at both designing and estimation stages to optimize gain in precision for estimating sensitive proportion using CRSS plan. This motivated us to fill up this gape in the literature and suggest a new improved procedure.

Let $\{(Y_{[i]j}, X_{(i)j}); i = 1, 2, \ldots, m; j = 1, 2, \ldots, n\}$ be a bivariate ranked set sample. The respondents are instructed to select one of the two aforesaid statements (a) and (b) by using a randomizing device and report 'yes'('no') according to the statement selected by the device and their actual status. In addition, each individual is advised to provide his/her true value of X.

Now, on the lines of [22] estimator, we propose the following estimator under CRSS plan:

$$\hat{\pi}_{A(crss)} = \frac{\hat{\lambda}_{r,rss} - (1-p)}{2p-1}, \tag{4.3}$$

where $\hat{\lambda}_{r,rss} = \hat{\lambda}_{[rss]}\bar{X}/\hat{\bar{X}}_{(rss)}$ is ratio estimator of λ, $\hat{\lambda}_{[rss]} = \frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{[i]j}$, $\hat{\bar{X}}_{(rss)} = \frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{(i)j}$. To derive bias and MSE of the suggested ratio estimator up to the first order of approximation, we proceed as follows:

Let

$$\xi_o = \frac{\hat{\lambda}_{[rss]} - \lambda}{\lambda} \quad \text{and} \quad \xi_1 = \frac{\hat{\bar{X}}_{(rss)} - \bar{X}}{\bar{X}}$$

such that $E(\xi_0) = 0 = E(\xi_1)$. Following [20, 21] and keeping in view randomized response model, we have

$$E(\xi_o^2) = \frac{1}{k}\left\{\frac{1}{\lambda^2}((2p-1)^2\pi(1-\pi) + p(1-p)) - \frac{1}{m}\sum_{i=1}^{m}\tau_{y[i]}^2\right\} E(\xi_1^2) = \frac{1}{k}\left\{C_x^2 - \frac{1}{m}\sum_{i=1}^{m}\tau_{x(i)}^2\right\}$$

and

$$E(\xi_0\xi_1) = \frac{1}{k}\left\{\frac{1}{\lambda}(2p-1)\pi C_{xy} - \frac{1}{m}\sum_{i=1}^{m}\tau_{xy[i]}\right\}$$

For proof, see S1 Appendix.

Now, expressing (4.3) in terms of ξ_i, $i = 0, 1$, we have

$$\begin{aligned}
\hat{\pi}_{A(crss)} &= \frac{1}{2p-1}\left\{\frac{\lambda(1+\xi_0)\bar{X}}{\bar{X}(1+\xi_1)} - (1-p)\right\} \\
&= \frac{1}{2p-1}\{\lambda(1+\xi_0)(1+\xi_1)^{-1} - (1-p)\} \\
&= \frac{1}{2p-1}\{\lambda(1+\xi_0)(1-\xi_1+\xi_1^2) - (1-p)\} \\
&= \frac{1}{2p-1}\{\lambda(1+\xi_0-\xi_1+\xi_1^2-\xi_0\xi_1) - (1-p)\} \\
&= \pi + \frac{\lambda}{2p-1}\{\xi_0 - \xi_1 + \xi_1^2 - \xi_0\xi_1\}
\end{aligned} \tag{4.4}$$

Applying expectation on both sides of (4.4), we have

$$\begin{aligned}
\mathrm{E}(\hat{\pi}_{A(crss)}) &= \pi + \frac{\lambda}{2p-1}\{0 + \mathrm{E}(\xi_1^2) - \mathrm{E}(\xi_0\xi_1)\} \\
&= \pi + \frac{\lambda}{(2p-1)k}\left\{C_x^2 - \frac{1}{m}\sum_{i=}^{m}\tau_{x(i)}^2 - \frac{(2p-1)\pi}{\lambda}C_{yx} + \frac{1}{m}\sum_{i=1}^{m}\tau_{yx[i]}\right\} \\
&= \pi + \frac{\lambda}{(2p-1)k}\left\{C_x^2 - \frac{(2p-1)\pi}{\lambda}C_{yx} - \frac{1}{m}\sum_{i=}^{m}\tau_{x(i)}^2 + \frac{1}{m}\sum_{i=1}^{m}\tau_{yx[i]}\right\}
\end{aligned}$$

So the bias expression in the final form is given by

$$\mathrm{Bias}(\hat{\pi}_{A(crss)}) = \frac{\lambda}{(2p-1)k}\left\{C_x^2 - \frac{(2p-1)\pi}{\lambda}C_{yx} - \frac{1}{m}\sum_{i=1}^{m}(\tau_{x(i)}^2 - \tau_{yx[i]})\right\} \tag{4.5}$$

We see that the bias of $\hat{\pi}_{A(crss)}$ approaches to zero as k becomes infinitely large, indicating $\hat{\pi}_{A(crss)}$ is a consistent estimator of π. To obtain MSE of $\hat{\pi}_{A(crss)}$ up to the first order of approximation, we extract the following expression from (4.4)

$$\hat{\pi}_{A(crss)} - \pi \approx \frac{\lambda}{2p-1}\{\xi_0 - \xi_1\} \tag{4.6}$$

By the definition of MSE, from (4.6), we have

$$\begin{aligned}
\mathrm{MSE}(\hat{\pi}_{A(crss)}) &= \mathrm{E}(\hat{\pi}_{A(crss)} - \pi)^2 \\
&= \frac{\lambda^2}{(2p-1)^2}\mathrm{E}\{\xi_0^2 + \xi_1^2 - 2\xi_0\xi_1\} \\
&= \frac{\lambda^2}{(2p-1)^2}\{\mathrm{E}(\xi_0^2) + \mathrm{E}(\xi_1^2) - 2\mathrm{E}(\xi_0\xi_1)\} \\
&= \frac{\lambda^2}{k(2p-1)^2}\left\{\frac{1}{\lambda^2}((2p-1)^2\pi(1-\pi) + p(1-p)) - \frac{1}{m}\sum_{i=1}^{m}\tau_{y[i]}^2 + \right. \\
&\qquad \left. C_x^2 - \frac{1}{m}\sum_{i=1}^{m}\tau_{x(i)}^2 - \frac{2}{\lambda}(2p-1)\pi C_{xy} + \frac{2}{m}\sum_{i=1}^{m}\tau_{xy[i]}\right\} \\
&= \left\{\mathrm{Var}(\hat{\pi}_{srs}) - \frac{\lambda}{k(2p-1)^2}\{2(2p-1)\pi C_{xy} - \lambda C_x^2\}\right. \\
&\qquad \left. - \frac{\lambda^2}{mk(2p-1)^2}\sum_{i=1}^{m}\left(\tau_{y[i]}^2 + \tau_{x(i)}^2 - 2\tau_{yx[i]}\right)\right\}
\end{aligned}$$

or

$$\mathrm{MSE}(\hat{\pi}_{A(crss)}) = \mathrm{MSE}(\hat{\pi}_{Y(srs)}) - \frac{\lambda^2}{mk(2p-1)^2}\sum_{i=1}^{m}(\tau_{y[i]} - \tau_{x(i)})^2 \tag{4.7}$$

Since the second term on the right side of (4.7) is always positive. Hence, $\mathrm{MSE}(\hat{\pi}_{A(crss)}) < \mathrm{MSE}(\hat{\pi}_{Y(srs)})$, provided $p \neq 0.5$. In other words, the expression (4.7) reveals that the proposed estimator $\hat{\pi}_{A(crss)}$ is more reliable (having less risk) than $\hat{\pi}_{Y(srs)}$.

The relative efficiency of $\hat{\pi}_{A(rss)}$ with respect to $\hat{\pi}_{Y(srs)}$ can be measured by examining

$$\mathrm{RE}[\hat{\pi}_{A(crss)}, \hat{\pi}_{Y(srs)}] = \frac{\mathrm{MSE}(\hat{\pi}_{Y(srs)})}{\mathrm{MSE}(\hat{\pi}_{A(crss)})} = \left\{1 - \frac{\lambda^2 \sum_{i=1}^{m} (\tau_{y[i]} - \tau_{x(i)})^2}{m^2(2p-1)^2 \mathrm{MSE}(\hat{\pi}_{Y(srs)})}\right\}^{-1} \tag{4.8}$$

The numerical results, for different choices of m and p when X follows (i) normal distribution (ii) uniform distribution, are computed by numerical integration technique using Mathematica Software. The RE results, obtained by using the expression (4.8), are reported in S5-S8 Tables in S1 File. Note that, for the choice $p = 0.5$, RE becomes undefined, so we have omitted RE values against $p = 0.5$. As expected, all results in S5-S8 Tables in S1 File are greater than 1, and RE is an increasing function of m i.e., more precise results can be obtained by increasing m. It can be observed from S5-S8 Tables in S1 File that there is no symmetry among the RE values, obtained under the interval $0.1 \leq p < 0.5$ and $0.5 < p \leq 0.9$, as was observed for the case of $\hat{\pi}_{(crss)}$ and $\hat{\pi}_{(srs)}$ (see S1-S4 Tables in S1 File). However, as all RE values are greater than unity, $\hat{\pi}_{A(crss)}$ can be considered as an efficient alternative to $\hat{\pi}_{Y(srs)}$.

5 An application to real data

Following the Reference [24], we have conducted a small scale survey to collect the primary data set of 500 male students in Quaid-i-Azam University, Islamabad. In this survey, each student was asked about his age and a sensitive attribute–whether he has a 'girl-friend' or not. On our request, the students spared themselves for this activity and promised to response truthfully via the Warner's [1] randomizing device with $p = 0.2$. We considered 'age' as a concomitant variable X and 'girl-friend' as a sensitive attribute Y. The purpose of this data gathering was to make known of the quantities such as mean and variance of X along with proportion of study attribute Y and correlation coefficient ρ, which are given by $\bar{X} = 24$, $\sigma_x^2 = 5$, $\pi = 0.30$ and $\rho = 0.35$ respectively.

Assuming the above population data, we took a concomitant-based ranked set sample of size $k = 5(2) = 10$ as follows: We selected $m^2 = 25$ students by simple random sample with replacement sampling and randomly partitioned them into 5 sets each of size 5. Furthermore, the students in each set are ranked with respect to X and then ith ranked student is selected from the ith set ($i = 1, 2, \ldots, 5$) to estimate π. A layout of CRSS method is given in S9. Table in S1 File, where $Y_{[i]jk}$ denotes ith judgment (imperfect) ordered statistic of the student in jth set at kth cycle and $X_{(i)jk}$ serves ith perfect ordered statistic of the student in jth set at kth cycle. In the final acquired data, we have omitted jth set information for the sake of brevity. On the other hand, under simple random sample plan, 6 out of 10 students reported 'yes', that is, $\hat{\lambda}_{srs} = 0.6$. From the data given in S9. Table in S1 File, we have computed some estimates and

their associated variances (MSEs) for illustration purpose as given below.

$$\hat{\lambda}_{[crss]} = \frac{1}{10}\sum_{i=1}^{5}\sum_{k=1}^{2} Y_{[i]k} = 0.60 \qquad \hat{\bar{X}}_{(rss)} = \frac{1}{10}\sum_{i=1}^{5}\sum_{k=1}^{2} X_{(i)k} = 28.00$$

$$\hat{\pi}_{(srs)} = \frac{\hat{\lambda}_{srs} - (1-p)}{2p-1} = 0.33 \qquad \hat{\pi}_{(crss)} = \frac{\hat{\lambda}_{rss} - (1-p)}{2p-1} = 0.33$$

$$\hat{\lambda}_{r,rss} = \hat{\lambda}_{[rss]}\left(\frac{\bar{X}}{\hat{\bar{X}}_{(rss)}}\right) = 0.51 \qquad \hat{\pi}_{A(crss)} = \frac{\hat{\lambda}_{r,rss} - (1-p)}{2p-1} = 0.48$$

From (3.1) and (3.6), we have $\widehat{\mathrm{Var}}(\hat{\pi}_{(srs)}) = 0.07$ and $\widehat{\mathrm{Var}}(\hat{\pi}_{(crss)}) = 0.06$. Similarly, from (4.2) and (4.7), we have $\widehat{\mathrm{MSE}}(\hat{\pi}_{Y(srs)}) = 0.074$ and $\widehat{\mathrm{MSE}}(\hat{\pi}_{Y(crss)}) = 0.070$.

As expected, both $\hat{\pi}_{(srs)}$ and $\hat{\pi}_{(crss)}$ estimates are very close to true π. It can be observed that $\widehat{\mathrm{Var}}(\hat{\pi}_{(crss)})$ is less than $\widehat{\mathrm{Var}}(\hat{\pi}_{(srs)})$. This supports $\hat{\pi}_{(crss)}$ instead of $\hat{\pi}_{(srs)}$ for the estimation of π. Similarly, $\widehat{\mathrm{MSE}}(\hat{\pi}_{A(crss)})$ is less than $\widehat{\mathrm{MSE}}(\hat{\pi}_{Y(srs)})$ indicates that proposed ratio method for estimating sensitive attribute is better than ordinary estimator given in the Reference [22]. Moreover, we can expect further improvement in these results by taking into account multiple-concomitants situation in the present study, as advised in the Reference [14], which is in progress.

6 Conclusion

In this study, we have suggested an efficient alternative to Warner's model [1] for estimating sensitive proportion under CRSS plan. Additionally, a new estimator that based on ratio method has also been proposed using CRSS and compared with its SRS counterpart given in [22]. Both mathematical and numerical results support our proposed estimators.

In future research, it would be interesting to explore effects on the results in Bayesian framework under ranked set sampling methods.

Finally, we would like to discuss the case of generalizing the proposed ratio estimator $\hat{\lambda}_{[r,rss]}$ of λ so as to incorporate concomitant information along with its known parameters for further enhancing accuracy of the results. It is worth-mention that one can also estimate λ via exponential ratio estimator [25]. Thus, two general families of estimators for λ are presented as

$$\hat{\lambda}_{[r,crss]} = \begin{cases} \hat{\lambda}_{[rss]}\left\{\frac{a\bar{X}+b}{a\bar{x}_{(rss)}+b}\right\} & \text{Ratio family} \\ \text{and} & \\ \hat{\lambda}_{[rss]}\exp\left\{\frac{a(\bar{X}-\bar{x}_{(rss)})}{a(\bar{X}+\bar{x}_{(rss)})+2b}\right\} & \text{Exponential family} \end{cases}$$

Where $a \neq 0$ and b are known parameters of X. For specific problem, any one of them can be selected to better estimate sensitive proportion as oppose to existing [25] procedure. Hence, this study has provided different options to the experimenter for obtaining precise measure of sensitive proportion.

Supporting information

S1 Appendix.
(PDF)

S1 Fig. Probability tree diagram of *i*th response.
(TIF)

S1 File. Relative efficiencies of the proposed methods and a layout of real data set.
(PDF)

Acknowledgments

The authors are thankful to an Academic Editor and two anonymous reviewers for providing useful comments that substantially improved the previous version of the article.

Author Contributions

Supervision: Muhammad Yousaf Shad.

Writing – original draft: Azhar Mehmood Abbasi.

References

1. Warner SL (1965). Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60: 63–69. https://doi.org/10.1080/01621459.1965.10480775 PMID: 12261830
2. Horvitz DG, Shah BV, and Simmons WR. (1967). The unrelated question randomized response model. Proceedings of the Social Statistics Section, Journal of the American Statistical Association: 65–72.
3. Greenberg BG, Abul-Ela A-LA, Simmons WR, and Horvitz DG (1969). The unrelated question randomized response model: Theoretical framework. Journal of the American Statistical Association 64: 520–539. https://doi.org/10.1080/01621459.1969.10500991
4. Kuk AYC (1990). Asking sensitive questions indirectly. Biometrika 77: 436–438. https://doi.org/10.1093/biomet/77.2.436
5. Mangat NS (1994). An improved randomized response strategy. Journal of the Royal Statistical Society: Series B (Methodological) 56: 93–95.
6. McIntyre GA (1952). A method for unbiased selective sampling, using ranked sets. Australian Journal of Agricultural Research 3: 385–390. https://doi.org/10.1071/AR9520385
7. Takahasi K and Wakimoto K (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. Annals of the Institute of Statistical Mathematics 20: 1–31. https://doi.org/10.1007/BF02911622
8. Stokes SL (1977). Ranked set sampling with concomitant variables. Communications in Statistics-Theory and Methods 6: 1207–1211. https://doi.org/10.1080/03610927708827563
9. Frey J (2011). A note on ranked set sampling using a covariate. Journal of Statistical Planning and Inference. 141: 809–816. https://doi.org/10.1016/j.jspi.2010.08.002
10. Zamanzade E and Vock M (2015). Variance estimation in ranked set sampling using a concomitant variable, Statistics & Probability Letters 105: 1–5. https://doi.org/10.1016/j.spl.2015.04.034
11. Zamanzade E and Mohammadi M (2016). Some Modified Mean Estimators in Ranked Set Sampling Using a Covariate. Journal of Statistical Theory and Applications 15: 142–152. https://doi.org/10.2991/jsta.2016.15.2.4
12. Zamanzade E and Mahdizadeh M (2018). Distribution function estimation using concomitant-based ranked set sampling. Hacettepe Journal of Mathematics and Statistics 47: 755–761. https://doi.org/10.15672/HJMS.201814420708
13. Ashour SK and Abdallah MS (2019). New distribution function estimators and tests of perfect ranking in concomitant-based ranked set sampling. Communications in Statistics-Simulation and Computation. 1–26. https://doi.org/10.1080/03610918.2019.1659360
14. Terpstra JT and Liudahl LA (2004). Concomitant-based rank set sampling proportion estimates. Statistics in Medicine 23: 2061–2070. https://doi.org/10.1002/sim.1799 PMID: 15211603
15. Zamanzade E and Mahdizadeh M (2017). A more efficient proportion estimator in ranked set sampling. Statistics & Probability Letters 129: 28–33.

16. Abbasi AM and Shad MY (2021). Estimation of population proportion using concomitant-based ranked set sampling. Communications in Statistics-Theory and Methods. https://doi.org/10.1080/03610926.2021.1916529

17. Chen H, Stasny EA, and Wolfe DA (2008). Ranked set sampling for ordered categorical variables. Canadian Journal of Statistics 36: 179–191. https://doi.org/10.1002/cjs.5550360201

18. Santiago A, Sautto JM, and Bouza CN (2019). Randomized estimation a proportion using ranked set sampling and Warners procedure. Investigacion Operacional 40: 356–361.

19. David HA and Nagaraja HN (2003). Order Statistics. 3rd Edition. New York, John Wiley & Sons.

20. Dell TR and Clutter JL (1972). Ranked set sampling theory with order statistics background. Biometrics 545–555. https://doi.org/10.2307/2556166

21. Samawi HM and Muttlak HA (1996). Estimation of ratio using rank set sampling. Biometrical Journal 38: 753–764. https://doi.org/10.1002/bimj.4710380506

22. Yan Z (2006). Ratio method of estimation of population proportion using randomized response technique. Model Assisted Statistics and Applications 1: 125–130. https://doi.org/10.3233/MAS-2005-1209

23. Diana G and Perri PF (2009). Estimating a sensitive proportion through randomized response procedures based on auxiliary information. Statistical Papers 50: 661–672. https://doi.org/10.1007/s00362-007-0107-y

24. Al-Sobhi MM, Hussain Z, Al-Zahrani B (2014) General Randomized Response Techniques Using Polya's Urn Process as a Randomization Device. PLoS ONE 9(12). https://doi.org/10.1371/journal.pone.0115612 PMID: 25541936

25. Bahl S and Tuteja RK (1991). Ratio and product type exponential estimators. Journal of Information and Optimization Sciences 12: 159–164. https://doi.org/10.1080/02522667.1991.10699058