Journal of Applied Statistical Science
Volume 11, Number 1, pp. 83-90

ISSN 1067-5817

Ranked Set Sampling from a Dichotomous Population

Herber Lacayo
Center for Environmental Information & Statistics
United States Environmental Protection Agency
401 M Street, S.E.
Washington, DC 20460

Nagaraj K. Neerchal
Department of Mathematics & Statistics
University of Maryland Baltimore County
Baltimore, MD 21250

Bikas K. Sinha
Stat-Math Division
Indian Statistical Institute
Calcutta, India, 700035

ABSTRACT

We propose to discuss certain aspects of Rank Set Sampling (RSS) *vis-a-vis* Simple Random Sampling (SRS) in the context of a dichotomous population. Basu's(1971) "Elephant Example" provides an effective means of demonstrating the superiority of RSS over SRS from the point of view of representativeness of a sample. Next, we turn our attention to the problem of estimation of the proportion, p, of individuals possessing a certain attribute in a population. It is well known that the sample proportion from an RSS has a smaller variance than the sample proportion from an SRS of the same size. However, it turns out that nonnegative unbiased estimate of the variance of sample proportion from an RSS is generally unavailable. A natural, albeit biased, estimator is proposed to overcome the problem.

Key Words and Phrases: Binomial distribution; Censoring; Detection limit; Environmental sampling; Simple random sampling.

1 Introduction

Ranked Set Sampling (RSS), like the better-known Simple Random Sampling (SRS), is useful in environmental applications. The purpose of this paper is to highlight its key features in the context of sampling from a dichotomous population. Below, we use Basu's (1971) "Elephant Example" to discuss our ideas.

Consider a herd of elephants consisting of Mothers and Babies in proportions $q(= 1-p)$ and p respectively. Let us examine the following two methods of estimating the average weight for the entire herd.

Simple Random Sampling (Case of $n = 2$): Draw a sample of elephants at random and use their average weight as an estimate of the average weight for the entire herd. Suppose the two elephants picked at random were both mother elephants (or both calves); it's folly to proceed to estimate the average weight of the herd from such a highly non-representative sample. Occurrence of such non-representative samples is more frequent than one would normally desire. For the case of $n = 2$, chance of a non-representative sample is $100(p^2+q^2)\%$ and the sample average would vary widely since $100p^2\%$ of the time it will be very high (the mother-mother case) and $100q^2\%$ of the time it will be very low (the calf-calf case).

Ranked Set Sampling (Case of $n = 2$): Sample two elephants randomly in the morning. Pick the smaller of the two and weigh it. Sample two elephants randomly in the afternoon. Pick the larger of the two and weigh it. Use the average weight of the elephants chosen in this way as an estimate. It is known that this estimate is again an unbiased estimate of the average weight for the herd. It can be shown that the probability of arriving at a non-representative sample under the RSS scheme is

$$100[q^2(1-p^2)+(1-q^2)p^2]\% = 100[p^2+q^2-2p^2q^2]\%, \qquad (1.1)$$

which is smaller than the corresponding probability under SRS ($n = 2$) given by $100[p^2 + q^2]\%$. A formula, applicable for an arbitrary sample size is given later in equation (11) of section 2. Table 1 in the Appendix gives these probabilities for $n = 2, 3, 4$, and 5, and for $p = 0.1, 0.2, 0.3, 0.4$, and 0.5.

The above example captures the salient feature of RSS, namely producing less variable and more representative samples. This is the reason why the sample mean from an RSS has a smaller variance than the sample mean from an SRS of the same size. We refer the readers to Patil, Sinha and Taillie (1994) and the references therein for a number of theoretical results comparing RSS to SRS for various standard estimation problems. Another useful reference is Ni Chuiv and Sinha (1998).

In this paper we focus on two issues. In section 2, we discuss the superiority of RSS over SRS in obtaining more representative samples. In section 3, some disadvantages of RSS are brought forth in the context of making inference regarding a Bernoulli parameter based on an RSS.

2 Inequalities Relating SRS and RSS

In this section, we first establish an inequality involving the probability of extreme values under RSS and SRS. Second, we derive the probability function of the number of calves in a ranked set sample of a given size. We give an algorithm to compute the probability function. Third, we point out a natural connection between this probability structure and censoring (see Remark 2, below). Finally, we conclude the section by giving a general result involving tail probabilities under the two sampling schemes.

A ranked set sample of size n is obtained by first picking n independent simple random samples, each of size n, and then visually identifying the i-th largest in size, $E_{(ii)}$, from the i-th simple random sample:

$$(E_{i1}, E_{i2}, \cdots, E_{in}), \quad 1 \leq i \leq n.$$

Within each SRS the ranking is done from the largest in size to the smallest. Thus, the ranked set sample consists of $(E_{(11)}, E_{(22)}, \cdots E_{(nn)})$, where $E_{(ii)}$ denotes the i-th largest unit in the i-th SRS. Subsequent analysis will be based on the measurements made on these units.

We now compare the probability of obtaining an extreme sample under RSS with the corresponding probability under SRS. In the case of a herd of elephants consisting of only two sizes, namely, mothers and calves, let p denote the proportion of calves in the herd. Let X denote the number of calves in a sample of size n, drawn either by SRS or RSS.

Under the SRS scheme, X is Binomial with parameters n and p. Thus,

$$P_{SRS}(X=x)=\binom{n}{x}p^x q^{n-x}, \qquad x=0,1,\ldots,n. \tag{2.1}$$

In particular, the probability of obtaining a sample containing only calves is

$$P_{SRS}(X=n)=p^n. \tag{2.2}$$

On the other hand, under the RSS scheme [see Neerchal and Lacayo (1998)],

$$P_{RSS}(X=n)=\prod_{i=1}^{n}\sum_{l=n-i+1}^{n}\binom{n}{l}p^l q^{n-l}. \tag{2.3}$$

Note that,

$$P_{RSS}(X=n) \;<\; \sum_{l=n}^{n}\binom{n}{l}p^l q^{n-l}=p^n=P_{SRS}(X=n).$$

Similarly, for the probability of obtaining a sample consisting of only the adults,

$$\begin{aligned} P_{RSS}(X=0) &= \prod_{i=1}^{n}\left\{\sum_{l=0}^{n-i}\binom{n}{l}p^l q^{n-l}\right\} \\ &< \sum_{l=0}^{0}\binom{n}{l}p^l q^{n-l}=q^n, \\ \text{or} \quad P_{RSS}(X=0) &< P_{SRS}(X=0). \end{aligned}$$

The above observation that RSS has a lower probability of producing the extreme observations is also true for sampling from an arbitrary continuous population as shown below.

Let $F_{(r)}(y)$ denote the cumulative distribution function (CDF) of the r^{th} order statistic in a simple random sample of size n from a population given by the cumulative distribution function $F(y)$. As noted by Dell and Clutter (1972) [see, equation(1), p. 547], we have

$$F(y)=\frac{1}{n}\sum_{r=1}^{n}F_{(r)}(y).$$

Let $Y_{(n)}$ denote the maximum from an SRS of size n. Then

$$\begin{aligned} P\left\{Y_{(n)}\le y\right\} &= \{F(y)\}^n=\left\{\frac{1}{n}\sum_{r=1}^{n}F_{(r)}(y)\right\}^n \\ &\ge \prod_{i=1}^{n}F_{(r)}(y). \end{aligned}$$

Therefore, if we let $Y_{((n))}$ denote the maximum from a ranked set sample of size n, then

$$P\left\{Y_{((n))}\le y\right\}=\prod_{i=1}^{n}F_{(r)}(y)\le P\left\{Y_{(n)}\le y\right\}. \tag{2.6}$$

The inequality

$$P\left\{Y_{(1)} \le y\right\} \le P\left\{Y_{((1))} \le y\right\} \tag{2.7}$$

involving the minimums, follows by a similar argument. We refer the reader to Theorem 2.8 of David (1981) for a general result of this type for the r^{th} order statistic.

Now, we derive a general expression for the probability function of the number of calves in a ranked set sample of size n. Let $X_{(i)}$ be the indicator variable defined as follows:

$$\begin{array}{rcll} X_{(i)} & = & 1 & \text{if the i-th largest in the i-th SRS is a calf} \\ & = & 0 & \text{otherwise} \end{array}$$

Since, each $X_{(i)}$ is based on an independent SRS, we have that

$$P_{RSS}\left\{X_{(1)} = x_{(1)}, \cdots, X_{(n)} = x_{(n)}\right\} = \prod_{i=1}^{n} P_{RSS}\left\{X_{(i)} = x_{(i)}\right\}$$

Now, for each i, it can be shown that

$$P_{RSS}\left\{X_{(i)} = 1\right\} = \sum_{l=n-i+1}^{n} \binom{n}{l} p^{l}(1-p)^{n-l} = \pi_i.$$

Therefore,

$$P_{RSS}\left\{X_{(i)} = x_{(i)}\right\} = \pi_i^{x_{(i)}}(1-\pi_i)^{1-x_{(i)}}, \quad x_{(i)} = 0 \text{ or } 1. \tag{2.9}$$

Thus, by summing over all possible combinations of $x_{(i)}$'s such that there are exactly k ones and $n-k$ zeros, we get

$$P_{RSS}(X = k) = \sum_{\mathbf{x}\in S_k} \prod_{i=1}^{n} \pi_i^{x_{(i)}}(1-\pi_i)^{1-x_{(i)}}, \tag{2.10}$$

where $S_k = \{(x_{(1)}, x_{(2)}, \ldots x_{(n)}) : x_{(1)} + x_{(2)} + \ldots + x_{(n)} = k\}$. Note that the number of terms in the above summation grows rather rapidly as the sample size increases. However, RSS is not recommended when n is large because of the practical difficulties of ranking a large number of sampling units (elephants in the above example).

There is an easily implemented algorithm for computing $P_{RSS}(X = k)$ for a given n, k and p. We start by computing

$$P_{RSS}(X = 0) = \prod_{i=1}^{n}(1-\pi_i),$$

$$\text{and } P_{RSS}(X = n) = \prod_{i=1}^{n} \pi_i.$$

For $0 < k < n$, use the recursion

$$P_{RSS}(X = k) =$$
$$P\left\{\textstyle\sum_{i=1}^{n-1} X_{(i)} = k-1\right\} P\left\{X_{(n)} = 1\right\} + P\left\{\textstyle\sum_{i=1}^{n-1} X_{(i)} = k\right\} P\left\{(X_{(n)} = 0\right\}.$$

Note that π_i is a binomial tail probability and is readily available. Tables of probabilities $P_{RSS}(X = k)$ can be constructed using the above recursion algorithm. Readers can obtain from the authors an Splus code which implements this algorithm. To give a flavor of the results, computations are reported for a few combinations of n and p in the Appendix (p=0.1,0.2,0.3,0.4 and 0.5, and n=2,3, and 4). Upon examining the results in Table 1 of the Appendix, we make the following remarks.

Remark 1. The calculated probabilities of extreme values are smaller under RSS than in SRS. For every n and p there seems to exist an interval (l, u) such that for $l \leq k \leq u$

$$P_{RSS}(X = k) > P_{SRS}(X = k)$$

and for k outside the interval (l, u),

$$P_{RSS}(X = k) < P_{SRS}(X = k)$$

Remark 2. In analyzing environmental data one often has to deal with measurements that are below detection limits (BDL). Suppose M denotes the detection limit and let p denote the probability that a randomly chosen sample leads to a measurement below the detection limit. Let X denote the number of BDL observations in a sample of size n. Then, under the SRS scheme, X has a Binomial distribution and under the RSS scheme, distribution of X is given by equation (11). It seems evident from Table 1 in the Appendix and extensive calculations not reported here that extreme values of X outside an interval (l, u) as in Remark 1, have lower probabilities under RSS.

We conclude this section by giving a general result relating the tail probabilities under the two sampling schemes. By applying a result of Hoeffding (1956) [also see Marshall and Olkin (1979, page 375)], one can obtain the following inequality involving the tail probabilities under the two sampling schemes.

$$\begin{array}{rcllll} P_{RSS}(X \leq k) & \leq & P_{SRS}(X \leq k) & \quad \text{if} & 0 \leq k \leq np - 1, \\ P_{RSS}(X \leq k) & \geq & P_{SRS}(X \leq k) & \quad \text{if} & np \leq k \leq n. \end{array}$$

3 Inference for the Bernoulli Parameter

In the previous section, we discussed sampling from a dichotomous population in order to demonstrate that under RSS, there is a higher probability of obtaining more representative samples than under SRS. We also note that the discussion leads to an interesting analogue of the binomial probability mass function where the successive trials are independent but the associated success probabilities are not identical. It is natural to investigate the usual inference problems such as estimation of p and its higher powers under the RSS scheme, and compare with the well known results under SRS. In this section we present these results mainly to point out that some of these results are somewhat different from those obtained in the case of sampling from a continuous population. This sheds further light on the nature of RSS.

Suppose the characteristic of interest is the proportion (p) of study area plots where the plutonium concentration (PC) exceeds a pre-determined level (PDL) based on health concerns. PC are measured using laboratory analyses and are expensive to obtain. However, prevalence of americium, can be measured on the field by hand-held instruments. In this example, the Americium Concentrations (AC) can be used to rank the sampling units (study area plots) and soil samples from the appropriate subset may be sent to the laboratory for determining the PC. Let $E_1, E_2, \ldots, E_n$ constitute an SRS of size n and let $X_i = 1$ if PC exceeds PDL in field sample E_i and $X_i = 0$, otherwise. Let T_n denote the sample sum. Then the "natural" estimator of p is given by:

$$\hat{p} = T_n/n = \bar{T}_n \text{ say.} \tag{3.1}$$

On the other hand, for applicability of the RSS of size n, we sample n sets of n study area plots $\{(E_{i1}, E_{i2}, \ldots, E_{in})\}$ and we order these study area plots in ascending order of their AC values measured by the hand held instrument. From the i-th SRS, we identify the i-th largest among the ordered plots and measure its PC value by a laboratory analysis. Then we let $X_{i:i} = 1$ if the PC value exceeds PDL in this study area plot and $X_{i:i} = 0$ otherwise. Then $\{X_{1:1}, X_{2:2}, \ldots, X_{n:n}\}$ is the RSS of size n and an unbiased estimator of p is given by:

$$\tilde{p} = \sum X_{i:i}/n = \tilde{T}_n/n, \text{ say}$$

With an SRS data, we are able to provide MVUE's of all powers of p up to and including p^n and these are based on factorial moments of T_n. Likewise, based on RSS data, we can easily provide *unique* unbiased estimators of all powers of p up to and including p^n. These are just linear in the observations!! Specifically, using factorial notations, we deduce that

$$E(T_n^{(k)}/n^{(k)}) = p^k = E[\sum\{i^{(k)} - (i-1)^{(k)}\}X_{i:i}]/n^{(k)}.$$

The second part in the above equation along with *completeness property* of the Family of Binomial Distributions ensures that the linear combination of $X_{i:i}$'s is unique for every $k = 1, 2, \ldots, n$. Further, it turns out that for $k = 1, 2, \ldots, n$

$$Var(\tilde{p}^k)|RSS(n) = \sum[\{(i-1)^{(k)} - i^{(k)}\}^2 F_{i:n}\bar{F}_{i:n}]/\{n^{(k)}\}^2, \tag{3.2}$$

where $F_{j:n} = P[T_n \leq j]$ and $\bar{F}_{j:n} = 1 - F_{j:n}, j = 0, 1, \ldots, n-1$. Also, the above expression is symmetric in p about 0.5. It would be interesting to check if, for all k, it is concave in p as well. Towards this, Li(1997) did some numerical computations for the case of $k = 1$ and the results are quite encouraging. An analytical proof has so far eluded us.

We now address the problem of variance comparison between SRS and RSS for unbiased estimation of $p^k, k = 1, 2, 3, \cdots$. It is well known that for all $n \geq 1$, uniformly in p,

$$Var_{RSS}(\tilde{p}) \leq Var_{SRS}(\hat{p})$$

with strict inequality for all $n > 2$. For $k = 2$ as well, we have verified that for all $n \geq 2$ and for all p,

$$Var_{RSS}(\tilde{p}^{(2)}) \leq Var_{SRS}(\hat{p}^2).$$

with strict inequality for $n > 2$. However, it is not *true* uniformly in p for higher values of k.

We conclude the article with a remark on the non-negativity of the estimator based on RSS. With an SRS sample and $n \geq 2$, it is well-known that

$$\hat{V}(\hat{p}) = X(n-X)/n^2(n-1).$$

Thus the variance estimate based on SRS data is trivially non-negative. For the RSS data, however, we observe that

$$\begin{aligned} E(\tilde{T}_n^2) &= E[(\sum X_{i:i})^2] = \sum F_{i-1:n} + \sum\sum F_{i-1:n}F_{j-1:n} \\ &= \sum F_{i-1:n}\bar{F}_{i-1:n} + (\sum F_{i-1:n})^2 = \sum F_{i-1:n}\bar{F}_{i-1:n} + (np)^2 \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{V}(\tilde{p}) &= \hat{V}[\tilde{T}_n/n] = [T_n^2 - n^2(\tilde{p}^2)]/n^2 \\ &= [T_n^2 - 2n\{\sum(i-1)X_{i:i}/(n-1)\}]/n^2 \end{aligned}$$

It now turns out that this variance estimate is *rarely* non-negative!! For example, when $T_n = k$, let us set

$$X_{2:2} = X_{3:3} = \ldots = X_{k+1:k+1} = 1 \ \& \ \text{all other } X_{i:i} = 0.$$

Then the expression in (11) is non-negative iff

$$\begin{aligned} & k^2 \geq 2n[k(k+1)/2]/(n-1) \\ \text{iff} \quad & (n-1)k \geq n(k+1) \text{ which is a contradiction!!} \end{aligned}$$

As a matter of fact, only when

$$X_{1:1} = X_{2:2} = X_{3:3} = \ldots = X_{k:k} = 1$$

we have non-negativity. To overcome this difficulty, one can use $\tilde{p}$ in place of p in the expression for $V(\tilde{p})$. Because of concavity of the variance as a function of p, it turns out that the variance estimator thus formed is positively biased and trivially positive-valued.

Acknowledgements

This work was partially supported by a cooperative agreement between the United States Environmental Protection Agency and the Department of Mathematics and Statistics, University of Maryland Baltimore County. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the above Agency.

References

[1] Basu, D., 1971, "An essay on the logical foundations of survey sampling. I. With comments by G. A. Barnard, V. P. Godambe, J. Hjek, J. C. Koop and R. Royall and a reply by the author", *Foundations of statistical inference (Proc. Sympos., Univ. Waterloo, Waterloo, Ont., 1970), Holt, Rinehart and Winston of Canada, Toronto, Ontario, Canada,* pp. 203–242.

[2] David, H. A., 1981, *Order Statistics,* John Wiley, New York.

[3] Dell, T. R. and Clutter, J. L., 1972, "Ranked set sampling theory with order statistics background", *Biometrics,* Vol. 23, pp. 545–555.

[4] Hoeffding, W., 1956, "On the distribution of the number of successes in independent trials", *Ann. Math. Stat.,* Vol. 27, pp. 713–721.

[5] Li, D., 1997, On Certain Aspects of Ranked Set Sampling, *Ph.D. thesis,* Department of Mathematics and Statistics University of Maryland Baltimore County, Baltimore, MD 21250.

[6] Marshall, A. W. and Olkin, 1979, *Inequalities: Theory of Majorization and its Applications,* Academic Press, New York.

[7] Neerchal, N. K. and Lacayo, H., 1998, Ranked Set Sampling of Basu's Elephants, *Technical Report,* Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250.

[8] Ni Chuiv, N. and Sinha, B. K., 1998, "On Some Aspects of Ranked Set Sampling in Parametric Estimation", *Order Statistics: Applications,* N. Balakrishnan and C. R. Rao, Eds., North-Holland, Amsterdam, pp. 337–377.

[9] Patil, G. P., Sinha, A. K. and Taillie, C., 1994, "Ranked Set Sampling" *Handbook of Statistics, Volume 12: Environmental Statistics,* G. P. Patil and C. R. Rao, Eds., North-Holland, Amsterdam, pp. 167–200.

Appendix

Table 1. Probability distribution of number of calves
(Italics: RSS, Roman: SRS)

		Proportion of calves in the herd				
Sample size	k	0.1	0.2	0.3	0.4	0.5
2	0	0.8100	0.6400	0.4900	0.3600	0.2500
		0.8019	*0.6144*	*0.4459*	*0.3024*	*0.1875*
	2	0.0100	0.0400	0.0900	0.1600	0.2500
		0.0019	*0.0144*	*0.0459*	*0.1024*	*0.1875*
3	0	0.7290	0.5120	0.3430	0.2160	0.1250
		0.7079	*0.4551*	*0.2617*	*0.1310*	*0.0547*
	3	0.0010	0.0080	0.0270	0.0640	0.1250
		0.0000	*0.0004*	*0.0038*	*0.0179*	*0.0547*
4	0	0.6561	0.4096	0.2401	0.1296	0.0625
		0.6194	*0.3259*	*0.1422*	*0.0493*	*0.0126*
	4	0.0001	0.0016	0.0081	0.0256	0.0625
		0.0000	*0.0000*	*0.0002*	*0.0021*	*0.0126*
5	0	0.5905	0.3277	0.1681	0.0778	0.0313
		0.5375	*0.2260*	*0.0718*	*0.0162*	*0.0023*
	5	0.0001	0.0003	0.0024	0.0102	0.0313
		0.0000	*0.0000*	*0.0000*	*0.0002*	*0.0023*