Thailand Statistician
April 2022; 20(2): 233-239
http://statassoc.or.th
Contributed paper

Some Inferential Aspects of Mixture Sampling Designs

Srijib Bhusan Bagchi [a] and Bikas K. Sinha*[b]
[a] Burdwan University, West Bengal, India.
[b] Indian Statistical Institute, Kolkata, India.
*Corresponding author; e-mail: bikassinha1946@gmail.com

Received: 6 April 2020
Revised: 2 June 2020
Accepted: 6 June 2020

Abstract

Starting with two well-known sampling designs viz., SRSWR and SRSWOR, in the usual manner, we introduce mixture designs with desirable properties. We review the literature and develop further results from two perspectives: Structural and Inferential. Permutation invariance of the sampling units under both the sampling designs plays a significant role in this study.

Keywords: SRSWR design, SRSWOR design, mixture designs, coherent mixtures, structural aspects, inferential aspects, permutation invariance.

1. Introduction

We start with a finite labelled population of size N and a fixed sample size of $n < N$. We are all aware of two well-known sampling designs/schemes viz., $SRSWR(N, n)$ and $SRSWOR(N, n)$. For each $k = 1, 2, \ldots, n$, let $D_{k;N}$ denote $SRSWOR(N, k)$ design. Let $Q_{k;n;N}$= $N^{-n}\boldsymbol{\Delta}^k 0^n N_{ck}$ where $\boldsymbol{\Delta}^k 0^n = \boldsymbol{\Delta}^k x^n$, evaluated at $x = 0$, $\boldsymbol{\Delta}$ being the 'difference operator'. Further, N_{c_k} is the usual co-efficient of x^k in the binomial expansion of $(1 + x)^N$ where both N and $k \leq N$ are positive integers. It is known [Vide Basu (1958)] that $\sum_{k=1}^{k=n} Q_{k;n;N} = 1$. Moreover, it is known that a mixture of the family of $SRSWOR(N, k)$ Designs with mixing proportions $Q_{k;n;N}$ is identified as $SRSWR(N, n)$ Design. These two classical sampling designs have been studied in great detail - both from structural perspectives and inferential perspectives. We refer to Sinha and Sen (1989) for details and many related results.

Our purpose is to introduce some such mixtures and study their structural and inferential aspects.

2. Special Mixtures

For a fixed $K < n$, define $D_{k;K;n;N}$ = $\sum_{k=1}^{k=K} Q_{k;n;N} SRSWOR(N; k) / \sum_{k=1}^{k=K} Q_{k;n;N}$. Naturally, this sampling design is based on the first K components of the SRSWR mixtures with normed mixing proportions in the truncated form. For $K = n$, this mixture identifies itself as $SRSWR(N, n)$ design. Vide Basu (1958), Korwar and Serfling (1974) and Hedayat and Sinha (1991).

A general mixture $D_{w_k;K;n;N}$= $\sum_{k=1}^{k=K} w_k SRSWOR(N; k) / \sum_{k=1}^{k=K} w_k$ is defined as usual with the non-negative weight functions $w_k \geq 0$, subject to $\sum_k w_k > 0$.
Note that under $SRSWOR(N; k)$, we provide the sample mean as an unbiased estimator of the

population mean with variance $S^2[1/k-1/N]$. This suggests the well-known fact: Under $SRSWR$ $(N;n)$ sampling scheme, the sample mean based on distinct units in a sample is unbiased for the population mean with variance $S^2[E(1/k)-1/N]$. Further to this, $E(1/k)$ has an analytical expression given by $[1^{n-1}+2^{n-1}+\ldots+(N-1)^{n-1}]/N^n$. Vide Hedayat and Sinha (1991). For higher order moments of $1/k$, we refer to Sinha and Sen (1989).

3. Coherent Mixtures

An $SRSWR(N,n)$ sampling design is easy to implement. This has been a strong practical point in its favor. However, it fails to provide any merit towards the estimation of a finite population mean. We denote by ν the number of distinct units in a sample underlying $SRSWR(N,n)$. It is known that $E(\nu)=N[1-(1-1/N)^n]=\nu_0(1-f)+(\nu_0+1)f$ where ν_0 is the largest integer contained in $E(\nu)$. This suggests that a 2-point mixture of $SRSWOR(N,\nu_0)$ and $SRSWOR(N,\nu_0+1)$ designs with mixing proportions $1-f$ and f would amount to the same expected sample size as $SRSWR(N,n)$. For this 2-point design, the variance of the mean of units so drawn would have the expression [ignoring the last part involving $1/N$]

$$(1-f)[1/\nu_0]+f[1/(\nu_0+1)].$$

Comparing this expression with $[1^{n-1}+2^{n-1}+\ldots+(N-1)^{n-1}]/N^n$ makes sense and we can possibly recommend a 2-point mixture design as against the $SRSWR(N,n)$ design. Here are some examples to this effect.

Example 1

(a) Consider $N=10,n=4$. It follows that $E(\nu)=3.44$, approx. This yields $\nu_0=3,f=0.44$ and further, $E(1/\nu)=0.3025$ while $E(1/k)=0.2967$ for the 2-point mixture.

(b) Consider $N=15,n=5$. It follows that $E(\nu)=4.376$, approx. This yields $\nu_0=4,f=0.376$ and further, $E(1/\nu)=0.2355$ while $E(1/k)=0.2312$ for the 2-point mixture.

(c) Consider $N=20,n=5$. It follows that $E(\nu)=4.5244$. This yields $\nu_0=4,f=0.5244$. Further, $E(1/\nu)=0.2262$ while $E(1/k)=0.2238$, for the 2-point mixture.

4. Other Comparable Mixtures

Under $SRSWR(N,n)$ design, units are likely to appear any number of times up until n. Let us now impose a pre-condition on the maximum number of appearances of the units in the sample. To start with, we choose a positive integer $K<n$ and demand that no unit should appear more than K times in $SRSWR(N,n)$ design. With this restriction, we denote the underlying sampling design by the notation $SRSWR(N,n;K)$. It is interesting to note that the 'invariance principle holds for all permutations of the units in the population and accordingly, for every possible value of the number of distinct units in a sample [under this restriction] the sample mean based on the distinct units is an unbiased estimator for the population mean and the conditional variance is given by $S^2[1/\nu-1/N]$. Therefore, the over-all variance of the estimate is obtained by averaging over all possible samples the reciprocal of ν. This part is a non-trivial task. Even in such a situation, we can work out a mixture of two suitably chosen SRSWOR designs - matching the expected sample size and possessing smaller average variance. We take up an example below.

Example 2

(a) Consider $N=10,n=5,K=3$. We have evaluated the distribution of ν, the number of distinct units in a sample. The total number of admissible samples of size 5 is given by 99, 540. And the decompositions are :

$$N[\nu=2]=900,N[\nu=3]=18000,N[\nu=4]=50,400,N[\nu=5]=30240.$$

It follows that $E(\nu) = 4.10488, E(1/\nu) = 0.25214$. Using a 2-point mixture distribution with $\nu_0 = 4$ and $\nu_0 + 1 = 5$, with mixing proportions $(1 - 0.10488)$ and 0.10488 respectively, we deduce that $E(\nu)$ remains the same while$E(1/\nu) = 0.24475 < 0.25214$.

(b) Consider $N = 15, n = 5, K = 3$. We have evaluated the distribution of ν, the number of distinct units in a sample. The total number of admissible samples of size 5 is given by 758310, excluding 1065 inadmissible samples. And the decompositions are :

$$N[\nu = 2] = 2100, N[\nu = 3] = 68250, N[\nu = 4] = 327600, N[\nu = 5] = 360360.$$

It follows that $E(\nu) = 4.38, E(1/\nu) = 0.2344$. Using a 2-point mixture distribution with $\nu_0 = 4$ and $\nu_0 + 1 = 5$, with mixing proportions $(1 - 0.38)$ and 0.38 respectively, we deduce that $E(\nu)$ remains the same while$E(1/\nu) = 0.231 < 0.2344$.

(c) Consider $N = 20, n = 5, K = 3$. As in the above, we have evaluated the distribution of ν, the number of distinct units in a sample. The total number of admissible samples of size 5 is given by 3198080. And the decompositions are :

$$N[\nu = 2] = 3800, N[\nu = 3] = 171000, N[\nu = 4] = 1162800, N[\nu = 5] = 1860480.$$

It follows that $E(\nu) = 4.5259, E(1/\nu) = 0.2257$. Using a 2-point mixture distribution with $\nu_0 = 4$ and $\nu_0 + 1 = 5$, with mixing proportions $(1 - 0.5259)$ and 0.5259 respectively, we deduce that $E(\nu)$ remains the same while$E(1/\nu) = 0.2237 < 0.2257$.

5. Combinatorial Properties of *SRSWR(N, n; K)* Design

We start with the total count $A(N, n; K)$ of samples underlying this sampling design.

Theorem 1 $A(N, n; K) = N \sum_{t=0}^{(K-1)} (n-1)_{c_t} A(N-1, n-t-1, K)$.

Proof: Consider a particular situation where unit number say, 1 is drawn first. In the next $(n-1)$ draws, unit number 1 can appear additional $t(0 \leq t \leq K-1)$ times and in the remaining $(n-t-1)$ draws, $(N-1)$ remaining units (other than the unit number 1) can appear at most K times each. This number is $A(N-1, n-t-1; K)$ for $t = 0, 1, \ldots, K-1$. Since this counting argument is invariant with respect to the label of the unit drawn at the first attempt, we obtain the above identity.

We may deduce results for some special cases.
Case 1. $A(N, K; K) = N^K$,
Case 2. $A(N, K+1; K) = N^{(K+1)} - N; K = 1, 2, \ldots,$
Case 3. $A(N, K+2; K) = NA(N, K+1; K) - N(N-1)(K+1)$.

Case 1 follows by definition. Case 2 is easy to establish. We prove Case 3 below. By 'inclusion-exclusion principle', we deduce

$$\begin{aligned} A(N, K+2; K) &= N[A(N-1, K+1; K) + \sum_{t=1}^{K-1} (K+1)_{ct} A(N-1, K+1-t, K)] \\ &= N[A(N-1, K+1; K) + N^{K+1} - (N-1)^{K+1} - (K+1)(N-1) - 1] \\ &= N[N^{K+1} - (K+2)(N-1) - 1] \\ &= NA(N, K+1; K) - N(N-1)(K+1)]. \end{aligned}$$

The above provides an explicit expression for $A(N, K+2; K)$ as also a recurrence relation between $A(N, K+2; K)$ and $A(N, K+1; K)$.

6. Inclusion Probabilities for an *SRSWR(N, n; K)* Design

Using the notation $A(N, n; K)$, we deduce that, for each unit i,

$$\pi_i = 1 - \frac{A(N-1, n; K)}{A(N, n; K)}.$$

Furher, for any two distinct units i, j,

$$\pi_{i,j} = \frac{A(N, n; K) - 2A(N-1, n : K) + A(N-2, n; K)}{A(N, n; K)}.$$

For $K = 2, 3$, it is easy to write down expressions for π_is and $\pi_{i,j}$s.

7. Number of Samples on Distinct units for an *SRSWR(N, n; K)* Design

Let $A(N, n; K; \nu)$ denote the total number of samples of size n under an $SRSWR(N, n; K)$ design having ν distinct units each. Here is a recurrence relation.

Theorem 2 $A(N, n; K; \nu) = \sum_{t=1}^{K} n_{c_t} A(N-1, n-t; K; \nu-1)$.

Proof: Consider any particular unit say i, which is drawn $t(\leq K)$ times. Therefore, in the remaining $(n-t)$ draws, $\nu - 1$ distinct units will be drawn out of $N-1$ units. Hence the relation follows.

We also have an alternative representation for $A(N, n; K; \nu)$.

$$A(N, n; K; \nu) = \nu \sum_{t=0}^{(K-1)} (n-1)_{c_t} A(N-1, n-t-1; K; \nu-1) \tag{1}$$

Consider any drawn distinct unit, say i. This unit may appear further in the sample t more times in the remaining $(n-1)$ trials $(t = 0, 1, 2, , K-1)$. This number is $(n-1)_{c_t} A(N-1, n-t-1; K; \nu-1)$. Hence the representation holds.

By replacing N and n respectively by $N-1$ and $n-1$ in the above, it also follows that

$$A(N-1, n-1; K; \nu) = \nu \sum_{t=0}^{(K-1)} (n-2)_{c_t} A(N-1, n-t-2; K; \nu-1) \tag{2}$$

Further to this, we also have a relation :

$$A(N, n; K) = \sum_{\nu} N_{c_\nu} A(N, n; K; \nu) \tag{3}$$

We now go back to $A(N, n; K; \nu)$ and simplify the RHS further.

$$\nu \sum_{t=0}^{(K-1)} (n-1)_{c_t} A(N-1, n-t-1; K; \nu-1)$$

$$= \nu A(N-1, n-1; K; \nu-1) + \nu[\sum_{t=1}^{(K-1)} (n-1)_{c_t} A(N-1, n-t-1; K; \nu-1)].$$

8. Finite Population Mean Estimation based on $SRSWR(N, n; K)$ Design

Once a random sample of size n has been drawn according to this design, we are aware of the restrictions in the nature and composition of the sample of size n so drawn. However, permutation invariance of the distinct units drawn prevails and we can argue as in Sinha and Sen (1989) for every number ν of distinct units. This justifies use of the mean of distinct units as an unbiased

estimator of the population mean with variance $S^2[E(1/\nu) - 1/N]$ where $E[1/\nu]$ is to be computed wrt $SRSWR(N, n; K)$ design. Also, permutation invariance holds for the entire set of samples of size n each, even with the restriction imposed through the parameter K. This means that the usual mean based on all units is also unbiased for the population mean. Further to this, it follows that the mean based on distinct units fares better than the mean based on all units. It must be noted that not all N^n so-called 'unrestricted' SRSWR samples of size n are under the purview of the $SRSWR(N, n; K)$ design. Accordingly, even though the mean based on all units [including repeats] is unbiased for the population mean, its variance computation is not straightforward. We examine this feature in a specific example below.

Before we proceed further, we note that for $K = n$, we have $SRSWR(N, n)$ sampling and it is well-known that the variance of the sample mean is σ^2/n. However, for $K < n$, this is far from being true. It would be interesting to work out an expression for the variance of the sample mean based on all units. Below we make an attempt to work it out when$N = 10, n = 5$ and $K = 3$.

We have already analyzed the sample compositions for the total of $M = 99540$ samples - each of size 5 - under $SRSWR(10, 5; 3)$ design.

We start with the sample mean

$$\bar{y}(s) = 1/5[\sum f_i(s)y_i]$$

where $f_i(s)$ = frequency count for y_i in the sample. It is known that $0 \leq f_i(s) \leq K = 3$ for each combination of (i, s).

We define $y_i^* = y_i - \bar{Y}, \bar{Y}$ being the population mean.

We will evaluate an explicit expression for $\bar{y}^*(s) = 1/5[\sum f_i(s)y_i^*]$ by looking into each value of ν. We will work out the coefficient of y_1^* and generalize it to the rest by invoking the invariance principle.

Different values of ν are listed below.

Case (i) : $\nu = 2, M(\nu) = 900, K = 3$

Sample frequencies are of the type : $(3, 2, 0, 0, 0, ...)$. There are two positive frequencies viz., 3 and 2.

Coefficient of y_1^* will be 3 with frequency 90 and 2 also with frequency 90.

Case (ii) : $\nu = 3, M(\nu) = 18000, K = 3$

Sample frequencies are of the type : $(a)(3, 1, 1, 0, 0, ...)$ OR $(b)(2, 2, 1, 0, 0, ...)$

For Type (a) : Coefficient of y_1^* will be 3 or 1 and for Type (b) : Coefficient of y_1^* will be 2 or 1.

The frequency counts are as follows :

Type (a): 3 with frequency 720 and 1 with frequency 1440

Type (b) : 2 with frequency 2160 and 1 with frequency 1080

Case (iii): $\nu = 4, M(\nu) = 50400, K = 3$

Sample frequencies are of the type : $(2, 1, 1, 1, 0, ...)$

Coefficient of y_1^* will be 2 or 1.

The frequency counts are as follows :

2 with frequency 5040 and 1 with frequency 15120

Case (iv) : $\nu = 5, M(\nu) = 50400, K = 3$

Sample frequencies are of the type : $(1, 1, 1, 1, 1, 0, ...)$

Coefficient of y_1^* will be 1.

The frequency count is as follows : 1 with frequency 15120.

Combining all the $M = 99540$ samples across all possible values of ν, we now find coefficient of y_1^* in the expression for $\sum_s [f_1(s)y_1^*]/5M$ as

$$[1/5 \times 99540][(270 + 180) + (3 \times 720 + 1440) + (2 \times 2160 + 1080) + (2 \times 5040 + 15120) + (15120)]$$

$$= 49770/(5 \times 99540) = 1/10.$$

This is true for all the $N = 10$ units in the population because of permutation invariance. Hence unbiasedness of the sample mean of all the 5 units is verified.

We will now proceed to compute an expression for the variance of the sample mean based on all units, with repeats. This will be more involved.

We have $M = 99540$ samples with decompositions according to different values of ν. Further, we already know from the above, contribution to coefficient of y_1^* from each type of sample compositions.

We will now derive expressions for (i) the coefficient of y_1^{*2} and (ii) the coefficient of the product $y_1^* y_2^*$ by looking at generic expressions like

$\sum_s[(ay_1^* + by_2^* + \dots)^2 + \dots]$ for each type of composition of the samples for various values of ν. We collect the terms denoted below by T_2, T_3, T_4 and T_5.

Case (i) $\nu = 2$: Sample frequency type $(3, 2)$

$$T_2 = 10(3y_1^* + 2y_2^*)^2 + 10(2y_1^* + 3y_2^*)^2$$

Case (ii) $\nu = 3$: Sample frequency types (a) $(2, 2, 1)$ and (b) $(3, 1, 1)$

$$T_3(a) = [144(2y_1^* + 2y_2^*)^2 + 144(2y_1^* + y_2^*)^2 + 144(y_1^* + 2y_2^*)^2];$$
$$T_3(b) = [160(3y_1^* + y_2^*)^2 + 160(y_1^* + 3y^*2)^2 + 160(y_1^* + y_2^*)^2]$$

Case (iii) $\nu = 4$: Sample frequency type $(2, 1, 1, 1)$

$$T_4 = 1680(2y_1^* + y_2^*)^2 + 1680(y_1^* + 2y_2^*)^2 + 3360(y_1^* + y_2^*)^2$$

Case (iv) $\nu = 5$: Sample frequency type $(1, 1, 1, 1, 1)$

$$T_5 = 6720(y_1^* + y_2^*)^2$$

Combining all the cases and subcases, we obtain

Coefficient of $y_1^* y_2^* = (120 + 120) + [(960 + 960 + 320) + (1152 + 576 + 576)] + (6720 + 6720 + 6720) + 13440 = 38384$.

Further, by looking at the expression $S1 = \sum_s[a(y_1^* + \dots) + \dots$ similar terms across all samples], we readily note that the coefficient of y_1^{*2} in $S2 = \sum_s[a(y_1^* + \dots)^2 + \dots$ similar terms across all samples] is the same as that of y_1^* in $S1$.

Because of permutation invariance, we may thus write

$$S2 = 49770 \sum_i y_i^{*2} + 19192 \sum_{i \neq j} y_i^* y_j^* = (49770 - 19192) \sum_i y_i^{*2} \quad (\text{since } \sum_i y_i^* = 0).$$

Next note that the population variance $S^2 = \sum_i y_i^{*2}/9$ so that $S2 = 30578 \times 9S^2 = 275202S^2$.

Hence Var$(\hat{\bar{Y}}) = [1/(5 \times 99540)][275202S^2] = 0.55295S^2$.

We have observed earlier that the variance of the mean of distinct units under $SRSWR$ $(N = 10, n = 5; K = 3) = E[(1/\nu) - 1/N]S^2 = (0.25214 - 0.1)S^2 = 0.15214S^2$.

9. Conclusions

We conclude from this study

1. $SRSWR(N = 10, n = 5)$ produces mean of distinct units with variance $E[1/\nu) - 1/N]S^2 = 0.1533S^2$.

2. $SRSWR(N = 10, n = 5; K = 3)$ produces mean of distinct units with variance $E[(1/\nu) - 1/N]S^2 = 0.15214S^2$.

3. $SRSWR(N = 10, n = 5; K = 3)$ produces mean of all units with variance $0.25295S^2$.

4. Recall $E = E_1E_2$ and $V = V_1E_2 + E_1V_2$. Here V corresponds to item 3 and V_1E_2 corresponds to item 2.

Therefore, as a byproduct, we derive $E_1V_2 = 0.10081S^2$.

Acknowledgements

We are thankful to two anonymous referees for careful reading of the manuscript and offering helpful suggestions towards improving our presentation. We acknowledge help from Dr. Sobita Sapam towards formatting of the manuscript in journal style.

References

Basu D. On sampling with and without replacement. Sankhya.1958; 20, 287-294.

Hedayat, AS, Sinha Bikas K. Design and Inference in Finite Population Sampling. New York: John Wiley & Sons; 1991.

Korwar, RM, Serfling, RJ On averaging over distinct units in sampling with replacement. Ann Math Stat. 1970; 41: 2132-2134.

Sinha Bikas K, Sen PK. On sampling with and without replacement. Sankhya. 1958; 51, 65-83.