ASA01HH.082101

THE HANSEN-HURWITZ ESTIMATOR REVISITED: PPS SAMPLING WITHOUT REPLACEMENT

Monroe G. Sirken, National Center for Health Statistics
5535 Belcrest Road,Room 700, Hyattsville, MD 20782

Key Words: network sampling and estimation, conventional sampling, finite population factor

1. Introduction - the question

Hansen and Hurwitz (1943) introduced the notion of sampling unequal size clusters with probabilities proportionate to size (pps) to estimate X, the sum of the x-variate over a finite population of M elements. Their procedure for sampling clusters with PPS and with replacement has been widely adopted in sample surveys. First, clusters are assigned consecutive intervals with lengths equal to cluster size (number of population elements in a cluster). Second, a target sample of consecutively numbered population elements is randomly drawn with replacement from a uniform distribution. Third, clusters within whose ranges the random numbers fall are selected with replacement.

Suppose the sampling procedure is changed so that the target sample of population elements is drawn without instead of with replacement. How would this sampling procedure change affect the precision of the Hansen-Hurwitz estimator of X? This question is addressed in this paper by applying the theory of network sampling.

2. Notation

M population elements (i = 1, 2, ..., M) are partitioned into R non overlapping clusters (j = 1, 2, ... , R), and M_j denotes the size of cluster j. Then

$$M = \sum_{j=1}^{R} M_j .$$

Let X_{ij} denote the value of the x-variate for population element i (i = 1, ..., M_j) in cluster j (j = 1, ..., R). Four parameters of interest are

$$X_j = \sum_{i=1}^{M_j} X_{ij}$$ = the sum of the x-variate over the M_j elements in cluster j (j = 1, ..., R);

$$\bar{X}_j = \frac{X_j}{M_j}$$ = the average value of the M_j population elements in cluster j (j = 1, ..., R);

$$X = \sum_{j=1}^{R} X_j$$ = the sum of the x-variate over the M elements; and

$$\bar{X} = \frac{X}{M}$$ = the average value of M element in R clusters.

3. Hansen-Hurwitz selection procedure

The Hansen-Hurwitz (HH) selection procedure for sampling unequal-size-clusters with pps and with replacement is basically a 3-step operation (Hansen, Hurwitz and Madow, 1953).

Step 1. M population elements (i = 1, 2, ..., M are assigned consecutive integers, 1 to M, and R clusters (j = 1, 2, ..., R) are assigned consecutive intervals with lengths equal to cluster size M_j.

Step 2. A fixed-size target sample of m population elements is drawn from a uniform population spanning integers 1 to M by simple random sampling with replacement.

Step 3. Clusters within whose ranges the random numbers of the target sample fall are drawn with replacement.

In the HH selection procedure, neither population elements in the clusters drawn in step 3 nor individual population elements drawn in step 2 are removed from the sampling frame after selection.

4. Hansen-Hurwitz estimator

Based on a sample of m clusters of population units selected with PPS and replacement of population elements and clusters, the well known unbiased HH estimator of X, also known as the pps estimator, is

$$X'_{HH} = \frac{1}{r}\sum_{j=1}^{r}\frac{M}{M_j}X_j = \frac{M}{r}\sum_{j=1}^{r}\bar{X}_j \qquad (1)$$

where

r = m = cluster sample size
= population element sample size.

This $\bar{X}_j$'s in equation (1) are summed over clusters (j = 1, . . ., r) encompassing the target sample of m sampled population element(s) selected in step 3. The same cluster is counted as often as populations elements falling within its ranges is selected in the sample of m population elements. This means, for example, that if exactly two of the m sampled population elements fall in the range of cluster j, then two terms in equation (1) are $\bar{X}_j$, whether the two selected population elements are different population elements or the same elements - a possibility because population elements and clusters of population elements are selected with replacement.

The variance of the HH estimator of X is

$$\text{Var}\,(X'_{HH}) = \frac{M}{m}\sum_{j=1}^{R} M_j(\bar{X}_j - \bar{X})^2. \quad (2)$$

The unbiased estimator of this variance is

$$\text{Var}\,(X'_{HH}) = \frac{M^2}{r(r-1)}\sum_{j=1}^{r} (\bar{X}_j - \bar{X}')^2 \qquad (3)$$

where

$$\bar{X}' = \frac{X'_{HH}}{M}.$$

5. Options for sampling without replacement

The variance given by equation (2) applies only when population elements and clusters are sampled with replacement. Two ways of converting the 3-step HH sample selection procedure with replacement into a 3-step sample selection procedure without replacement are:

Option 1. In step 3, M_j population elements in clusters (j = 1, 2, ..., r) are not replaced. Steps 1 and 2 are unchanged.

Option 2. In step 2, m population elements (i = 1, 2, ..., m) are not replaced. Steps 1 and 3 are unchanged.

When clusters are selected without replacement (Option 1) and m > 1, the HH estimator given by equation (1) is not a pps estimator, and typically the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) or adaptations of it (Brewer and Hanif, 1983) are used instead of the HH estimator. When the target sample of m population elements is selected without replacement (Option 2), the HH estimator is essentially a pps estimator, and its variance in derived in the following section.

6. PPS sampling without replacement of population elements

First, we show that the HH estimator of X [equation (1)] is equivalent to a network sampling (NS) estimator of X with the following attributes. The NS estimator is based on a simple random sample design in which population elements are selection units, and $\bar{X}_j$'s are countable at every population element i (i = 1, 2, ..., M_j) in cluster j (1, 2, ..., R) because clusters are networks. [See (Sirken, 1997)] for a description of network sampling].

Let $X_j(i)$ = the sum of the x-variate over the M_j population elements in cluster j containing population element i . Then, $X_j(i) = X_j$, and $\bar{X}_j(i) = \bar{X}_j$. Rewriting equation (1) in this notation

$$X'_{HH} = \frac{M}{r}\sum_{j=1}^{r} \bar{X}_j = \frac{M}{m}\sum_{i=1}^{m} \bar{X}_j(i) = X'_{NS} \quad (4)$$

where X'_{NS} is an unbiased NS estimator of X. The NS estimator sums the $\bar{X}_j(i)$'s over the target sample of m population elements (i= 1, 2, ..., m) selected by simple random sampling. The HH estimator sums the $\bar{X}_j$'s over r sample clusters (j = 1, 2, ..., r) selected with pps.

If the target sample of m population elements is selected with replacement, variances of the HH and NS estimators are equivalent.

$$Var_w(X'_{NS}) = \frac{M}{m}\sum_{i=1}^{M} [\bar{X}_j(i) - \bar{X}]^2 =$$

$$\frac{M}{m}\sum_{j=1}^{R} M_j\,(\bar{X}_j - \bar{X})^2 = Var\,(X'_{HH}). \qquad (5)$$

Because the m target population elements are selected by simple random sampling, the variance of the NS estimator without replacement of target population elements is

$$Var_{wo}(X'_{NS}) = \frac{M-m}{M-1}\,Var_w(X'_{NS}). \quad (6)$$

And because $Var_w(X'_{NS}) = Var\,(X'_{HH})$ and m = r, it follows that

$$Var_{wo}(X'_{NS}) = \frac{M-r}{M-1}Var\,(X'_{HH}). \qquad (7)$$

7. Conclusion - answer to the question

This is the answer to the question that is posed in the Introduction: In pps sampling of clusters by the HH selection procedure, the design effect on the HH estimator of selecting m population elements without instead of with replacement from a finite population of M elements is equal to the finite population factor, fpc = (M - r) / (M - 1), where the cluster sample size r = m..

References

Brewer, K.R.W. and Hanif M. (1983). *Sampling With Unequal Probabilities*. New York: Springer-Verlag.

Hansen, M.M. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* **14** 333-362.

Hansen, M.M. and Horwitx, W.N. (1953). *Sample Survey Methods and Theory* Vol. 1 341-345. New York: Wiley.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47** 663-685.

Sirken, M.G. (1997). Network sampling. *Encyclopedia of Biostatistics* **4** 2977-2986. John Wiley & Sons.

.