Journal of science, education and technology
Volume 1 Number 2

April 2008
ISSN 1596-9770

On Variances when Data Arise From Two-Stage Sampling with Replacement

L. A. Nafiu and M.D. Shehu
Department of Mathematics / Computer Science,
Federal University of Technology, Minna, Nigeria

Abstract

In this paper, we present a general purpose equal probability sampling and unequal probability sampling with replacement plans using fixed sample size. It compares their resulting standard errors using data of Nafiu (2007) in a survey to estimate the population of diabetic patients in Niger State, Nigeria. It was observed that estimators under unequal probability with replacement are always better than those ones under equal probability with replacement.

Keywords: Unequal probability sampling, equal probability sampling, two-stage, one-stage and with replacement.

Introduction

The problem of selecting a with replacement sample of fixed size and with unequal probabilities and such that certain conditions on the joint selection probabilities of pairs of units are satisfied has a simple solution provided the sampler is willing to change the selection probabilities of a (usually) relatively small number of units. Furthermore, if the sampler is willing to accept some variation, which may be shown to be small, in sample size then the constraints on both probabilities and joint probabilities may be satisfied exactly by a simple selection procedure. The procedure if particularly convenient for the sequential selection of units by computer from a list maintained in machine-readable form.

Simple random sampling is a method of selecting n units out of the N such that every one of the ${}^{N}C_{n}$ distinct samples has an equal chance of being drawn. In practice, a sample random sample is drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. The units that bear these n numbers constitute the sample.

It is easily verified that all ${}^{N}C_{n}$ distinct samples have an equal chance of being selected by this method. Consider one distinct sample, that is, one set of n specified units. At the first draw the probability that some of the n specified units are selected is n/N. At the second draw the probability that some of the remaining $(n-1)$ specified units is drawn is $(n-1)/(N-1)$ and so on. Hence, the probability that all n specified units are selected in n draws is

$$\frac{n}{N}\cdot\frac{(n-1)}{(N-1)}\cdot\frac{(n-2)}{(N-2)}\cdots\frac{1}{(N-n+1)}=\frac{n!(N-n)!}{N!}=\frac{1}{{}^{N}C_{n}}.$$

Random sampling with replacement is entirely feasible: at any draw, all N members of the population are given an equal chance of being drawn, no matter how often they have already been drawn. The formulas for the variances and the estimated variances of estimates made from the sample are often simpler when sampling is with replacement than when it is without replacement. For this reason, sampling with replacement is sometimes used in the more complex sampling plans.

Sample Designs and Notations

Suppose that the population was stratified into L independent strata, indexed by $s = 1,2,3 \ldots, L$ and that the members of the $s-th$ stratum was grouped into N_{1s} primary sampling units indexed by $i = 1,2,\Lambda, N_{1s}$ and the $i-th$ primary sampling unit included N_{2si} members by $j = 1,2,\ldots\ldots, N_{2si}$. The corresponding symbols for sample are denoted by the lower case n with the same subscripts as shown in table 1 below. X_{hij} and Y_{hij} are variables for X and Y characteristics, respectively

Table 1: Symbols for two-stage clustered sample data.

	Population	Sample
1st –stage units 2nd-stage units	N_1 N_{2i}	n_1 n_{2i}
1st –stage index 2nd-stage index Index for cells	$i = 1,2,\Lambda, N_1$ $j = 1,2,\Lambda, N_{2i}$ $h = 1,2,\Lambda, q$	$i = 1,2,\Lambda, n_1$ $j = 1,2,\Lambda, n_{2i}$ $h = 1,2,\Lambda, q$
Totals	$N = \sum_{i\ 1=}^{N_1} N_{2i}$	$n = \sum_{i\ 1=}^{n_1} n_{2i}$
Cell Counts	$Y_{h2} = \sum_{i=1}^{N_1}\sum_{j=1}^{N_{2i}} y_{hij}$	$y_{h2} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_{2i}} y_{hij}$
Cell Proportions	Y_{h2}/N	y_{h2}/n
Ratios	$R_{h2} = X_{h2}/Y_{h2}$	$r_{h2} = x_{h2}/y_{h2}$

Table 2 below also shows the variances of aggregates by the types of sampling designs.

Table 2: Variances by types of designs and aggregates.

Design Types	Aggregate From	
	One-stage sampling	Two-stage sampling
Equal probability with replacement	$\text{var}(y_{h1e})$	$\text{var}(y_{h2e})$
Unequal Probability with replacement	$\text{var}(y_{h1u})$	$\text{var}(y_{h2u})$

Variances

Sampling could be done with equal or unequal probability (probability proportional to the size) with replacement. Within each stage, we may consider any combination of these options. We consider a generalized variance formula for any estimate $\hat{\Theta}_h$ in the $h-th$ cell based on completely arbitrary probabilities of selection. The total variance is then the sum of the variances for all strata. The symbol E is used for the operator of expectation, var for the variance and $\hat{\text{var}}$ for the unbiased estimate of the var. We may write:

$$\text{var}(\hat{\Theta}_h) = \underset{1}{\text{var}}(\underset{>1}{E}(\hat{\Theta}_h)) + \underset{1}{E}(\underset{>1}{\text{var}}(\hat{\Theta}_h)) \quad 3.1$$

where ">1" is the symbol to represent all stages of sampling after the first sampling.

If $\hat{\Theta}_h = y_{hi}$ as defined in table 2, its unbiased estimate can be written as:

$$\hat{\text{var}}(\hat{\Theta}_h) = \underset{1}{\hat{\text{var}}}(\hat{\Theta}_h) + \sum_{i=1}^{n_1} \pi_i^{(1)} \underset{>1}{\hat{\text{var}}}(y_{hi}) \quad 3.2$$

where $\pi_i^{(1)}$ is the probability of the $i-th$ unit included among the n_1 primary sampling units.

The expression (3.1) may be written into three components as:

$$\text{var}(\hat{\Theta}_h) = \underset{1}{\text{var}}(\underset{2}{E}(\underset{>2}{E}(\hat{\Theta}_h))) + \underset{1}{E}(\underset{2}{\text{var}}(\underset{>2}{E}(\hat{\Theta}_h))) + \underset{1}{E}(\underset{2}{E}(\underset{>2}{\text{var}}(\hat{\Theta}_h)))$$

3.3

Substituting $y_{hi} = \sum_{j=1}^{n_{2i}} y_{hij}$ in (3.2), the unbiased estimate of (3.3) can be written as:

$$\hat{\text{var}}(\hat{\Theta}_h) = \underset{1}{\hat{\text{var}}}(\hat{\Theta}_h) + \sum_{i=1}^{n1} \pi_i^{(1)} \underset{2}{\hat{\text{var}}}(y_{hi}) + \sum_{i=1}^{n1} \pi_i^{(1)} \sum_{j=1}^{n_{2i}} \pi_{ij}^{(2)} \underset{>2}{\hat{\text{var}}}(y_{hij})$$

3.4

where $\pi_{ij}^{(2)}$ is the probability of selecting the $j-th$ second stage unit in the selected $i-th$ first stage unit. This extension is now obvious for further stages of sampling.

We may summarize the above formula in words: an unbiased estimator of sampling variance in two-stage sampling, when the first stage sampling is with replacement, is obtained as the sum of two components. The first component estimates the variance as if only the first stage sampling had taken place. The second component is the weighted sum of the estimates, within the selected first stage units, of the variance due to the second stage of sampling(the first stage units being regarded as fixed); the weights are the probabilities of selection of these first stage units (Durbin, 1953).

If the sampling is done with replacement at the first stage, only the first term remains in (3.4) regarded as the limit of $\pi_i^{(1)} \to 0$. In this case, it is simple to estimate variances in multistage sampling with any number of stages when the first stage, with replacement, uses the same unequal probabilities at each drawing, while other stages are arbitrary, but carried out independently in different selected first stage units.

Consider variances for various sampling situations:

i. $\text{var}(y_{hle})$ for Equal Probability with Replacement in One-stage Sampling.

In this case,

$$\pi_i = nP_i \text{ and } \pi_{ii'} = n(n-1)P_iP_{i'} \qquad (3.5)$$

where P_i is the probability that the $i-th$ element is included in the replacement sampling at any draw and the variance is:

$$\mathrm{var}(Y_{h1e}) = \frac{N_1}{(N_1-1)}\sum_{i=1}^{N_1}(Y_{hi} - \bar{Y}_{h1_e})^2$$

(3.6)

The unbiased estimator of (3.6) can be expressed as:

$$\hat{\mathrm{var}}(y_{h1e}) = \frac{n_1}{(n_1-1)}\sum_{i=1}^{n_1}(y_{hi} - \bar{y}_{h1e})^2$$

(3.7)

ii. $\mathrm{var}(y_{h1u})$ for Unequal Probability with Replacement in One-stage Sampling.

Let ${}_rP_i$ be the probability that the $i-th$ individual is selected at the $r-th$ drawing, then

$$\sum_{i=1}^{N_1} {}_rP_i = 1, \quad \pi_i = \sum_{r=1}^{n_1} {}_rP_i, \quad \pi_{ii'} = \sum_{r\neq s}^{n_1}\sum {}_rP_{i\,s}P_j$$

(3.8)

Kendall and Stuart (1968) show that the variance for unequal probability with replacement in one-stage sampling is given as:

$$\mathrm{var}(Y_{h1u}) = \sum_{i=1}^{N_1}\pi_i(1-\pi_i)Y_{hi}^2 + \sum_{i\neq i}^{N_1}(\pi_{ii'} - \pi_i\pi_{i'})Y_{hi}Y_{hi'}$$

(3.9)

From $E(\sum_{i=1}^{n} g(y_i)) = \sum_{i=1}^{N}\pi_i g(Y_i)$ and

$$E(\sum_{i\neq i'}^{n}\sum g(y_i y_{i'})) = \sum_{i\neq i'}^{N}\sum \pi_{ii'} g(Y_i Y_{i'})$$

for any function g of observations, the unbiased estimator of (3.9) is given as:

$$\hat{\mathrm{var}}(y_{h1u}) = \frac{1}{2}\sum_{i\neq i'}^{n_1}\sum \frac{(\pi_i\pi_{i'} - \pi_{ii'})}{\pi_{ii'}}(y_{hi} - y_{hi'})^2$$

(3.10)

iii. $\mathrm{var}(y_{h2e})$ for Equal Probability with Replacement in Two-stage Sampling.

The m sample subunits in the $i-th$ unit are chosen by simple random sampling. The unbiased population variance is given as:

$$\mathrm{var}(Y_{h2e}) = \frac{N_1^2 M^2(1-f_1)S_1^2}{N_1} + \frac{N_1 M^2(1-f_2)}{N_1 m}\sum_{i=1}^{N_1} S_{wi}^2$$

(3.11)

where

$$S_1^2 = \frac{1}{N_1-1}\sum_{i=1}^{N_1}(Y_{hi} - \bar{Y}_h)^2, S_{wi}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_{hij} - \bar{Y}_{hi})^2, f_1 = \frac{n_1}{N_1}, f_2 = \frac{m}{M}$$

An unbiased sample estimator of (3.11) is given as:

$$\mathrm{var}(y_{h2e}) = \frac{N_1^2 M^2(1-f_1)s_1^2}{n_1} + \frac{N_1 M^2(1-f_2)}{n_1 m}\sum_{i\ 1=}^{N_1} s_{wi}^2$$

(3.12)

where

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_{hi} - \bar{y}_h)^2, s_{wi}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_{hij} - \bar{y}_{hi})^2, f_1 = \frac{n_1}{N_1}, f_2 = \frac{m}{M}$$

iv. $\mathrm{var}(y_{h2u})$ for Unequal Probability with Replacement in Two-stage Sampling.

The m_i sample subunits in the ith unit are chosen by simple random sampling. Primary units are selected with probabilities proportional to P_i with replacement. Result for $P_i = \frac{M_i}{M}$ where $M = \sum_{i\ 1=}^{N} M_i$. The unbiased estimator of the population variance is given as:

$$\mathrm{var}(Y_{h2e}) = \frac{1}{N_1}\sum_{i=1}^{N_1} P_i(\frac{Y_{hi}}{P_i} - Y_{h2e})^2 + \frac{1}{N_1}\sum_{i=1}^{N_1}\frac{M_i^2(1-f_{2i})}{m_i P_i}S_{wi}^2$$

(3.13)

where

$$S_{wi}^2 = \frac{1}{M_i - 1}\sum_{i=1}^{M_i}(Y_{hij} - \bar{Y}_{hi})^2, \quad f_{2i} = \frac{m_i}{M_i}$$

An unbiased sample estimator of (3.13) is given as:

$$\operatorname{var}(\hat{y}_{h2e}) = \frac{1}{n_1}\sum_{i=1}^{n_1} P_i\left(\frac{y_{hi}}{P_i} - y_{h2e}\right)^2 + \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{M_i^2(1-f_{2i})}{m_i P_i} s_{wi}^2 \qquad (3.14)$$

where

$$s_{wi}^2 = \frac{1}{m_i - 1}\sum_{i=1}^{m_i}(y_{hij} - \bar{y}_{hi})^2, \quad f_{2i} = \frac{m_i}{M_i}$$

Empirical Study and Results

In this section, empirical study is carried out using the data from Nafiu (2007), Unpublished M.Sc. thesis, Department of Statistics, University of Ilorin, Ilorin, Nigeria. The sampling variances for each of the designs were obtained with the aid of computer program written in Microsoft Visual C++ programming language (Hubbard, 2000). The results of the output generated are presented in tables 3, 4, 5 and 6.

Table 3: Standard Errors for the data in the year 2000

Design Types	One-stage Sampling	Two-stage Sampling
Equal Probability with Replacement	22884	18021
Unequal Probability with Replacement	22515	17711

Table 4: Standard Errors for the data in the year 2001

Design Types	One-stage Sampling	Two-stage Sampling
Equal Probability with Replacement	26396	24500
Unequal Probability with Replacement	25502	24400

Table 5: Standard Errors for the data in the year 2002

Design Types	One-stage Sampling	Two-stage Sampling
Equal Probability with Replacement	36033	35086
Unequal Probability with Replacement	35612	32262

Table 6: Standard Errors for the data in the year 2003

Design Types	One-stage Sampling	Two-stage Sampling
Equal Probability with Replacement	32055	30267
Unequal Probability with Replacement	32002	30144

Discussion of Results

The results presented in tables 3, 4, 5 and 6 indicate that standard error of unequal probability sampling with replacement is always less than standard error of equal probability sampling with replacement. It was also observed that two-stage sampling has its standard error less than that of one-stage sampling.

Conclusion and Recommendation

When an unbiased estimator of high precision and an unbiased sample estimate of its standard error are required, the sampling system employing unequal probabilities at each stage of sampling is particularly appropriate. Hence, unequal probability sampling with replacement under two-stage is recommended for any complex sampling designs and estimations.

References

Durbin, J. (1953). "Some Results in Sampling Theory When the Units are Selected with Unequal Probabilities", Journal of the Royal Statistical Society, 15,254-262.

Hubbard, J.R. (2000). *Programming with C++*, Second Edition. Schaum's Outlines, New Delhi: Tata McGraw-Hill Publishing Company Limited.

Kendall, M.G. and Stuart, A.S. (1986). *The Advanced Theory of Statistics*, New York: Hafner Publishing Company.

Nafiu, L.A. (2007). "Comparison of Four Estimators under Sampling without Replacement", Unpublished M.Sc. Thesis, University of Ilorin, Ilorin.

Okafor, F. C. (2002). *Sample Survey Theory with Applications*, Nigeria: Afro-Orbis Publications.

Thompson, S.K. (1992). *Sampling*. New York: John Wiley and Sons Inc.