Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 17, No. 1, 2019 (New Series), pp 63-72

Pursuing Further with an Innovative Approach the Issue to Settle the Size of a Sample to Draw from a Finite Survey Population

Arijit Chaudhuri[1] **and Tathagata Dutta**[2]
[1] *Indian Statistical Institute, Kolkata*
[2]*Indian Institute of Technology, Kanpur*

Received: January 15, 2018; Revised: October 09, 2018; Accepted: October 10, 2018

Abstract

It is considered desirable that with a high probability an unbiased estimator for a finite population total should have its error within a preassigned magnitude.Keeping in addition its estimated coefficient of variation within given limits a procedure is presented for a suitable size of the sample required to draw.

Key words: Coefficient of variation; Estimation error; Finite population; General sampling; Sample-size specification.

1 Introduction

Determining the size of a probability sample from a finite survey population is a classically recognized problem. Chaudhuri (2010, 2014, 2018) considered a selection criterion applying Chebyshev's inequality. Keeping the approach intact a slight amendment is proposed. In Section 2 below the essential criteria are presented. Then Section 3 illustrates alternative sampling strategies. Section 4 presents a brief account of the literature on earlier recommendations by eminent predecessors.

2 Fixing the Size of a Sample to Take from a Finite Survey Population

Suppose from a finite population $U = (1, \ldots, i, \ldots, N)$ of a known number N of identifiable units a sample s of a size n is to be selected according to a design p with a probability $p(s)$. The purpose is to unbiasedly estimate $Y = \sum_{1}^{N} y_i$, the population total (or the mean $\bar{Y} = \frac{Y}{N}$) of a variable y bearing values y_i for the units i in U. Suppose a statistic $t = t(s, y_i | i \in s)$ satisfying $E(t) = Y$, writing E to denote the Expectation operator, for every $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$ is to

Corresponding Author: Tathagata Dutta
E-mail: tathagata2.dutta@gmail.com

be employed for this. Suppose t has the variance, $V(t) = E(t-Y)^2$ and the design p is such that $\pi_i = \sum_{s \ni i} p(s) > 0 \,\forall\, i$ and $\pi_{ij} = \sum_{s \ni i,j} p(s) > 0 \,\forall\, i \neq j$. Also, $\sum_{\substack{j=1 \\ \neq i}} \pi_{ij} = (n-1)\pi_i$ in case we restrict to p for which every sample s contains a common number n of units, each distinct.

The problem we address here is to choose the number n in a rational way. For this a criterion as follows needs to be specified.

Chaudhuri (2010, 2014) employed Chebyshev's theorem giving us

$$\text{Prob}\left[|t-Y| \leq \lambda\sqrt{V(t)}\right] \geq 1 - \frac{1}{\lambda^2} \tag{2.1}$$

for an arbitrary positive number λ.

He also chose f & α such that,

$$\text{Prob}\left[|t-Y| \leq fY\right] \geq 1-\alpha \tag{2.2}$$

with $\alpha(0 < \alpha < 1)$ suitably specified. Here $f(0 < f < 1)$is to be proposed suitably to keep the estimation error within a fraction of the unknown total sought to be estimated.

Combining (2.1) with (2.2) one needs

$$\alpha = 1/\lambda^2 \;\text{ and }\; \lambda\sqrt{V(t)} = fY \tag{2.3}$$

If $V(t)$ may be expressed in terms of $\underline{Y}, N$ and n one may tabulate a choice of n in terms of N, f, α and other parameters involving $V(t)$ and $\underline{Y}$ suitably chosen.

Since, V(t) & $\underline{Y}$ are unknown, we propose the following criterion.

Since for (2.1) and (2.2) we need

$$100\frac{\sqrt{V(t)}}{Y} = 100f\sqrt{\alpha}$$
$$\text{or } 100\frac{\sqrt{V(t)}}{\bar{Y}} = 100f\sqrt{\alpha} \quad [when, E(t) = \bar{Y}]$$

and $100\frac{\sqrt{V(t)}}{Y}$ is the coefficient of variation (CV) of the statistic t as an unbiased estimator for Y the chosen value of $100f\sqrt{\alpha}$ is to equal this CV. But this CV is a function of all the co-ordinates of $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$ which can never be known. So we propose the following steps to follow in choosing n:

Let us

(1) express $V(t)$ explicitly and obtain an unbiased estimator, say, $v(t)$ for $V(t)$;

(2) choose a sample of an arbitrarily suitable sample-size n;

(3) calculate for the sample so chosen the value of

$$cv = \frac{100\sqrt{v(t)}}{t} \qquad \text{and}$$

(4) obtain a table of values of cv against $100f\sqrt{\alpha}$;

(5) repeat this procedure keeping $T = 100f\sqrt{\alpha}$ fixed and try to obtain a value of n , say n_0, which achieves cv nearest to this target T;

(6) treat n_0, rounded up to the nearest positive integer, as the required solution.

Let us illustrate several combinations of p and t to try this procedure.

3 Illustration of Strategies, that is Combination (p, t) with Right Sample Sizes

Let $x_i(> 0 \forall i)$ be the known size-measures of the respective units i of a population $U(= 1, \ldots, i, \ldots, N)$ with the total $X = \sum_1^N x_i$. Let $p_i = \frac{x_i}{X}(0 < p_i < 1 \forall i \in U), \sum_1^N p_i = 1)$ be the normed size-measures. Here, we try (cf. Chaudhuri and Dutta (2018)) postulating a simple model so as to write

$$y_i = \beta x_i + \epsilon_i, \ i \in U \tag{3.1}$$

with β as a constant, taken throughout as 10 in the examples below and ϵ_i's as independent standard normal random variables, i.e, $\epsilon_i \sim N(0, 1)$.
We illustrate a population of size N=67 & take n separately as 6,11,17,23 & 27 mostly, unless mentioned otherwise. We shall calculate cv's & T's taking separately $f = 0.1, 0.05$ & $\alpha = 0.05$ and observe how far the two, namely cv & T 'match' or 'mismatch'. Now, the strategies illustrated are as follows.

3.1 Simple Random Sampling with Replacement (SRSWR)

For this simplest sampling design

$$\bar{y} = \frac{1}{n}\sum_{r=1}^{n} y_r$$

the sample mean in draws with y_r as the y-value for the unit chosen on the r-th draw the following are well-known

$$E(\bar{y}) = \bar{Y}, \quad V(\bar{y}) = \frac{(N-1)S^2}{N}, \text{ where}$$

$$S^2 = \frac{1}{N-1}\sum_{1}^{N}(y_i - \bar{Y})^2 \, ;$$

$$s^2 = \frac{1}{(n-1)}\sum_{r=1}^{n}(y_r - \bar{y})^2 \text{ has } E(s^2) = S^2 \, ;$$

$$\frac{N-1}{N}\frac{s^2}{n} \text{ has } E\left(\frac{N-1}{N}\frac{s^2}{n}\right) = V(\bar{y})$$

and $$cv = 100\frac{s}{\bar{y}}\frac{1}{\sqrt{n}}\sqrt{\frac{N-1}{N}} \text{ should be close to } T = 100f\sqrt{\alpha}\,.$$

The table below shows the calculations for SRSWR sampling design.
In this and in the subsequent tables we show N/n to satisfy ourselves that the sampling fraction n/N remains reasonable in magnitude.

Table 1: Table for SRSWR

N	α	f	n	cv	T	N/n
67	0.05	0.1	6	0.79	2.236	11.167
67	0.05	0.1	11	0.82	2.236	6.1
67	0.05	0.05	17	0.59	1.118	3.94
67	0.05	0.05	23	0.5	1.118	2.91
67	0.05	0.05	27	0.52	1.118	2.48

From the table, we may regard the appropriate sample size as 11, in case $\alpha = 0.05$ & $f = 0.1$ and 17, in case $\alpha = 0.05$ & $f = 0.05$.

3.2 Simple Random Sampling without Replacement (SRSWOR)

For this $\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$,
the sample mean has $E(\bar{y}) = \bar{Y}$, $V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$,
$s^2 = \frac{1}{(n-1)}\sum_{i\in s}(y_i - \bar{y})^2$ has ,
$E(s^2) = S^2$ and
$cv = 100\frac{s}{\bar{y}}\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)}$ should be close to $T = 100f\sqrt{\alpha}\,.$

Table 2: Table for SRSWOR

N	α	f	n	cv	T	N/n
67	0.05	0.1	6	0.76	2.236	11.167
67	0.05	0.1	11	0.79	2.236	6.1
67	0.05	0.05	17	0.52	1.118	3.94
67	0.05	0.05	23	0.44	1.118	2.91
67	0.05	0.05	27	0.38	1.118	2.48

The table below shows the calculations for SRSWOR sampling design.

In this case, appropriate sample size n turns out to be 11, in case $\alpha = 0.05$ & $f = 0.1$ and 17, in case $\alpha = 0.05$ & $f = 0.05$.

3.3 SRSWOR with $\mathbf{t = X\bar{r} + \frac{(N-1)n}{(n-1)}(\bar{y} - \bar{r}\,\bar{x})}$, the Unbiased Estimator for Y due to Hartley & Ross (1954)

Here as also subsequently $x_i (> 0$ for every i in $U)$ are known values with $X = \sum_{1}^{N} x_i$, $\bar{X} = \frac{X}{N}$, $\bar{x} = \frac{1}{n}\sum_{i\in s} x_i$, $\bar{r} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{x_i}$.

Since $Y = E(t)$, it follows that $V(t) = E(t^2) - Y^2$. Since this t is based on SRSWOR in n draws, it is known that $\pi_i = \frac{n}{N}\ \forall\ i$ in U and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}\ \forall\ i, j (i \neq j) \in U$. So,

$$v = t^2 - \left[\frac{N}{n}\sum_{i\in s} y_i^2 + \frac{N(N-1)}{n(n-1)}\sum\sum_{i\neq j\in s} y_i y_j\right]$$

is an unbiased estimator for $V(t)$ and $cv = 100\frac{\sqrt{v}}{t}$ should come close to $T = 100f\sqrt{\alpha}$.

The table below illustrates the sample size for Hartley-Ross estimator.

Table 3: Table for Hartley-Ross estimator

N	α	f	n	cv	T	N/n
67	0.05	0.1	6	15.2	2.236	11.167
67	0.05	0.1	11	18.75	2.236	6.1
67	0.05	0.05	17	4.82	1.118	3.94
67	0.05	0.05	23	14.78	1.118	2.91
67	0.05	0.05	27	6.07	1.118	2.48

Hence, from the table, it can be seen that the appropriate sample size is $n = 6$, in case $\alpha = 0.05$ & $f = 0.1$ and $n = 17$, in case $\alpha = 0.05$ & $f = 0.05$.

3.4 Sampling with Probability Proportional to Size with Replacement (PPSWR) Employing Hansen & Hurwitz (1943) Estimator

Using the x_i's of Section 3.3 as the size-measures of the units and $p_i = \frac{x_i}{X}$ as normed size measures $(0 < p_i < 1, \sum_1^N p_i = 1)$ a PPSWR sample s may be taken in n draws. Then, Hansen-Hurwitz unbiased estimator for Y is

$$t_{HH} = \frac{1}{n}\sum_{r=1}^{n}\frac{y_r}{p_r};$$

here y_r, p_r are the y, p-values of the units chosen on the rth draw, $r = 1, \ldots, n$.

An unbiased estimator v for $V(t_{HH})$ is

$$v = \frac{1}{2n^2(n-1)}\sum_{r\neq r'}^{n}\sum^{n}\left(\frac{y_r}{p_r} - \frac{y_{r'}}{p_{r'}}\right)^2.$$

So, $cv = 100\frac{\sqrt{v}}{t_{HH}}$ for a sample may be checked against the target $T = 100f\sqrt{\alpha}$.

The table below illustrates the sample size for PPSWR sampling, employing Hansen & Hurwitz estimator.

Table 4: Table for Hansen & Hurwitz estimator

N	α	f	n	cv	T	N/n
67	0.05	0.1	6	0.433	2.236	11.167
67	0.05	0.1	11	0.258	2.236	6.1
67	0.05	0.05	17	0.213	1.118	3.94
67	0.05	0.05	23	0.231	1.118	2.91
67	0.05	0.05	27	0.197	1.118	2.48

Hence, from the table, it can be seen that the appropriate sample size is $n = 6$, in case $\alpha = 0.05$ & $f = 0.1$ and $n = 23$, in case $\alpha = 0.05$ & $f = 0.05$.

3.5 Horvitz & Thompson's Estimator (H-T estimator, 1952 Combined with a Sampling Scheme Specified for Illustration

The HT estimator is

$$t_{HT} = \sum_{i\in s}\frac{y_i}{\pi_i}$$

with an unbiased estimator for $V(t_{HT})$ as

$$v = \sum_{i<j\in s}\sum \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}$$

taking $\pi_{ij} > 0 \,\forall\, i \neq j$ in U.

Then, $cv = 100\frac{\sqrt{v}}{t}$ is to be checked via $T = 100f\sqrt{\alpha}$.

Our illustrated sampling scheme is as follows:

On the 1st draw take a PPS sample; on the second draw take one PPSWOR sample, leaving aside the unit chosen in the 1st draw; now, leaving aside the sampled units in the 1st two draws take an SRSWOR sample in (n-2) draws to have a final sample of size n. First calculate the $\pi_i(2)$ & $\pi_{ij}(2)$ for the 1st 2 draws; then using these values, compute $\pi_i(n)$ and $\pi_{ij}(n)$ for the realized scheme with sample size n, as shown by the formulae below.

The formulae for $\pi_i(2), \pi_i(n), \pi_{ij}(2), \pi_{ij}(n)$ are as follows :

$$\begin{aligned}
\pi_i(2) &= p_i + \textstyle\sum_{k\neq i} p_k \dfrac{p_i}{1-p_k} \\
\pi_i(n) &= \pi_i(2) + (1-\pi_i(2))\frac{n-2}{N-2} \\
\pi_{ij}(2) &= \frac{p_ip_j}{(1-p_i)} + \frac{p_ip_j}{(1-p_j)} \\
\pi_{ij}(n) &= \pi_{ij}(2) + \frac{n-2}{N-2}(\pi_i(2) + \pi_j(2) - 2\pi_{ij}(2)) \\
&\quad + \frac{n-2}{N-2}\frac{n-3}{N-3}(1 - \pi_i(2) - \pi_j(2) + \pi_{ij}(2))
\end{aligned}$$

With the help of these expressions, we calculate the variance, thereby the cv of the estimator t_{HT}.

The table below illustrates the sample size for H-T estimator.

Table 5: Table for H-T estimator

N	α	f	n	cv	T	N/n
67	0.05	0.1	4	51.18	2.236	16.75
67	0.05	0.1	5	42.77	2.236	13.4
67	0.05	0.1	6	34.46	2.236	11.167

Hence we see that, while using H-T estimator, $n = 6$ is the most appropriate sample size among the three choices, in case $\alpha = 0.05$ & $f = 0.1$.

3.6 Rao-Hartley-Cochran (RHC, 1962) Strategy of Sampling and Estimation

Using the same p_i's as in Section 3.5 RHC sampling consists in forming n non-overlapping random groups of the units of U by SRSWOR taking N_i units in the ith group, $i = 1, \ldots, n$ and independently choosing PPS sample of 1 unit from each of these n groups. Writing $\sum_n$ as the sum over these n groups and $\sum_n \sum_n$ as the sum over the disjoint pairs of (with no duplication) these groups and Q_i as the sum over the N_i units of p_i-values falling in the ith group the RHC's unbiased estimator for Y is

$$t_{\text{RHC}} = \sum_n y_i \frac{Q_i}{p_i}$$

and RHC's unbiased estimator for the variance of this t_{RHC} is

$$\begin{aligned} v &= \frac{(\sum_n N_i^2 - N)}{(N^2 - \sum_n N_i^2)} \sum_n Q_i \left(\frac{y_i}{p_i} - t_{\text{RHC}} \right)^2 \\ &= \frac{(\sum_n N_i^2 - N)}{(N^2 - \sum_n N_i^2)} \sum_n \sum_n Q_i Q_i \left(\frac{y_i}{p_i} - \frac{y_{i'}}{p_{i'}} \right)^2 \end{aligned}$$

Then, $cv = 100 \frac{\sqrt{v}}{t_{\text{RHC}}}$ should be checked versus $T = 100 f \sqrt{\alpha}$.

The table below illustrates the sample size for Rao-Hartley-Cochran estimator.

Table 6: Table for RHC estimator

N	α	f	n	cv	T	N/n
67	0.05	0.1	6	16.92	2.236	11.167
67	0.05	0.1	11	11.28	2.236	6.1
67	0.05	0.1	17	26.24	2.236	3.94

Hence we see that, while using RHC estimator, $n = 11$ is the most appropriate sample size among the three choices, in case $\alpha = 0.05$ & $f = 0.1$.

3.7 Lahiri's (1951) Sampling with the Unbiased Ratio Estimator

Using the same x_i's and p_i's, $i \in U$, let a sample s be chosen in n draws with

$$p(s) = \frac{1}{\binom{N-1}{n-1} X} \sum_{i \in s} x_i$$

and the ratio estimator

$$t_R = X \left(\sum_{i \in s} y_i \right) \Big/ \left(\sum_{i \in s} x_i \right)$$

which is unbiased for Y be employed. Then, writing $a_{ij} = p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$

$$v(t_R) = \sum_{i<j\in s}\sum a_{ij} \left(\frac{X}{\sum_{i\in s} x_i}\right)\left(\frac{N-1}{n-1} - \frac{X}{\sum_{i\in s} x_i}\right)$$

is an unbiased estimator for $V(t_R)$ (cf. Chaudhuri (2014, page 65)). So $cv = 100\frac{\sqrt{v(t_R)}}{t_R}$ may be checked against $T = 100f\sqrt{\alpha}$.

Calculation for this sampling scheme is not easy and hence not tried.

4 A Short Account of Earlier Literature on Exercises in Sample-size Determination

Som (1973) to our gratification, mentioned size of a sample, variability, sample selection method and estimation procedure as important in survey sampling. Cochran (1953, 1963,1977), to our surprise, in his chapter 4 had his topic as 'The estimation of sample size', though covered SRSWR and SRSWOR theories quite earlier and more surprisingly considered sample size as important when a finite population proportion P of an attribute is to be estimated by its sample analogue p from an SRSWR and its modification for SRSWOR, he considered the error $|p - P|$ to be bounded by a number d (positive) with a high probability $1 - \alpha$. In this and in general case also he needed 'normal distribution' for the standardized deviate $e = (estimator - the\,estimand\,parameter)/(standard\,deviation\,of\,the\,estimator)$. Thus, for SRSWR, when estimator is the sample mean and the parameter is the population mean, the sample-size should be taken as $n = (\frac{tS}{d})^2$, where S is the population standard deviation and t is the value of a standard normal deviate above which lies the area α of the normal distribution. For SRSWOR n' modifies as $n'' = \dfrac{n'}{(1 + \frac{n'}{N})}$. For double sampling, two-stage sampling, stratified sampling involving SRSWR and SRSWOR solutions have been derived. But,for general unequal probability sampling Cochran offers no solutions, to our knowledge.But he considers variance function in conjunction with cost function to derive solutions for sample-size, further improved upon by Yates (1960). His illustrious followers like Sukhatme and Sukhatme (1970), Murthy (1967), Des Raj (1968) , Singh and Choudhary (1986), and others have also given similar results with little novelty in approach. Since the present work is not a review of the literature on sample-size determination exercises we refrain from elaborating further.

Acknowledgements

Both of us are grateful to an anonymous referee for his/her appreciative support for our exercise and especially the suggestion leading to our adding the Section 4 which hopefully has led to an improved write-up.

References

Chaudhuri, Arijit (2010). *Essentials of Survey Sampling.* Prentice Hall of India, Delhi, India.

Chaudhuri, Arijit (2014). *Modern Survey Sampling.* Taylor & Francis, CRC, Boca Raton, Florida, USA.

Chaudhuri, Arijit (2018). *Survey Sampling.* Taylor & Francis, CRC , Boca Raton, Florida, USA.

Chaudhuri, Arijit and Dutta, Tathagata (2018). Determining the size of a sample to take from a finite population. *Statistics and Applications*, (ISSN 2452-7395, on line), **16(1)**, New Series, 37–44.

Cochran, W.G. (1953, 1963, 1977). *Sampling Techniques.* 1st, 2nd, 3rd Ed.,John Wiley & Sons, N.Y., USA

Des Raj (1968). *Sampling Theory.* Tata McGraw-Hill Publishing Company Ltd. New Delhi.

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

Hartley, H.O. and Ross, A. (1954). Unbiased ratio estimators. *Nature*, **174**, 270–271.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, **77**, 89–96.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimators. *Bulletin of International Statistical Institute*, **33**(2), 133–140.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of Royal Statistical Society* B, **24**, 482–491.

Singh, Daroga and Chaudhary, F.S. (1986). *Theory and Analysis of Sample Survey Designs.* Wiley Eastern Limited, New Delhi.

Som, R.K. (1973). *A manual of sampling techniques.* Heinemann, London.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*, Charles Griffin and co., London, third edition.