Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16, No. 1, 2018 (New Series), pp 37-44

Determining the Size of a Sample to take from a Finite Population

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

Received: September 15, 2017; Revised: November 24, 2017; Accepted: December 10, 2017

Abstract

Size of a sample to choose from a finite survey population is known to be ascertainable by Chebyshev's inequality together with anticipated magnitude of population coefficient of variation in case of simple random sampling. To fix it for unequal probability sampling and for complex estimators, additional model-postulations are needed as illustrated.

Key words: Chebyshev's inequality; Finite population; Sample-size; Simple random sampling; Varying probability sampling.

1. Introduction

In laying down a rule to prescribe how many units of a finite survey population are to be selected in a sample, a rational approach should be followed. In Cochran (1977) and other classically eminent text books on survey sampling this topic has of course been covered.

Chaudhuri (2010, 2014) suggested an alternative approach of applying the Chebyshev's inequality in the following way.

Suppose t is an unbiased estimator for a population total $Y = \sum_{i=1}^{N} y_i$ for a real variable y taking values y_i for the respective units i of $U = (1, \ldots, i, \ldots, N)$.

Writing $V(t)$ as the variance of t one gets the Chebyshev's inequality as

$$\text{Prob}\left[|t - Y| \leq \lambda\sqrt{V(t)}\right] \geq 1 - \frac{1}{\lambda^2} \tag{1.1}$$

for a fixed positive number λ. If one intends to employ t to estimate Y with a requirement that, for a suitably chosen positive proper fraction $f(0 < f < 1)$ and a positive number α quite close to zero, say, α=0.01 or 0.05 etc in (0,1), that

$$\text{Prob}\left[|t - Y| < fY\right] \geq 1 - \alpha, \tag{1.2}$$

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

then (1.1) may be equated to (1.2) so as to get

$$\lambda\sqrt{V(t)} = fY$$

and

$$\alpha = \frac{1}{\lambda^2}$$

Fixing f and α and anticipating possible values of

$$\frac{V(t)}{Y^2}$$

and especially formulae for $V(t)$ involving the population and sample sizes respectively N and n, reasonable choices of n vis-a-vis known or anticipated values for other parameters involved may be rationally proposed.

Illustrative examples are suggested in section 2 below.

2. Sampling and Estimation Procedures with Sample-sizes Tabulated

(i) Simple Random Sampling With Replacement (SRSWR) and Estimating Population Mean by Sample Mean

Let $\bar{Y} = \frac{1}{N}\sum_1^N y_i = \frac{Y}{N}, \quad S^2 = \frac{1}{N-1}\sum_1^N (y_i - \bar{Y})^2.$

Letting n be the sample-size i.e the number of draws in an SRSWR from the population of size N, and $\bar{y}$ be the sample mean we have

$$\text{Prob}\left[|\bar{y} - \bar{Y}| \leq \lambda\sqrt{V(\bar{y})}\right] \geq 1 - \frac{1}{\lambda^2},$$
$$\text{Prob}\left[|\bar{y} - \bar{Y}| \leq f\bar{Y}\right] \geq 1 - \alpha,$$

with $\lambda > 0,\ 0 < \alpha < 1,\ V(\bar{y}) = \frac{(N-1)S^2}{Nn}$.

So, taking $\alpha = \frac{1}{\lambda^2},\ \lambda S\sqrt{\frac{N-1}{Nn}} = f\bar{Y},\ CV = 100\frac{S}{\bar{Y}}$, the population coefficient of variation, it is advisable to take

$$n = \frac{(N-1)(CV)^2}{N\alpha f^2} \tag{1.3}$$

So, given N, α, f, CV, the Table 1 gives n, rather, its approximation by the nearest higher integer.

Table 1: Calculation of sample-size for SRSWR

N	α	f	CV	n (cf. (1.3))
80	0.05	0.1	0.1	20
60	0.05	0.1	0.08	13
100	0.05	0.1	0.1	20
50	0.05	0.1	0.05	5

N.B. The sampling fraction $\frac{n}{N}$ seems fine.

(ii) Simple Random Sampling Without Replacement (SRSWOR) and Estimating Population Mean by Sample Mean

For a sample s of size n by SRSWOR, the sample mean $\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$ has $V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$. So, with $\alpha(0 < \alpha < 1),\ f(0 < f < 1)$ we may note

$$\text{Prob}\left[|\bar{y} - \bar{Y}| \leq \lambda\sqrt{V(\bar{y})}\right] \geq 1 - \frac{1}{\lambda^2},$$
$$\text{Prob}\left[|\bar{y} - \bar{Y}| \leq f\bar{Y}\right] \geq 1 - \alpha;$$

So, taking $\alpha = \frac{1}{\lambda^2},\ \lambda S\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} = f\bar{Y}$, it is reasonable to take

$$n = \frac{N}{1 + N\alpha\frac{f^2(100)^2}{(CV)^2}} \tag{1.4}$$

So, fixing N, α, f, CV we may tabulate n, rather the nearest higher integer from (1.4) to tabulate as follows

Table 2: Calculation of sample-size for SRSWOR

N	α	f	CV	n (cf. (1.4))
80	0.05	0.1	0.1	16
60	0.05	0.1	0.08	11
100	0.05	0.1	0.1	17
50	0.05	0.1	0.05	5

N.B. Sampling fraction $\frac{n}{N}$ seems OK and commensurate with Table 1.

(iii), (iv), (v)

Sample-size determination for
(iii) Hansen and Hurwitz (1943) Estimator for Population total from Probability Proportional to Size With Replacement (PPSWR) sampling in n draws, (iv) H-T estimator (1952) for Population Total by ΠPS or IPPS sampling with sample-size n to estimate the total of a population of size N and (v) Rao, Hartley, Cochran (RHC, 1962) Estimator by RHC sampling in n draws to estimate the total of a population of size N. Here IPPS means Inclusion Probability Proportional to Size.

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 and a PPSWR sample be chosen in n draws. The Hansen and Hurwitz estimator for $Y = \sum_1^N y_i$ based on such a sample is

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

writing y_r, p_r for the y and p-value of a unit of U chosen on the rth draw. Its variance is

$$V(t_{HH}) = \frac{1}{n}\sum_{\substack{i=1 \\ i<j}}^{N}\sum_{j=1}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2.$$

By (1.1) and (1.2) we then need

$$\text{Prob}\left[|t_{HH} - Y| \leq fY\right] \geq 1 - \alpha = 1 - \frac{V(t_{HH})}{f^2 Y^2} \tag{1.5}$$

so that $\alpha = \frac{V(t_{HH})}{f^2 Y^2}$.

Unlike in cases of SRSWR and SRSWOR both employing $\bar{y}$ here appealing to the concept of the coefficient of variation does not yield a useful solution to the problem of finding an appropriate value for n.

Instead, we try postulating a simple model so as to write

$$y_i = \beta x_i + \epsilon_i,\ i \in U \tag{1.6}$$

with β as an unknown constant and ϵ_i's as independent random variables with expectations $E_m(\epsilon_i) = 0\,\forall\, i$ and variances $V_m(\epsilon_i) = \sigma^2 x_i^g$ with an unknown $\sigma(> 0)$ and an unknown constant $g(0 \leq g \leq 2)$. We propose to approximate α by

$$\hat{\alpha} = \frac{E_m V(t_{HH})}{f^2 E_m(Y^2)}$$

With simple algebraic exercises it follows that

$$E_m V(t_{HH}) = \frac{\sigma^2}{n}\left[X\sum_{1}^{N} x_i^{g-1} - \sum_{1}^{N} x_i^g\right] \tag{1.7}$$

and

$$E_m(Y^2) = \beta^2 X^2 + \sigma^2 \sum_{1}^{N} x_i^g \tag{1.8}$$

So, combining (1.5) – (1.8) one solves n from

$$\alpha = \frac{\frac{\sigma^2}{n}\left[X\sum_{1}^{N} x_i^{g-1} - \sum_{1}^{N} x_i^g\right]}{f^2\left(\beta^2 X^2 + \sigma^2 \sum_{1}^{N} x_i^g\right)}$$

To work out n as an integer just exceeding the solution on choosing $\sigma, f, \alpha, \beta, g$ and taking x as a random sample from the negative exponential distribution with unity as its mean is a simple task. Hence the Table 3 below.

Table 3: Calculation of sample-size for PPSWR sampling

N	f	α	σ^2	β	g	n
100	0.2	0.05	1	5	1	22
100	0.2	0.05	1	5	1.5	19
100	0.2	0.05	1	5	2	20
50	0.2	0.05	1	5	1	16
50	0.2	0.05	1	5	2	20
50	0.1	0.05	1	10	0.5	12

N.B. The sampling fraction comes out rather okay.

While employing for Y the H-T estimator (1952) to be based on a suitable ΠPS or IPPS sampling scheme for which $\pi_i = nP_i$, for i in U and every sample contains a fixed number n of units, each distinct, so that $\pi_i = \sum_{s \ni i} p(s), \left(0 < np_i < 1 \,\forall\, i \text{ and } \sum_1^N \pi_i = n\right)$ and $\pi_{ij} = \sum_{s \ni i,j} p(s)$ satisfies $\sum_{\substack{j=1 \\ \neq i}}^{N} \pi_{ij} = (n-1)\pi_i$ for every i in U we may work out n as follows.

As is well-known, the H-T estimator is $t_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$ with the variance $V(t_{HT}) = \sum_{i<j}^{N}\sum^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$. Now, as before, we have

$$\text{Prob}\left[|t_{HT} - Y| \le fY\right] \ge 1 - \alpha = 1 - \frac{V(t_{HT})}{f^2Y^2}$$

Postulating the same model as in case of PPSWR, it follows that

$$E_m V(t_{HT}) = \frac{\sigma^2}{n}\left[X\sum x_i^{g-1} - n\sum_1^N x_i^g\right]$$

and

$$E_m(Y^2) = \beta^2X^2 + \sigma^2\sum_1^N x_i^g.$$

So, as in case of PPSWR for an IPPS sampling scheme we get

$$\alpha = \frac{\frac{\sigma^2}{n}\left[X\sum_1^N x_i^{g-1} - n\sum_1^N x_i^g\right]}{f^2\left[\beta^2X^2 + \sigma^2\sum_1^N x_i^g\right]} \tag{1.9}$$

So, fixing N, α, β, f, g and taking $\sigma^2 = 1$ and x as sampled from the negative exponential distributions with mean as unity we may work out n as the minimum positive integer just exceeding the value of n from the equation (1.9) as tabulated below

Table 4: Calculation of size of a ΠPS or IPPS sample

N	f	α	σ^2	β	g	n
100	0.2	0.05	1	10	1	7
100	0.1	0.05	1	10	1	19
100	0.1	0.05	1	10	2	15
50	0.1	0.05	1	10	1.5	11
50	0.1	0.05	1	10	2	12
50	0.1	0.05	1	10	1	13

N.B. In this case also the sampling fraction $\frac{n}{N}$ works alright.

Keeping x_i, p_i, α, f, N same as in the above two cases, the Rao, Hartley, Cochran (RHC, 1962) sampling works also similarly if RHC unbiased estimator is employed to estimate Y. We need as before

$$\text{Prob}\left[|t_{RHC} - Y| \leq fY\right] \geq 1 - \alpha = 1 - \frac{V(t_{RHC})}{f^2 Y^2};$$

here

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

Here by Q_i we mean the sum of the p_i-values for the units falling in the ith group of N_i units of U assigned by choosing or applying SRSWOR method from the population such that $\sum_n N_i = N$. Here $\sum_n$ means sum over the n groups formed by SRSWOR each from U taking N_i units in the ith group, $i = 1, \ldots, n$. Obviously, $\sum_n Q_i = \sum_1^N p_i = 1$. By the method we are following we need to note

$$V(t_{RHC}) = \frac{(\sum_n N_i^2 - N)}{N(N-1)} \sum_{i<j}^{N} \sum_{i<j}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

To work out n, as in the previous two cases we need to postulate the same model as we have introduced and calculate $E_m V(t_{RHC})$ which simplifies to

$$\sigma^2 \frac{(\sum_n N_i^2 - N)}{N(N-1)} \left[X \sum_1^N x_i^{g-1} - \sum_1^N x_i^g\right]; \text{ also}$$

$$E_m(Y^2) = \beta^2 X^2 + \sigma^2 \sum_1^N x_i^g.$$

This leads to

$$\alpha = \sigma^2 \frac{(\sum_n N_i^2 - N)}{N(N-1)} \frac{\left[X \sum_1^N x_i^{g-1} - \sum_1^N x_i^g\right]}{f^2 \left(\beta^2 X^2 + \sigma^2 \sum_1^N x_i^g\right)} \tag{1.10}$$

Following RHC's recommendation we take $N_i = \left[\frac{N}{n}\right] = K$ for $i = 1, \ldots, m$ and $N_i = \left[\frac{N}{n}\right] + 1 = (K+1)$ for $i = m+1, \ldots, n$ designating thus the n groups formed such that $\sum_n N_i = N$;

incidentally, K is thus uniquely determined. Then, as in earlier two cases, allowing the condition of taking n as the nearest integer just exceeding the value of n as as solved from the equation (1.10) above we find the solution for n starting with $x, N, f, \alpha, \sigma^2, \beta$ and g as below

Table 5: Calculations for sample-size determination in employing RHC estimator for population total to be based on RHC scheme of sampling

N	f	α	σ^2	β	g	n
200	0.1	0.05	1	10	1	19
200	0.05	0.05	1	10	1	60
100	0.1	0.05	1	10	2	17
100	0.1	0.05	1	10	1.5	16

N.B. Here also the sampling fraction is seen to work out quite satisfactorily.

3. Conclusion

Some other strategies of estimating a population total employing suitable schemes of sampling may also be similarly tried to yield right choices of sample size in practice in the context of Finite Survey Population inferencing in deriving suitable empirical results.

In deference to a recommended suggestion from the referee we add the following remark by him verbatim :
In community level surveys in health, nutrition etc., it is most common to use the following Cochran's formula for determining the sample size for estimating a parameter,

$$n = \frac{(z_\alpha * z_\alpha)pq}{d * d} \tag{1.11}$$

where n is the sample size, p is the likely value of the parameter(say prevalence of a disease, in proportions), $q = 1 - p$, d is the margin of error and z_α is the value of the normal deviate corresponding to level of significance α. d is usually taken as $d = ep$, where e is the relative permissible margin in error. For $z_\alpha = 2, \alpha = 0.05$ and 10% relative error margin the sample sizes for $p = 0.50$ and 0.05 are respectively 400 and 7600 respectively.
Our present work was, however, intended to avoid usual normality assumption and to take note of the relationship between n and N for the usual designs except SRSWR design .

Acknowledgments
The authors are pleased to gratefully acknowledge the support and suggestion from an anonymous referee for his/her helpful comments and advice.

References

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

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

Cochran, W.G. (1977). *Sampling Techniques.* John Wiley & Sons.

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

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.

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.