International Journal of Scientific Research in Mathematical and Statistical Sciences **Research Paper**
Vol.6, Issue.1, pp.122-130, February (2019) **E-ISSN:** 2348-4519
DOI: https://doi.org/10.26438/ijsrmss/v6i1.122130

An Optimum Sample Size in Cross Sectional Studies

J. Sreedharan[1*], S. Chandrasekharan[2], A. Gopakumar[3]

[1]Department of Community Medicine, College of Medicine, Gulf Medical University, Ajman, UAE
[2] Department of Statistics, Annamalai University, Annamalai Nagar, Tamilnadu, India
[3] Research Scholar, Annamalai University, Annamalai Nagar, Tamilnadu, India

[*]*Corresponding Author: drjayadevans@gmail.com, Tel.: +97-15079-61051*

Available online at: www.isroset.org

Received: 30/Jan/2019, Accepted: 18/Feb/2019, Online: 28/Feb/2019

***Abstract*—**Sample size plays a vital role in any research as very less sample and very large sample may lead to false conclusions. There are many formulas available for the calculation of sample size according to the design of the study or/and statistical tool. These formulas generally give a minimum sample size but not provide solution for finding the required upper limit. Studies which include large sample give estimates with higher precision but may lead to large sample fallacy with statistical significance even for insignificant effects. This paper aimed to develop a formula for optimum sample size in the context of cross sectional study design. Pattern of changes in the results at minimum, optimum, large and extreme large sample size is discussed. Also we compare the result of cross sectional study with the result of a case control study. This study observed that an additional factor (m-factor) to be multiplied with the existing formula, that gives an optimum large sample in a cross-sectional study. The m-factor is a function of Z_α and Z_β.

Keywords*— *Minimum sample size, Optimum sample size, Large sample size, Cross sectional study, Power

I. INTRODUCTION

Statistical methods have greater application in conduct of a research from the stage of planning through designing, collecting, analysing and interpretation of data. One of the important and challenging steps in a research is the determination of sample size (n). If the sample size is inadequate, results cannot be generalized to the whole population. Calculation of sample size is mainly depends upon the objective of the research and design of the study. Various formulas are derived for determining sample size in accordance with the design of the study which is descriptive or analytic [1-2]. Analytical studies are broadly classified into observational and experimental studies [3-4]. Therefore, calculation of sample size also changes according to different types of study designs such as cross-sectional study, case-control study, cohort study, intervention study, studies involving animals and so on [5-7].

Cross sectional study is one of the observational studies that produces descriptive results and helps to generate a hypothesis for further research. The purpose of conducting cross sectional study is to describe the characteristics of a population with descriptive statistics. The design is suitable to estimate the prevalence of an outcome/exposure and hence can be used as a baseline for other type of extensive studies [8]. In current practice, test of hypothesis is performed in cross sectional studies for comparing the groups and thereby to generate and strengthen the hypothesis. It tests the hypothesis about exposure and outcome relationships with a measure "Prevalence Odds Ratio (OR)". OR can be tested for its statistical significance with the help of suitable inferential techniques [9]. According to the variable type, a regression model can be used to test the significance of exposure effect on the outcome variable. Since cross sectional studies focus on the objectives such as estimating prevalence and further finding associated factors of the outcome variable, sample size should be calculated in such a way that to achieve both these study objectives. Generally in cross sectional studies, sample size can be calculated with proportion of the outcome, level of significance and the margin of error (L). Since cross sectional studies at existing sample size are not that powerful to test a hypothesis for valid inferences, larger samples are required to provide statistically valid and reliable research conclusions about the null hypothesis. But, if the number of study subjects is beyond a particular level, the statistical test results lead to fallacious conclusions [10-11]. At larger samples, statistical tests are more likely to identify important as well as unimportant smaller differences as significant [12].

Research objectives are also important factors in choosing appropriate formula for sample size calculation. Generally, objectives of the study are converted to measurable hypotheses and sample size formulas are derived according to the type of statistical tests used to prove or disprove this

hypothesis. In other words, sample size calculation is performed according to the different objectives such as estimating population parameters [13] or testing of hypothesis such as equality of means, proportions, coefficient of correlation/regression etc [14-16]. In relation to this, the main factors to be considered while calculating sample size are Type I error (α), Type II error (β), Power (1-β), Confidence level, Standard Deviation and the effect size [17-18]. A study should have at least 80% power to detect the real effect [19]. Implies, sample size should be increased to get an increased power of 80% and above. Since it is difficult to identify minor clinically significant differences at small samples, a larger sample size is required if difference in the groups is small. The sample size calculated using any formula gives the minimum number of samples which is required to conduct a study; the size can be chosen more than this minimum, but not less. Even then, the calculated sample size is often compromised, looking into feasibilities such as funds available for the research, study period and availability of the samples [20].

A study with high power indicates high chance of detecting a real exposure effect [21]. A larger sample is required for getting a higher power [22]. When the sample size formula includes power term, size of samples will get larger compared to the size obtained at the existing formula. At this optimal 'n', means or proportions can be estimated in cross sectional studies and estimates can be compared among exposed & non-exposed groups for finding association. As a part of ensuring sample representativeness and drawing valid test results, power term (Z_β) to be included in the in the sample size formula which helps to obtain adequate sample size [23-24].

There are many formulas derived for the calculation of sample size, specifically for estimation of parameters and test of hypothesis [25-26]. This paper discusses in section II about finding an adequate sample size for estimation of prevalence and test of association in cross sectional studies. Further optimum sample size formula is derived with 'm-factor' in section III, with which statistical test results leads to reasonably valid conclusion. In Section IV, optimum sample size and degree of risk is calculated on a real dataset for cross sectional study as well as case control study. The statistical significance of OR at various samples is discussed in section V and conclusive statements are given in section VI. In summary, this optimum sample size will help to avoid the chances of large sample fallacies to an extent and reduces the cost incurred for the study due to selection of unnecessary larger samples. In this context, the paper compares the pattern of changes in statistical test results (significant/Non Significant) in terms of p values at minimum, optimal and larger samples.

II. Sample size calculation in Cross Sectional Study – Existing Method and Proposed Method

The existing formula of minimum sample size in cross-sectional study for a binary exposure includes Z value (Z_α), variance and margin of error. The formula for estimating proportion [27-31] is

$$n = \frac{Z_\alpha^2 pq}{L^2} \quad \text{--------} \ (1)$$

Where p is the Expected population proportion based on previous studies/pilot studies and q= 1 – p. L is the Margin of Error (absolute measure of error or relative precision which can be considered up to 10%), Z is the standard normal variate at the chosen level of significance

Statistically, the above mentioned formula is appropriate for sample size calculation in estimating prevalence, but not adequate enough for testing a hypothesis.
If no test of hypothesis is involved, the existing formula gives adequate sample size to conduct a cross sectional study. If the cross sectional study is aimed at testing the significance of association, the above formula does not give sufficient samples to generate powerful inferential test results. In order to achieve adequate samples power of the test to be incorporated in the sample size formula. Since type II error is the factor that detects false negative differences, Power has greater role in testing of a hypothesis which helps to accept/reject the null hypothesis accurately. If power is not included in the sample size formula, an adequate number of samples will not be reflected on the selected sample to test a hypothesis. Lack of power in test of prevalence among exposure or non-exposure group may have further impact in finding other determinants of the outcome variable. The formula which lack power leads to the risk factors of the disease that may not be the real risk factors. Implies, the test of association to find the exposure effect is valid when sample is adequate and representative. Though the sampling technique has major role in ensuring selection of representative sample, sample size and test of sample representativeness are also important steps for accurate reflection of the population in the selected sample. With reference to this, an adequate sample size formula is suggested in cross sectional studies and the formula is derived from the definition of type I error.

Type I error, α = probabilty of rejecting null hypothesis (H_0) when null hypothesis is true. Null hypothesis is rejecting when the value of the test statistic value is greater than the critical value (cv). This can be expressed in the notation as follows,

$\alpha = P(X \geq cv/H_0) = 1 - P(X \leq cv/H_0)$

$= 1 - P(\frac{X-\mu_0}{\sigma} \leq \frac{cv-\mu_0}{\sigma} / H_0)$ (by standerdization)

$= 1 - F(\frac{cv-\mu_0}{\sigma}) = F(\frac{\mu_0-CV}{\sigma})$ since 1-F(x) = F(-x)

$Z_\alpha = \frac{\mu_0-CV}{\sigma}$

Therefore critical value, $cv = \mu_0 - Z_\alpha\sigma$ ……(2)

A researcher invests a fair amount of money, time and resources for entire conduct of the study based on a sample and therefore power analysis should be reflected on the sample size formula to get valid results. Power calculation helps to seek evidence against the null hypothesis and gives proper generalization of sample results to the population. In the below session, power formula is derived to incorporate in the sample size formula for drawing valid conclusions.

Power = 1-β = Probability of rejecting null hypothesis (H_0) when the alternative hypothesis (H_1) is true. This statement can be notated as follows,

$1-\beta = P(X \geq cv/H_1)$

$= 1 - P(X \leq cv/H_1)$

$= 1 - P(\frac{X-\mu}{\sigma} \leq \frac{CV-\mu}{\sigma}/H_1)$

(by standerdization)

$= 1 - P(\frac{X-\mu}{\sigma} \leq \frac{\mu_0-\mu}{\sigma} - Z_\alpha/H_1)$

(by substituting equation 2, $cv = \mu_0 - Z_\alpha\sigma$)

$= 1 - F(\frac{\mu_0-\mu}{\sigma} - Z_\alpha)$

$1-\beta = F(Z_\alpha - \frac{\mu-\mu_0}{\sigma})$ since 1-F(x) = F(-x)

$\beta = 1 - F(Z_\alpha - \frac{\mu-\mu_0}{\sigma}) = F(\frac{\mu_0-\mu}{\sigma} - Z_\alpha)$

$Z_\beta = \frac{\mu_0-\mu}{\sigma} - Z_\alpha$

$Z_\alpha + Z_\beta = \frac{\mu_0-\mu}{\sigma}$

$\frac{Z\alpha+Z_\beta}{\mu_0-\mu} = \frac{1}{\sigma}$ ……(3)

If $X_1, X_2 \ldots X_n$ follows Normal distribution N (μ,σ), then the test statistic $X' = \sum X_i/n$ follows normal distribution N(μ, $\sigma/\sqrt{n}$).

$Z = \frac{(X'-\mu)}{\sigma/\sqrt{n}}$ follows N (0,1)

Implies when testing mean in normal data, SD is $\sigma/\sqrt{n}$ and therefore σ in equation (3) can be replaced by $\sigma/\sqrt{n}$.

$\frac{Z\alpha+Z_\beta}{\mu_0-\mu} = \frac{1}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}$

Squaring both sides to solve for sample size n, is the number of subjects needed for a attaining a power of 1-β to detect a significant difference.

$n = \sigma^2 (\frac{Z\alpha+Z_\beta}{\mu_0-\mu})^2$

therefore, sample size for testing mean for a continious variable is $n = \sigma^2 (\frac{Z\alpha+Z_\beta}{\mu_0-\mu})^2$ ……(4)

Similarly for testing proportion in a binomial data,
If $y_1, y_2 \ldots y_n$ are n independent binary outcomes with probability of success p, then 'p' follows N $(p, \sqrt{\frac{pq}{n}})$ and equation 3 becomes,

$\frac{Z\alpha+Z_\beta}{p_0-p} = \frac{1}{\sqrt{pq}/\sqrt{n}} = \frac{\sqrt{n}}{\sqrt{pq}}$ (By replacing σ_n by $\sqrt{\frac{pq}{n}}$)

$n = pq (\frac{Z\alpha+Z_\beta}{p_0-p})^2$

Hence sample size for testing proportion for a binary variable in single sample test is

$n = pq (\frac{Z\alpha+Z_\beta}{p-p_0})^2$ ……(5)

III. A FACTOR FOR OPTIMUM LARGE SAMPLE SIZE IN CROSS SECTIONAL STUDIES

Relevant In cross sectional studies, minimum Sample size for estimation of population parameters is

$n = \frac{Z_\alpha^2 pq}{L^2}$ (as in equation 1)

$\frac{n}{Z_\alpha^2} = \frac{pq}{L^2}$ …… (6)

In cross sectional study, basically the researcher is dealing with one sample. Objective of the cross sectional study is to find the prevalence and comparison of proportion of the outcome across exposed and non-exposed group, to strengthen a related hypothesis. The design includes selection of one sample during data collection and later while analyzing the data, the sample is classified into exposed/non-exposed and diseased/non-diseased. Implies, it is a study with one sample. The effect size in this scenario is the difference between proportion of diseased and non-diseased.

This difference is logically equivalent to the margin of error that indicates shift from the proportion of referenced group.

Therefore, adequate sample size for testing a hypothesis of proportion in one sample design (equation 5) can be written as,

$$n = \frac{(Z_\alpha + Z_\beta)^2 pq}{(p-p_0)^2} = \frac{(Z_\alpha + Z_\beta)^2 pq}{L^2} \quad \ldots\ldots (7)$$

Equation (7) gives a larger sample compared to the existing method provided in equation (1)
In this context, Equation (1) & (7) can be expressed in the following inequality,

$$\frac{Z_\alpha^2 pq}{L^2} < \frac{(Z_\alpha + Z_\beta)^2 pq}{L^2}$$

$$\frac{pq}{L^2} < \frac{(Z_\alpha + Z_\beta)^2 pq}{Z_\alpha^2 L^2} \quad \ldots\ldots\ldots (8)$$

Replacing $\frac{pq}{L^2}$ of the inequality (8) by $\frac{n}{Z_\alpha^2}$ of equation (6)
Then inequality (8) becomes,

$$\frac{n}{Z_\alpha^2} < \frac{(Z_\alpha + Z_\beta)^2 pq}{Z_\alpha^2 L^2}$$

$$n < \frac{(Z_\alpha + Z_\beta)^2 Z_\alpha^2 pq}{Z_\alpha^2 L^2}$$

Implies, $n < \frac{(Z_\alpha + Z_\beta)^2}{Z_\alpha^2} \frac{Z_\alpha^2 pq}{L^2}$

$$n < m \left(\frac{Z_\alpha^2 pq}{L^2}\right) \text{ where } m = \frac{(Z_\alpha + Z_\beta)^2}{Z_\alpha^2}$$

Compared to the existing method, a larger sample size can be obtained by $n = m \left(\frac{Z_\alpha^2 pq}{L^2}\right)$ ….. (9)
which can be considered as optimal maximum sample size by the inclusion of power in the formula.

Equation (1) is the existing method for calculating sample size in a cross sectional study for finding the prevalence. Researchers are using the same sample size if they are also interested in testing the significance of disease rate among exposed and non-exposed. But compared to existing method, equation (9) gives an adequate optimum number of samples to detect an effect as it includes power term in the formula. Since two sample proportion test and an association test has similar effect [32-33], proposed formula (equation 9) can be used to achieve both the objectives of a cross sectional study. Thereby hypothesis can be strengthened to conduct further extended studies aiming at proving of the hypothesis. Additional factor (m) included in equation (9), ultimately gives higher samples for better precision of the estimates and mitigates the chance of large sample fallacy to an extent. In the upcoming part, sample size by the proposed method is proved as an adequate optimum 'n' for conducting a cross sectional study.

IV. Calculation of Optimum Sample Size – Application on a Real Data Set

Considering a large-scale, multi-round survey [34] called National Family Health Survey (NFHS) conducted during the year 2015, under the stewardship of the Ministry of Health and Family Welfare, Government of India (DHS 2015). For this study, samples are collected from the households throughout the states of India.

Minimum sample size and an adequate larger sample are calculated for a cross sectional study based on this real data that includes binary outcome variable 'Anemia'. If the objective is to find the prevalence and determinants of Anemia, minimum sample size can be calculated based on level of significance, margin of error and proportion of disease that can be identified from a previous study/pilot study. Based on the proportion of Anemia (p) from NFHS of 2015, minimum sample size is calculated for the selected states of India. And an optimum larger sample is also determined using the proposed method (m factor) for α level 0.05 and Margin of error as 10% of the Prevalence. Compared to the existing method, sample size formula that includes power term (preferably above 80%) gives larger number of samples. Results are presented in the table 1.

Moreover adequate sample size is calculated for higher powers such as 1-β=0.90, 0.95, 0.99. In Table 2, values of 'm factor' are presented at different power. Six states of India are selected from NFHS dataset and 'm' factor is calculated at required power. Then optimal larger sample size is determined for corresponding values of 'm'. Z_β values for the power (1-β) 0.80, 0.90, 0.95, 0.99 are 0.84, 1.28, 1.65 and 2.33 respectively. Accordingly, values of 'm' factor vary from 2.0, 2.7, 3.4 and 4.8. Minimum sample size required to conduct a cross sectional study is already calculated and presented in table 1 for α=5%, which is multiplied with the corresponding values of factor 'm' to determine the optimal larger 'n'.

Table 1: Minimum Sample Size and adequate sample size in Cross Sectional Study at 80% power

Selected States of India	Proportion Anemic (p)	L = 10% of p	Minimum 'n' in cross sectional study	Adequate 'n' in cross sectional study (1-β=.80)
Kerala	0.35	0.035	713	1456
Karnataka	0.46	0.046	451	920
Maharashtra	0.46	0.046	451	920
Delhi	0.53	0.053	341	696
Uttar Pradesh	0.53	0.053	341	696
Tamil Nadu	0.57	0.057	290	592

Table 2: Adequate sample size in cross sectional study at various power values (at 1-β=0.80, 1-β=0.90, 1-β=0.95, 1-β=0.99 and α=5%)

Power (1-β)	Z_β	m = $(Z_\alpha+z_\beta)2 / Z_\alpha^2$	Kerala	Karnataka & Maharashtra	Delhi & Uttar Pradesh	Tamil Nadu
			p=35%, minimum sample size =713	p=46%, minimum sample size =451	p=53%, minimum sample size =341)	p=57%, minimum sample size =290)
			Optimum 'n' in Cross Sectional Study			
0.80	0.84	2.0	1456	920	696	592
0.90	1.28	2.7	1948	1232	932	792
0.95	1.65	3.4	2419	1530	1157	984
0.99	2.33	4.8	3416	2161	1634	1389

If 35% is considered as the outcome proportion (p) from the previous study (NFHS 2015), then around 720 (n=713) samples are minimum required to estimate the prevalence of anemia (for margin of error 4% & α level 5%). Since cross sectional studies use to find the exposure effect, sample size can be go up to 1460 (n=1456) subjects which may give a better representative sample and valid test results. The additional factor 'm' takes the values 2.04, 2.73, 3.39 & 4.79 for various power 80%, 90%, 95% & 99% respectively. An optimum larger sample is the 'm' times of the minimum sample size. This gives an optimum size for sample, below which reliable conclusion from statistical testing procedures could be derived. If a researcher wanted to choose larger samples than the minimum required looking at better precision, they can collect a sample with size up to this optimum level. Above this level, chance of large sample fallacy is high. An attempt is done to prove this statement in table 3.

Each subject is selected by Systematic sampling method. Odds Ratio and its statistical significance on the minimum, optimum, large and extreme large samples are presented in table No. 3. The tables 3.1 to 3.3, depict changes in the inferential test results at various samples. Minimum sample size required for conducting a case-control study is also calculated and presented in table 4. Since exposure is considered as participant's Wealth Index and outcome variable as Anemia, proportion of exposure (Not Rich Group/poor) is compared among anemic (p_1) and non-anemic group (p_2). In both case-control and cross sectional design, binary logistic regression model is used to estimate and test the significance of OR (OR is equal to 1 or not) (Table 5).

V. STATISTICAL SIGNIFICANCE OF ODDS RATIO AT VARIOUS SAMPLE SIZE

In Table 3.1; 3.2; 3.3, the results of simple binary logistic regression is presented in detail. Association found between participant's wealth status and their anemic status. Anemic status of the study subjects is collected as severe, moderate, mild anemic and non-anemic. Participants' 'wealth status' (Poor, Middle & Rich) is assumed to be one of the exposure associated with Anemia. Based on this, a simple logistic regression model is created with exposure as 'Wealth status' and outcome as 'Anemia'. Results are given across three selected states of India. Minimum sample size and its optimum 'n' is already calculated and shown in table 1 & 2, at which association between participants' wealth status and Anemia (Anemic cases includes severe, moderate & mild) is presented with percentage of anemic across the ''rich and not rich' group. Here 'non-rich group' is the category that includes subjects of poor and middle class group. The degree of association is presented with OR and tested using simple logistic regression method. Statistical significance of OR is identified from 'p' values. Precision of OR is presented with Confidence Interval (CI). Precision of the estimates among various study designs (Case-control study & cross sectional study) are also assessed in terms of Confidence Interval of OR. The focus of this research was to identify the pattern of changes in the statistical test results at various samples in the context of cross sectional study design. In Table 3 & 5, degree of association (OR) is compared across sample sizes of cross sectional study and Case Control study.

Table 3: Odds Ratio and its statistical significance in cross sectional study at minimum, optimum, large and extreme large samples

Table 3.1: Karnataka State

Sample Size (n)	Wealth Index	% Anemic	%Not Anemic	P value	OR	CI	
						Lower	Upper
At Minimum 'n' (451)	Not Rich	50.5%	49.5%	0.18 (NS)	1.3	0.88	1.96
	Rich	43.7%	56.3%		1	--	--
At adequate 'n' (2.04 x 451)	Not Rich	47.7%	52.3%	0.93 (NS)	1.0	0.77	1.33
	Rich	47.4%	52.6%		1	--	--
At large 'n' (2.5 x 451)	Not Rich	45.4%	54.6%	0.05 (p≤0.05)	1.3	1.01	1.65
	Rich	39.4%	60.6%		1	--	--
At large 'n' (3 x 451)	Not Rich	49.8%	50.2%	0.007 (p≤0.01)	1.4	1.09	1.72
	Rich	42.0%	58.0%		1	--	--
At large 'n' (3.5 x 451)	Not Rich	49.9%	50.1%	0.03 (p≤0.05)	1.2	1.01	1.49
	Rich	44.7%	55.3%		1	--	--
At large 'n' (4 x 451)	Not Rich	48.9%	51.1%	0.009 (p≤0.01)	1.3	1.07	1.56
	Rich	42.6%	57.4%		1	--	--
At large 'n' (4.5 x 451)	Not Rich	47.7%	52.3%	0.008 (p≤0.01)	1.3	1.07	1.55
	Rich	41.5%	58.5%		1	--	--
At large 'n' (5 x 451)	Not Rich	49.0%	51.0%	0.000 (p≤0.001)	1.4	1.16	1.64
	Rich	41.0%	59.0%		1	--	--
At Extreme large 'n' (NFHS-2015 data)	Not Rich	46.9%	53.1%	0.000 (p≤0.001)	1.1	1.01	1.13
	Rich	45.4%	54.6%		1	--	--

Table 3.2: Tamil Nadu State

Sample Size (n)	Wealth Index	% Anemic	% Not Anemic	P value	OR	CI	
						Lower	Upper
At Minimum 'n' (290)	Not Rich	62.3%	37.7%	0.16 (NS)	1.4	0.88	2.25
	Rich	54.0%	46.0%		1	--	--
At adequate 'n' (2.04 x 290)	Not Rich	58.6%	41.4%	0.28 (NS)	1.2	0.87	1.67
	Rich	54.2%	45.8%		1	--	--
At large 'n' (2.5 x 290)	Not Rich	59.3%	40.7%	0.03 (p≤0.05)	1.4	1.03	1.86
	Rich	51.3%	48.7%		1	--	--
At large 'n' (3 x 290)	Not Rich	60.0%	40.0%	0.03 (p≤0.05)	1.3	1.03	1.76
	Rich	52.8%	47.2%		1	--	--
At large 'n' (3.5 x 290)	Not Rich	60.9%	39.1%	0.001 (p≤0.001)	1.5	1.19	1.96
	Rich	50.5%	49.5%		1	--	--
At large 'n' (4 x 290)	Not Rich	58.0%	42.0%	0.024 (p≤0.05)	1.3	1.04	1.64
	Rich	51.5%	48.5%		1	--	--
At large 'n' (4.5 x 290)	Not Rich	58.0%	42.0%	0.029 (p≤0.05)	1.3	1.03	1.59
	Rich	52.0%	48.0%		1	--	--
At large 'n' (5 x 290)	Not Rich	62.9%	37.1%	0.000 (p≤0.001)	1.5	1.19	1.83
	Rich	53.4%	46.6%		1	--	--
At Extreme large 'n' (NFHS-2015)	Not Rich	59.5%	40.5%	0.000 (p≤0.001)	1.2	1.19	1.29
	Rich	54.2%	45.8%		1	--	--

Table 3.3: Kerala State

Sample Size (n)	Wealth Index	% Anemic	% Not Anemic	P value	OR	CI	
						Lower	Upper
At Minimum 'n' (713)	Not Rich	36.4%	63.6%	0.62 (NS)	1.1	0.74	1.64
	Rich	34.1%	65.9%		1	--	--
At adequate 'n' (2.04 x 713)	Not Rich	37.5%	62.5%	0.43 (NS)	1.1	0.85	1.48
	Rich	34.9%	65.1%		1	--	--
At large 'n' (2.5 x 713)	Not Rich	36.5%	63.5%	0.24 (NS)	1.2	0.91	1.49
	Rich	33.1%	66.9%		1	--	--
At large 'n' (3 x 713)	Not Rich	37.9%	62.1%	0.11 (NS)	1.2	0.96	1.51
	Rich	33.7%	66.3%		1	--	--

At large 'n' (3.5 x 713)	Not Rich	37.9%	62.1%	0.11 (NS)	1.2	0.97	1.42
	Rich	34.3%	65.7%		1	--	--
At large 'n' (4 x 713)	Not Rich	38.4%	61.6%	0.05 ($p \leq 0.05$)	1.2	1.01	1.49
	Rich	33.8%	66.2%		1	--	--
At large 'n' (4.5 x 713)	Not Rich	37.7%	62.3%	0.02 ($p \leq 0.05$)	1.3	1.04	1.50
	Rich	32.6%	67.4%		1	--	--
At large 'n' (5 x 713)	Not Rich	38.4%	61.6%	0.01 ($p \leq 0.05$)	1.2	1.04	1.48
	Rich	33.4%	66.6%		1	--	--
At Extreme large 'n' (NFHS-2015)	Not Rich	37.0%	63.0%	0.003 ($p \leq 0.01$)	1.1	1.04	1.24
	Rich	34.1%	65.9%		1	--	--

Table 4: Sample Size for conducting Case-Control Study

States	P1	P2	No. of cases for case-control study	No. of controls to cases (1:2)	No. of controls to cases (1:3)
Kerala	0.182	0.202	3949	7898	11848
Tamil Nadu	0.496	0.55	894	1788	2683
Karnataka	0.621	0.654	1962	3924	5886

Larger samples are required if difference between the exposure rates is minute

Table 5: Risk Estimation (OR) in Case-Control Study

Type of Study	States of India	Wealth Index	Not Rich		Rich		p value	OR	CI	
			No.	%	No.	%			Lower	Upper
Case-Control Study	Karnataka (1962:5886) (1:3 ratio)	Anemic	1269	64.7	693	35.3	0.30 (NS)	1.0	0.94	1.16
		Not Anemic	3733	63.4	2153	36.6		1	--	--
	Tamil Nadu (894:2683) (1:3 ratio)	Anemic	1121	54.9	921	45.1	≤0.01	1.2	1.08	1.41
		Not Anemic	749	49.7	758	50.3		1	--	--

Many trials were generated for various samples and at different states of NFHS data. The findings resulted from Simple Binary logistic Regression are listed below,

- Large sample fallacy is evident in p-values for increasing samples. As sample size increases, p values decreases and therefore small effects are detected as statistically significant.
- OR is almost same at minimum 'n', optimum 'n' and large 'n'. Specifically, OR obtained at optimum 'n' (proposed method) is equal or more close to the OR obtained at extreme large sample of NFHS which can be considered as a population estimate. Compared to existing method, sample estimate calculated based on the proposed method is more likely to be the true estimate.

  Moreover, OR obtained at proposed sample size of cross sectional study is same as that of case-control study. As cross sectional study & case-control study aiming at associated factors of outcome variable, similarity in OR at optimum sample size of cross sectional design and minimum sample size of case control design shows that proposed method is effective in finding exposure effect; this finding is ensured specifically when the presence of exposure comes first. Though small difference in exposure rates shows statistical significance in case control study, these subtle effects are not turned to be statistically significant in cross sectional study.
- From minimum to its optimum sample size (n_1, n_2) of cross sectional study, p values of table 3 shows 'effects are not statistically significant (OR=1)'. But at larger samples ($>n_2$), the same OR turned to be statistically significant when tested for OR=1 against OR $\neq$ 1. Implies, the same effect gets as statistically significant for larger samples. In the case of Kerala state, difference in the proportion of anemia among rich and non-rich group is small compared to that of other states. Therefore the statistical significance of OR among Kerala cohort is identified at larger samples. It is due to small effect size. If the effect size is less, larger samples are required to detect these smaller effects. This shows effect size has an important role in determining the limit of the optimum sample size and its effect will be studied in further extended studies.
- Consistent Pattern of change in CI (CI becoming stringent) observed within the optimum sample size interval (n_1, n_2). Compared to existing method, stringent Confidence Interval for the Odds Ratio is observed at the proposed sample size. Implies, more précised results are available at optimum sample size compared to existing method of cross sectional study. CI is more stringent at sample size of case control study, though it is obvious when samples getting larger.

- Overall, measure of exposure effect is almost same at various sample sizes with any study designs. Mainly statistical test results are misleading for larger samples. Therefore determination of optimum sample size has great role in any kind of studies.

VI. CONCLUSION

This research proposed a formula for an optimum large sample when a study uses a cross sectional design; the formula for calculation of optimum 'n' includes power term and produces a better large sample. The estimation showed that estimates are almost same irrespective of the sample size, but optimum sample size produces the sample estimates which are close to population estimates. Large sample may mislead the statistical test results and therefore the proposed method is beneficial to fix an optimum sample size in cross sectional studies. An extended research is in progress for fixing an optimum sample size (n) for linear regression coefficient in a cross sectional study.

ACKNOWLEDGMENT

Authors would like to acknowledge Annamalai University, India and Gulf Medical University, UAE for providing approval and support to conduct the research.

REFERENCES

[1] H. Shah, *"How to calculate sample size in animal studies?",* National Journal of Physiology, Pharmacy and Pharmacology, Vol.**1**,Issue.**1**,pp.**35–39**, **2011**.

[2] P. Patra, *"Sample size in clinical research, the number we need",* International Journal of Medical Science and Public Health, Vol.1,Issue.1,pp.5–9, 2012.

[3] J. Cai, D. Zeng, "*Sample size/power calculation for case-cohort studies*", Biometrics, Vol.60,Issue.4,pp.1015–24, 2004.

[4] V. Kasiulevicius, V. Sapoka, R. Filipaviciute, "*Sample size calculation in epidemiological studies,* Gerontology, Vol.7,Issue.4,pp.225–31, 2006.

[5] M. Puopolo, *"Biostatistical approaches to reducing the number of animals used in biomedical research",*Annali dell'Istituto Superiore di Sanità, Vol.40,pp.157–63, 2004.

[6] P.R. Burton, A.L. Hansell, I. Fortier, T.A. Manolio, M.J. Khoury, J. Little, P. Elliott, *"Size matters: Just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology",* International Journal of Epidemiology, Vol.38,Issue.1,pp.263–73, 2009.

[7] J. Charan, T. Biswas, *"How to Calculate Sample Size for Different Study Designs in Medical Research?",* Indian Journal of Psychological Medicine, Vol.35,Issue.2,pp.121-126, 2013.

[8] M.S. Setia, *"Methodology Series Module 3: Cross-sectional Studies",* Indian Journal of Dermatology, Vol.61,Issue.3,pp.261-264, 2016.

[9] M. Szklo, F.J. Nieto, "*Epidemiology: Beyond the Basics",* MA: Jones and Bartlett Publishers, Sudbury, 2004.

[10] L. Naing,, T. Winn, B.N. Rusli, *"Practical issues in calculating the sample size for prevalence studies",* Archives of Orofacial Sciences, Vol.1,pp.9–14, 2006.

[11] T. Dahiru, T.S. Kene, A.A. Aliyu, *"Statistics in medical research: misuse of sampling and sample size determination*", Annals of African Medicine, Vol.5,Issue.3,pp.158-61, 2006.

[12] M. Noordzij, G. Tripepi, F.W. Dekker, C. Zoccali, M.W. Tanck, K.J Jager, *"Sample size calculations: basic principles and common pitfalls*", Nephrology Dialysis Transplant, Vol.25,Issue.5,pp.1388-93, 2010.

[13] R.K. Malhotra, A. Indrayan, *"Simple nomogram for estimating sample size for sensitivity and specificity of medical tests*"', Indian Journal of Ophthalmology, Vol.58,Issue.6,pp.519–22, 2010.

[14] T.J. Naduvilath, R.K. John, L. Dandona, *"Sample size for ophthalmology studies",* Indian Journal of Ophthalmology, Vol.48,Issue.3,pp.245-50, 2000.

[15] S.B. Hulley, S.R. Cumming, W.S. Browner, D. Grady, N. Hearst, T.B. Newman. "*Designing Clinical Research*", 2nd ed., Lippincot Williams & Wilkins, Philadelphia, USA, 2001.

[16] R.B Dell, S. Holleran, R. Ramakrishnan, *"Sample Size Determination",* Institute for Laboratory Animal Research Journal, Vol.43,Issue.4,pp.207-213, 2002.

[17] D.J. Biau, S. Kerneis, R. Porcher, *"Statistics in brief: The importance of sample size in the planning and interpretation of medical research",* Clinical Orthopaedics and Related Research, Vol.466,Issue.9,pp.2282–8, 2008.

[18] P. Kadam, S. Bhalerao, *"Sample size calculation*", International Journal of Ayurveda Research, Vol.1,Issue.2,pp.55-57, 2010.

[19] P. Bacchetti, *"Current sample size conventions: Flaws, harms, and alternatives*", BMC Medicine, Vol.8,Issue.17, 2010.

[20] G.T Fosgate, *"Practical sample size calculations for surveillance and diagnostic investigations*", Journal of Veterinary Diagnostic Investigation, Vol.21,Issue.1,pp.3–14, 2009.

[21] E. Whitley, J. Ball, *"Statistics review 4: Sample size calculations*", Critical Care, Vol.6,Issue.4,pp.335–41, 2002.

[22] Hazra, N. Gogtay, *"Biostatistics series module 5: Determining sample size",* Indian Journal of Dermatology, Vol.61,Issue.5,pp.496-504, 2016.

[23] F. Farrokhyar, D. Reddy, R.W. Poolman, M. Bhandari. "*Why perform a priori sample size calculation?*", Canadian Journal of Surgery, Vol.56,Issue.3,pp.207-213, 2013.

[24] K. Malterud, V.D. Siersma, A.D. Guassora, *"Sample Size in Qualitative Interview Studies: Guided by Information Power"* Qualitative health research, Vol.26,pp.1753-60, 2015.

[25] N. Gogate, "*Principles of sample size calculation,*" Indian Journal of Ophthalmology, Vol.58, Issue.6, pp.517–8, 2010.

[26] K. Sivasakthi, M.L. William, *"A Class of Super-Efficient Estimators of the Normal Variance: A Study on Sample Size Preference"*, International Journal of Scientific Research in Mathematical and Statistical Sciences, Vol.5, Issue.6, pp.212-221, 2018.

[27] S. Lemeshow, D.W. Hosmer, K. Janelle, K.L. Stephen, "*Adequacy of sample size in health studies*", John Wiley & Sons Ltd., World Health Organization, Englan, 1990. Available at: http://www.who.int/iris/handle/10665/41607

[28] M.A. Pourhoseingholi, M. Vahedi, M. Rahimzadeh. *"Sample size calculation in medical studies",* Gastroenterology and Hepatology from Bed to Bench, Vol.6,Issue.1,pp.14-17, 2013.

[29] N.N. Naing, *"Determination of Sample Size",* The Malaysian Journal of Medical Sciences, Vol. 10,Issue.2,pp.84-86, 2003.

[30] G.T. Fosgate, *"Practical sample size calculations for surveillance and diagnostic investigations*", Journal of Veterinary Diagnostic Investigation, Vol.21,Issue.1,pp.3–14, 2009.

[31] D.I. Glenn, "*Determining Sample Size",* Institute of Food and Agricultural Sciences, University of Florida, IFAS extension (PEOD6; Reviewed 2003), 1992. Available at: https://www.tarleton.edu/academicassessment/documents/Samplesize.pdf

[32] N Pandis, "*Sample calculations for comparing proportions",* American Journal of Orthodontics and Dentofacial Orthopedics, Vol.141,Issue.5,pp.666-7, 2012.

[33] R Rana, Singhal, R, *"Chi-square test and its application in hypothesis testing*", Journal of the Practice of Cardiovascular Sciences, Vol.1,Issue.1, pp.69-71, 2015.
[34] Demographic and Health Surveys (DHS), *"National Family Health Survey (NFHS-4)*, Latest Publication, 2015-2016. India. Available at: http://rchiips.org/nfhs/

AUTHORS PROFILE

Prof. Jayadevan Sreedharan pursued Ph.D in Biostatistics from Kerala University, India and second Ph.D in Epidemiology from Tampere University, Finland. He was Fellow (UICC) and received Diploma in Cancer Prevention (NCI, USA) and GradDipHPE (GMU, UAE).

He is currently working as Professor of Epidemiology & Statistics in the Department of Community Medicine, College of Medicine, GMU, UAE since 2009. He is a member of Kerala Statistical Association and has membership in research Committees. He has published more than 150 research papers in reputed journals and conducted more than 20 medical researches. He has 20 years of teaching experience and more than 15 years of research supervision experience for medical students (UG & PG) as well as PhD scholars.

Dr. Subramanian Chandrasekharan pursued Ph.D degree in Statistics from Annamalai Univrsity, Chidambaram, India in 2003. He did his M.Sc and M.Phil degree in statistics from Bharathiar University, Coimbatore, India in 1989 and 1990 respectively.

He is currently working as Assistant Professor in the Department of Statistics, Annamalai University, Tamilnadu since 2003. He has many years of experience in teaching and research supervision for students of statistics Department. His current research areas are stochastic modelling, Biostatistics and Generalized Distribution.

Ms. Aji Gopakumar, completed UG and PG in statistics & currently pursuing PhD in Statistics from Annamalai University, Tamil Nadu, India. At present, she is working for Gulf Medical University as Jr. Research Analyst/ Adjunct Instructor in Institutional Planning & Research (IPR), for past 7 years.

She has 5 years teaching experience in 'Biostatistics' for Medical undergraduate students of GMU. Also, consultation is provided in statistical analysis of various medical research projects and assisted for research report preparation & its publication. She has 19 publications in national and international journals.

.

