Chapter 2

Determination of appropriate Sample Size

Discussion of this chapter is on the basis of two of our published papers

"Importance of the size of sample and its determination in the context of data related to the schools of Guwahati" which was published in the *Bulletin of the Gauhati University Mathematics Association* Vol. 12, 2012

&

"An investigation on effect of bias on determination of sample size on the basis of data related to the students of schools of Guwahati" which was published in the *International Journal of Applied Mathematics and Statistical Sciences* Vol. 2, Issue 1, 2013

In survey studies, once data are collected, the most important objective of a statistical analysis is to draw inferences about the population using sample information. "How big a sample is required?" is one of the most frequently asked questions by the investigators. If the sample size is not taken properly, conclusions drawn from the investigation may not reflect the real situation for the whole population.

So, in this chapter we have discussed

- Importance of the size of sample and the method of determination of a sample size along with the procedure of sampling in relation to our study.
- If there is any effect of bias on determination of sample size

2.00 Introduction:

In spite of the application of scientific method and refinement of research techniques, tools and designs, educational research has not attained the perfection and scientific status of physical sciences. Therefore, there is a great necessity to study properly about different tools and techniques of research methodology. While studying a particular phenomenon, the researchers of this field face a problem at the beginning as

what may be the representative sample. Very few research articles are there which deals with the issue of determination of sample size.

Sample size calculation for a study, from a population has been shown in many books e.g. Cochran (1977), Mark (2005) and Singh and Chaudhury (1985). The aim of the calculation is to determine an adequate sample size which can estimate results for the whole population with a good precision. In other words, one has to draw inference or to generalize about the population from the sample data. The inference to be drawn is related to some parameters of the population such as the mean, standard deviation or some other features like the proportion of an attribute occurring in the population. It is to be noted that a parameter is a descriptive measure of some characteristics of the population whereas if the descriptive measure is computed from the observations in the sample it is called a statistic. Parameter is constant for a population, but the corresponding statistic may vary from sample to sample. Statistical inference generally adopts one of the two techniques, namely, the estimation of population parameters or testing of a hypothesis.

The process of obtaining an estimate of the unknown value of a parameter by a statistic is known as estimation [39, 71, 86]. There are two types of estimations viz. point estimation and interval estimation.

If the inference about the population is to be drawn on the basis of the sample, the sample must conform to certain criteria: the sample must be representative of the whole population [7, 64]. The question arises as to what is a representative sample and how such a sample can be selected from a population.

The computation of the appropriate sample size is generally considered to be one of the most important steps in statistical study. But it is observed that in most of the studies this particular step has been overlooked. The sample size computation must be done appropriately because if the sample size is not appropriate for a particular study then the inference drawn from the sample will not be authentic and it might lead to some wrong conclusions [49].

Again, when we draw inference about parameter from statistic, some kind of error arises. The error which arises due to only a sample being used to estimate the population parameters is termed as sampling error or sampling fluctuations. Whatever may be the degree of cautiousness in selecting sample, there will always be a difference between the parameter and its corresponding estimate. A sample with the smallest sampling error will always be considered a good representative of the population. Bigger samples have lesser sampling errors. When the sample survey becomes the census survey, the sampling error becomes zero. On the other hand, smaller samples may be easier to manage and have less non-sampling error. Handling of bigger samples is more expensive than smaller ones. The non-sampling error increases with the increase in sample size [116].

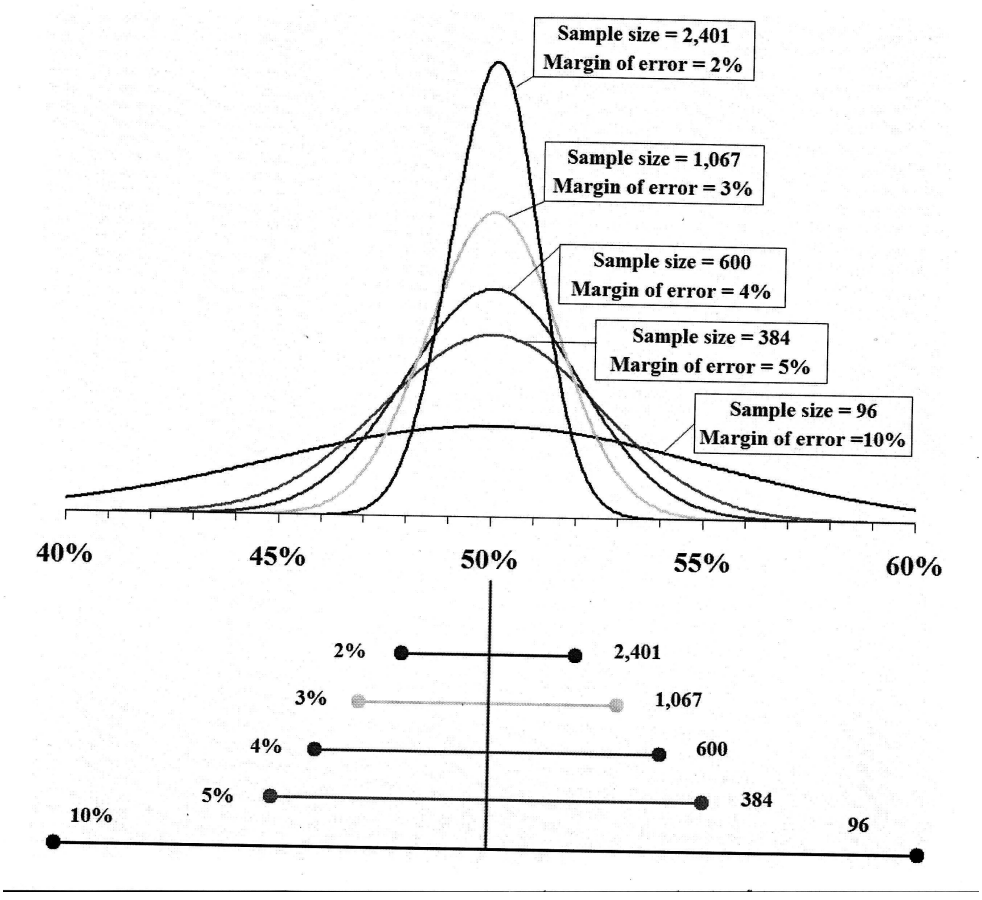


Fig 2.1, 2.2: Figures showing relationship between sampling error and sample size

There are various approaches for computing the sample size [5, 57, 117]. To determine the appropriate sample size, the basic factors to be considered are the level of precision required by users, the confidence level desired and degree of variability.

i) **Level of Precision** :

Sample size is to be determined according to some pre assigned 'degree of precision'. The 'degree of precision' is the margin of permissible error between the estimated value and the population value. In other words, it is the measure of how close an estimate is to the actual characteristic in the population. The level of precision may be termed as sampling error. According to W.G.Cochran (1977), precision desired may be made by giving the amount of errors that are willing to tolerate in the sample estimates. The difference between the sample statistic and the related population parameter is called the sampling error. It depends on the amount of risk a researcher is willing to accept while using the data to make decisions. It is often expressed in percentage. If the sampling error or margin of error is ±5%, and 70% unit in the sample attribute some criteria, then it can be concluded that 65% to 75% of units in the population have attributed that criteria.

High level of precision requires larger sample sizes and higher cost to achieve those samples.

ii) **Confidence level desired** :

The confidence or risk level is ascertained through the well established probability model called the normal distribution and an associated theorem called the Central Limit theorem.

The probability density function (p. d. f) of the normal distribution with parameters μ and σ is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

where, μ is the mean and σ is the standard deviation.

In general, the normal curve results whenever there are a large number of independent small factors influencing the final outcome. It is for this reason that many practical distributions, be it the distribution of annual rainfall, the weight at birth of babies, the heights of individuals etc. are all more or less normal, if sufficiently large number of items are included in the population. The significance of the normal curve is much more than this. It can be shown that even when the original population is not normal, if we draw samples of n items from it and obtain the distribution of the sample means, we notice that the distribution of the sample means become more and more normal as the sample size increases. This fact is proved mathematically in the Central Limit theorem. The theorem says that if we take samples of size n from any arbitrary population (with any arbitrary distribution) and calculate $\bar{x}$, then sampling distribution of $\bar{x}$ will approach the normal distribution as the sample size n increases with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$

$$\text{i.e. } \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

A sample statistic is employed to estimate the population parameter. If more than one sample is drawn from the same population, then all the sample statistics deviate in one way or the other from the population parameter. In the case of large samples, where n>30, the distribution of these sample statistic is a normal distribution. Generally, a question arises that how much should a sample statistic miss the population parameter so that it may be taken as a trustworthy estimate of the parameter. The confidence level tells how confident one can be that the error toleration does not exceed what was planned for in the precision specification.

Usually 95% and 99% of probability are taken as the two known degrees of confidence for specifying the interval within which one may ascertain the existence of population parameter (e.g. mean). 95% confidence level means if an investigator takes 100 independent samples from the same population, then 95 out of the 100 samples will provide an estimate within the precision set by him. Again, if the level of

confidence is 99%, then it means out of 100 samples 99 cases will be within the error of tolerances specified by the precision.

In case of normal distribution, the curve is said to extend from -3σ distance on the left to +3σ distance on the right.

A well known result of the distribution theory says that

$$\text{if } X \sim N(\mu, \sigma^2) \text{ then } Z = \frac{X - \mu}{\sigma} \text{ is a standard normal variate i.e. } Z \sim N(0,1).$$

While calculating the sample size, the desired confidence level is specified by the z value. The z-value is a point along the abscissa of the standard normal distribution. It is known from the table of normal curve that 95 percent of the total area of the curve falls within the limits ±1.96σ, where σ is the standard deviation of the distribution and 99 percent of that fall within the limits ±2.58σ. In other words, 95% of the area under the normal curve is specified by the z-value of 1.96 and z- value of 2.58 will specify 99% of the cases under the normal curve. These will represent confidence levels of 95% and 99% respectively.

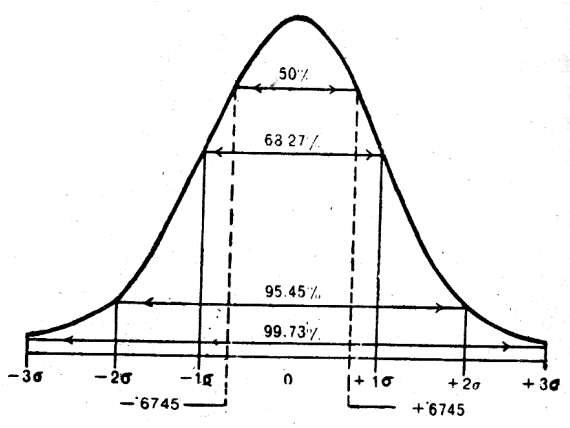


Fig 2.3: Standard Normal Curve

iii) Degree of variability:

The degree of variability in the attributes being measured refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the

sample size required to be, to obtain a given level of precision. For less variable (more homogeneous) population, smaller sample sizes works nicely. Note that a proportion of 50% indicates a greater level of variability than that of 20% or 80%. This is because 20% and 80% indicate that a large majority do not or do, respectively, have the attribute of interest. Because a proportion of 0.5 indicates the maximum variability in a population, it is often used in determining a more conservative sample size.

2.01 Strategies for determining sample size:

To determine a representative sample size from the target population, different strategies can be used according to the necessity of the research work.

Use of various formulae for determination of required sample sizes under different situations is one of the most important strategies.

There are different formulae for determination of appropriate sample size when different techniques of sampling are used. Here, we will discuss about the formulae for determining representative sample size when simple random sampling technique is used. Simple random sampling is the most common and the simplest method of sampling. Each unit of the population has the equal chance of being drawn in the sample. Therefore, it is a method of selecting n units out of a population of size N by giving equal probability to all units.

(a) Formula for proportions:

i) Cochran's formula for calculating sample size when the population is infinite:

Cochran (1977) developed a formula to calculate a representative sample for proportions as

$$n_0 = \frac{z^2 pq}{e^2} \qquad (2.1)$$

where, n_0 is the sample size, z is the selected critical value of desired confidence level, p is the estimated proportion of an attribute that is present in the population, $q = 1 - p$ and e is the desired level of precision [22].

For example, suppose we want to calculate a sample size of a large population whose degree of variability is not known. Assuming the maximum variability, which is equal to 50% (p =0.5) and taking 95% confidence level with ±5% precision, the calculation for required sample size will be as follows--

p = 0.5 and hence q =1-0.5 = 0.5; e = 0.05; z =1.96

So, $n_0 = \frac{(1.96)^2 (0.5)(0.5)}{(0.05)^2}$ =384.16=384

Again, taking 99% confidence level with ±5% precision, the calculation for required sample size will be as follows--

p = 0.5 and hence q =1-0.5 = 0.5; e = 0.05; z =2.58

So, $n_0 = \frac{(2.58)^2 (0.5)(0.5)}{(0.05)^2} = 665.64 = 666$

Following table shows sample sizes for different confidence level and precision.

Table 2.1

Sample size calculated for different confidence level and precision

Confidence level	Sample size (n_0)		
	e =.03	e =.05	e = .1
95%	1067	384	96
99%	1849	666	166

ii) Cochran's formula for calculating sample size when population size is finite:

Cochran pointed out that if the population is finite, then the sample size can be reduced slightly. This is due to the fact that a very large population provides proportionally more information than that of a smaller population. He proposed a correction formula to calculate the final sample size in this case which is given below

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}} \qquad (2.2)$$

Here, n_0 is the sample size derived from equation (2.1) and N is the population size.

Now, suppose we want to calculate the sample size for the population of our study where, population size is $N = 13191$. According to the formula (2.1), the sample size will be 666 at 99% confidence level with margin of error equal to (0.05). If $\frac{n_0}{N}$ is negligible then n_0 is a satisfactory approximation to the sample size. But in this case, the sample size (666) exceeds 5% of the population size (13191). So, we need to use the correction formula to calculate the final sample size.

Here, $N = 13191$, $n_0 = 666$ (determined by using (2.1))

$$n = \frac{666}{1 + \frac{(666 - 1)}{13191}} = 634.03 = 634$$

But, if the sample size is calculated at 95% confidence level with margin of error equal to (0.05), the sample size become 384 which does not need correction formula. So, in this case the representative sample size for our study is 384.

iii) Yamane's formula for calculating sample size :

Yamane (1967) suggested another simplified formula for calculation of sample size from a population which is an alternative to Cochran's formula. According to him, for a 95% confidence level and $p = 0.5$, size of the sample should be

$$n = \frac{N}{1 + N\left(e^2\right)} \tag{2.3}$$

where, N is the population size and e is the level of precision [131].

Let this formula be used for our population, in which N =13191 with ±5% precision. Assuming 95% confidence level and p =0.5, we get the sample size as

$$n = \frac{13191}{1 + 13191(.05)^2} = 388$$

To see which formula gives a better measure of the sample size, we calculated sample sizes for different schools from their respective population which we gathered during our investigation. Table 2.2 and 2.3 respectively shows the sample values which were calculated by Yamane's formula and Cochran's formula and we have plotted those values in fig. 2.3. The figure 2.3 shows that values calculated through both the formulae are in quite good agreement.

Table 2.2:

Sample sizes calculated by Yamane's formula

Sl. no. of schools	Population size,N	Sample size, *n* for 95% confidence level:		
		±5%	±7%	±10%
1	450	212	136	82
2	582	229	150	85
3	693	254	158	87
4	799	266	163	89
5	806	267	163	89
6	845	272	164	89
7	858	273	165	90
8	892	276	166	90
9	909	278	167	90
10	922	279	167	90
11	9 85	285	169	91
12	1009	287	170	91

13	1058	290	171	91
14	1073	292	171	91
15	1115	294	173	92
16	1167	299	174	92
17	1184	299	174	92
18	1256	303	176	93
19	1298	305	176	93
20	1322	307	177	93
21	1584	319	181	94
22	1908	330	184	95

Table 2.3

Sample sizes calculated by Cochran's formula

Sl.no. of schools	Population size,N	Sample size, *n* at 95% confidence level:			Sample size, *n* at 99% confidence level:		
		±5%	±7%	±10%	±5%	±7%	±10%
1	450	208	137	79	269	194	121
2	582	231	146	83	311	215	130
3	693	248	153	84	340	228	134
4	799	259	158	86	364	239	137
5	806	259	158	86	364	239	137
6	845	265	159	86	372	243	138
7	858	265	161	86	374	243	139
8	892	269	161	86	381	246	141
9	909	270	162	87	385	248	141
10	922	270	162	87	387	248	141
11	9 85	276	163	87	396	253	142
12	1009	278	165	88	398	253	143
13	1058	282	166	88	409	256	143
14	1073	282	166	88	411	256	144
15	1115	286	168	88	416	262	144
16	1167	289	168	89	424	264	146
17	1184	291	169	89	427	264	146
18	1256	295	169	89	435	268	147
19	1298	295	170	90	441	270	147
20	1322	298	170	90	444	270	148
21	1584	310	175	91	469	281	151
22	1908	320	178	91	493	288	152

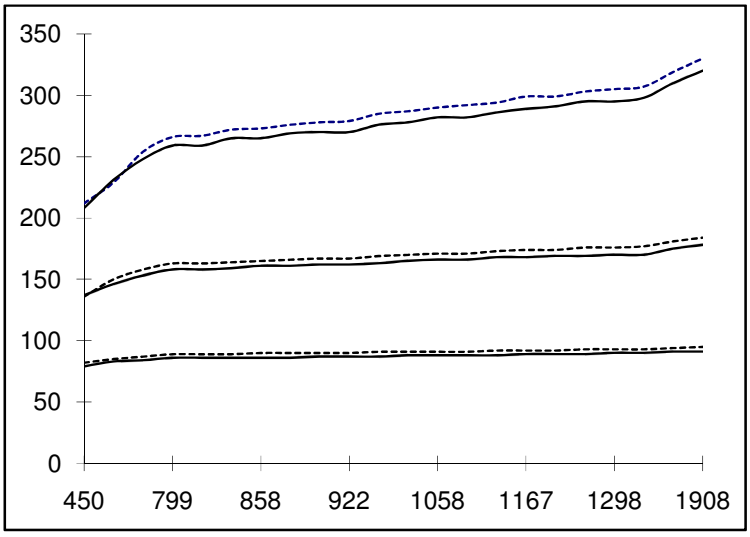


Fig. 2.4

x axis—population size, y axis—sample size

……. Values are calculated according to Yamane's formula and _____ values according to Cochran's formula. The uppermost pair is for 5%, middle one for 7% and the lower one for 10% level of significance

We want to mention here that though other formulae are also available in different literatures, the above two formulae are used extensively in comparison to the others.

After calculating the representative sample size the main aim of an investigator is to find the proper method of selecting samples. Sampling is simply the process of learning about the population on the basis of sample collected from the population. Sample is constituted by a part or fraction of the population. Thus, in the sampling technique, instead of every unit of the population, only a part of it is studied and the conclusions are drawn for the entire population on the basis of the sample.

2.02 Comparative study of two different methods of allocation:

In our study, for selection of samples, stratified random sampling technique has been adopted. The three categories of schools such as Government and Government Provincialised schools under SEBA (Secondary Education Board of Assam),

Permitted private schools under SEBA, Affiliated private schools under CBSE (Central Board of Secondary Education) of Guwahati were considered as the three strata. The sample from each stratum is taken through simple random sampling technique. The stratification is done to produce a gain in precision in the estimates of characteristics of the whole population.

The stratification was done following the principles that –

i) The strata (i.e. categories of schools) are non-overlapping and together comprise the whole population.
ii) The strata (i.e. categories of schools) are homogeneous within themselves with respect to the characteristics under study

All the VIII standard students of government, private including SEBA and CBSE schools of Guwahati formed the population of the study. Initially, we estimated the size of sample from a total of 13191 students of class VIII at 95% confidence level with $\pm 5\%$ level of precision which was found to be 384. Thus, the sample size of 384 students of 13 selected schools to examine performance of students in mathematics is considered under the present study. This sample can be considered representative of the student population of Guwahati, with students coming from a wide range of socio-economic backgrounds and from each of the four types of schools such as normal Co-Educational, Co-Educational segregated by gender, only Boys and only Girls schools. The allocation of the samples to the different categories of schools was carried out through both the proportional allocation method and optimum allocation method of stratified random sampling.

A. Sample size through proportional allocation method :

The proportional allocation method was originally proposed by Bowley (1926). In this method, the sampling fraction, $\frac{n}{N}$ is same in all strata. This allocation was used to obtain a sample that can estimate size of the sample with greater speed and a higher degree of precision. The allocation of a given sample of size n to different stratum was done in proportion to their sizes. i.e. in the i^{th} stratum,

$$n_i = n\frac{N_i}{N} \qquad \text{i =1, 2, 3.}$$

Where n represents sample size, N_i represents population size of the i[th] strata and N represents the population size. In our study, $N = 13191$; $n = 384$.

B. The sample size through optimum allocation method :

The allocation of the sample units to the different stratum is determined with a view to minimize the variance for a specified cost of conducting the survey or to minimize the cost for a specified value of the variance. The cost function is given by

$$C = a + \sum_{i}^{k} n_i c_i$$

Where, a is the observed cost which is constant, c_i is the average cost of surveying one unit in the i^{th} stratum.

Therefore, the required sample size in different stratum is given by

$$n_i = n\frac{\frac{N_i S_i}{\sqrt{c_i}}}{\sum_{i}^{k}\frac{N_i S_i}{\sqrt{c_i}}} \qquad (2.4)$$

Where, n= sample size for the study, N_i = population size for the study, S_i =variance of the i^{th} stratum.

If the average cost of surveying per unit (i.e. c_i) is the same in all the strata, then, the optimum allocation becomes the Neyman allocation. As cost of expenditure such as printing of questionnaires, sending and collecting of questionnaires etc. for different categories of schools during the survey by us are almost the same, therefore, we can use Neyman allocation in order to determine size of sample for each categories of school. So, in our case, the sample size in different categories of schools is given by a simplified form of (2.4) which is given by

$$n_i = \frac{nN_iS_i}{\sum N_iS_i} \ ,$$

Where, $S_i^2 = \frac{N_i}{N_i - 1} P_iQ_i$ is the population variance of the i^{th} stratum·

N_i = population size of i^{th} stratum,

P_i = proportion of students who secured 50% or more mark in annual examination in i^{th} stratum

$$= \frac{number\ of\ students\ in\ i^{th}\ category\ of\ school\ who\ secured\ 50\%\ or\ more\ marks\ in\ mathematics}{total\ number\ of\ students\ in\ i^{th}\ category\ of\ school}$$

and $Q_i = 1 - P_i$.

Following table illustrates the distribution of the sizes of samples in different strata for proportional and optimum allocation methods which were calculated on the basis of above discussion.

Table 2.4:

Distribution of sample students by category of schools

Categories of school	Total students		
	N_i	n_i (Prop)	n_i (Opt)
Govt.(SEBA)	5609	163	181
Private(SEBA)	3498	102	106
Private(CBSE)	4084	119	97
TOTAL	13191	384	384

2.03 Calculation of variances:

The formula to calculate variances of mean for different sampling methods are given below:

i) For simple random sampling:

$$Var(\hat{\mu})_R = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

where, $s^2 = \frac{n}{n-1}pq$

p = proportion of Mark in annual examination who secured 50% and above in mathematics in all the selected schools, $q = 1-p$, N = population size, n = sample size.

ii) For stratified random sampling:

$$Var(\hat{\mu})_{St} = \frac{1}{N^2}\sum N_i(N_i - n_i)\frac{S^2{}_i}{n_i}$$

where, $S_i^2 = \frac{N_i}{N_i - 1}P_iQ_i$

N =Total population size, N_i =population size of ith stratum, n_i =sample size of i^{th} stratum,

a) For proportional allocation:

$$Var(\hat{\mu})_{St(prop)} = \sum\frac{N_i}{N}\frac{S_i^2}{n}\left(\frac{N-n}{N}\right),$$

b) For optimum allocation :

$$Var(\hat{\mu})_{St(opt)} = \frac{(\sum w_iS_i)^2}{n} - \frac{\sum w_iS_i{}^2}{N} \quad \text{where} \quad w_i = \frac{N_i}{N}$$

Following table shows the variances of all the schools through different methods.

Table 2.5

Table showing variances:

Method	$Var(\hat{\mu})_R$	$Var(\hat{\mu})_{St(prop)}$	$Var(\hat{\mu})_{St(opt)}$
Variances	0.00060839	0.0004673	0.00046

2.04 Gain in efficiency (GE) in stratified random sampling over simple random sampling without replacement :

In order to observe how the sample size gets affected due to different types of allocation, an analysis on gain in efficiency (GE) due to different types of allocations is utmost required.

1) Gain in Efficiency (GE) due to proportional allocation :

$$GE_{prop} = \frac{Var(\hat{\mu})_R - Var(\hat{\mu})_{(St)prop}}{Var(\hat{\mu})_{(St)prop}} = \frac{0.00060839 - 0.0004673}{0.0004673} = 0.3017333 = 0.30$$

2) Gain in Efficiency (GE) due to optimum allocation :

$$GE_{opt} = \frac{Var(\hat{\mu})_R - Var(\hat{\mu})_{(St)opt}}{Var(\hat{\mu})_{(St)opt}} = \frac{0.00060839 - 0.00046}{0.00046} = 0.3223913 = 0.32$$

From the above results it can be said that optimum allocation provides little better estimates as compared to proportional allocation. But, the most serious drawback of optimum allocation is the absence of the knowledge of the population variances i.e. S_is of different strata in advance. In that case, the calculations are carried out by performing a pilot survey and by drawing simple random samples without replacement from each stratum as suggested by P. V. Sukhatme (1935) [51].

Due to the above mentioned drawback, the allocation of sample size to different strata for our study has been calculated by proportional allocation method. As shown above, by using this method we have gained an efficiency of 0.30 over the simple random sampling.

After examining the gain in efficiency (GE) for allocation of sample size to each category of school, students were selected randomly from different schools within that category. In the present study, students were selected from each school by using Cochran formula at 95%confidence level with ±15% margin of error. Out of these 13 schools, 6 are from Government SEBA; 3 are from Private. SEBA and 4 are from Private CBSE schools. In case of Private CBSE schools total sample size is 119. But when students of 4 schools are taken into consideration, it becomes 131. Hence, to make it 119, from each of the 4 schools three students were not taken into account. Following table illustrates the distribution of the sample by gender and category of schools.

Table 2.6

The distribution of sample size for class VIII students of different schools of Guwahati

Category of schools	**Sl. No.**	Name of school	Population size	Sample size (max)	Allotted sample size		
					Boys	Girls	Total
SEBA (Govt.)	1	Ulubari H.S.	95	30	16	14	30
	2	Dispur Vidyalaya	88	29	16	13	29
	3	Ganesh Mandir Vidyalaya	112	31	17	14	31
	4	Noonmati M.E. School	79	28	12	16	28
	5	Uzan Bazaar Girls' School	43	22	_	22	22
	6	Arya Vidyapeeth High School	46	23	23	_	23
SEBA (Pvt.)	7	Nichol's School	125	32	22	10	32
	8	Asom Jatiya Vidyalaya	200	36	26	10	36
	9	Holy Child School	170	34	_	34	34
CBSE(Pvt.)	10	Gurukul Grammar School	154	34	14	17	31
	11	Maharishi Vidya Mandir School	160	34	17	14	31

	12	Sarala Birla Gyan Jyoti	115	31	13	15	28
	13	Shankar Academy	118	32	17	12	29
Total					193	191	384

2.05 Comparative study of effect of bias in the context of data of our study:

It is well known that during the collection of sample units, both sampling and non-sampling errors creep into the process. The non sampling errors occur because the procedures of observation (data collection) may not be perfect and their contributions to the total error of survey may be substantially large, which may affect survey results adversely. On the other hand, the sampling errors arise because a part (sample) from the whole (population) is taken for observation in the survey. Since in our study sample size is 384, which is quite large, hence, by virtue of the Central Limit Theorem (CLT) we can use normal probability table to calculate the effect of bias for the questionnaires used in order to collect the data.

The total error is expressed as:

$$Total\ \ Error\ (TE) = \sqrt{Mean\,Square\,Error\,(MSE)} = \sqrt{Variance\,of\ mean + Square\,of\ \ Bias}$$

Again, Bias is the difference between the estimated value of population mean and sample mean.

Even with estimators that are un-biased in probability sampling, errors of measurement and non response may produce biases in the numbers that we compute from the data.

To examine the effect of bias, let us suppose that the estimate $\hat{\mu}$ is normally distributed about a mean m and is at a distance B from the true population value μ. Therefore, the amount of bias is $B = m - \mu$. As a statement about the accuracy of the estimate, we declare that the probability that the estimate $\hat{\mu}$ is in error by more than 1.96σ is 0.05.

This can be calculated with the help of the following transformation

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu+1.96\sigma}^{\infty} e^{\frac{-(\mu-m)}{2\sigma^2}} d\mu = \phi(\mu + 1.96\sigma)$$

Now putting

$$\mu - m = \sigma t$$

in above integral, we get lower limit of the range of integration for 't', as

$$\frac{\mu - m}{\sigma} + 1.96 = 1.96 - \frac{B}{\sigma},$$

where, $B = m - \mu$ is the amount of bias that occurs for adjusting the sample size for each strata.

Thus, we require to calculate bias by consulting the normal probability table with the help of the following:

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{1.96-\frac{B}{\sigma}}^{\infty} e^{\frac{-t^2}{2}} dt = \phi\left(1.96 - \frac{B}{\sigma}\right)$$

In table 2.7, effect of a bias B on the probability of an error greater than 1.96σ has been shown in tabular form. The calculations were carried out using the normal probability table. Data provided in table 2.7 are plotted in figure 2.4 where Praobability of error (less than -1.96 σ) and (greater than 1.96 σ) are plotted against B/σ values (x-axis).

Table 2.7

Effect of a Bias B on the probability of an error greater than 1.96σ

B/σ	Probability of error		Total
	<-1.96σ	>1.96σ	
0.01	0.0244	0.0256	0.0500

0.03	0.0233	0.0268	0.0501
0.05	0.0222	0.0281	0.0503
0.07	0.0212	0.0294	0.0506
0.09	0.0202	0.0307	0.0509
0.10	0.0197	0.0314	0.0511
0.25	0.0136	0.0436	0.0572
0.40	0.0091	0.0594	0.0685
0.55	0.0060	0.0793	0.0853
0.70	0.0039	0.1038	0.1077
0.85	0.0025	0.1335	0.1360
1.00	0.0015	0.1685	0.1700
1.50	0.0003	0.3228	0.3231

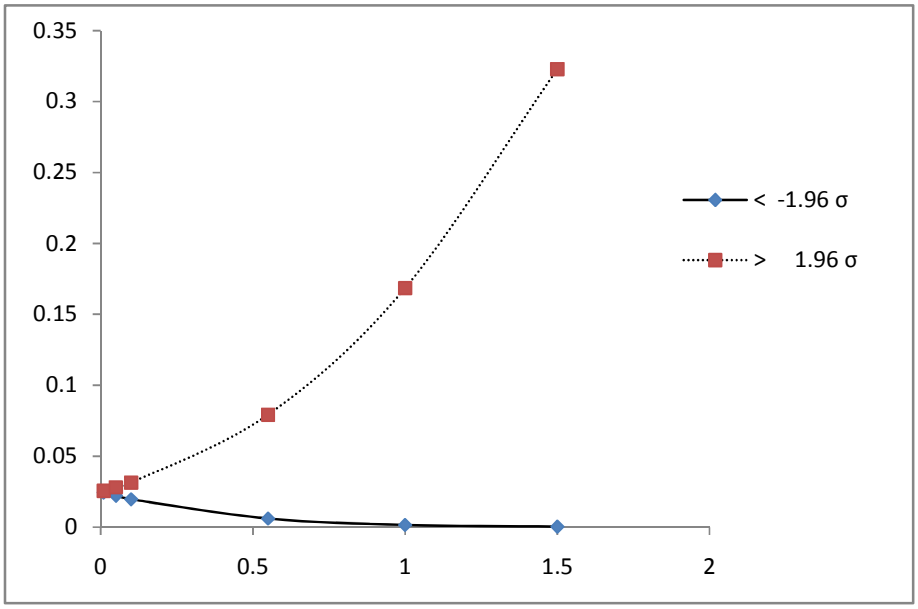


Fig. 2.5:

Praobability of error (less than -1.96 σ) and (greater than 1.96 σ) vs B/σ values (x-axis)

(Generated from the above table)

It is known that in order to compare a biased estimator with an unbiased estimator, or two estimators with different amounts of bias, a useful criterion is the mean square error (MSE) of the estimates, measured from the population values that are being estimated.

The relationship between MSE and Bias is given by

$$MSE(\hat{\mu}) = Variance\ of\ (\hat{\mu}) + (Bias)^2$$

In the following tables variances for different categories of schools of Guwahati, included in the sample are shown. Total sample size and sample sizes in different strata has been calculated with margin of error ±0.05. But, while calculating the sample sizes in the 13 selected schools, the margin of error was taken to be ±0.15; because greater precision requires larger sample sizes, which is not practicable in case of selection of sample from different schools. For this difference in precision, some bias may occur in the process and hence it becomes very important to calculate the bias and its effect.

Table 2.8

Variances for different categories of schools

Strata	Sample size, n	No. of students securing 50 or more	P	Q	Variances
SEBA Govt.	163	55	.34	.66	.00134493
SEBA Pvt.	102	74	.73	.27	.00189458
CBSE Pvt.	119	102	.86	.14	.000990608

In the following tables probability of an absolute error $\geq 1\sqrt{MSE}$ and $1.96\sqrt{MSE}$ for different categories of schools are given. Below each table, graphs of MSE, $1\sqrt{MSE}$ and $1.96\sqrt{MSE}$ versus B/σ values (in x axis) are shown

Tables showing probability of an absolute error ≥ 1√MSE and 1.96√MSE

Table 2.9: For SEBA Govt.:

V=0.00134493, p=0.34, q=0.66

$\frac{B}{\sigma}$	MSE	$1\sqrt{MSE}$	$1.96\sqrt{MSE}$
0.01	0.00384493	0.0620075	0.121535
0.03	0.00384493	0.0620075	0.121535
0.05	0.00384493	0.0620075	0.121535
0.07	0.00394493	0.0628087	0.123105
0.09	0.00394493	0.0628087	0.123105
0.10	0.00394493	0.0628087	0.123105
0.25	0.00444493	0.0666703	0.130674
0.40	0.00604493	0.0777492	0.152388
0.55	0.00864493	0.0929781	0.182237
0.70	0.0129449	0.113776	0.223001
0.85	0.0198449	0.140872	0.276109
1.00	0.0302449	0.173911	0.340865
1.50	0.105745	0.325184	0.637362

Fig 2.6 : For SEBA Govt.:

$\frac{B}{\sigma}$ values (x-axis) vs MSE, $1\sqrt{MSE}$ and $1.96\sqrt{MSE}$

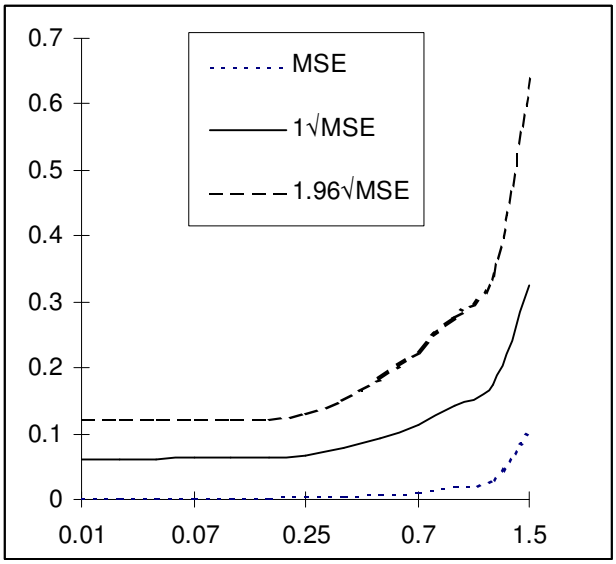


Table 2.10: For SEBA Pvt.

V=0.00189458, p=0.73, q=0.27

$\frac{B}{\sigma}$	MSE	$1\sqrt{MSE}$	$1.96\sqrt{MSE}$
0.01	0.00439458	0.0662916	0.129932
0.03	0.00439458	0.0662916	0.129932
0.05	0.00439458	0.0662916	0.129932
0.07	0.00449458	0.0670416	0.131402
0.09	0.00449458	0.0670416	0.131402
0.10	0.00449458	0.0670416	0.131402
0.25	0.00499458	0.0706723	0.138518
0.40	0.00659458	0.081207	0.159166
0.55	0.00919458	0.0958884	0.187941
0.70	0.0134946	0.116166	0.227686
0.85	0.0203946	0.14281	0.279907
1.00	0.0307946	0.175484	0.343948
1.50	0.106295	0.326028	0.639016

Fig 2.7 : For SEBA Pvt.

$\frac{B}{\sigma}$ values (x-axis) vs MSE, $1\sqrt{MSE}$ and $1.96\sqrt{MSE}$

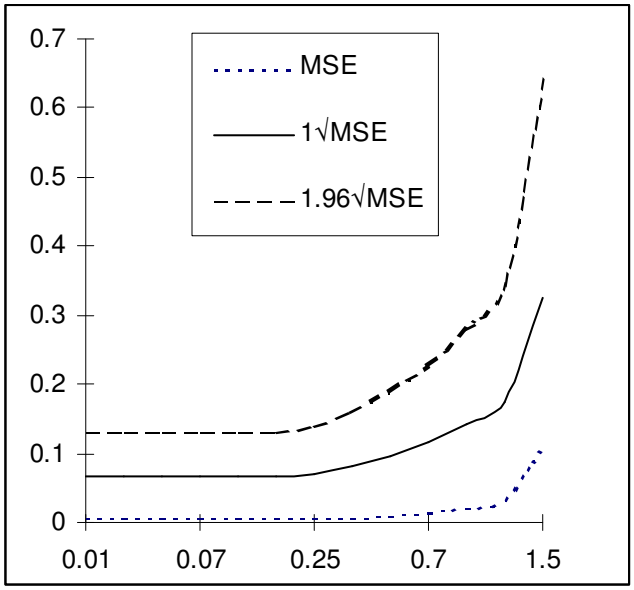


Table 2.11: For CBSE Pvt.

V=0.000990608, p=0.86, q=0.14

$\frac{B}{\sigma}$	MSE	$1\sqrt{MSE}$	$1.96\sqrt{MSE}$
0.01	0.00349061	0.0590814	0.115799
0.03	0.00349061	0.0590814	0.115799
0.05	0.00349061	0.0590814	0.115799
0.07	0.00359061	0.0599217	0.117447
0.09	0.00359061	0.0599217	0.117447
0.10	0.00359061	0.0599217	0.117447
0.25	0.00409061	0.0639579	0.125357
0.40	0.00569061	0.0754361	0.147855
0.55	0.00829061	0.0910528	0.178463
0.70	0.0125906	0.112208	0.219927
0.85	0.0194906	0.139609	0.273633
1.00	0.0298906	0.172889	0.338862
1.50	0.105391	0.324639	0.636293

Fig 2.8 : For CBSE Pvt.:

$\frac{B}{\sigma}$ values (x-axis) vs MSE, $1\sqrt{MSE}$ and $1.96\sqrt{MSE}$

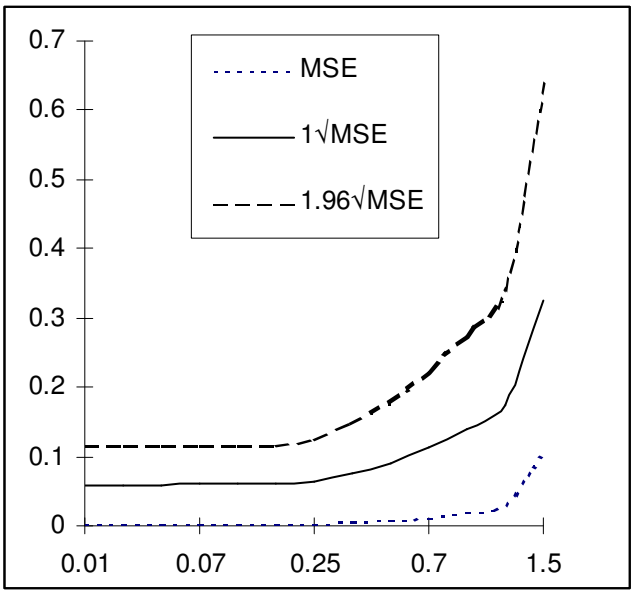


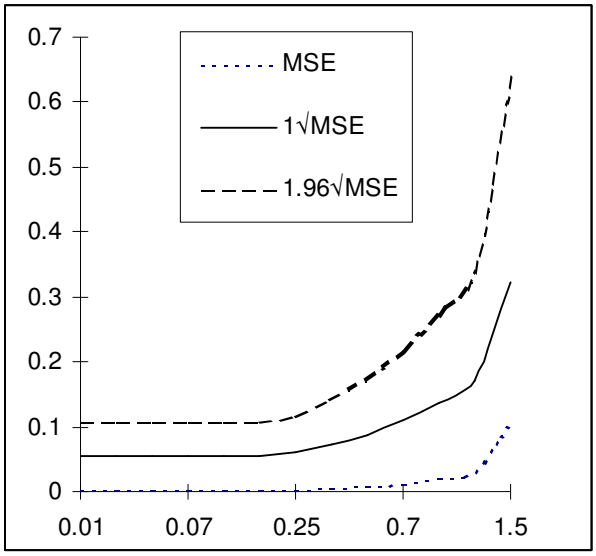
Table 2.12:

For All the schools: V=0.0004673 , p=0.60, q=0.40

$\frac{B}{\sigma}$	MSE	$1\sqrt{MSE}$	$1.96\sqrt{MSE}$
0.01	0.0029673	0.0544729	0.1067668
0.03	0.0029673	0.0544729	0.1067668
0.05	0.0029673	0.0544729	0.1067668
0.07	0.0030673	0.0553832	0.108551
0.09	0.0030673	0.0553832	0.108551
0.10	0.0030673	0.0553832	0.108551
0.25	0.0035673	0.0597268	0.1170645
0.40	0.0051673	0.0718839	0.1408924
0.55	0.0077673	0.0881322	0.1727391
0.70	0.0120673	0.1098512	0.2153083
0.85	0.0189673	0.1377218	0.2699347
1.00	0.0293673	0.1713689	0.335883
1.50	0.1048673	0.3238322	0.6347111

Fig 2.9 : For All the schools

$\frac{B}{\sigma}$ values (x-axis) vs MSE, $1\sqrt{MSE}$ and $1.96\sqrt{MSE}$

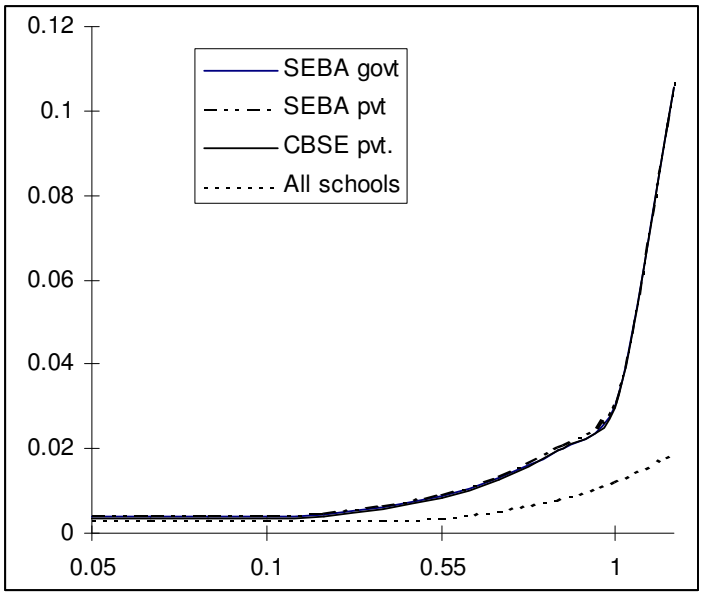


The following figure shows the comparison between the MSE of different categories mentioned above.

Fig 2.10:

Comparision for all the schools:

$\frac{B}{\sigma}$ values (x-axis) vs MSE of all categories of schools



Use of the MSE as criterion to determine the accuracy of an estimator amounts to regarding two estimates that have the same MSE are equivalent. It has been shown by Hansen, Hurwitz and Madow that if for $\frac{B}{\sigma}$, MSE is less than one half, then the estimator can be considered almost identical with its true value [52]. The tables 2.9, 2.10, 2.11, 2.12 and their corresponding graphs in figures 2.5, 2.6, 2.7 and 2.8 highlights this criterion in case of our study. So, we can conclude that the effect of bias in our study is negligible and the estimations derived from the selected samples will be in good agreement with their corresponding values for the whole population.

2.06 Conclusions:

There are different formulae given by different educationists for the determination of appropriate sample sizes. The researcher should choose the formula according to their needs and convenience. In choosing the right one, the researcher has to take into consideration about the maximum budget, time limit, nature of the study along with desired level of precision, confidence level and variability within the population of interest. Using an adequate sample along with high quality data collection will result in more reliable and valid results.