INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

The Impact of Sample Size on the Probability Samples to Estimate the Total population Number

Mohammedelameen Eissa Qurashi[1], **Mubarak H Elhafian**[2]*

1 Sudan University of Science & Technology, Faculty of Science, Department of Statistic, Sudan
2 King Abdul-Aziz University, Collage of Science and Arts Department of Mathematics, Rabigh, Saudi Arabia

OPEN ACCESS

Received: 05-02-2022
Accepted: 22-09-2022
Published: 17-10-2023

Citation: Qurashi ME, Elhafian MH (2023) The Impact of Sample Size on the Probability Samples to Estimate the Total population Number . Indian Journal of Science and Technology 16(39): 3316-3324. https://doi.org/10.17485/IJST/v16i39.303

* **Corresponding author**.

hafian10@yahoo.com, mhelhafian@kau.edu.sa

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: to determine the best sampling technique based on the allowable error. **Method:** Data of people infected with malaria in Khartoum State during the year 2019 was considered. The data were taken from the Ministry of Health, Khartoum State Sudan, where the population is divided into four stratums. The mean, the total number, and the confidence interval were estimated. A comparison between simple, stratified and systemic random sampling in terms of the accuracy of the estimate $V(\bar{y}_{\text{ran}}) = 1.59 \geq V(\bar{y}_{\text{prop}}) = 0.71 \geq V(\bar{y}_{\text{opt}}) = 0.04$, and the effect of marginal error in determining the sample size was known. **Findings:** The study found that when a smaller allowable error is used, the sample size increases and the estimates are more accurate. **Novelty:** Through the study, we recommend the researchers for using stratified sampling with the optimum distribution to get better accuracy.

Keywords: Allowable Error; Stratified Random Sampling; Population Total Number; Mean Variance; Sample Size

1 Introduction

Samples have become the basic importance for many theoretical and experimental studies. The most important question for the researchers is "what is the appropriate sample size do I need?" [(1)]. Calculation of sample size has a lot of application in the fields, assuring validity, accuracy, reliability [(2)]. In addition, industrial establishments use samples to monitor the quality of their production to check the progress of production according to the required specifications, and to study random behavior such as the demand for their products compared to other competing products [(1)].

The research aims to find out the effect of the sample size on estimating estimators of random samples and the effect of marginal error in determining the sample size.

2 Methodology

This type of study relies on identifying the problem mainly, then identifying the causes of the problem and making appropriate recommendations that include solutions. For the purpose of applying the effect of the sample size on the quantities of probability

samples (simple, stratified, and systemic), the mean, the total number, and the construct the confidence interval were estimated, and then a comparison between simple, stratified and systemic random sampling in terms of the accuracy of the estimate, and the effect of marginal error in determining the sample size was known.

In order to identify the main goals of the study problem, the following hypotheses must first be tested: The sample size effect on the probability sampling estimators, using higher allowable error gives smaller the sample size and vice versa and the relative efficiency of the sample decreases as the permissible error increases.

2.1 Estimation of Sample Size

To specify the sample size is one of important things in research [(2)]. If the sample size is very large, the costs of data size and its organization will be high, and the large sample requires great efforts [(3)]. On the other hand, if we use a relatively small sample, we may sacrifice the accuracy of the results, just as our use is limited. Therefore, we are concerned with determining the size of the sample and this is only done by knowing the following things, the most important of which are: What do we expect in terms of the limits of error we want and the areas of use of their results. If the community is made up of different sections and information is required for the different departments, then the sample size can be calculated for each section separately by summing, so we get the total sample size [(3)].

2.2 Sample size to estimate the population mean

We will assume that the arithmetic mean of the sample which is an unbiased estimate of the population mean $(\bar{Y})$ [(4)]. By making use of the concept of the limits of confidence, we can obtain a probabilistic expression that relates to $(\bar{Y})$with $(\bar{Y})$and let us suppose that we allow a measure of (d) of error in our estimation of the population mean (α), so the probabilistic expression of error is

$$P_r[|\bar{y}+\bar{Y}| \geq d] = \alpha \tag{1}$$

This equation can be expressed in another form:

$$P_r[|\bar{y}+\bar{Y}| \geq d] = 1-\alpha$$

$$p_r\left[-d < \bar{y}-Y^{-} < d\right] = 1-\alpha$$
$$p_r\left[\bar{y}+d < -Y^{-} < -\bar{y}+d\right] = 1-\alpha$$
$$p_r\left[\bar{y}-d < Y^{-} < \bar{y}+d\right] = 1-\alpha$$

That is, the probability that the population mean falls between the arithmetic mean of the sample ± allowable error is equal to one minus the allowable error probability. As we know from our study, $\bar{Y}$ is roughly distributed according to the normal distribution with a mean $(\bar{y})$ and a variance $V(\bar{y})$then [(5)].

$$d = t\sqrt{V(\bar{y})} \tag{3}$$

Where (t) is the tabular value to the normal distribution which corresponds to the level of significance $(1-\alpha)$ and the variance of the population mean is:

whereas:

$f = \frac{n}{N}$ (sampling fraction)

$$d^2 = t^2\frac{\sigma^2}{n}(1-f) = \frac{t^2\sigma^2}{n} - \frac{t^2\sigma^2}{N} = \frac{Nt^2\sigma^2 - nt^2\sigma^2}{nN}$$
$$= nNd^2 + nt^2\sigma^2 = Nt^2\sigma^2 = n\left(Nd^2 + t^2\sigma^2\right) = Nt^2\sigma^2$$

$$n = \frac{Nt^2\sigma^2}{Nd^2 + t^2\sigma^2} = \frac{\left(\frac{t\sigma}{d}\right)^2}{1+\frac{1}{N}\left(\frac{t\sigma}{d}\right)^2} \tag{4}$$

This formula provides us with sample size (n) that allows for allowable error (d) with probability (α). If the population size is large then we will assume that:

$$\frac{1}{N}\left(\frac{t\sigma}{d}\right)^2 \approx 0$$

The initial sample size is

$$n_0 = \frac{(t\sigma)^2}{d^2} \tag{5}$$

In practice, if it is not possible to make this assumption, i.e. if (n) is large, then the final sample size required is

$$n = \frac{n_0}{1+\frac{n_0}{N}} \tag{6}$$

2.3 Simple random sampling [SRS]

SRS can be defined with a size (n) of a population of (N) which is the selection of (n) from the items without replacement the item drawn from the sum of (N) from the items and the number of samples is C_n^N therefore, each sample of size (n) has the same chance in testing.

The probability of selecting any sample is $\frac{1}{C_n^N}$ this type of sampling is called simple random sampling, and sometimes it is called unrestricted random sampling [6]

- **The mean:**

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{7}$$

- **The Mean variance:**

$$V(\bar{y}) = \frac{\sigma^2}{n}(1-f) \tag{8}$$

Where, Population total number

$$\hat{Y} = N\bar{y} \tag{9}$$

2.4 Stratified Random Sampling

Stratified random sampling, in which the population is divided into stratums, and a random sample of a certain size is withdrawn from each stratum, meaning that we consider each stratum as an independent population and call these sections into which the study population is divided into strata. This method gives confirmation of where the sample represents all stratums of population as well as the selection. A sample of each layer requires a frame for each layer separately, and notes this information was not required in the simple random sampling [7]

In general, in the Stratified sampling, we divide the population whose size is (N) to (L) from stratums of sizes (N_1, N_2,…, N_L) in order and all of these stratums are considered homogeneous non-overlapping societies, all of which are the original population, that is:

$$N = N_1 + N_2 + \ldots + N_L = \sum_{h=1}^{L} N_h$$

Then we draw simple random samples inside the stratums, their sizes $n_1, n_2, \ldots, n_L$ in a row so that

$$n = n_1 + n_2 + \ldots + n_L = \sum_{h=1}^{L} n_h$$

There are several methods of distribution the sample:

- **Proportional Allocation**

The allocation of the sample on each stratum shall be on the basis that it is proportional to the number of the total sampling units in the stratum, i.e.:

$$\frac{n_h}{n} = \frac{N_h}{N}$$

$h = 1, 2, \ldots, L$

- **The mean**

$$\bar{y}h = \frac{\sum_{i=1}^{nh} yh_i}{nh} \tag{11}$$

- **The Mean variance**

$$V_{p\,\mathrm{rop}}\ (\bar{y}_{st}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h \sigma_h^2 \tag{12}$$

Where: $W_h$strata weight

- **Optimum Allocation**

The allocation of the sample to the different stratums is done according to the following formula:

$$n_h = n \frac{w_h \sigma}{\sum w_h \sigma_h} \tag{13}$$

This is called Neyman Allocation

- **The Mean variance**

$$V\left(\bar{y}_{\mathrm{opt}}\right) = \frac{\sum_{h=1}^{L} \left(w_h \sigma_h\right)^2}{n} - \frac{\sum_{h=1}^{L} w_h {\sigma_h}^2}{N} \tag{14}$$

2.5 Comparison of precision of proportional stratified sampling, optimum and Simple random sampling

If we use the stratified random sampling accurately, it gives more accurate results than those of the simple random sampling, where we obtain a variance of the estimated mean or the estimated total value less than the variance of the mean or the estimated total value in the simple random sampling, but it should be noted that not every stratified sample leads to a variance. Less than simple random sampling as if differs significantly from the optimum distribution. The stratified sampling may lead to greater variance than the simple random sampling. Then compare the stratified sampling proportional to the stratified for the optimal distribution, then compare the three different methods, and accordingly[8]:

- **Simple random sampling variance**

$$V(\bar{y}) = \frac{\sigma^2}{n}(1-f) \tag{15}$$

- **Variation of proportional stratified mean**

$$V_{p\,\mathrm{rop}}\ (\bar{y}_{st}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h \sigma_h^2 \tag{16}$$

- **Variation of proportional stratified mean**

$$V\left(\bar{y}_{\mathrm{opt}}\right) = \frac{\sum_{h=1}^{L}\left(w_h \sigma_h\right)^2}{n} - \frac{\sum_{h=1}^{L} w_h {\sigma_h}^2}{N} \tag{17}$$

As

$$V_{\mathrm{prop}}\ (\bar{y}_{st}) \leq V_{\mathrm{ran}}\ (\bar{y})$$

$$V_{opt}\,(\bar{y}_{st}) \leq V_{pro\ p}\,(\bar{y}_{st}) \tag{18}$$

$$\therefore V_{\mathrm{ran}}\ (\bar{y}) \geq V_{\mathrm{prop}}\ (\bar{y}_{st}) \geq V_{\mathrm{opt}}\ (\bar{y}_{st}) \tag{19}$$

2.6 Systematic Sampling

It gives a sample with equal distances between the elements, and therefore it is expected to give a more accurate estimate of the population mean if we use a random sample, unless the units that make up the sample are equal or related to each other [9]. The Systematic sample is widespread and widely used in practical applications due to its low costs and ease of conducting even easier than simple random sampling as well as fewer errors in testing the sample units [10].

- **The mean**

$$\bar{y}_{ij} = \frac{1}{k}\sum_{i=1}^{k} y_{ij} \tag{20}$$

- **The Mean variance**

$$V\left(\bar{y}_{sy}\right) = \frac{1}{k}\sum_{i=1}^{k}\left(\overline{y_i} - \bar{Y}\right)^2 \tag{21}$$

2.7 Comparison between Systematic Sampling, Stratified sampling and Simple random sampling

The comparison depends on the size of the stratum where it is often for large samples the variance of is greater than the mean variance of the simple random sample [11].

3 Result and Discussion

3.1 Data Set

The data of the study is the numbers of people infected with malaria in Khartoum State Sudan during the year 2019 were taken from the Ministry of Health, Khartoum State, where the population infected with malaria in Khartoum State is divided into four stratums as follows:

- **First strata**: the numbers of people with malaria registered in the health centers of charitable organizations.
- **Second strata:** the numbers of people infected with malaria registered in government health centers.
- **Third strata:** the numbers of people infected with malaria registered in central government hospitals.
- **Fourth strata**: the numbers of people infected with malaria registered in terminal government hospitals.

Table 1. Description of the study population

Strata	n	Y	$\bar{Y}_h$	σ_h	W_h
First	84	69557	828.06	678.27	0.170
Second	84	34671	412.75	356.96	0.170
Third	180	11182	62.12	96.46	0.370
Fourth	144	6623	45.99	75.51	0.290
Total	492	122033	248.03	437.26	1.00

3.2 Description of the study population (strata)

The study population represented by the stratums will be described, and the stratum size (N_h), the sum for each stratum (Y), the mean ($\bar{Y}_h$), the standard deviation of the stratum (σ_h) and stratum weight (W_h) in following table:

3.3 Estimate sample size

Using equation (3) and assume that the probability of 5% and an allowable error (0.01,0,0.05,0.10,0.15,0.20) from the mean of population the results are as in the following table

Table 2. The estimated sample size according to allowable error

Allowable error d	$\mathbf{n}_0$	Sample size
0.01	119423.08	490
0.05	4776.92	446
0.10	1194.23	348
0.15	1383.84	363
0.20	298.48	185

From Table 2, we note that when the allowable error increase in the sample size will increase,

3.4 Estimation by simple random sampling method

Sample size of (490 is selected to estimate the mean and the total number

Table 3. Description Measures of SRS method

SRS y_i	$\bar{y}_h$	$S^2{}_h$	Mean variance	population total number	Upper limit	Lower limit
121895	248.77	191852.76	1.57	123494.84	123610.8	121178.88

From Table 3 the total number of people infected with malaria in population is not less than (123610.8) nor greater than (121178.88).

3.5 Estimation by stratified random sampling method

For estimation by stratified sampling, we use proportional and optimal distribution as follows:

3.5.1 ***Optimal distribution***

Sample is selected with a proportional method as follows

$N = 492 \quad h = 1, 2, 3, 4$

Table 4. Measures of stratified random sampling Optimal distribution

Stratified random sampling	$\bar{y}_{st}$	Mean variance	Population total number	Upper limit	Lower limit
Optimal distribution	248.31	0.71	122168.52	122978.55	121358.49

From the above table estimate, we find that the total number of people infected with malaria is neither less than (122978.55) nor greater than (121358.49). And also we find the mean estimated by stratified sampling with proportional distribution (248.31)

very close to the population mean (248.03).

Table 5. Description of the sample selected by proportional distribution method

Strata	N_h	$\bar{y}_h$	S_h	W_h	n_h
First	83	269.96	336.61	0.170	82
Second	83	203.95	233.19	0.170	75
Third	180	264.91	309.11	0.370	195
Fourth	144	204.82	273.35	0.290	

3.5.2 ***Optimum Distribution (Neyman)***

A sample size is chosen by the optimal distribution method as follows:

From the above estimate, we find the mean estimated by stratified sampling with proportional distribution (248.08) very close to the population mean (248.03).

Table 6. Estimation of populatn total number

Simple random sampling Optimum Distribution (Neyman)	$\bar{y}_{st}$	Mean variance	population total number	Upper limit	Lower limit
	248.08	0.04	122055.36	122248.22	121862.50

From Table 6, we find that the total number of people infected with malaria is neither less than (122248.22) nor greater than (121862.50).

3.6 Comparison of the accuracy of stratified sampling with proportional, optimal and simple random distribution

$$V\left(\bar{y}_{\text{rean}}\right) = 1.59 \geq V\left(\bar{y}_{\text{prop}}\right) = 0.71 \geq V\left(\bar{y}_{\text{opt}}\right) = 0.04$$

We notice that the variance of the random sampling is the largest variance, followed by the variance of the stratified sampling by proportional distribution method, then stratified by the optimal distribution method. This means that the stratified sampling is more accurate than the random sampling. As we conclude, the theory is achieved.

3.7 Estimation by the method of systematic random sampling

A systematic sample of size (490) is chosen according to a proportional distribution, where a sample of size (83) is chosen from the first strata, (83) from the second, (181) from the third, and (143) from the fourth, and each strata is calculated as follows:

- **First strata**

$$K_1 = \frac{N_1}{n_1} = \frac{84}{83} = 1$$

- **Second strata**

$$K_2 = \frac{N_2}{n_2} = \frac{84}{83} = 1$$

- **Third strata**

$$K_3 = \frac{N_3}{n_3} = \frac{180}{181} = 1$$

- **Fourth strata**

$$K_4 = \frac{N_4}{n_4} = \frac{144}{143} = 1$$

Table 7. Estimating Total number of population using systematic random sampling

systematic random sampling	$\bar{y}_{sy}$	Mean variance	population total number	Upper limit	Lower limit
	244.62	0.39	120353.04	120950.92	119755.16

3.8 Comparison between systematic, stratified, and simple random sampling

For a comparison between stratified, systematic and random sampling, we use the ANOVA table as follows:

Table 8. Analysis of Variance

S.O.V	d.f	S.S	M.S
Between strata	3	674711.41	$69032.76 = {S_{wst}}^2$
Within strata	1956	135235175.39	
Total	1959	135909886.79	$69377.18 = S^2$

We can find the

$$V(\bar{y}) = \left(1 - \frac{490}{492}\right)\frac{69377.18}{490} = (0.0041)(141.59) = \underline{\underline{0.58}}$$

$$V(\bar{y}_{st}) = \left(1 - \frac{490}{492}\right)\frac{69032.76}{490} = (0.96)(140.08) = \underline{\underline{0.57}}$$

We notice that both systematic and stratified sampling are more accurate than simple random sampling, and systematic sampling is more accurate than stratified sampling.

3.9 Relative efficiency of the estimated variances

Table 9. Relative efficiency of the estimated variances of the sample types

allowable error (d)	n	$V(\bar{y}_{\text{ran}})$	$V(\bar{y}_{\text{prop}})$	$V(\bar{y}_{opt})$	$V(\bar{y}_{\text{sy}})$	$e_1 = \frac{V(\bar{y}_{opt})}{V(\bar{y}_{\text{ran}})}$	$e_2 = \frac{V(\bar{y}_{opt})}{V(\bar{y}_{\text{prop}})}$	$e_3 = \frac{V(\bar{y}_{opt})}{V(\bar{y}_{\text{sy}})}$
0.01	490	1.59	0.71	0.67	1.26	0.42	0.94	0.53
0.05	446	36.66	15.37	13.06	1.36	0.36	0.85	9.60
0.10	348	97.94	59.54	50.69	2.15	0.52	0.85	23.58
0.15	363	101.54	137.24	117.32	4.21	1.16	0.85	27.87
0.20	185	156.24	248.69	194.03	6.87	1.24	0.87	28.24

From Table 9 and Figure 1, we can notice that as allowable error increases, the relative efficiency of the sample decreases, and vice versa.

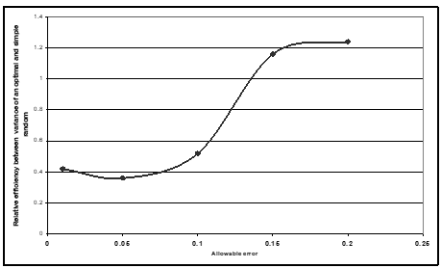

Fig 1. Shows compare the allowable error and relative efficiency of optimal variance and simple random

4 Conclusion

From the previous result we can conclude that increasing the allowable error gives a small sample size and the results of the estimation are less accurate, Large samples require a lot of time and effort, but give more accurate results. The estimation by the stratified random sampling method is more accurate than simple random sampling if the population is non homogeneous. In the optimal distribution, the less homogeneity in the stratum increases the number of units chosen from the stratum, meaning the size of the sample taken from the stratum in the optimum distribution depends on the variance of the stratum. Stratified sampling with optimal distribution is more accurate than simple random stratified sampling. Stratified sampling is more accurate than simple random sampling.

The study recommended that it is preferable to use stratified sampling if the community is not homogeneous; the larger the sample size, the greater the probability of the sample representing the study population.

References

1) Casteel A, Bridier NL. Describing Populations and Samples in Doctoral Student Research. *International Journal of Doctoral Studies*. 2021;16:339–362. Available from: http://ijds.org/Volume16/IJDSv16p339-362Casteel7067.pdf.
2) Memon MA, Ting H, Cheah JH, Thurasamy R, Chuah F, Cham TH. Sample Size for Survey Research: Review and Recommendations. *Journal of Applied Structural Equation Modeling*. 2020;4(2):i. Available from: https://jasemjournal.com/wp-content/uploads/2020/08/Memon-et-al_JASEM_-Editorial_V4_Iss2_June2020.pdf.
3) Wiśniowski A, Sakshaug JW, Ruiz DAP, Blom AG. Integrating Probability and Nonprobability Samples for Survey Inference. *Journal of Survey Statistics and Methodology*. 2020;8(1):120–147. Available from: https://doi.org/10.1093/jssam/smz051.
4) Naing L, Nordin RB, Rahman HA, Naing YT. Sample size calculation for prevalence studies using Scalex and ScalaR calculators. *BMC Medical Research Methodology*. 2022;22(1):209. Available from: https://doi.org/10.1186/s12874-022-01694-7.
5) Raifman S, Devost MA, Digitale JC, Chen YH, Morris MD. Respondent-Driven Sampling: a Sampling Method for Hard-to-Reach Populations and Beyond. *Current Epidemiology Reports*. 2022;9(1):38–47. Available from: https://doi.org/10.1007/s40471-022-00287-8.
6) Gumpili SP, Das AV. Sample size and its evolution in research. *IHOPE Journal of Ophthalmology*. 2022;1:9–13. Available from: https://doi.org/10.25259/IHOPEJO_3_2021.
7) Kovacs M, Van Ravenzwaaij D, Hoekstra R, Aczel B. SampleSizePlanner: A Tool to Estimate and Justify Sample Size for Two-Group Studies. 2022. Available from: https://doi.org/10.1177/25152459211054059.
8) Lin N, Rusli N, Hanif A, Rahman. Sample size calculation for prevalence studies using Scalex and ScalaR calculators. 2022. Available from: https://doi.org/10.1186/s12874-022-01694-7.
9) Lohr SL. Sampling: design and analysis. Chapman and Hall/CRC.. 2021. Available from: https://doi.org/10.1201/9780429298899.
10) Sharma G. Pros and cons of different sampling techniques. *International journal of applied research*. 2017;3(7):749–752. Available from: https://www.allresearchjournal.com/archives/2017/vol3issue7/PartK/3-7-69-542.pdf.
11) Wongwilai S, Phudetch P, Saelek P, Khuptawatin A, Wongcharoensin K, Chaitongrat S, et al. The role of innovative ideas in business sustainability: Evidence from textile industry. *Uncertain Supply Chain Management*. 2022;10(1):285–294. Available from: https://doi.org/10.5267/j.uscm.2021.8.011.