*J Bangladesh Agril Univ 22(2): **146-157**, 2024* https://doi.org/10.3329/jbau.v22i2.74547

ISSN 1810-3030 (Print) 2408-8684 (Online)

Journal of Bangladesh Agricultural University

Journal home page: http://baures.bau.edu.bd/jbau

Research Article
Determining Adequate Sample Size for Social Survey Research

Md Kamrul Hasan[1]✉ and Lalit Kumar[2]

[1]Department of Agricultural Extension and Rural Development, Patuakhali Science and Technology University, Dumki 8602, Patuakhali, Bangladesh

[2]EastCoast Geospatial Consultants, Armidale, NSW 2350, Australia

ARTICLE INFO

Article history
Received: 18 December 2023
Accepted: 25 June 2024
Published: 30 June 2024

Keywords
Research,
Survey,
Sample size,
Sample control ratio,
Sampling applet

Correspondence
Md Kamrul Hasan
✉: kamrulext@pstu.ac.bd

ABSTRACT

Determination of a valid sample size is a fundamental step in research. This paper explains how existing formulas are tied in a single thread by applying the concept of standard error, margin of error, *Z* and *t* scores, confidence interval and sampling distribution. Bringing the concept of *sample control ratio*, we suggest a unified formula which is $n = \frac{Nt^2\rho^2}{N+t^2\rho^2}$ where I is the sample size, *N* is the population, *t* is the *t*-value at a desired level of probability with *df = (N-1)* and ρ is the sample control ratio to be estimated by $\frac{1}{6\varepsilon}$ for continuous variables and $\frac{\sqrt{p(1-p)}}{\varepsilon}$ for categorical variables where ε is the proportion of acceptable error and *p* is the proportion of presence of an attribute in the population. This formula does not need the finite population correction, and it has been derived from and consistent with existing formulas. A researcher does not need to calculate the error margin in absolute terms for this formula, and it is sufficient to provide only the proportion of error (e.g., 0.03 or 0.05). This paper should help social scientists, researchers, academicians and students determine the appropriate sample size for their research with greater confidence and clarity.

Introduction

Determining sample size is a common and very important step for any survey research work because a larger sample size increases cost, while a smaller sample size reduces precision. Determining adequate sample size from a population is one of the most fundamental tasks in research. A valid generalization of findings is always dependent on the sample size and sampling techniques used. Inferences are made based on the samples about the population without observing the entire population (Upton and Cook, 1996). Sample size should be large enough to ensure the minimum possible risk of accepting false hypothesis within an acceptable limit (Diamond, 1989). An appropriate sample size coupled with an appropriate sampling technique is the foundation of statistical manipulation of the gathered data in social surveys. A sample is a representative fraction of the population (an entire collection of subjects or units of study) (Moore et al., 2014). The subjects or units could be people, crops, animals, areas or anything else whose characteristics are studied in research. Sample size indicates the number of samples to be drawn from the population to make a fair generalization about the population after statistical analysis (Witte and Witte, 2017).

Attempts to determine the appropriate sample size is an age-old statistical strategy (e.g., Cochran, 1953; Haldane, 1945; Seelbinder, 1953; Stein, 1945). Most of the statistical books have touched on the formula of calculating sample size (e.g., Cochran, 1977; Kothari, 2004; Sampath, 2001; Yamane, 1967). However, individual books explain their own styles that lack the linkage among different formulas with different levels of accuracy. Karimnezhad and Parsian (2018) and Martin and Elster (2020) have applied Bayesian approach that results in a higher sample size with better accuracy. In a case-control study where relationship is measured based on odds-ratio, sample size formula would be different (Sambucini, 2000). Even more sophisticated approaches are sometimes suggested that can outperform usual random selection

Cite This Article
Hasan, M.K. and Kumar, L. 2024. Determining Adequate Sample Size for Social Survey Research. *Journal of Bangladesh Agricultural University,* 22(2): 146-157. https://doi.org/10.3329/jbau.v22i2.74547

or Bayesian approach (Heller et al., 2015). For example, tail distribution of sample distribution is checked against the corresponding population distributions to justify sample adequacy (Chou and Johnson, 1987). Besides, sample size estimation focusing on a single variable in a multivariate study could be misleading, and therefore multiple variables should be considered (Benedetti et al., 2019; Liu, 2013). The estimation of sample size for categorical variables are different from that for continuous ones (Huschens, 1990; Laga and Likeš, 1975). Therefore, selection of the appropriate formula for sample size determination is often confusing.

Various formulas are different, for example, the use of *Z* or *t* values, estimating the population variance, deciding on which type of variables should be used in determining the sample size. Bartlett II et al. (2001) and Israel (1992) explained the application of Cochran's (1977) and Yamane's (1967) formulas that provide a solid understanding of the formulas. Another widely used formula for estimating the sample size has been suggested by Krejcie and Morgan (1970) who did not provide a clear explanation of the derivation of their equation. Therefore, it is easy to get confused about which formula a researcher should use in their research. This paper explains the formulas of sample size determination in a consistent manner and layman's language so that researchers and students do not require the complex mathematical proof of the derivation of the formulas. This paper does not aim to invent a new formula, which is fundamentally different from existing ones, but to propose a unified formula, which is consistent with the available formulas. The proposed unified formula can be used for determining the minimum sample size, which will confine the sampling errors within the accepted limits and be applicable to social surveys having continuous or binary variables and small or large populations. This paper will help social scientist, researchers and students determine the appropriate sample size with more confidence and clear understanding of the process that builds up the formulas.

Sample size derived from standard error

When we draw a sample size of *n* from a population of *N*, specific statistics (e.g., mean, standard deviation and coefficient of variation) can be calculated from this sample. If we have *k* number of samples of n_k sizes, we can calculate *k* number of means from the *k* number of samples. These *k* number of means will allow us to create a distribution plot (Fig. 1) which is termed as the sampling distribution. Thus, the sampling distribution is a hypothetical distribution that represents the distribution of a statistic for an infinite number of samples. The sample, which we use in research, is just one of those infinite number of possible samples that could have been selected.

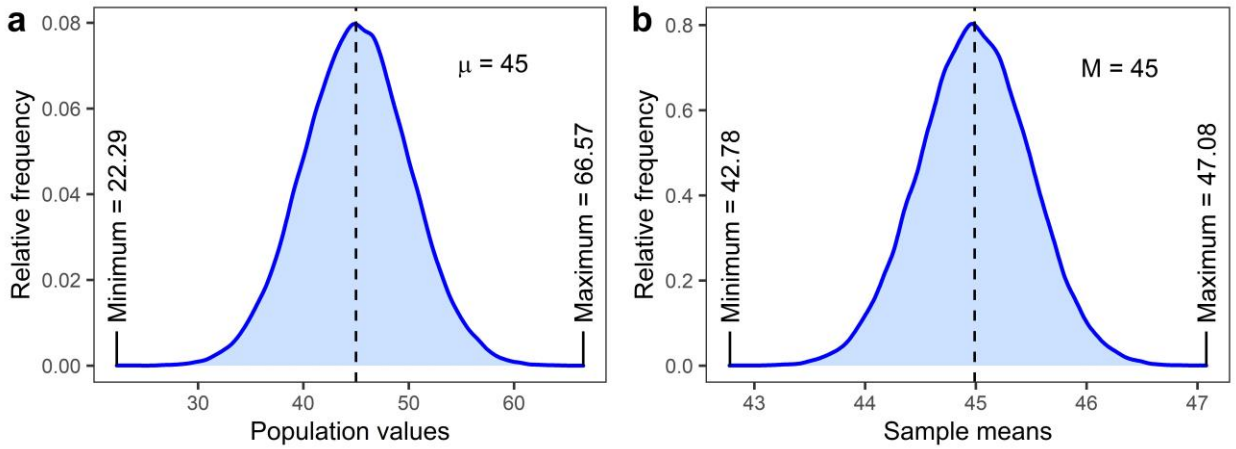

Fig. 1. Population *(N)* values and sample means, **a.** Normally distributed *N* values generated using R codes [set.seed(1); rnorm(100000, mean=45, sd=5)], **b.** Distribution of means of 100,000 samples (each of which has *n* = 100) randomly drawn from the population

It is very unlikely that the sample mean *(M)* will be equal to the population mean (μ). Fig. 1a has been constructed using a population *(N)* of randomly generated 100,000 values [$\sim N\ (\mu = 45, \sigma = 5)$], where, μ and σ have been assumed to be, for example, 45 and 5, respectively. From the *N*, we have drawn 100,000 samples each of which has 100 units (*n* = 100) and plotted the calculated 100,000 sample means in Fig. 1b. Figure 1 shows that the population mean (μ = 45) and the mean (*M* = 45) of the means of the randomly drawn 100,000 samples are equal. However, all the sample means are not equal to 45 rather they vary between 42.78 and 47.08. For example, the mean and standard deviation of the first sample is 45.26 and 5.09,

respectively, which are different from the population parameters. Therefore, it is very unlikely that a sample mean out of these 100,000-sample means will be equal to the population mean. A researcher wants to determine the centre of this sampling distribution (Fig 1b) where the best estimate of the true population mean should be. Figure 1 also illustrates that the average of the sampling distribution represents the population parameter (μ here for example). We can also calculate the standard deviation of the sampling distribution which is termed as the standard error (*SE*). In this example, the standard deviation of the sampling distribution is 0.5 which should be equal to the *SE* calculated from the sample standard deviation using Eq. 1 (Witte and Witte, 2017).

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{5.09}{\sqrt{100}} = 0.509 \approx 0.5 \quad \text{..... Eq. 1}$$

We have created 100,000 samples each with size 100, and 5.09 is the standard deviation of the first sample of these total samples. Although σ denotes the standard deviation of the population, we have used sample standard deviation in a practical sense because σ is unknown for most of the cases. However, this shows how *SE* can be estimated from sample values. Calculation in Eq. 1 depicts that the standard deviation of the sampling distribution (i.e., *SE* = 0.5) is almost equal to the *SE* calculated from population values. This standard error indicates a rough estimate of how the sample means deviate from the population mean. This implication of standard error helps make inferences about the population using the sample values.

In statistics, the standard deviation of a sampling distribution is called the *SE*. The concept of "Standard error" (in sampling, it is termed as "Sampling error") is central to the sampling theory and determination of sample size (Cochran, 1977; Kothari, 2004). Standard deviation is the dispersion of scores with respect to the average in a single sample, and standard error is the deviation of averages from the average of averages in a particular sampling distribution. We never actually create the sampling distribution. All we have to deal with is the sample standard deviation. The higher the standard deviation is, the higher will be the standard error (sampling error).

In our example, *SE* calculated from the population and sample standard deviations are almost equal. However, this will not be equal if the samples are not representative of the population. To make the sample representative, two criteria must be fulfilled: (a) sample size should be large enough, and (b) sample drawn should be random enough. This paper specifically focuses on the determination of sample size which will be large enough to draw a valid inference about the population with a given level of accuracy and small enough to avoid unnecessary expenses of time and resources.

A close examination of Eq. 1 shows that *SE* is inversely proportional to the square root of the sample size (*n*). Figure 2 shows the effect of sample size on the *SE*. Any increases in *n* results in a decrease in the *SE*, which is calculated from sample standard deviation as we do not have true population values. The *SE* decreases rapidly in the initial increases of *n*, which is very optimistic that a small increase in the *n* should give us a far better inference about the population. In practice, we do not need the *SE* (calculated from the sample) which is less than the true *SE* (calculated from the population). Therefore, we need to determine the optimum size of the sample.

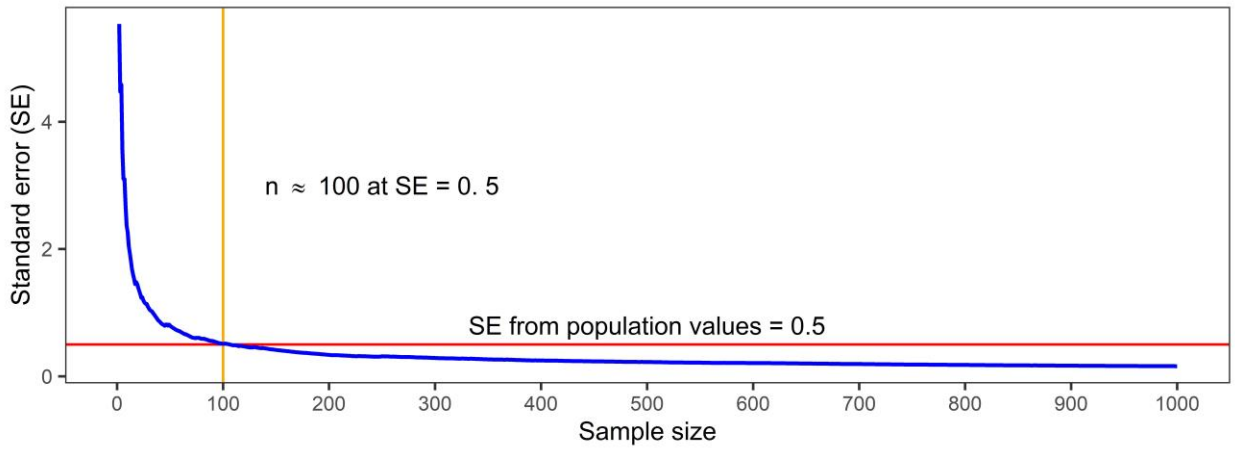

Fig. 2. Sensitivity of standard error to sample size

Estimating sample size from standard error

From Eq. 1, we can determine *n* by solving *n* as shown in Eq. 2.

$$SE = \frac{\sigma}{\sqrt{n}} \Longrightarrow n = \frac{\sigma^2}{SE^2} \quad \text{..... Eq. 2}$$

As we know all these values, $\sigma^2 = 5^2 = 25$ and $SE^2 = 0.5^2 = 0.25$, therefore *n* will be 100. This seems very simple but a problem arises when we do not know *SE* and σ. This issue can be solved by collecting data in two steps (Cochran, 1977; Seelbinder, 1953; Stein, 1945). In the first step, data is collected from a small fraction of the population (for example, n_0 = 30) that allows the researcher to estimate the σ and *SE*. The *n* is then calculated using these σ and *SE*. In the second stage, data is collected from the remaining (n-n_0) number of samples. So far, we have formulated the equations considering continuous variables where σ represents the standard deviation of the population. To use Eq. 2 for categorical variables (e.g., dichotomous or binary variables), σ has to be replaced by $\sqrt{p(1-p)}$ which is the standard deviation of a binary variable where *p* is the proportion of presence of a response. For example, if there are 45% businesspersons in a population, *p* of the businesspersons will be 0.45. Therefore, the *n* for a binary variable will follow Eq. 3.

$$SE = \frac{\sqrt{p(1-P)}}{\sqrt{n}} \Longrightarrow n = \frac{p(1-p)}{SE^2} \quad \text{..... Eq. 3}$$

However, this formula (Eq. 2 or Eq. 3) does not take population size into account which must be considered to make the samples representative of the population. Therefore, the margin of error needs to be incorporated in the equation of sample size determination.

Margin of error, confidence interval and choice between Z or t *scores*

The estimated sample size has crucial importance on the accuracy of estimated confidence intervals (Liu, 2013). Therefore, determination of population parameters from sample statistics requires the concept of confidence interval (*CI*) and margin of error (δ). The δ is subtracted from and added to sample statistic to obtain the *CI*. The *CI* tells that a statistic (e.g., mean and standard deviation) falls between a certain interval produced by the δ. The *CI* is expressed as Eq. 4 (Moore et al., 2014).

$$CI = Statistic \pm Z_{1-\alpha/2}SE \; or$$

$$t_{1-\alpha/2}SE \Longrightarrow Statistic \pm \delta \quad \text{..... Eq. 4}$$

In Eq. 4, $Z_{1-\alpha/2}SE$ or $t_{1-\alpha/2}SE$ is called the margin of error (δ) which depends on the *SE* as well as on the *Z* or *t* values. Although *Z* score of a standard normal curve is constant for a constant α (level of probability of confidence), *t* scores vary with population parameters such as σ and *N*. However, we can calculate *CI* of mean by adding δ to and subtracting δ from the mean value. In our example (see Section 2), mean of the first sample (of the randomly generated 100,000 samples) = 45.26, *SE* = 0.5 and $Z_{1-\alpha/2}$ = 1.96 at α = 95% level of probability for a standard normal distribution. Therefore, δ = 1.96 × 0.5 = 0.98 and *CI* = 45.26 ± 0.98 = [44.28, 46.24] which means that the true population mean value falls outside this range from 44.28 to 46.24 for 5% of the times. In our example dataset, the range of population values is 44.28 (= 66.57 – 22.29) and δ = 0.98 (in absolute proportion) which is 2.21% of 44.28 (in percentage of the population scale points).

A choice between *Z* or *t* scores depends on the population size. When the population is small (<30), $t_{1-\alpha/2}$ value is used to calculate the δ. Figure 3 shows that increasing *n* (degree of freedom or *df*) has decreased the *t*- values because *Z*-score calculation does not need the *df* parameter. Therefore, *Z*-score is constant at a specific α for a standard normal distribution with μ = 0 and σ = 1. However, *t*-values depend on the *df*, and for larger *df* they tend to coincide with *Z*-scores. A larger sample than 30 is traditionally considered as a large sample and assumed to follow a normal distribution, and hence the central limit theorem can be applied. Figure 3 also shows that the gap between *t* and *z* scores increases at lower α levels. Hereafter, for simplicity, $t_{1-\alpha/2}$ and $Z_{1-\alpha/2}$ have been denoted as *t* and *Z*, respectively. To apply the formula for smaller samples, it is recommended to use *t*-scores instead of *Z*-values (Bartlett II et al., 2001).

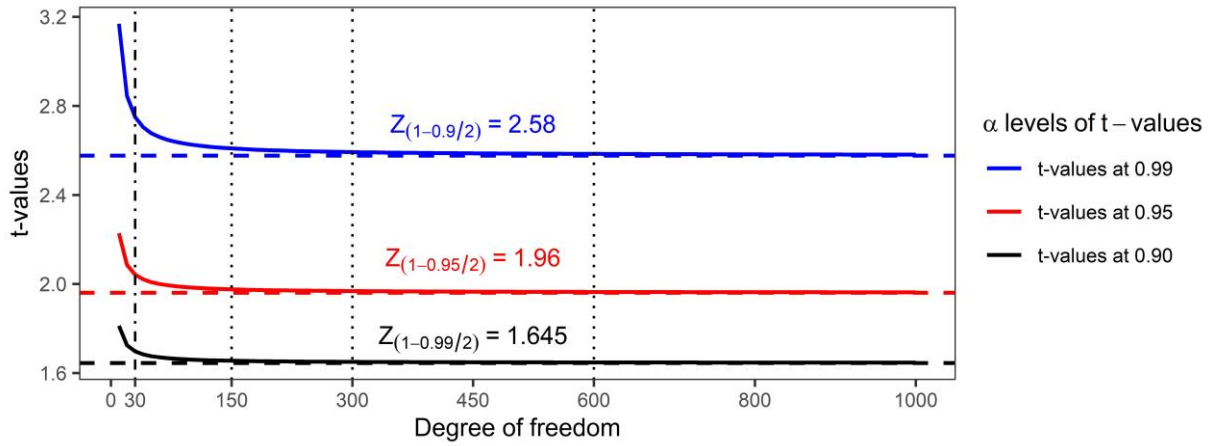

Fig. 3. Changes in *t*-scores (solid lines) relative to the degree of freedom and corresponding constant *Z*-scores (dash lines)

Derivation of formulas for sample size determination from margin of error

If we extract the margin of error part from Eq. 4 it looks as Eq. 5 for continuous variables and Eq. 6 for categorical variables.

$$\delta = t \times SE \implies \delta = t\frac{\sigma}{\sqrt{n}} \quad \text{..... Eq. 5}$$

$$\delta = t \times SE \implies \delta = t\frac{\sqrt{p(1-p)}}{\sqrt{n}} \quad \text{..... Eq. 6}$$

Equation 5 and Eq. 6 are the same where σ is replaced with $\sqrt{p(1-p)}$ and p is the probability of the proportion of positive response in the population. Now, we can use Eq. 5 and Eq. 6 to create formulas for the determination of *n*.

$$\text{Eq.5} \implies n = \frac{t^2\sigma^2}{\delta^2} \implies n_0 = \frac{t^2\sigma^2}{\delta^2} \quad \text{..... Eq. 7}$$

$$\text{Eq. 6} \implies n = \frac{t^2p(1-p)}{\delta^2} \implies n_0 = \frac{t^2p(1-p)}{\delta^2} \quad \text{.... Eq. 8}$$

This is how Cochran (1977: 75, 78) devised the formula of obtaining *n* for continuous data (Eq. 7) and categorical data (Eq. 8) where n_0 denotes the initial sample size. Here, the margin of error (δ) is also known as absolute error or precision. However, for a larger population when $\frac{n_0}{N}$ ratio is appreciable or not negligible (>0.05, for example, as mentioned by Bartlett II et al. (2001) a finite population correction formula (Eq. 9) has to be applied to obtain the final sample size (*n*) (Cochran, 1977).

$$n = \frac{n_0}{1+\frac{n_0-1}{N}} = \frac{n_0}{1+\frac{n_0}{N}} \quad \text{..... Eq. 9}$$

When $n = N$, Eq. 9 can result in the final *n* which is 50% of the n_0 and gradually decreases as $\frac{n_0}{N}$ increases. Again the difference between the use of $\frac{n_0-1}{N}$ and $\frac{n_0}{N}$ can be ignored (Cochran, 1977, p. 76). Here, (n_0-1) may be important to make the formula applicable for small samples. The term (n_0-1) very slightly increases the *n* in Eq. 9 to make it adjusted for a smaller population and smaller samples. It does not affect the calculation of *n* for a larger population. It could be analogous to Bessel's correction in the calculation of variance from sample values where *(n-1)* is used as the divisor (similar to *df = n-1* to get an unbiased estimate of population variance (Upton and Cook, 1996; Warne, 2017).

If we substitute the value of *n* from Eq. 8 in Eq. 9 it provides us with another formula of estimation of *n* (Eq. 10) via the following calculation.

$$n = \frac{n_0}{1+\frac{n_0-1}{N}}$$
$$= \frac{n_0}{\frac{N+n_0-1}{N}}$$
$$= \frac{n_0}{\frac{N-1+n_0}{N}}$$
$$= \frac{N \times n_0}{(N-1)+n_0}$$
$$= \frac{N \times \frac{t^2p(1-p)}{\delta^2}}{(N-1)+\frac{t^2p(1-p)}{\delta^2}}$$
$$= \frac{\frac{t^2Np(1-p)}{\delta^2}}{\frac{\delta^2(N-1)+t^2p(1-p)}{\delta^2}}$$

$$= \frac{\frac{t^2Np(1-p)}{\delta^2}}{\frac{\delta^2(N-1)+t^2p(1-p)}{\delta^2}}$$

$$= \frac{t^2Np(1-p)}{\delta^2} \times \frac{\delta^2}{\delta^2(N-1)+t^2p(1-p)}$$

$$\therefore n = \frac{t^2Np(1-p)}{\delta^2(N-1)+t^2p(1-p)} \quad \text{..... Eq. 10}$$

Although *t* values vary with *df*, it is customary to use *t* = 1.96 at α = 95% (Bartlett II et al., 2001) which is true around *df* ≈ 300. Cochran (1977) used *t* = 2 which is appropriate at α = 5% and *df* ≈ 60. However, *Z* is always equal to 1.96 at α = 95% for a normal curve. Thus, Z^2 = 1.96 × 1.96 = 3.8419 ≈ 3.841459 = X^2 at α = 95% and *df* = 1 (for a binary variable). Therefore, Krejcie and Morgan (1970) used X^2 in their widely used formula of sample size (Eq. 11) that does not need population correction because finite population correction has already been embedded in this formula.

$$n = \frac{\chi^2Np(1-p)}{\delta^2(N-1)+\chi^2p(1-p)} \quad \text{..... Eq. 11}$$

Kothari's (2004) derivation of the formula for *n* has a subtle difference in the sense that it used the finite population multiplier, which is $\sqrt{(N-n)/(N-1)}$. This multiplier was multiplied with δ that resulted in a formula for the determination of *n* (Eq. 12). In this formula, Kothari (2004) used *Z*-values instead of t-values and we know that $\sigma = \sqrt{p(1-p)}$ for categorical variables.

$$n = \frac{Z^2N\sigma^2}{\delta^2(N-1)+Z^2\sigma^2}$$

$$= \frac{Z^2Np(1-p)}{\delta^2(N-1)+Z^2p(1-p)} \quad \text{..... Eq. 12}$$

Similar to the derivation of Eq. 11, we can also derive Eq. 13 from Eq. 8 and Eq. 9. In this case, we have considered $n = \frac{n_0}{1+\frac{n_0}{N}}$ and $\delta^2(N-1)$ in Eq. 11 has now become δ^2N in Eq. 13.

$$n = \frac{\chi^2Np(1-p)}{\delta^2N+\chi^2p(1-p)} \quad \text{..... Eq. 13}$$

If Eq. 13 takes the value of X^2 = 3.84 at α = 95%, *df* = 1 and *p* = 0.5, Eq. 14 emerges assuming 0.96 ≈ 1.

$$n = \frac{\chi^2Np(1-p)}{\delta^2N+\chi^2p(1-p)}$$

$$= \frac{N \times 3.84 \times 0.5(1-0.5)}{\delta^2N+3.84 \times 0.5(1-0.5)}$$

$$= \frac{N \times 0.96}{\delta^2N+0.96}$$

$$\therefore n = \frac{N}{\delta^2N+1} \quad \text{..... Eq. 14}$$

This formula (Eq. 14) has been suggested by Yamane (1967) which is only applicable for categorical variables with *p*= 0.5, α = 0.95 and *df* = 1, i.e., for binary variables. Here, *p* = 0.5 has been used because it provides the highest variance and maximum sample size for categorical variables (Krejcie and Morgan, 1970). As a proof, we can differentiate the *variance* = *np(1-p)* for a binary variable with respect to *p*. The variance is the largest when the first-order derivative is zero. The following solution shows that this condition is fulfilled only when *p* = 0.5.

$$\frac{d\big(np(1-p)\big)}{d(p)} = 0$$

$$\Longrightarrow \frac{d(np-np^2)}{d(p)} = 0$$

$$\Longrightarrow \frac{d(np)}{d(p)} - \frac{d(np^2)}{d(p)} = 0$$

$$\Longrightarrow n - 2np = 0$$

$$\Longrightarrow n(1-2p) = 0$$

$$\Longrightarrow 1-2p = 0$$

$$\Longrightarrow 2p = 1$$

$$\therefore p = 0.5$$

This can also be visualized as in Fig. 4 to show the maximum variance at *p* = 0.50.

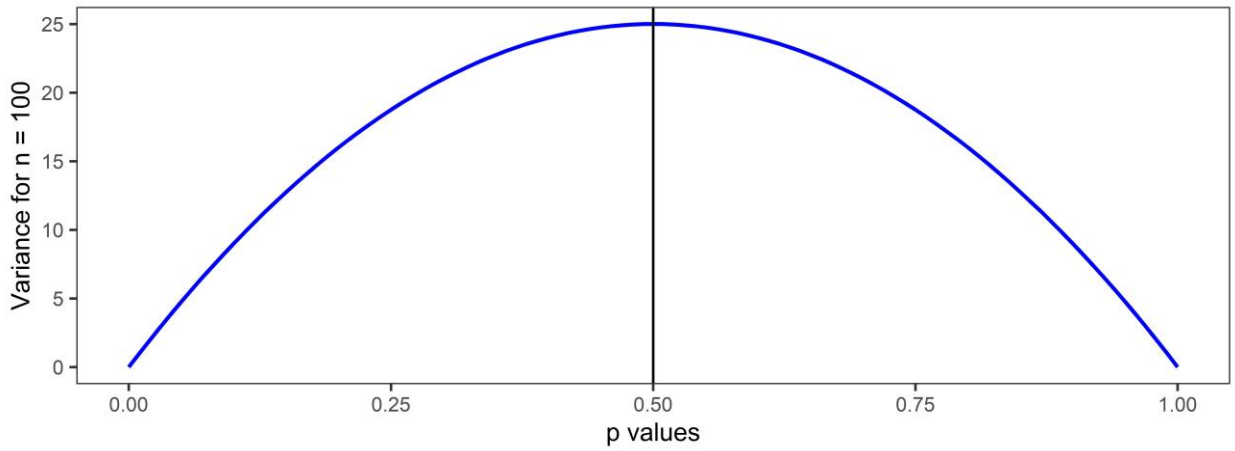

Fig. 4. Variances for different *p* values

Application of the formulas to estimate sample size

It is clear from Section 5 that the formulas (Eq. 7 and Eq. 8) suggested by Cochran (1977) have the maximum number of tuning parameters and applicable to both continuous and categorical variables as well as small and large populations. Several estimations are required to apply these formulas to determine the appropriate sample size, which are the total number of sampling units in the population *N*, t scores with population *df = (N-1)* at the desired level of α, population variance σ^2 or *p(1-p)* and squared margin of error in absolute term δ^2.

The population size is determined based on the total sampling units in the sampling frame in a specified area. For example, if a researcher wishes to study the effect of age, monthly income and training on homestead gardening on the nutritional status of married women in a district the sampling frame will be composed of only the married women. Total population (male, female both married and unmarried) of the district could be one million but the sampling population or sampling frame is much lower that contains only married women. Let us assume the total number of married women in the district is 250,000 which is the sampling population *(N)*. Determination of *t*-score at a specified α and *df* is straightforward. A higher α produces higher accuracy but generates a higher *n*. We will consider α = 0.95, *df* = (250000 – 1) = 249999 and t_1*-0.95/2 (df=249999)* = 1.96. Estimation of σ^2 and δ^2 depends on the variable types and their range of values. In this example, we have several major variables such as age, monthly income, homestead gardening (Yes = 1 or No = 0) and nutritional status (7-point Likert type scale: Extremely poor = 1, very poor = 2, poor = 3, moderate = 4, high = 5, very high = 6, extremely high = 7). We assume the age of the married women in the population varies between 21 to 60 years and therefore the number of unique scale points in round numbers will be *(maximum – minimum +1)* is 40. If the monthly income varies from $501 to $2000 the number of scale points in round numbers will be 1500. For homestead gardening, the number of scale points will be (1 – 0 + 1) = 2 and for nutritional status, it will be 7 – 1 + 1 = 7.

Estimation of variance for continuous variables needs to account for how many standard deviations (σ) from mean are required to include all the observations in a population. For a normal distribution (Fig. 5), $\mu \pm 1\sigma$ covers 68.27%, $\mu \pm 2\sigma$ contains 95.45% and $\mu \pm 6\sigma$ includes almost all (99.99966%) of the observations. This 6σ (six-sigma) is the basis of lean management where 6σ is expressed to represent 3.4 defects per million items (Pepper and Spedding, 2010). This 6σ is not only applicable to continuous variables but also for normally distributed any variables with more than two scale points. Bartlett II et al. (2001) mentioned that 6σ would contain 98% of all the responses in a 7-point Likert type scale. Thus, we can estimate population σ by dividing the total number of scale points by 6 because of $6\sigma \approx$ 100% of scale points $\Rightarrow \sigma$ = number of scale points ÷ 6. Now we can apply this strategy to estimate variances for different non-binary variables in our example.

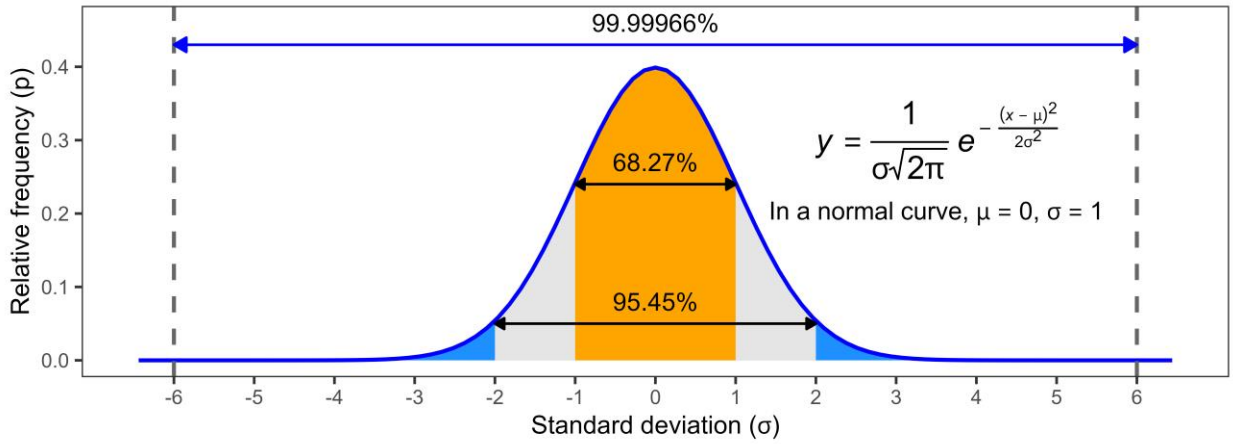

Fig. 5. Standard normal curve showing μ ± 6σ includes 99.99966% of the observations

We have three non-binary variables, namely age (40 scale points), monthly income (1500 scale points) and nutritional status (7 scale points). Therefore, for age, σ = 40/6 = 6.67; for monthly income, σ = 1500/6 = 250; and for nutritional status, σ = 7/6 = 1.167.

We have one last tuning parameter in Eq. 7 and Eq. 8, that is the margin of error *(δ)* in absolute terms. According to Krejcie and Morgan (1970), acceptable δ for continuous and categorical data are, respectively, 3% and 5% of the total scale points. This means that the δ depends on the scale points and the proportion of error (ε), a researcher is willing to accept while making inferences about the population, which is 0.03 (3%) for continuous variables and 0.05 (5%) for categorical variables. However, researchers may choose any other ε as they consider appropriate for their specific research works. Therefore, δ in absolute term will be the number of scale points (*SP*) multiplied by ε expressed as proportion. Thus, for age, $\delta = SP \times \varepsilon = 40 \times 0.03 = 1.2$; for monthly income, δ = 1500 × 0.03 = 45; and for nutritional status, $\delta = 7 \times 0.03 = 0.21$.

We have one binary variable that is participation in homestead gardening. We have already mentioned that *p* is required only for dichotomous or binary categorical variables and the maximum sample size is obtained at maximum variance when *p* = 0.5. However, *p* is the proportion of presence of an attribute in the population. This proportion can be obtained from previous studies, reports or pilot study (Bartlett II et al., 2001; Cochran, 1977). If this is impractical to figure out, it is suggested to use *p* = 0.5 to avoid the risk of accepting the false hypothesis. In our example, we assume that 35% of the married women are engaged in homestead gardening, so *p* = 0.35 for this variable. Again, for this categorical variable, we are willing to accept 5% margin of error and therefore $\delta = 0.05$.

Now we have all the parameters needed to calculate the sample size by plugging these values in Eq. 7 and Eq. 8 followed by the population correction using Eq. 9. The calculated sample sizes (rounding up, such as 52.3 → 53) for different variables are shown in Table 1.

Table 1. Sample size for different variables as an example

Variables	Initial sample (n_0)	Final sample with population correction (n)
Age	119	119
Monthly income	119	119
Homestead gardening	350	350
Nutritional status	119	119

Three important features can be seen from Table 1. Firstly, $n_0 = n$ because $\frac{n_0}{N}$ = 119/250000 to 350/250000 which is negligible. Secondly, sample size estimated from the binary variable (homestead gardening) is larger than the continuous variables (age, monthly income, or nutritional status). Thirdly, all the continuous variables have the identical sample size. This happens because of the formation of Eq. 7 and Eq. 8 which can be expressed as Eq. 15.

$$n_0 = \frac{t^2\sigma^2}{\delta^2} = t^2\rho^2 \quad \text{..... Eq. 15}$$

Where $\rho = \frac{\sigma}{\delta}$ and $\sigma = \sqrt{p(1-p)}$ for binary variables. This ρ can be termed as 'sample control ratio' that controls the estimation of the sample size to a greater extent. In our example, this ρ for age = 6.67/1.2 = 5.56, monthly income = 250/45 = 5.56, nutritional status = 1.167/0.21 = 5.56 and homestead gardening = $\sqrt{0.35(1-0.35)}$ /0.05 = 9.54. For a binary variable,

when p = 0.5 and δ = 0.05, the highest ρ can be obtained which is 10 and of course ρ = 0 for p = 0.

Now, we can examine how the sample size formula (Eq. 8) is sensitive to N (Eq. 9) as shown in Fig. 6. If population correction factor is not applied, sample size will always be 385 when t = 1.96, p = 0.5 and δ = 0.05. If this 385 is larger than 5% of the total population, the correction formula (Eq. 9) is applied to reduce the sample size which happens when N < 7700. Therefore, the formula of Cochran (1977) is sensitive to N only when N < 7700, and the sample size will never be larger than 385 at α = 0.95. Therefore, finite population correction can adjust the sample size substantially to a lower size for only small populations (Israel, 1992).

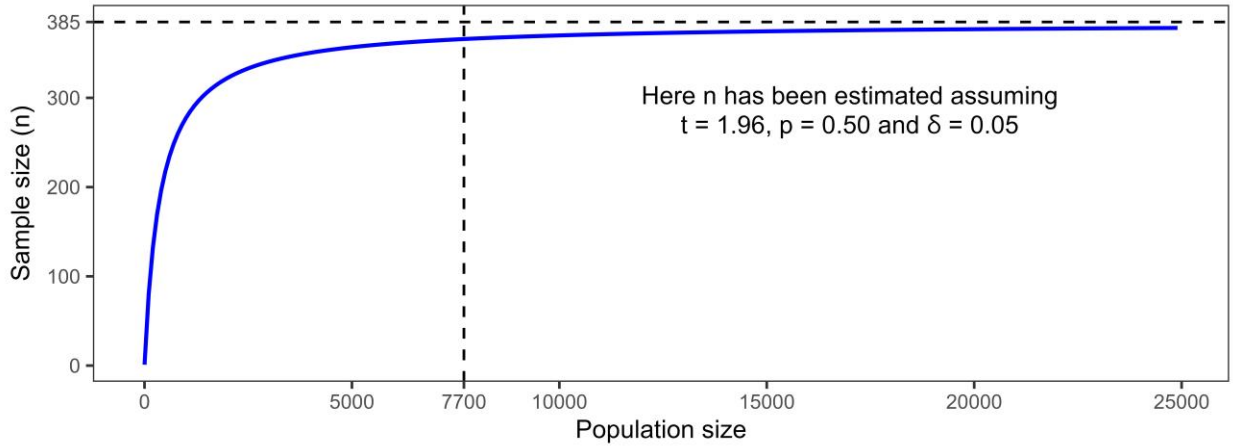

Fig. 6. Sample size according to Cochran (1977). Without population correction, the sample size is always 385 (rounding up of 384.16) which is 5% of 7700

A unified formula for both continuous and categorical data and its application

We have seen that Eq. 7 (for continuous data) and Eq. 8 (for categorical data) combined with Eq. 9 have the highest number of tuning parameters which should provide a more accurate sample size. To generate a unified formula, which is applicable to both continuous and categorical data, we need to solve for σ (in Eq. 7) from p (in Eq. 8) and we know that $\sigma = \sqrt{p(1-p)}$. From Eq. 9 and Eq. 15, we can derive Eq. 16 after the finite population correction.

Eq. 15 $\Longrightarrow n_0 = t^2\rho^2$

Eq. 9 $\Longrightarrow n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{t^2\rho^2}{1+\frac{t^2\rho^2}{N}} = \frac{Nt^2\rho^2}{N+t^2\rho^2}$

$\therefore n = \frac{Nt^2\rho^2}{N+t^2\rho^2}$ ..... Eq. 16

This unified formula (Eq. 16) has three parameters, which are N, t and ρ where ρ is the sample control ratio as explained in Eq. 15. Specifications of the parameters are the same as we have done in Section 6. As finite population correction works only when N < 7700, we have included this in Eq. 15 to make it applicable for both small and large populations. We have also suggested to use t-values instead of constant Z-values at a specific α to reflect the population df. The only parameter remaining, which is sample control ratio (ρ), has to be estimated by using the following procedure:

Eq. 15 $\Longrightarrow \rho = \frac{\sigma}{\delta}$

Section 6 $\Longrightarrow \sigma = \frac{SP}{6}$ and $\delta = SP \times \varepsilon$, where SP is the number of scale points.

$\therefore \rho = \frac{\frac{SP}{6}}{SP\times\varepsilon} = \frac{SP}{6} \times \frac{1}{SP\times\varepsilon} = \frac{1}{6\varepsilon}$, where ε is the proportion of error willing to accept.

For binary variable, $\delta = \varepsilon$ and $\sigma = \sqrt{p(1-p)}$, therefore, $\rho = \frac{\sigma}{\delta} = \frac{\sqrt{p(1-p)}}{\varepsilon}$.

Therefore, it is not necessary to estimate the number of scale points or variances in the continuous and Likert-type scale variables to determine the sample size. For binary variables, estimation of p is important but if it is not practical it is suggested to assume p = 0.5. During estimating the sample size, it is always advisable to consider the variable that would play major roles in the research. It is always better to estimate the sample size considering several important variables and select the highest number to avoid accepting false hypotheses. Setting the α level at higher level, for example 99%, will result in a higher sample size with better accuracy.

For a categorical variable, if $p \leq 10\%$, it is recommended to continue sampling until the desired number of rare items have been included in the sample and this process is known as inverse sampling (Cochran, 1977; Haldane, 1945). In many situations, response rate may not be 100% where oversampling would be necessary (Bartlett II et al., 2001). For example, if an investigator foresees that they can obtain data from 85% of the samples, they should set the sample size as $n/0.85$ to have the required number of responses from the population. Although the suggested formula would yield the minimum sample size that would be sufficient at a given level of confidence interval and accepted error limits, there could be some other conditions that play a vital role in the sampling calculation. For example, to be able to apply central limit theorem in parametric tests with several groups of data, sample size in each group should be larger than 30 (Kothari, 2004) or 25 (Witte and Witte, 2017). Similarly, specific statistical analysis may require minimum number of samples per explanatory variable, for example, factor analysis or multiple regression demands at least ten observations per independent variable and not less than 100 observations in total (Bartlett II et al., 2001).

There could be some situations where researchers may arbitrarily set sample sizes (e.g., 100, 110 and 85) though this is not a recommended practice. It could happen due to constraints of time, manpower, funds and communication to reach the sampled respondents. The survey cost always plays an important role in the sampling design, and it can result in the exclusion of some of the population units from the selected samples (Chang et al., 2004). In such scenarios, it should be revealed to the audience about how much error (ε) could be associated with the inferences. Let us consider a researcher has investigated 76 sample adults (18 to 75 years old) from a total of 1234 adults in a village to explore the effect of age on their marital status (1 = married, 0 = otherwise). In their sample, they have found that standard deviations for age and marital status are 0.46 and 9.5 years, respectively. In this case, the error can be estimated from Eq. 17 which has been derived from Eq. 16 considering $\rho = \frac{\sigma}{\delta}$ and $t_{1-0.95/2}$ = 1.96 at *df* = 1234 – 1 = 1233.

$$\delta = t\sigma\sqrt{\frac{N-n}{N\times n}} \quad \text{..... Eq. 17}$$

Therefore, δ for age = 2.07 and δ for marital status = 0.10. We know from Section 6 that $\delta = SP \times \varepsilon$ (for continuous variables) and $\delta = \varepsilon$ (for categorical variables). In this case, *SP* = 75 – 18 + 1 = 58. Therefore, proportion of error (ε) could have been for age = δ/SP = 2.07/58 = 0.0357 = 3.57% and for marital status 0.10 = 10% in the inferences made based on 76 samples out of 1234 adults.

We also have developed an applet which can be run in a web browser using the URL <https://kamrulext.shinyapps.io/sample/>, or running an R command <if(!require("shiny")) install. packages ("shiny"); require(" shiny"); runGist("d53e8ff6845ea1a6c92199c75ab5d43e")>. Three R packages have been used to create this applet which are *'shiny'* (Chang et al., 2020), *'tidyverse'* (Wickham, 2017; Wickham et al., 2019) and *'ggpmisc'* (Aphalo, 2020) that work in the R Version 4.0.2 (R Core Team, 2020) and RStudio Version 1.3.1056. A screenshot of the applet is presented in Fig. 7 showing that users can specify different parameters and obtain the minimum sample size for both categorical and continuous variables as well as for small and large populations.

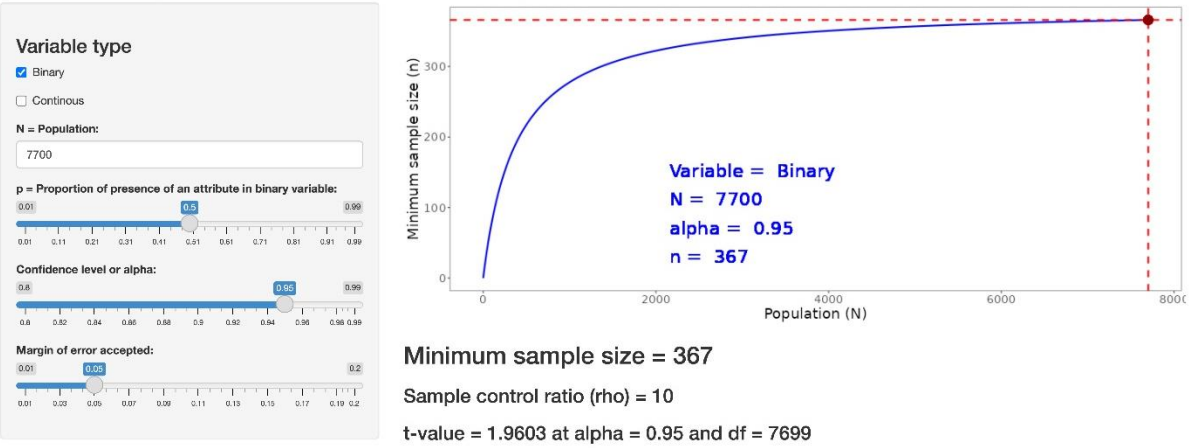

Fig. 7. Screenshot of the web applet for sample sizes estimation which is independent of operating systems and browsers

Comparison of sample sizes obtained from different methods

Mostly used parameters for determining the sample size are p = 0.5, α = 0.05 and ε = 0.05. Samples sizes for binary variables using these parameters have been calculated using different formulas (Table 2). Binary variables have been used for demonstration because they produce the largest number of samples. Table 2 shows that sample sizes computed by the formulas of Cochran (1977) and Kothari (2004) are almost equal but the formula of Krejcie and Morgan (1970) gives larger sample sizes. The differences in sample sizes, in this case, are produced due to the use of t, z or X^2 values though the structural formations of the formulas are identical. Yamane's (1967) simplified formula overestimates sample size that would increase time and cost of sample surveys. The proposed formula by this paper provides sample sizes consistent with the formula of Krejcie and Morgan (1970), and in all cases is within the range obtained by the other four comparing formulas. In addition, it takes the population into account while estimating *t*-values for the sample size determination. Smaller sample sizes can minimize expenditures of surveys but they may fail to confine the sampling error within the limit of errors that the researcher is willing to accept.

Table 2. Sample sizes obtained from various formulas

N	Sample size *(n)*					*t*-value at *df = (N-1)* and *α = 0.05*
	Cochran[1]	Kothari[2]	Krejcie[3]	Yamane[4]	Proposed[5]	
20	19	19	20	20	20	1.729133
50	43	43	45	45	45	1.676551
100	74	74	80	80	80	1.660391
200	116	116	132	134	133	1.652547
300	143	143	169	172	170	1.649966
400	163	162	197	200	197	1.648682
500	177	176	218	223	218	1.647913
600	188	187	235	240	235	1.647401
700	196	196	249	255	249	1.647036
800	203	203	260	267	261	1.646763
900	209	209	270	277	270	1.646550
1000	214	214	278	286	279	1.646380
1500	230	230	306	316	307	1.645871
2000	239	239	323	334	323	1.645616
4000	254	254	351	364	351	1.645235
6000	259	259	362	375	362	1.645108
8000	262	262	367	381	367	1.645044
10000	264	264	370	385	371	1.645006
100000	270	270	383	399	383	1.644869

[1]Cochran's (1977) formula (Eq. 10): $n = \frac{t^2 Np(1-p)}{\delta^2(N-1)+t^2 p(1-p)}$

[2]Kothari's (2004) formula (Eq. 12): $n = \frac{Z^2 Np(1-p)}{\delta^2(N-1)+Z^2 p(1-p)}$

[3]Krejcie and Morgan's (1970) formula (Eq. 11): $n = \frac{\chi^2 Np(1-p)}{\delta^2(N-1)+\chi^2 p(1-p)}$

[4]Yamane's (1967) formula (Eq. 14): $n = \frac{N}{\delta^2 N+1}$

[5]Proposed formula (Eq. 16): $n = \frac{Nt^2\rho^2}{N+t^2\rho^2}$ where $\rho = \frac{\sqrt{p(1-p)}}{\delta}$ for binary variables

Conclusion

The aim of this paper is to discuss the various formulas to estimate the sample size in social survey research and to propose a simple methodology to determine sample sizes. It has been shown here that all the widely used equations for sample size share the same principles with different styles of notations and simplifications. A modified formula, with a freely available Applet, has been suggested which is based on the well-known equations of Cochran (1977). All other formulas discussed in this paper can fundamentally be derived from these well-known equations. Application of the proposed formula has been made easier by providing examples that do not need any advanced statistical knowledge. The proposed formula in this paper has some clear advantages. Firstly, users do not need to estimate the error margin in absolute terms, instead they only need to provide the proportions of error they are willing to accept while making inferences. Secondly, this formula has only one tuning parameter

(the sample control ratio) that needs to be estimated from the given proportions of error. Lastly, it does not require the finite population correction which is already embedded in this formula. A comparison with four other common sample size determination methods shows that the proposed formula produces sample sizes within the range determined by the other four widely used methods. This paper has not generated any tables for sample size determination because the formula suggested is very easy to apply using the applet in various situations, such as surveys consisting of binary or continuous variables and small or large populations, with greater transparency and lesser confusion. It is expected to be highly useful for the researchers and students interested in social survey research.

References

Aphalo, P. J. 2020. ggpmisc: Miscellaneous Extensions to 'ggplot2'. R Package Version 0.3.5. https://CRAN.R-project.org/package=ggpmisc.

Bartlett II, J. E., Kotrlik, J. W., Higgins, C. C. 2001. Organizational research: determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal, 19*(1), 43-50.

Benedetti, R., Andreano, M. S., Piersimoni, F. 2019. Sample selection when a multivariate set of size measures is available. *Statistical Methods & Applications, 28*(1), 1-25. https://doi.org/10.1007/s10260-018-00433-x

Chang, H. J., Wang, C. L., Huang, K. C. 2004. Simple random sample equivalent survey designs reducing undesirable units from a finite population. *Statistical Papers, 45*(2), 287-295. https://doi.org/10.1007/BF02777229

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J. 2020. shiny: Web Application Framework for R. *R Package Version 1.5.0*. https://CRAN.R-project.org/package=shiny.

Chou, Y.-M., Johnson, G. M. 1987. Sample sizes for strong two-sided distribution-free tolerance limits. *Statistische Hefte, 28*(1), 117-131. https://doi.org/10.1007/BF02932595

Cochran, W. G. 1953. *Sampling Techniques.* 1st ed. John Wiley & Sons, London.

Cochran, W. G. 1977. *Sampling Techniques.* 3rd ed. John Wiley & Sons, London.

Diamond, W. J. 1989. *Practical Experimental Designs for Engineers and Scientists.* 3rd ed. John Wiley & Sons, London.

Haldane, J. 1945. On a method of estimating frequencies. *Biometrika*, 33(3), 222-225.

Heller, M., Hannig, J., Leadbetter, M. R. 2015. Optimal sample planning for system state analysis with partial data collection. *Stat, 4*(1), 69-80. https://doi.org/10.1002/sta4.79

Huschens, S. 1990. Necessary sample sizes for categorial data. *Statistical Papers, 31*(1), 47-53. https://doi.org/10.1007/BF02924673

Israel, G. D. 1992. *Determining Sample Size.* Institute of Food and Agricultural Sciences, University of Florida, Florida, USA.

Karimnezhad, A., Parsian, A. 2018. Most stable sample size determination in clinical trials. *Statistical Methods & Applications, 27*(3), 437-454. https://doi.org/10.1007/s10260-017-0419-6

Kothari, C. R. 2004. *Research Methodology: Methods and Techniques.* 2nd Revised ed. New Age International (P) Ltd.

Krejcie, R. V., Morgan, D. W. 1970. Determining sample size for research activities. *Educational and Psychological Measurement, 30*(3), 607-610. https://doi.org/10.1177/001316447003000308

Laga, J., Likeš, J. 1975. Sample sizes for distribution-free tolerance intervals. *Statistische Hefte, 16*(1), 39-56. https://doi.org/10.1007/BF02923053

Liu, X. S. 2013. Sample size determination for the confidence interval of mean comparison adjusted by multiple covariates. *Statistical Methods & Applications, 22*(2), 155-166. https://doi.org/10.1007/s10260-012-0212-5

Martin, J., Elster, C. 2020. The variation of the posterior variance and Bayesian sample size determination. *Statistical Methods & Applications, 30*, 1135-1155. https://doi.org/10.1007/s10260-020-00545-3

Moore, D. S., McCabe, G. P., Craig, B. A. 2014. *Introduction to the Practice of Statistics*. 8th ed. W. H. Freeman and Company, New York.

Pepper, M. P., Spedding, T. A. 2010. The evolution of lean Six Sigma. *International Journal of Quality & Reliability Management, 27*(2), 138-155. https://doi.org/10.1108/02656711011014276

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sambucini, V. 2000. Sample size determination for inferences on the odds ratio. *Journal of the Italian Statistical Society, 9*(1), 219-243. https://doi.org/10.1007/BF03178967

Sampath, S. 2001. *Sampling Theory and Methods.* CRC Press, Narosa Publishing House, New Delhi.

Seelbinder, B. M. 1953. On Stein's two-stage sampling scheme. *The Annals of Mathematical Statistics, 24*(4), 640-649.

Stein, C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics, 16*(3), 243-258.

Upton, G., Cook, I. 1996. *Understanding Statistics.* Oxford University Press, Oxford.

Warne, R. T. 2017. Statistics for the Social Sciences: A General Linear Model Approach. Cambridge University Press, Cambridge. https://books.google.com.au/books?id=-c9CDwAAQBAJ

Wickham, H. 2017. tidyverse: Easily Install and Load the 'Tidyverse'. R Package Version 1.2.1. https://CRAN.R-project.org/package=tidyverse. https://CRAN.R-project.org/package=tidyverse

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. 2019. Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Witte, R. S., Witte, J. S. 2017. *Statistics.* 11th ed. John Wiley & Sons, London.

Yamane, T. 1967. *Statistics: An Introductory Analysis.* 2nd ed. Harper & Row, John Weatherhill, Inc, New York, Tokyo. https://books.google.com.au/books?id=Wr7rAAAAMAAJ

.