Journal of Scientific Research & Reports

26(5): 90-97, 2020; Article no.JSRR.58400
ISSN: 2320-0227

Sample Size Determination in Survey Research

Anokye M. Adam[1*]

[1]*Department of Finance, School of Business, University of Cape Coast, Ghana.*

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JSRR/2020/v26i530263
Editor(s):
(1) Dr. Suleyman Korkut, Duzce University, Turkey.
Reviewers:
(1) Sonal Kapoor, India.
(2) Tanmoy Dasgupta, The University of Burdwan, India.
Complete Peer review History: http://www.sdiarticle4.com/review-history/58400

Short Research Article

Received 18 April 2020
Accepted 22 June 2020
Published 25 June 2020

ABSTRACT

Obtaining a representative sample size remains critical to survey researchers because of its implication for cost, time and precision of the sample estimate. However, the difficulty of obtaining a good estimate of population variance coupled with insufficient skills in sampling theory impede the researchers' ability to obtain an optimum sample in survey research. This paper proposes an adjustment to the margin of error in Yamane's (1967) formula to make it applicable for use in determining optimum sample size for both continuous and categorical variables at all levels of confidence. A minimum sample size determination table is developed for use by researchers based on the adjusted formula developed in this paper.

Keywords: Sample size determination; Yamane formula; survey research; Likert scale.

1. INTRODUCTION

One of the key challenges that social science researchers face in survey research is the determination of appropriate sample size which is representative of the population under study. This is to ensure that findings generalized from the sample drawn back to the population are with limits of random error [1]. It is impossible to make accurate inferences about the population when a test sample does not truly represent the population from which it is drawn due to sample bias [2]. This makes the appropriate sample size important in survey research. However, researchers continue to incorrectly estimate sample size due to misuse or inappropriate use

*Corresponding author: E-mail: aadam@ucc.edu.gh;

of sample size determination tables and formulas [3]. Wunsch [4] identified two most consistent flaws when determining sample size as disregard for sampling error and nonresponse bias. In addition, disregard for sample variance and treatment of all estimand as dichotomous (population proportion) is a common flaw in survey research. As pointed out by Bartlett, et al. [1], a simple survey reveals numerous errors and questionable approaches to sampling size selection in published manuscript surveyed. These errors emanate from an insufficient statistical understanding of these sample size selection methods and quest to use the simplest method in survey research [3]. As noted by Israel [5], the difficulty of obtaining a good estimate of population variance has increased the popularity of sample size based on proportion. Taro Yamane (1967) formula which is a simplified formula for proportion has become popular with researchers for these reasons. Denoting by n the sample size, Taro Yamane formula is given by $n = \frac{N}{1+\mathrm{N}\mathrm{d}^2}$, where N is the population size and d is the margin of error. Strictly speaking, Yamane formula is an approximation of known sample size formulas such as Krejcie and Morgan [6] and Cochran [7] formulas for proportion at 95% confidence level and population proportion of 0.5. Yamane formula in its present state is, therefore, best suited for categorical variables and only applicable when the confidence coefficient is 95% with a population proportion of 0.5.

Bartlett, Kotrlik and Higgins [1] argued for the different sample size for dichotomous (categorical) variables and continuous variables. Though sample based on proportion is conservative, it has a cost implication for data collection and processing.

This paper proposes an adjustment to the margin of error in Yamane to allow it to be applicable for use in determining sample size for both continuous and categorical variables at all levels of confidence. Besides, a minimum sample size determination table is developed for use by researchers based on the adjusted formula developed in this paper. The paper contributes to the existing literature by removing the restriction of the use of Yamane formula.

The paper is structured as follows: Section 2 looks at the mathematical derivation of the proposed adjusted formula from Krejcie and Morgan [6] and Cochran [7] formulae. Section 3 presents the estimation of variance for both categorical and continuous variables.

2. MATHEMATICAL DERIVATION

We begin by considering the formula used by Krejcie and Morgan in their 1970 article “Determining Sample Size for Research Activities”

$$\mathrm{s} = \frac{\chi^2 NP(1-P)}{\mathrm{d}^2(N-1)+\chi^2 p(1-p)} \qquad (1)$$

s= required sample size
χ^2= the table value of chi-square for 1 degree of freedom at the desired confidence level.
N= the population size
P= the population proportion
d= the degree of accuracy expressed as a proportion

From equation (1), we can write that

$$S = \frac{N}{\frac{N\,\mathrm{d}^2}{\chi^2 P(1-P)} - \frac{\mathrm{d}^2}{\chi^2 P(1-P)} + 1} \qquad (2)$$

$$\Longrightarrow \mathrm{S}_{\mathrm{d}} \rightarrow 0 = n = \frac{N}{1+\frac{N\,\mathrm{d}^2}{\chi^2 P(1-P)}} \qquad (3)$$

Krejcie and Morgan [6] recommended the use of .50 as an estimate of the population proportion to maximize variance, which will also produce the maximum sample size. So at 95% confidence level, P = 0.5, $\chi^2 P(1-P) \approx 1$

$n = \frac{N}{1+N\,\mathrm{d}^2}$ which is Slovin or Yamane formula.

Again, given Cochran’s [7] formula $s = \frac{Z^2 P(1-P)}{\mathrm{d}^2}$ and finite correction factor $= \frac{S}{1+\frac{S}{N}}$, Tejada and Punzalan [2] had proved that for P=0.5 and at 95% confidence level, $s = \frac{1}{\mathrm{d}^2}$ and $n = \frac{N}{1+N\,\mathrm{d}^2}$. This implies that Yamane formula is a special case of Krejcie and Morgan [6] formula or Cochran’s [7] formula. Hence, Krejcie and Morgan’s, Cochran’s and Yamane’s formulas coincide when estimating sample size using a 95% confidence coefficient and P = 0.5.

In effect, when $\chi^2 P(1-P) = t^2 \sigma^2$, the general formula for determining sample size becomes

$$\Longrightarrow n = \frac{N}{1+\mathrm{N}(\frac{d}{t\sigma})^2} \qquad (4)$$

This allows the adjusted Yamane’s formula applicable at different population proportion levels and confidence levels.

3. ESTIMATION OF VARIANCE

Cochran [7] listed four ways of estimating population variances for sample size determinations: (1) take the sample in two steps, and use the results of the first step to determine how many additional responses are needed to attain an appropriate sample size based on the variance observed in the first step data.; (2) use pilot study results: (3) use data from previous studies of the same or a similar population; or (4) estimate or guess the structure of the population assisted by some logical mathematical results. Bartlett, et al. [1] observed that the first three ways are logical and produce valid estimates of variance; but not feasible to use them because of technical difficulty to implement. The fourth option is rather likely to be used by survey researchers due to its flexibility.

Bartlett, et al. [1] showed that the standard deviation of survey research using Likert-type items is estimated as the ratio on the inclusive range of the scale to the number of standard deviations that would include all positive values in the range. Let, λ= the inclusive range, ρ = the number of standard deviations that would include all possible values in the range and □ = the degree of accuracy expressed as a proportion, then $\sigma = \frac{\lambda}{\rho}$ and mean margin of error $d = e\lambda$

The adjusted Yamane's formula in equation (4) becomes

$$n = \frac{N}{1+N\varepsilon^2} \quad (5)$$

Where,

n= minimum returned sample size
N = the population size
ε = adjust margin of error [$\varepsilon = (\frac{\rho e}{t})$]
e = the degree of accuracy expressed as a proportion
ρ= the number of standard deviations that would include all possible
t= t-value for the selected alpha level of confidence level

Park and Jung [8] argues that respondents tend to avoid choosing extreme responses categories, proportion choosing the middle option of Likert-type is larger than extreme responses, so that the coefficient of variation is smaller than 1 (about 0.3-0.5). This coefficient of variation and its associated mean values imply that the standard deviation of Likert-type item rounds to 1 point. Given that the standard deviation is 1 point, the number of standard deviations that would include all possible values in the range is one less the number of inclusive ranges for an odd number of points and equal to the number of inclusive ranges for an even number. For example, the number of standard deviations that would include all possible (ρ) for five-point Likert-type scale is four (i.e. two to each side of the mean) and for six-point Likert-type scale is six (i.e three to each side of the mean). The number of inclusive ranges for all survey research ranges from 2 for dichotomous responses to 10 for 11-point Likert-type scale. The minimum returned sample size varies inversely with the number of standard deviations that would include all possible (ρ). Inferences from Rasmussen [9], Owuor [10] and Norman [11] suggest that scales with 5 or more points can be treated as continuous data and be treated with parametric statistics. A 2-point and 5-point scales are recommended as least for categorical and continuous variables respectively. The number of standard deviations that would include all possible of 2-point scale, 2, and 5-point scale, 4, respectively yield maximum sample size for categorical and continuous variables. This is consistent with Cochran [7] that for a range of sample size which is relatively close, the researcher can settle on the largest sample size to be confident of achieving the desired accuracy. Thus, ρ = 2 is recommended for categorical variables and ρ =4 for continuous variable.

The choice of either number of standard deviations that would include all possible values in the range for categorical or continuous variable depends on whether a categorical variable will play a primary role in the data analysis or not [1]' if categorical variable will play a primary role in the data analysis, use number of standard deviations that would include all possible values in the range for categorical else use the number of deviations that would include all possible values in the range for continuous estimand. Krejcie and Morgan [6] recommended 5% as an acceptable margin of error for categorical data and 3% for continuous data.

To illustrate the use of the two examples, let us consider the following two examples:

Example 1:

Assume a researcher wants to examine the gender disparity in financial literacy among the

Cape Coast Metropolis in Ghana. If the researcher set the alpha level a priori at.05 and the population is 1500 retirees, what minimum returned sample size is required at 95% confidence level and margin of error of 0.05? The number of standard deviations that would include all possible categorical variables should be used because of gender.

$$n = \frac{N}{1+N\varepsilon^2}$$

Where

n= minimum returned sample size
N= Population size=1500
e= the degree of accuracy expressed as a proportion=0.05
ρ= the number of standard deviations that would include all possible values in the range =2
t= t-value for the selected alpha level or confidence level at 95% =1.96
ε = adjust margin of error [$\varepsilon = (\frac{\rho e}{t})$]

$$\Rightarrow \varepsilon = (\frac{2(0.05)}{1.96}) = 0.051$$
$$\Rightarrow n = \frac{1500}{1+1500\varepsilon^2}$$

$$n = \frac{1500}{1+1500(0.051)^2} = 306$$

Example 2:

A researcher examines how financial literacy, financial behaviour, and retirement planning to influence on the financial well-being of retirees in Cape Coast Metropolis of Ghana. A cross-sectional survey strategy was employed and a seven-point Likert-type scale was employed to measure financial literacy, financial behaviour, and retirement planning and financial well-being of retirees. If the researcher set the alpha level a priori at.05 and the population is 1500 retirees, what minimum returned sample size is required at 95% confidence level and margin of error of 0.05?

Unlike example 1, the number of standard deviations that would include all possible continuous variables should be used (i.e. 4).

$$n = \frac{N}{1+N\varepsilon^2}$$

Where

n= minimum returned sample size
N= Population size=1500
e= the degree of accuracy expressed as a proportion=0.05
ρ= the number of standard deviations that would include all possible values in the range =2
t= t-value for the selected alpha level or confidence level at 95% =1.96
ε = adjust margin of error [$\varepsilon = (\frac{\rho e}{t})$]

$$\Rightarrow \varepsilon = (\frac{4(0.03)}{1.96}) = 0.06218$$
$$\Rightarrow n = \frac{1500}{1+1500(0.06218)^2} = 226$$

The advantage of this adjustment hinges not only on its simplicity but its ability to determine the sample size of both continuous and categorical survey variables.

4. SAMPLE SIZE DETERMINATION TABLE

We present in Table 1 the minimum sample size values for many common sampling problems based on our adjusted formula for both continuous and categorical data. The table assumed one of the three commonly used confidence levels in survey research: 90%, 95% or 99% and used a margin of error of 3% for continuous data and 5% for categorical data. The Table is recommended for use by researchers if the indicated margin of error is appropriate for their study.

To validate the ability of the adjustment formula in estimating the required minimum sample size, we compared the sample obtained to sample size obtained from the frequently used sample size determination approaches such as Krejcie and Morgan [6] and Bartlett, Kortlik and Huggins [1]. Fig. 1 shows the plot of sample size obtained from the proposed adjusted margin of error, SS1, and Sample size obtained from Krejcie and Morgan [6], SS2, for categorical estimand versus Population at 5% significance level. The plot shows that sample sizes obtained from the two approaches are virtually the same with correlation $r = 0.9992, p < 0.001$. Similarly, a plot of sample size obtained from the proposed adjusted margin of error, CSS1, and Sample size obtained from Bartlett, Kortlik and Huggins [1], SS2, for continuous estimand versus Population is shown by Fig. 2. The plot shows that the new approach provides a more conservative sample size with $r = 0.97, p < 0.001$.

Table 1. Table for determining minimum returned sample size for a given population size for continuous and categorical data

Population size	**Sample size**					
	Categorical data (margin of error=.05), ρ=2			**Continuous data (margin of error=.03), ρ=4**		
	90% confidence Level $t = 1.645$	**95% confidence Level $t = 1.9\ 6$**	**99% confidence Level $t = 2.58$**	**90% confidence Level $t = 1.645$**	**95% confidence Level $t = 1.9\ 6$**	**99% confidence Level $t = 2.58$**
10	10	10	10	10	10	10
15	15	15	15	14	15	15
20	19	20	20	19	19	20
25	23	24	25	23	23	24
30	28	28	29	26	27	29
35	31	33	34	30	31	33
40	35	37	38	33	35	37
50	43	45	47	40	43	46
60	50	52	56	46	49	54
70	56	60	64	52	56	61
80	62	67	72	57	62	69
90	68	73	80	61	68	76
100	74	80	87	66	73	83
110	79	86	95	70	78	89
120	84	92	102	74	83	96
130	88	98	109	77	88	102
140	93	103	116	81	92	108
150	97	108	123	84	97	114
160	101	113	129	87	101	119
170	105	118	136	90	104	125
180	109	123	142	92	108	130
190	112	128	148	95	111	135
200	116	132	154	97	115	140
220	122	140	166	102	121	150
250	130	152	182	108	130	163
300	143	169	207	116	142	182
350	153	184	230	123	152	200
400	162	196	250	128	161	215

Popula-tion size	**Sample size**					
	Categorical data (margin of error=.05), ρ=2			**Continuous data (margin of error=.03), ρ=4**		
	90% confidence Level $t = 1.645$	**95% confidence Level $t = 1.9\ 6$**	**99% confidence Level $t = 2.58$**	**90% confidence Level $t = 1.645$**	**95% confidence Level $t = 1.9\ 6$**	**99% confidence Level $t = 2.58$**
450	169	208	269	133	168	229
500	176	218	286	137	174	241
600	187	235	316	144	185	262
700	196	249	342	149	194	279
800	203	260	364	153	201	293
900	209	270	383	156	206	306
1000	213	278	400	159	211	317
1200	221	292	429	163	219	334
1500	230	306	462	167	227	354
2000	239	323	500	172	236	376
3000	249	341	545	177	245	401
5000	257	357	588	182	254	424
8000	262	367	615	184	259	437
10000	264	370	625	185	260	442
20000	267	377	645	187	264	452
50000	270	382	657	188	266	459
100000	270	383	662	188	267	461
150000	271	384	663	188	267	461
200000	271	384	664	188	267	462
>1000000	271	385	666	188	267	463

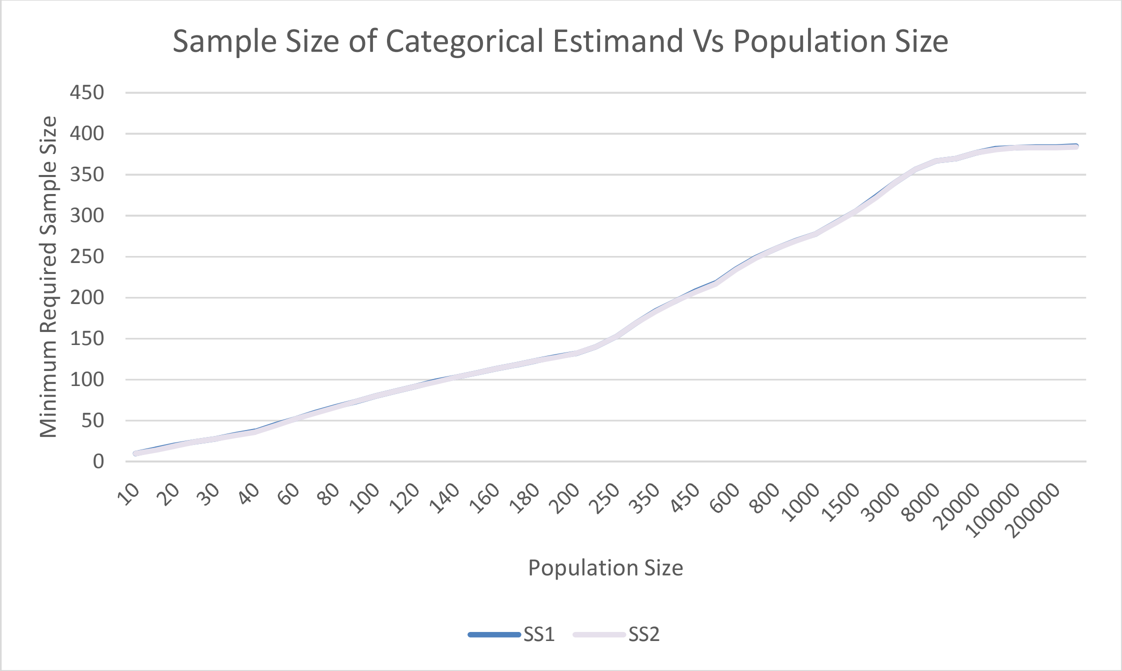

Fig. 1. Sample size of categorical estimand Vs population

Note: CSS1 is sample size obtained from the proposed approach and CSS2 is sample size obtained from the approach proposed by Krejcie and Morgan [6]

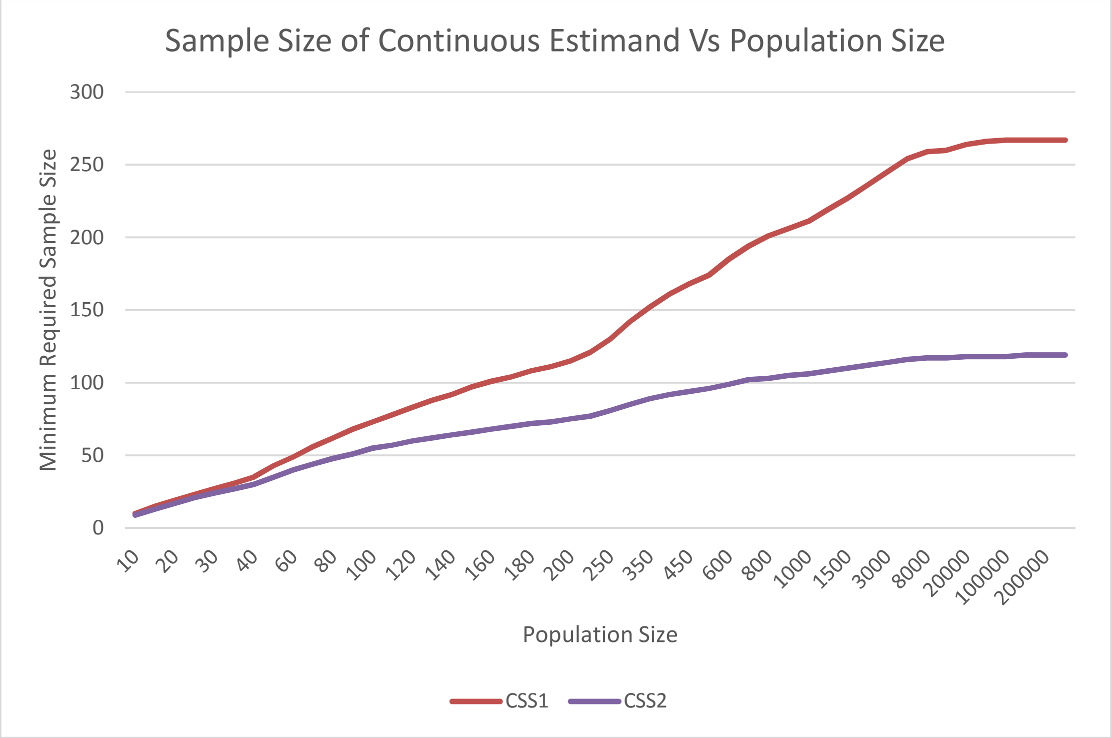

Fig. 2. Sample size of continuous estimand vs population

Note: CSS1 is sample size obtained from the proposed approach and CSS2 is sample size obtained from the approach proposed by Bartlett, Kortlik and Huggins [1]

5. CONCLUSION

In this paper, we propose an adjustment to the margin of error in Yamane (1967) formula to make it applicable for use in determining optimum sample size for both continuous and categorical variables at all levels of confidence. It has been shown that the degree of accuracy expressed as a proportion (margin of error in Yamane formula), $d,$ be adjusted by a factor of the ratio of the number of standard deviations that would include all possible values in the range to the t-value for the selected alpha level or confidence level, $\frac{\rho}{t}$,. Accordingly, ρ= 2 was recommended for categorical variables and ρ=4 for continuous variable. A minimum sample size determination table is developed for use by researchers based on the adjusted formula when certain assumptions are met.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Bartlett JE, Kotrlik JW, Higgins CC. Organizational research: Determining appropriate sample size in survey research. Information Technology, Learning, and Performance Journal. 2001; 19(1):43-50.
2. Taherdoost H. Determining sample size; how to calculate survey sample size. International Journal of Economics and Management Systems. 2017;2:237–239.
3. Tejada JJ, Punzalan JRB. On the misuse of Slovin's formula. The Philippine Statistician. 2012;61(1):129–136.
4. Wunsch D. Survey research: Determining sample size and representative response. Business Education Forum. 1986;40(5): 31-34.
5. Israel GD. Determining sample size, University of Florida Cooperative Extension Service, Institute of Food and Agriculture Sciences, EDIS; 1992.
6. Krejcie RV, Morgan DW. Determining sample size for research activities. Educational and Psychological Measurement. 1970;30:607-610.
7. Cochran WG. Sampling techniques (3rd ed.). New York: John Wiley & Sons; 1977.
8. Park J, Jung M. A note on determination of sample size for a Likert scale. Communication of the Korean Statistical Society. 2009;16:669-673.
9. Rasmussen J. Data transformation, type 1 error rates and power. British Journal of Mathematical and Statistical Psychology. 1989;42:203-213.
10. Owuor CO. Implications of using Likert data in multiple regression analysis. University of British Columbia: Doctoral dissertation; 2001.
11. Norman G. Likert scales, levels of measurement and the “laws” of statistics. Adv Health Sci Educ Theory Pract. 2010;15(5):625–632.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/58400