Minimum Sample Size Calculation Using Cumulative Distribution Function

Louangrath. P.I. ★ & Sutanapong, Chanoknath★★

About the author

★Louangrath, P.I. is an Assistant Professor in Business Administration at Bangkok University, Bangkok, Thailand. For correspondence, he could be reached by email at: Lecturepedia@gmail.com
★★Sutanapong, Chanoknath is an independent researcher. She may be reached by email at: chanoknath.sutanapong@gmail.co

ABSTRACT

Minimum sample size is a requirement in most experimental designs. Research in social science requires minimum sample size calculation in order to support the claim that the sample represents the population. If the sample does not adequately represent the population, generalizability could not be achieved. In this study, we present a minimum sample size calculation method by using the cumulative distribution function of the normal distribution. Since most quantitative data in social science research employ surveys with responses in the form of Likert or non-Likert scales, the CDF of the normal distribution curve is an appropriate tool for sample size determination. We use binary data in a form of (0,1), and continuous data, in a form of quantitative non-Likert (0,1,2,3), and Likert (1,2,3,4,5), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10) scales as the bases for our modeling. We used Monte Carlo simulation to determine the number of repetition for each scale to achieve normality. The minimum sample size was determined by taking the natural log of the Monte Carlo repetition multiplied by *pi*. We found that in all cases, the minimum sample size is about 30 where we maintain the confidence interval at 95%. For non-parametric case, the new sample size calculation method may be used for discrete and continuous data. For parametric modeling, we employed the entropy function for common distribution as the basis for sample size determination. This proposed sample size determination method is a contribution to the field because it served as a unified method for all data types and is a practical tool in research methodology.

Keywords: Cumulative distribution function (CDF), Likert scale, normal distribution, sample size.

JEL Code: C02, C12, C15, C65, C85

CITATION:

Louangrath, P.I. and Sutanapong, Chanoknath (2019). "Minimum sample size calculation using cumulative distribution function." *Inter. J. Res. Methodol. Soc. Sci.*, Vol., 5, No. 1: pp. 100-113. (Jan. – Mar. 2019); ISSN: 2415-0371. DOI: 10.5281/zenodo.2667494

1.0 INTRODUCTION

The purpose of this paper is to provide a new method for calculating sample size based on the cumulative distribution function (CDF) of normally distributed data. We defined data as the quantitative scale used in the survey commonly employed in social science research. Data could generally be categorized into two types based on their distributions, namely discrete and continuous data. Demographic data may be classified as discrete data. Response scales in a form of Likert scale, such as (1,2,3,4,5), (1,2,3,4,5,6,7,), and (1,2,3,4,5,6,7,8,9,10) or non-Likert quantitative scale, such as (0,1,2,3), may be analyzed under continuous probability. This paper asserts that survey employing any of the aforementioned scale types may be analyzed by the CDF of a normal distribution.

Under Monte Carlo simulation, we can estimate the number of repetitions in order to achieve normality under the law of large number. For quantitative data, we expect the data to achieve normal distribution according to:

$$f(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{1}$$

For binary or discrete data, we expect to achieve normal distribution under the deMoivre-Laplace Theorem under the following condition:

$$\lim_{n\to\infty} \Pr\left(\frac{X-np}{\sqrt{npq}} \right) \geq Z \tag{2}$$

where X = total success of the category of interest, n = number of observation, $p = (s+1\_/(n+2)$, and Z = critical value whose corresponding percentage probability may be found in the Z table. We are aware that certain discrete data, despite the increase in the number of observations will remain discrete or binary. Such exceptional case would be tested and treated separately and would not fall within the generalized case cover in this paper. If the data is not normally distributed, as we observed in the later part of this paper, we propose to use the entropy function for each distribution to augment our sample size equation in order to determine the minimum sample size for each case. We intend the proposed sample size determination method to have practical application in all cases whether the data is discrete or continuous, or whether the investigator is using parametric or non-parametric method in the investigation.

The motivation for this paper comes from the gap in the literature in sample size calculation based on response space. Most literature discussed sample size calculation in the context of the relationship between sample and population of respondents or participants. In this paper, we focus on the scale of the response used in the survey as the probability space. One researcher wrote that:

> "The estimation of the minimum sample size required for any study is not a single unique method, but the concepts underlying most methods are similar. The determination of the sample size is critical in planning clinical research because this is usually the most important factor determining the time and funding to perform the experiment." (Gogtay, 2010).

We treat this assertion as logical, but not practical because it lacks certainty. This uncertainty was shared by Israel who wrote that sample size determination is influenced by the "purpose of the study, population size, the risk of selecting a "bad" sample, and the allowable sampling error." (Israel, 1992). The intent of this paper is to look for a unified method to calculate sample size that would overcome the limitations saw by Israel and Gotay.

The National Institute of Science and Technology (NIST) in the US proposed a more definite sample size calculation method based on sample-population proportion. According to NIST, sample size is a proportion of the population; that proportion could fairly represent the population at a specified level of precision defined as:

$$\Pr\left(\left|\hat{p}-P\right|\geq\sigma\right)=\alpha \qquad (2)$$

where $\hat{p}$ = estimated proportion; P = unknown population parameter; σ = specified precision of the estimate; α = probability value. The sample size of the approximately normally distributed population is:

$$n=Z_a^2\left(\frac{PQ}{\sigma^2}\right) \qquad (3)$$

where Z is a critical value at a; p = probability of the current process determined by $p=(s+1)/(k+2)$ and $q=1-p$. We found this method of sample size calculation limited in application and unstable in outcome because it could only be used for binary data; hence, the presence of P and Q in the equation. In addition, if P and Q is used, the variance should have also come from the binary data set, i.e. variance should have been written as $\sqrt{nPQ}$, not σ^2. A more efficient sample size formula should be able to accommodate both discrete and continuous data. One objective of this paper is to address this inadequacy.

Since the function of the sample is to represent the population, the precision of the equality between the sample and population may be written as:

$$\Pr\left(\left\|\frac{\bar{y}-\mu}{\mu}\right\|\geq\sigma\right)=\alpha \qquad (4)$$

where μ is the unknown population mean and $\bar{y}$ is the sample mean. The sample size is determined by:

$$n\approx\frac{Z_a^2\sigma^2}{\sigma^2\mu^2} \qquad (5)$$

where σ^2 is population variance. NIST's approach to determine sample size is based on proportional representation. The proportion is the ratio between the sample and the population.

In this paper, we assert that where response scale is used in survey, the observed value from the sample and the estimated value for the population should be confined to the probability space of the scale itself. No matter how large the population, the answer to the survey would fall within the value range of the scale. For example, if the Likert scale is used, for a scale of (1,2,3,4,5), the value of the sample and the population will fall between 1 and 5, similarly for a Likert scale of (1,2,3,4,5,6,7), the sample and population values would fall within the range of 1 and 7. Thus, in this paper, we propose the use of the probability space of the scale to estimate sample size for the studies using quantitative scale as response choice.

The intended contribution of this paper is to provide a practical tool for calculating sample size to researchers in social science. By so doing, we hope to dispel many questions about how many sample elements must be taken in a given research. If the research employs a survey with

response scales that are classified as quantitative data, we assert that under the calculation method proposed by this paper, the minimum sample size is about 30.

The scope of this paper is confined to sample size calculation for quantitative research in social science. In qualitative method, the content of the research generally is tainted with subjectivity of the investigator; as for sample size determination, it is also unsettled and is highly influenced by the investigator's subjectivity (Sandelowski, 1995). The lack of systematic guidance in qualitative research is evidenced by the practice of keep adding more participants until a saturation point is reached (Glaser, 1965). Although saturation point has been studied, definitive guidance on sample size determination in qualitative research remains unclear (Francis *et al.*, 2010; Guest *et al.*, 2006; and Wright *et al.*, 2011). Only suggestions on how many samples should be taken, but there is no definitive method for calculating sample size for qualitative research (Onwuegbuzie and Leech, 2007; Fugard and Pott, 2015). For instance, it has been suggested that samples should be collected to the point where a "theme" may be observed and a "theme analysis" may be used (Galvin, 2015). Due this lack of guidance in the research method itself, qualitative research is outside the scope of this paper. Nevertheless, despite such inadequacy in qualitative research, so long as the qualitative data could be coded in binary data, the proposed sample size determination method in this paper could also be useful for qualitative research. To that end, this is an additional contribution of our proposed sample size determination method because it eliminates the uncertainty in sample size calculation in qualitative research.

2.0 LITERATURE REVIEW

There are several scenarios where sample size becomes an issue in research in social science, (i) types of data, i.e. whether the data is time series or non-time dependent, and (ii) parametric of non-parametric testing. When dealing with the type of data, we are face with data that may be classified according to their type of distribution, such as discrete or continuous. Discrete distribution describes data that came from binary or categorical domain which are coded 1 = yes and 0 = no. Continuous data deals with quantitative scale which are commonly used in survey; for instance, (0,1,2,3) or (1,2,3,4,5) are common quantitative scales used in social science research survey.

The second situation where minimum sample size plays a role in assessing the adequacy of research methodology may deal with parametric and non-parametric cases. In parametric modeling, we are asking: "what is the minimum sample required for testing the proposed model?" A model is a mathematical function, also known as the predictive function. Minimum sample size in parametric studies allows the researcher to propose a reliable predictive function for a given construct or group of constructs to explain a given phenomenon. How reliable and valid that proposed predictive function may depend on whether in constructing such function, did the research use adequate sample. For instance, it has been suggested that in multiple regression modeling, each variable should have at least ten observations. Thus, a multiple regression function with three variables, x1, x2, and x3, should have at least 30 sample counts.

2.1 Sample size for non-parametric case

Conventional sample size determination may be divided into two scenarios: finite population and non-finite population. In finite population, the population size is known. Assuming that the population is large enough and normally distributed, sample size may be obtained through the Yamane equation:

$$n = \frac{N}{1 + N(e^2)} \tag{6}$$

where N = population size and e is the error level (Yamane, 1967; p. 886). We found this method of sample size calculation limited in application because in order to use the Yamane equation, the population size must be known. This requirement is not practicable because in most cases in real

life, the population size is either unknown or unstable. Even if the population is known, at 95% confidence interval, the Yamane equation tends to produce a fixed sample size at about 400. This number may be too large and costly; thus, making it unpractical. We also note that the Yamane method had been misunderstood and misused to mean that sample size in all cases is 400. This misunderstanding comes from the misuse of the standard error formula: $SE = \sigma / \sqrt{n}$ by allowing $\sigma = 1$, $SE = 0.05$ and solve for *n*. The answer is $n = 400$. We reject this approach to sample size calculations as erroneous and does not conform to the function and purpose of sample size. The assumption that $\sigma = 1$ is erroneous because such an assumption is may be true when the data is normally distributed. The setting of the error to 0.05 is a error is the misuse of misunderstanding of the standard *beta* where the error level is fixed at 0.05 for 0.95 confidence interval. However, the use of $SE = 0.05$ in this case has no empirical support. Both the assumptions of $\sigma = 1$ and $SE = 0.05$ are not supported by the data in each case.

A second scenario involved unknown population size. In the non-finite population case, a test sample must be taken to learn the approximate variance of the population. For non-finite population, Smith provides the following formula:

$$n = \frac{Z^2 \sigma^2}{e^2} \tag{7}$$

where Z is the critical value, sigma is the estimated standard deviation, and e is the error level (Smith, 1983). We found this method of sample size calculation limited in application and unstable in outcome because in order to use equation (7), a test sample must be used. Different sizes of the test sample change the outcome of *n*. For example, test sample sizes 10, 20 or 30 would have different required sample size *n*. These differences are evidence to prove that this sample size calculation method is not reliable. Reliability is defined as consistency in outcome. Equation (7) fails this requisite.

The formula given by Yamane and Smith may be classified as general approaches to sample size determination based on population size (Yamane's approach) and distribution of the test sample (Smith's approach). We reviewed the literature in more specific cases on the basis of data type, i.e. continuous or discrete data. The literature for the general case based on population size, and the specific cases based on data types, provides us a context for our introduction of sample size calculation method based on response space or instrument scale-based approach.

2.1 Sample size determination for categorical data

Sample size determination of categorical data or binary data for non-finite population was discussed by Cochran (1963: 75) who provides the following formula:

$$n = \frac{Z^2 PQ}{e^2} \tag{8}$$

This formulation is limited to binary data as indicated by *p* and *q* where $p = (s+1)/(n+2)$ and $q = 1 - p$.

For categorical data, minimum sample size may also be determined by two proportions of the categorical data:

$$n = \frac{2\left(Z_\alpha + Z_\beta\right)^2}{d^2} \tag{9}$$

where $d = (p_1 - p_2)\sqrt{p(-1p)}$ and $p = (p_1 + p_2)/2$.

The use of normal distribution as the reference or ideal condition to determine sample size is well documented (Devane and *et al.*, 2004). The element of sample size determination include the type of data and it distribution (Julios, 2004). While data distribution is not under the control of the researcher, certain terms are under the researcher's control. These uncontrollable factors include: (i) detectable effect size; (ii) probability of falsely rejecting the null hypothesis (alpha error); (iii) probability of rejecting false null hypothesis (beta error); and (iv) estimated standard deviation (sigma) (Karlson *et al.*, 2003). So what?

Sample size determination in social science allows the studies to make a generalization about the population through inference. However, in medical science sample size may allow the researcher to verify treatment effect. Noordzij *et al.* (2010) wrote that: "The main aim of a sample size calculation is to determine the number of participants needed to detect a clinically relevant treatment effect." Four components are considered when calculating sample size: (i) type I error; (ii) power; (iii) minimal clinically relevant difference; and (iv) variability (Noordzij, 2010). For medical science, Noodzij provides two methods for sample size determination where the data is continuous and discrete.

For discrete data, the formula was given:

$$n = \frac{\left|(a+b)^2(p_1q_1 + p_2q_2)\right|}{x^2} \qquad (10)$$

n = sample size of each group; p_1 = proportion of subjects with factor of interest in group 1; q_1 = proportion of subjects without factor of interest in group 1; p_2 = proportion of subjects with factor of interest in group 2; q_2 = proportion of subjects without factor of interest in group 2 $(1 - p_2)$; x = the difference the investigator wishes to detect; a = multiplier for alpha = 0.05; and b = multiplier for power = 0.80.

We found these two methods (equations 8, 9 and 10) of sample size calculation limited in application because the presence of P and Q allows the formula to be used only with discrete or binary data. In order to be efficient, the sample size formula must be able to accommodate both discrete and continuous data.

2.2 Sample size determination for continuous data

Gotay discussed two scenarios involving two means comparison in continuous and discrete data. In continuous data, for means comparison, the formula was given:

$$n = \frac{(Z_a + Z_b)^2\sigma^2}{d^2} \qquad (11)$$

where d = effect size. The effect size d of continuous data is given by Cohen (Cohen, 1988; p. 67):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} = \frac{\mu_1 - \mu_2}{s} \qquad (12)$$

The pooled standard deviation (s) is obtained by: $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$.

For two group comparison with continuous data, the formula for calculating sample size is given by:

$$n = \frac{2\left[(a+b)^2\sigma^2\right]}{(\mu_1 - \mu_2)^2} \tag{13}$$

n = sample size of each group; μ_1 = population mean of treatment group 1; μ_2 = population mean of treatment group 2; $\mu_1 - \mu_2$ = the difference that the investigator wants to detect; σ^2 = population variance; a = multiplier for alpha = 0.05; and b = multiplier for power = 0.80.

In these two methods (equations 12 and 13), we found the same inadequacy as we observed in equations 8, 9 and 10 , because they could only be used for continuous data and cannot accommodate binary data.

In this paper, we present a new method for calculating sample size based on the probability space in the survey's scale using the cumulative distribution function (CDF) as the basis. We note that discrete data does not have CDF, but the mass distribution function (MDF); nevertheless, the data set of (1,0) is converted to its continuous equivalence by using the DeMoivre-Laplace Theorem in order to obtain unified approach in sample size determination. The new sample size calculation can accommodate both discrete and continuous data of the response space. Whether the response space comes from binary data (1,0) or continuous data in a form of Likert or non-Likert scales, the new sample size calculation method could accommodate both types of data because we use the probability of the response element or component as the basis for sample size calculation. In all other sample size calculation methods hitherto, population parameter estimate was used as the building block. There is a lack of standard formula for sample size calculation because different types of modeling require different types of parameter estimation. The new method introduce in this paper overcomes this weakness of the traditional method. To the extent that the new method is an improvement over those found in the literature, the new method for sample size calculation is a contribution to the field.

2.3 Sample size determination for parametric modeling

Parametric modeling involves the use of predictive function to explain the data. The predictive function may depend on the type of the distribution of the data. In this paper, we present five common distribution found in social science research where psychometric scales are used in opinion surveys. These common distributions are: (i) normal, (ii) logistic, (iii) beta, (iv) gamma, and (v) Weibull distributions. For each distribution, we use the entropy as the threshold for which sample size determination may be determined. Entropy is defined as the function explaining the point at which information break down commences, i.e. the point where the stability of the distribution starts to deteriorate.

2.3.1 Normal distribution and its entropy function

When the data is obtained through quantitative scale, ideally if the data behaves according to the law of large number, the data would manifest normal distribution where the plot of the distribution curve resembles perfect bell shape. The PDF and CDF of the normal distribution are given by:

$$PDF(N) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-((x-\mu)^2)/2\sigma^2} \tag{14}$$

$$CDF(N) = \frac{1}{2}\left[1 - erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{15}$$

Alternatively, the CDF for the normal distribution may be obtain by the percentage probability function:

$$F(Z) = \frac{1}{1+\exp\left(-\sqrt{\pi}\left(\beta_1 Z^5 + \beta_2 Z^3 + \beta_3 Z\right)\right)} \tag{16}$$

where $Z = (x - \bar{x})/s$, $\beta_1 = 0.0004406$, $\beta_2 = 0.0418198$, and $\beta_3 = 0.90000000$. The entropy function for the normal distribution is given as:

$$E(N) = \frac{1}{2}\log\left(2\pi e\sigma^2\right) \tag{17}$$

2.3.2 Logistic distribution and its entropy function

In many instances, psychometric scales used in social science research aim to measure the pattern of emotion, judgment, loyalty or other forms of mental outlook, the data does not behave in a straight line as we might expect in normal distribution case. In many instances, these types of psychometric data will manifest a sigmoid pattern which may be captured by a logistic function. The data with this sigmoid pattern may be explained by logistic distribution. The PDF and CDF of logistic function are given by:

$$PDF(\log) = \frac{e^{-z}}{s(1-e^{-z})} \tag{18}$$

where $z = (x - \mu)/s$. The point-by-point percentage probability of x is determined by the CDF. The CDF of the logistic distribution is obtained by:

$$CDF(\log) = \frac{1}{1+e^{-z}} \tag{19}$$

The point at which the information of the logistic distribution starts to under go instability or dissipation is called entropy. Information entropy for the logistic distribution is given by:

$$E(\log) = \ln s + 2 \tag{20}$$

2.3.3 Beta distribution and its entropy function

When the data is negatively skewed, the distribution of the data may be classified as beta distribution. Negatively skewed data may provide information about the gravitation or change which the population is moving away from the mean to the upper range of the tail of the curve. In many instances, this change may signify an improvement or a pattern of adoption of new technology. The PDF and CDF of beta distributions are given by:

$$PDF(\beta) = \frac{X^{\alpha-1}(1-X)^{\beta-1}}{B(\alpha,\beta)} \tag{22}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. What is gamma ? what is a? what is b?

$$CDF(\beta) = I_x(\beta,\beta) \tag{23}$$

This is known as an incomplete beta function where $B(x;a,b) = \int_0^x t^{a-1}(1-t)dt$ for $x=1$ the incomplete beta function coincides with the complete beta function. The regularized incomplete beta function is obtained by: $I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}$.

The entropy function where the information loss is observed for the beta distribution is obtained by:

$$E(\beta) = \ln B(\alpha,\beta) - (\alpha-)\psi(\alpha+\beta) \tag{24}$$

2.3.4 Gamma distribution and its entropy function

The opposite of beta distribution is the gamma distribution. In gamma distribution, the data is positively skewed. Positively skewed data in psychometric measurement may provide information that the population is lagging behind a certain change or is slow to adopt new idea or innovation. The right-side of the tail of the curve may indicate introduction of new ideas, innovation or people's opinion starts top erode from the old mean. The PDF and CDF of the gamma distribution function are give by:

$$PDF(\gamma) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \tag{25}$$

$$CDF(\gamma) = \frac{1}{\Gamma(k)} \gamma\left(k, \frac{x}{\theta}\right) \tag{26}$$

where $k > 0 = shape$, $\theta > 0 = scale$, and $\alpha > 0 = shape$; $\beta > 0 = rate$.

The significant erosion from the old mean may be indicated by the breaking down of information. This information breakdown may be indicated by entropy. The entropy as a measure of information loss for the gamma distribution is given by:

$$E(\gamma) = k + n\theta + \ln\Gamma(k) + (1-k)\psi(k) \tag{27}$$

In cases where predictive model was used, the distribution function for each data set in the modeling may be used f

Table 1. Sample size calculation for parametric modeling

Distribution Tye	Entropy	Sample size equation
Normal	$E(N) = \frac{1}{2}\log\left(2\pi e\sigma^2\right)$	$n_\phi = \ln(n)\pi$ $n_{\phi*} = n_\phi(1+E)$ where n_ϕ is nonparametric sample size, $n_{\phi*}$ is parametric sample size, and E is the entropy function of the distribution.
Logistic	$E(\log) = \ln s + 2$	
Beta	$E(\beta) = \ln B(\alpha,\beta) - (\alpha-)\psi(\alpha+\beta)$	
Gamma	$E(\gamma) = k + n\theta + \ln\Gamma(k) + (1-k)\psi(k)$	

3.0 DATA AND M ETHODOLOGY

3.1 Data derived from content of response scale

The data used for calculating sample size is based on the component of the scale used in the survey. We assume that a good survey is one that employs a single scale throughout its survey. For instance, if a Likert scale (1,2,3,4,5) is used, all questions soliciting quantitative data in that survey would use only (1,2,3,4,5) for also questions. The characteristics of each scale type are reported in Table 1.

Table 2. Response scales and their characteristics

Scales	Type	$\overline{X} \pm S$	$\mu \pm \sigma$	Skew	Kurtosis
(1,0)	Discrete	0.50 ± 0.71	0.07 ± 0.37	-	-
(0,1,2,3)	Quantitative	1.50 ± 1.29	0.44 ± 1.28	0.00	-1.20
(1,2,3,4,5)	Likert	3.00 ± 1.58	1.84 ± 1.57	0.00	-1.20
(1,2,3,4,5,6,7)	Likert	4.00 ± 2.16	2.66 ± 2.15	0.00	-1.20
(1,2,3,4,5,6,7,8,9,10)	Likert	5.50 ± 3.03	3.93 ± 3.01	0.00	-1.20

The response scale is used as the observation set; however, the components used for purposes of calculating sample are the CDF of the scale elements. The CDF of each scale component reports the percentage probability of each component. These percentage probabilities are used to determine sample size.

With known CDF of each component of the scale, we then use Monte Carlo simulation to obtain the number of repetitions to have data points fill the probability space of a unit circle. In the final step, the sample size is obtained by multiplying the log number of Monte Carlo repetition by π. The rationale for taking the log of the number of repetition is to force fit the number of repetition into a unit circle. The rationale for using π comes from the ratio of the circumference to the diameter of the circle.

3.2 Proposed new method for minimum sample size calculation

Traditionally, power calculation is one indicator used in assessing the adequacy of sample size. Power is defined as the percentage probability of event defined by F(Z) as defined in by equation 16, *supra*. Power is the probability of not committing Type II error (failing to reject the null hypothesis despite the empirical evidence dictating otherwise). The probability of Type II error is $\beta = 1 - F(Z)$.

In this paper, we introduce a new way of calculating the power of the sample size through the use of probability space of the unit circle of the assumed normal distribution curve. We define power of the sample as the observed or empirical sample over the theoretical sample of the log number of Monte Carlo repetition (N) that would have fitted into the unit circle. This new power calculation may be obtained through the following steps:

Step 1: Determine the expected number of repetition ($\hat{N}$) for Monte Carlo simulation that would fit data points into a unit circle's probability space assuming that the data is normally distributed $N(0,1)$ with $E = 0.05$ for 95% confidence interval and $\sigma = 1.00$:

$$\hat{N} = \left(\frac{6\sigma}{E}\right)^2 \rightarrow LN(\hat{N}) \qquad (28)$$

The number 6 is used to represent 6 standard units of the sigma or standard deviation in the normal distribution curve. The error level (E) is the precision level determined by α or the

significance level in the distribution curve. The natural log value of the repetition ($\hat{N}$) is used because we need to descale the value to fit the unit circle. The expected sample size at various level of E is listed in Table 3. For our evaluation in this paper, we use $E = 0.05$ with the corresponding sample size of $n^* = 30.07$.

Step 2: Determine the empirical sample size by using the CDF of the scale as the basis:

$$n_\phi = \ln(N_o)\pi \tag{29}$$

where $N_{obs} = (3\sigma / E)^2$ with the inferential standard deviation σ calculated from $\{x_1, x_2, x_3\}$ with $x_1 = \max$, $x_2 = \min$, and $x_1 = (\max + \min)/2$, and $E_o = (\max - \min)/2)/50$.

Step 3: Power (Π) is determined by calculation the ratio of the empirical or test sample size (n_ϕ) to the theoretical sample size $(\hat{n})$ in step 1 by:

$$\Pi = 1 - \left(\frac{n_\phi - n^*}{n^*} \right) \tag{30}$$

where n_ϕ = sample size estimated by CDF of the scale components; $\hat{n}$ is the ideal sample size obtained by fitting value points into the unit circle's probability space under equation (29).

Table 3. Power calculation for sample size determined by CDF of scale components

Scale	n_ϕ	n^*	$n_\phi - n^*$	Π	Π%
(1,0)	33.61	30.07*	3.54	0.8823	88.23%
(0,1,2,3)	30.58	30.07	0.51	0.9830	98.30%
(1,2,3,4,5)	30.20	30.07	0.13	0.9957	99.57%
(1,2,3,4,5,6,7)	29.79	30.07	0.28	0.9907	99.07%
(1,2,3,4,5,6,7,8,9,10)	29.51	30.07	0.56	0.9814	98.14%

*Using E = 0.05 and $\sigma = 6$.

The power we proposed differs from the tradition power calculation for sample because we do not depend on the value of β , but based power on the ratio between the data points that can fit into the probability space of the unit circle. A theoretical or expected sample size is defined as the saturation of data points that would fit into a unit circle obtained under Monte Carlo simulation. This approach is more practical because we based our sample calculation on the CDF of the response scale used in the survey, not on parameter estimate in the population. We assert that in social science research where survey is used and quantitative response is employed, the expected value for the entire survey should lie within the range of the max and min values of the scale used in the survey. Thus, the basis for calculating sample size should be the survey scale, not some population parameter estimate. The observed value for the sample and the expected value in the population will fall within the max and min values of the scale used.

4.0 FINDINGS AND ANALYSIS

4.1 Standard sample size for quantitative survey using probability function

Using the Monte Carlo simulation to fit data points into the probability space of a unit circle, we found that for 6-sigma with error level of 0.05, the ideal sample size is 30.07. This fitted value is

used as a threshold value against which sample size obtained through survey response scales are compared. Table 3 tabulates various threshold sample size of –sigma with various levels of error.

Table 4. Threshold sample size under 6-sigma under various error level

σ	E	$\hat{N}$	$\ln(N)$	$n^* = \ln(N)\pi$
6.00	0.01	360,000	12.79	40.17
6.00	0.02	90,000	11.41	35.82
6.00	0.03	40,000	10.60	33.27
6.00	0.04	22,500	10.02	31.47
6.00	0.05	14,400	9.57	30.07
6.00	0.06	10,000	9.21	28.92
6.00	0.07	7,346.94	8.90	27.95
6.00	0.08	5,625	8.63	27.11
6.00	0.09	4,444.44	8.40	26.37
6.00	0.10	3,600	8.19	25.71

By using the cumulative probability of each element of the scale as the basis to calculate the sample size, we are able to determine the sample size for the various scales commonly used in quantitative research that employ written survey. Note that in this new approach, we also include discrete or binary data. In all cases, regardless of discrete or continuous and regardless of whether Likert or non-Likert scales used in the survey, we found that the minimum sample size is approximately 30. Table 4 shows the estimated sample size based on probability function for each survey response space.

Table 5. Using cumulative function of response scales to obtain sample size

Scales	$\overline{X}$	S	$\overline{Z}$	E	N	$n_\phi = \ln(n)\pi$
(1,0)	0.50	0.71	0.50	0.01	16,970	33.61
(0,1,2,3)	1.50	1.29	0.50	0.01	15,045	30.58
(1,2,3,4,5)	3.00	1.58	0.50	0.01	13,172	30.20
(1,2,3,4,5,6,7)	4.00	2.16	0.50	0.01	12,067	29.79
(1,2,3,4,5,6,7,8,9,10)	5.50	3.03	0.50	0.01	44,456	29.51
					Mean ± SD	30.74 ± 1.65

The use of the CDF as the basis for sample size calculation had been done in the past. However, the approach taken in the past was different than what we proposed in this paper. In the past, sample size calculation under CDF approach is obtained through:

$$n_{old} \geq \left(\frac{Z_\alpha + \Phi^{-1}(1-\beta)}{\mu^* / \sigma} \right)^2 \qquad (31)$$

Assume that for 95% confidence interval, $Z_\alpha = 1.65$ and $\beta = 0.05$, and μ and σ are inferred from an observation set. In this study, we test the power of this equation by using the response scale as the observation set. We find this approach inefficient because the use of depends on a test sample to obtain the inferential statistics of μ and σ. The values of these two components are not stable as the test sample size changes.

The comparative findings of sample size under these two methods using various response scale types are reported in table 5. Note that when use the response scale as the observation set to determine sample size, the old sample size calculation method failed to produce adequate sample size and also failed in the power test under equation.

Table 6. Comparison of two types of CDF-based sample size calculation

Scale	n_{old}	n_{new}	Π_{old}	Π_{new}
(1,0)	196.20	33.61	-4.52	0.8823
(0,1,2,3)	59.43	30.58	0.02	0.9830
(1,2,3,4,5)	5.11	30.20	0.17	0.9957
(1,2,3,4,5,6,7)	4.59	29.79	0.15	0.9907
(1,2,3,4,5,6,7,8,9,10)	4.12	29.51	0.14	0.9814

Note: Π_{old} is equation (31) and Π_{new} is equation (30). The n_{new} is n_{ϕ} that we proposed in this paper.

5.0 CONLCUSION

There are many sample size methods in the literature. These methods may be categorized into two types: (i) population-based, or (ii) data-based whether the data is discrete or continuous. Under the population-dependent approach to sample size calculation, there are at least two types of sample size calculation method depending on whether the population is known or unknown. For the data-based approach to sample size calculation, there are also many formulas used to calculate sample size depending on whether the data is discrete or continuous. These two lines of literature with their varied formulation caused uncertainty and confusion in sample size calculation. In this paper, we provide a general case of sample size calculation based on the CDF of the scale of the survey used in the research. By employing the scale as the data for Monte Carlo simulation, we are able to obtain the expected sample level at about 30 counts for both discrete and continuous data. In addition to new sample size calculation method, we also propose a new tool for power calculation. The new power calculation uses the actual sample size obtained from the survey scale and the expected sample size obtained from Monte Carlo simulation of 6-sigma case. This power calculation method is more efficient and practical.

REFERENCES

Cochran, W. G. (1963). *Sampling Techniques*, 2nd Ed., New York: John Wiley and Sons, Inc.; p. 75.

Devane D, Begley CM, Clarke M. (2004). "How many do I need? Basic principles of sample size estimation." *J. Adv. Nursing*. **47**: 297–302. https://www.ncbi.nlm.nih.gov/pubmed/15238124

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). "What is an adequate sample size? Operationalising data saturation for theory-based interview studies." *Psychology and Health*, **25**, 1229–1245. doi:10.1080/08870440903194015

Fugard A.J.B.; Potts H.W.W. (10 February 2015). "Supporting thinking on sample sizes for thematic analyses: A quantitative tool." *International Journal of Social Research Methodology*. 18 (6): 669–684. doi:10.1080/13645579.2015.1005453.

Galvin R (2015). "How many interviews are enough? Do qualitative interviews in building energy consumption research produce reliable knowledge?" *Journal of Building Engineering*, **1**: 2–12.

Glaser, B. (1965). "The constant comparative method of qualitative analysis." *Social Problems*. **12**, 436–445.

Gogtay, Nithya J. (2010). "Principles of sample size calculation." *Indian J. Ophthalmol*. 2010 Nov-Dec; **58**(6): 517–518. doi: 10.4103/0301-4738.71692

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2993982/

Guest, G., Bunce, A., & Johnson, L. (2006). "How many interviews are enough?: An experiment with data saturation and variability." *Field Methods*, **18**, 59–82. doi:10.1177/1525822X05279903

Israel, Glen D. (1992). "Determining Sample Size." Fact Sheet PEOD-6; November 1992 http://sociology.soc.uoc.gr/socmedia/papageo/metaptyxiakoi/sample_size/samplesize1.pdf

Julios S.A. (2004). "Sample sizes for clinical trials with normal data." *Stats Med.* **23**: 1921–86. https://www.ncbi.nlm.nih.gov/pubmed/15195324

Karlsson J, Engebretsen L, Dainty K. (2003). *ISAKOS scientific committee Considerations on sample size and power calculations in randomized clinical trials. Arthroscopy*; **19**: 997–9. https://www.ncbi.nlm.nih.gov/pubmed/14608320

Noordzij, M. , Tripepi, G., Friedo W., Zoccali, D.C., Tanck, M.W. and Jager, K.J. (2010). "Sample size calculations: basic principles and common pitfalls." *Nephrol Dial Transplant*, **25**: 1388–1393. CME Series. doi: 10.1093/ndt/gfp732 https://academic.oup.com/ndt/article-pdf/25/5/1388/5213790/gfp732.pdf

Onwuegbuzie, A. J., & Leech, N. L. (2007). "A call for qualitative power analyses." *Quality & Quantity*. **41**, 105–121. doi:10.1007/s11135-005-1098-1

Sandelowski, M. (1995). "Sample size in qualitative research." *Research in Nursing & Health*, **18**, 179–183.

Smith, M. F. (1983). "Sampling Considerations," *In* Evaluating Cooperative Extension Programs. *Florida Cooperative Extension Service Bulletin PE-1*. Institute of Food and Agricultural Sciences. University of Florida.

Wright, A., Maloney, F. L., & Feblowitz, J. C. (2011). "Clinician attitudes toward and use of electronic problem lists: a thematic analysis." *BMC Medical Informatics and Decision Making*, **11**, 36. doi:10.1186/1472-6947-11-36

Yamane, Taro (1967). *Statistics, An Introductory Analysis*, 2nd Ed., New York: Harper and Row; p. 886.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences*; P. 67. Routledge. ISBN 978-1-134-74270-7.