Scholarly Research Journal for Interdisciplinary Studies,
Online ISSN 2278-8808, SJIF 2021 = 7.380, www.srjis.com
PEER REVIEWED & REFEREED JOURNAL, JAN-FEB, 2023, VOL- 10/75
10.21922/srjis.v10i75.123021

CALCULATING THE OPTIMAL SAMPLE SIZE IN DIFFERENT SITUATIONS

Deepa Anwar[1]

Research Scholar, School of Mathematics and Allied Sciences

Rajshree Mishra[2]

Associate Professor (Dept.of Maths) Government Model Science College, Gwalior

J. P. Verma[3]

Vice Chancellor, Sri Sri Aniruddhadeva Sports University Chabua, Dibrugarh, Assam

Paper Received On: 25 FEBRUARY 2023

Peer Reviewed On: 28 FEBRUARY 2023

Published On: 01 MARCH 2023

Abstract

The belief is wide spread that studies are unethical if their sample size is not large enough to ensure adequate power .An important step when designing an empirical study is to justify the optimum sample size that will be required. Inappropriate, inadequate, or excessive sample sizes continue to influence the quality and accuracy of research. In order to generalize from a random sample and avoid sampling errors or biases, a random sample needs to be of adequate size .The aim of a sample size justification is to explain how the collected data is expected to provide valuable information given the inferential goals of the researcher. Two distinct investigations conducted on a same sample with the same methodology and achieving equivalent results are different only in terms of sample size .This manuscript describes the procedures for determining sample size using different formulas, a table is provided that can be used to select the sample size for a research problem based on three alpha levels (1%, 5% and10%)and a set of error rate. There are a number of practical issues in selecting the values for the parameters required in the sample size calculation formula. This study presents a summary of how to calculate the survey sample size in social research, information system research, industry, agriculture, and medical studies, to name just a few .In this context, sample size formulae given by different authors have also been discussed in the present article.

Keywords- *Statistical Power, Sample Size, Level of Significance, Sample Size Table*

INTRODUCTION

A common goal of survey research is to collect data representative of a population to generalize findings from a drawn sample back to a population. Samples should not be either too big or too small since both have limitations that can compromise the conclusions drawn from the studies.

There are some questions for which an experiment can't help us find the answer. For example, suppose we wanted to know what percentage of Indian smoke cigarettes, or what percentage of local market chicken is contaminated with salmonella bacteria. There is no experiment that can be done to answer these types of questions. We could test every chicken on the market, or ask every person if they smoke. This is a census, a count of each and every item in a population. The Census Method is also known as a Complete Enumeration Survey Method. . It provides intensive and in-depth information covering many facets of the problems. This method is the most commonly used by the government in connection with the national population, housing census, agriculture census, etc. where vast knowledge about these fields is required. . Same trend goes in the share market, every day market value is as important as the previous one.

For example, in a population census, not only the number of persons is counted, but the information is also collected on various other parameters like the number of males and females, age, education, marital status, occupational level, income health conditions, etc. Since, in this type of investigation, every item of the universe is taken into account, the conclusions are more accurate and reliable. This method is meaningless in the case of an infinite universe where the number of items is unlimited.

It seems like a census would be a straightforward way to get the most accurate, thorough information. But taking an accurate census is more difficult than it seems. There is an alternative to a census, and that is a sample. While a census is an attempt to gather information about every member of the population, sampling gathers information only about a part of population, the sample, to represent the whole. A sample survey is a type of method that is used for collecting data from or about the members of a population so that inferences about the entire population can be obtained from a subset, or sample, of the population members. Sampling is widely used in a variety of areas such as industry, manufacturing, agriculture, and medical studies, to name just a few.

"The larger the Sample Size the better the study" is not always true. One of the aims of applying appropriate sample size calculation formula is to get an optimum or adequate sample size. Availability of resources sets upper limit of the sample size and required accuracy sets lower limit of sample size. If sample size is too small it may fail to detect important effects or associations and if the sample size is too large it may result in loss of accuracy.

"How big sample size is sufficient for any type of qualitative /quantitative research" is one of the most frequently asked question by researcher. It can be calculated using a simple formula as the calculation needs only a few simple steps. However, the decision to select the appropriate values of parameters required in the formula is not simple as the sample size formula always. Sample size calculation for a study, estimating a population has been shown in many books e.g., Cochran (1977), Singh and Chaudhury (1985)., Mark (2005) and J.P. verma (2020)

DEFINTIONS

Population Size

This is the total number of distinct individuals in the population. If the population is large, but the researcher doesn't know how large, they can conservatively use 100,000. The sample size doesn't change much for populations larger than 100,000.

Sample Size

The size of the sample determines the probability of errors in the outcome, i.e. the larger the size of sample, the less are the chances of errors and the smaller the size; the higher are the chances of errors.

This is the minimum sample size need to estimate the true population proportion with the required margin of error and confidence level. Note that if some people choose not to respond they cannot be included in sample and so if no response is a possibility, then the sample size will have to be increased accordingly. In general, the higher the response rate the better the estimate, as no response will often lead to biases in estimation.

Population proportion (p)

It is a fraction of the population that has a certain characteristic. For example, suppose there are 1,000 people in the population and 235 of those people have blue eyes. The fraction of people who have blue eyes is 235 out of 1,000, or 235/1000. In statistics, a population proportion is a parameter that describes a percentage value associated with a population. The letter p or π is used for the population proportion, so p=235/1000 Or 0.235 or 23.5%

$$p = x/n \qquad (1)$$

Where: "x" is the number of items the researcher is interested in, and "n" is the total number of items in the population.

Sample proportion

The sample proportion is the expected results. This can often be determined by using the results from a previous survey, or by running a small pilot study. If unsure, use 50%, which is conservative and gives the largest sample size. If, the sample proportion is closer to 0 or 1 then this approximation is not valid and the researcher needs to consider an alternative method. The sample proportion is denoted by pˆ. Thus if 23.5% of people have blue eyes , p = 0.235 ; so, if in a sample of 200 people , 78 have blue eyes , then sample proportion pˆ=78/200=0.39.

$$\hat{p}=x/n \qquad (2)$$

Where:

"x" is the number of items the researcher is interested in, and

"n" is the total number of items in the sample.

Confidence level

A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of time, or it is the probability that the margin of error contains the true proportion. So, in statistics, it is another way to describe probability. For example, if we construct a confidence interval with a 95% confidence level, we are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval .If the study was repeated and the range is calculated each time, it would expect the true value to lie within these ranges on 95% of occasions, this means you have a 5% probability of incorrectly detecting a significant difference when one does not exist, i.e., a false positive result (otherwise known as type I error). The higher the confidence level the more certain it can be that the interval contains the true proportion.

2.6 Alpha Level (Level of significance, P-Value) or (prob. of Type I Error)

The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. For example, a significance level of .05 indicates a 5% risk of concluding that a difference exists when there is no actual difference. In social science, p =< .10 considers at 10% level of significant due to more heterogeneity of data set where as in medical science it will be p=.01 (1% level or less).The alpha level used in determining sample size in most educational research studies is either 5% (that is, P = 0.05) or 1% (P = 0.01) (Ary, Jacobs, & Razavieh, 1996). . In the field experiments, p-value is generally calculated at 5% level of significance, in general, an alpha level of 5% (P=0 .05) is acceptable for most research. An alpha level of .10 or lower may be used if the researcher is more interested in identifying marginal relationships, differences, or other statistical phenomena as a precursor to further studies. An alpha level of .01 may be used in those cases where decisions based on the research are critical and errors may cause substantial financial or personal harm.

Margin of Error.

A margin of error tells how many percentage points results will differ from the real population value. For example If the survey results show that 1,875 (75%) out of 2,500 students prefer e Books, a statistician would say that your margin of error, at a 95% confidence level, would be ±2%. In other words, the statistician is 95% confident that 73-77% of students would prefer eBooks if we surveyed population (all students). The margin of error can be calculated in two ways, depending on whether we have parameters from a population or statistics from a sample:

Margin of error (parameter) = Critical value x Standard deviation for the population.

Margin of error (statistic) = Critical value x Standard error of the sample.

The general rule relative to acceptable margins of error in educational and social research is as follows: For categorical data, 5% margin of error is acceptable, and, for continuous data, 3% margin of error is acceptable (Krejcie & Morgan, 1970).

Variance Estimation

Two sets of data can have the same mean but be distributed very differently. Variance measures how spread out a set of data is. More specifically, variance measures how far each number in the set is from the mean (average), A critical component of sample size formulas is the estimation of variance in the primary variables of interest in the study. The researcher does not have direct control over variance and must incorporate variance estimates into research design. Cochran (1977) 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.

The first three ways are logical and produce valid estimates of variance; therefore, they do not need to be discussed further. However, in many educational and social research studies, it is not feasible to use any of the first three ways and the researcher must estimate variance using the fourth method.

Power: (1-β)

To discuss and understand power, one must be clear on the concepts of Type I and Type II errors. Doug Rush provides a refresher on Type I and Type II errors (including power and effect size) in the Spring 2015 issue of the Statistics Teacher Network, but, briefly, a Type I Error is rejecting the null hypothesis when it is true, and a Type II Error is failing to reject a null hypothesis when it is false. The probability of a Type I error is typically known as Alpha (α), while the probability of a Type II error is typically known as Beta(β). Power is the probability of rejecting the null hypothesis when, in fact, it is false. Hence Power is the probability of avoiding a Type II error. Or the probability of obtaining a value of t (or Z) that is large enough to reject Ho when Ho is actually false . Mathematically, power is 1 –β. The power of a hypothesis test is between 0 and 1; if the power is close to 1, the hypothesis test is very good at detecting a false null hypothesis. Extremely low alpha leads to very high beta errors thus low power of test. similarly smaller the standard deviation, the higher the power and larger the sample .Beta is commonly set at 0.2 (power ≥0·8), experimental cannot control beta since it is dependent on the alpha .

Z statistic (Z)

For the level of confidence of 95%, which is conventional, Z value is 1.96. In these studies, researcher present their results with 95% confidence intervals (CI). Researcher who wants to be more confident (say 99%) about their estimates, the value of Z is set at 2.58.

Precision (d or e)

It is very important for researcher to understand the value of d very well**. Precision is how close two or more measurements are to each other.** If, wants to know which set of data is more precise, find the range (the difference between the highest and lowest scores). Sample A: 32.56 – 32.48 = .08.

Sample B: 15.38 – 15.32 = .06.

Sample B has the lowest range (.06) and so is the more precise.

From the formula, it can be conceived that the sample size varies inversely with the square of the precision (d2).

HOW TO CALCULATE THE SAMPLE SIZE

Strategies for Determining Sample Size (Glenn 1992, Rao 1985 and Sudman 1976, Singh and Masuku 2013) There are many approaches to determine the sample size. This includes, using a census for small populations, imitating a sample size of similar studies, using published tables, and also applying formulas to calculate a sample size.

Using a Census for Small Populations

Consideration of entire population as a sample is one of the mean approaches. Although cost considerations make this impossible for large populations, a census is more attractive for small populations (e.g., 200 or less). A census eliminates sampling error and provides data on all the individuals in the population. In addition, some costs such as questionnaire design and developing the sampling frame are "fixed," that is, they will be the same for samples of 50 or 200. Therefore, entire population will have to be sampled in small populations to achieve a desirable level of precision.

Using a Sample Size of a Similar Study

Another approach is to use the same sample size as those of studies similar to the plan. Without reviewing the methods used in these studies may run the risk of repeating errors that were made in determining the sample size for another study. However, a review of the literature in this discipline can provide supervision about typical sample sizes that are used (Glenn1992).

Using Published Tables

A third way to determine sample size is to rely on published tables, which provide the sample size for a given set of criteria. Sample sizes that would be necessary for given combinations of precision (d or e), confidence level and variability. Glenn (1992), presented three tables for the selection of sample size (Table-1 , Table-2 and Table-3). Please note two things. First, these sample sizes reflect the number of obtained responses and not necessarily the number of surveys mailed, or interviews planned. Second, the sample sizes in Table 2 presume that the attributes being measured are distributed normally or nearly so. If this assumption cannot be met, then the entire population may need to be surveyed.

Table 1

Table 1: Table for Determining Minimum Returned Sample Size for a Given Population Size for Continuous and Categorical Data

	Sample size					
	Continuous data (margin of error = .03)			Categorical data (margin of error = .05)		
Population size	alpha = .10 t = 1.65	alpha = .05 t = 1.96	alpha = .01 t = 2.58	p = .50 t = 1.65	p = .50 t = 1.96	p = .50 t = 2.58
100	46	55	68	74	80	87
200	59	75	102	116	132	154
300	65	85	123	143	169	207
400	69	92	137	162	196	250
500	72	96	147	176	218	286
600	73	100	155	187	235	316
700	75	102	161	196	249	341
800	76	104	166	203	260	363
900	76	105	170	209	270	382
1,000	77	106	173	213	278	399
1,500	79	110	183	230	306	461
2,000	83	112	189	239	323	499
4,000	83	119	198	254	351	570
6,000	83	119	209	259	362	598
8,000	83	119	209	262	367	613
10,000	83	119	209	264	370	623

NOTE: The margins of error used in the table were .03 for continuous data and .05 for categorical data. Researchers may use this table if the margin of error shown is appropriate for their study; however, the appropriate sample size must be calculated if these error rates are not appropriate. Table developed by Bartlett, Kotrlik, & Higgins.

Table 2 Sample Size for ±5% and ±10% Precision Levels where Confidence Level is 95% and P(population proportion)=0.5 when population size is less than equal to 500

Size of Population	Sample Size (n) for Precision (e) of:	
	±5%	±10%
100	81	51
125	96	56
150	110	61
200	134	67
250	154	72
300	172	76
350	187	78
400	201	81
450	212	82

Table 3. Sample Size for ±5% and ±10% Precision Levels where Confidence Level is 95% and P(population proportion)=0.5 when population size is more than equal to 500

Size of Population	Sample Size (n) for precision (e)	
	±5%	±10%
500	222	83
1,000	286	91
2,000	333	95
3,000	353	97
4,000	364	98
5,000	370	98
7,000	378	99
9,000	383	99
10,000	385	99
15,000	390	99
20,000	392	100
25,000	394	100
50,000	397	100
100,000	398	100

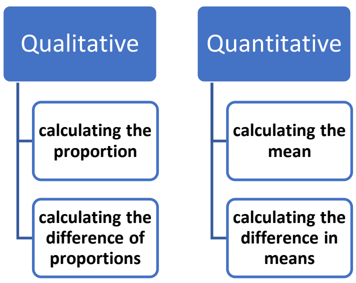
Critical values (Z value) at important level of significance are given below

Table 4

Level of significance	**1%**	**5%**	**10%**
Two tailed test	**2.58**	**1.96**	**1.645**
One tailed test	**2.33**	**1.645**	**1.282**

Using formulas to calculate a sample size.

Based on what we are measuring, there are four types of studies



Sample size for the mean

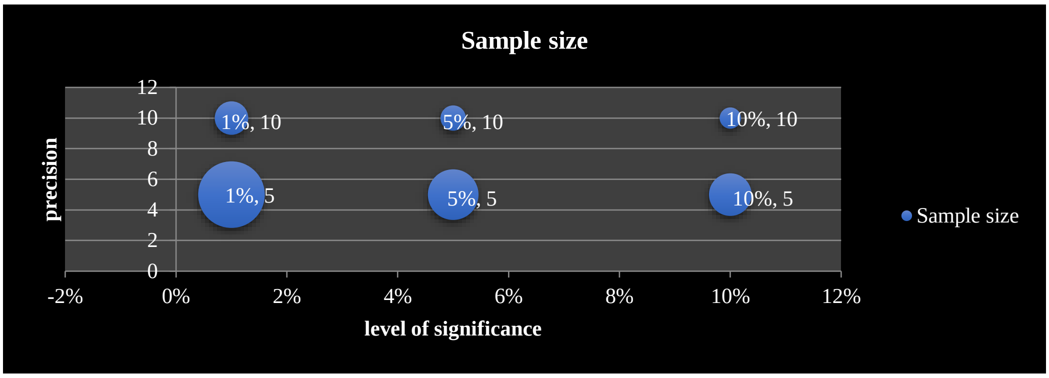
$$n=(Z^2 (SD)^2)/(e)^2 \quad (3)$$

where Z is critical value of confidence level(at 95% it is 1.96), SD is Standard deviation of variable. Value of standard deviation can be taken from previously done study or through pilot study, e (or d) is allowable error (or precision) .

Thus, the minimum sample size (n) required for estimating a population mean is directly proportional to the level of confidence and the standard deviation, whereas it is inversely proportional to the absolute error that is allowed in the estimation.

Table 5 Sample Size for ±5% Precision Levels where Confidence Level is 90%, 95% and 99% and SD is 25(assumed)

Level of significance	Z	SD	e(or d)	n (Calculated Sample size)	Sample size
10%	1.645	25	5	67.65	68
5%	1.96	25	5	96.04	96
1%	2.58	25	5	166.41	166
10%	1.645	25	10	16.91	17
5%	1.96	25	10	24.01	24
1%	2.58	25	10	41.60	42



Following results are showing that as the level of significance increases, the sample size decreases .similarly when the precision increases, the sample size decreases.so maximum level of significance with maximum precision require small sample size.

Sample size for proportions and prevalence (Cochran's formulas for categorical data when the population is infinite:): Danial (1999)

$$n=(Z^2 P(1-P))/d^2 \quad (4)$$

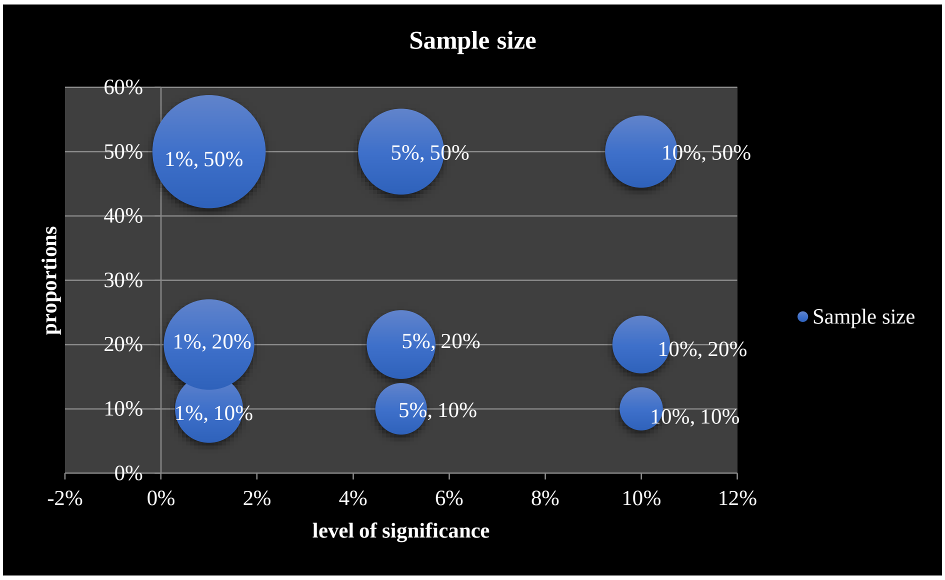
Where, n is sample size , Z is statistic for a confidence level according to the standard normal distribution (for the level of confidence of 95%, which is conventional, Z value is 1.96 for 0.025 in each tail), P is estimated prevalence or proportions of project area(in proportion of one; if 20%, P = 0.2, p is considered 0.5, when unknown), d is precision , the acceptable margin of error for proportion being estimated, (in proportion of one; if 5%, d = 0.05 to produce good precision and smaller error of estimate).

For example, calculation of sample size of a large population whose degree of variability is not known. Assuming the maximum variability, which is equal to 50% (p =0.5) and taking 95%

confidence level with ±5% precision, the calculation for required sample size will be as follows—

Table 6 Sample Size for ±5% Precision Levels where Confidence Level is 90%, 95% and 99% and estimated proportion is 10%, 20% and 50% .

Level of significance	Z	P	d	n(Calculated Sample size)	Sample size
10%	1.645	10%	0.05	97.41	97
10%	1.645	20%	0.05	173.18	173
10%	1.645	50%	0.05	270.60	271
5%	1.96	10%	0.05	138.29	138
5%	1.96	20%	0.05	245.86	246
5%	**1.96**	**50%**	**0.05**	**384.16**	**384(Standard)**
1%	2.58	10%	0.05	239.63	240
1%	2.58	20%	0.05	426.00	426
1%	2.58	50%	0.05	665.64	666



Following results are showing that as the level of significance increases, the sample size decreases .similarly when the proportions increases, the sample size also increases.so maximum level of significance with minimum proportion require small sample size.

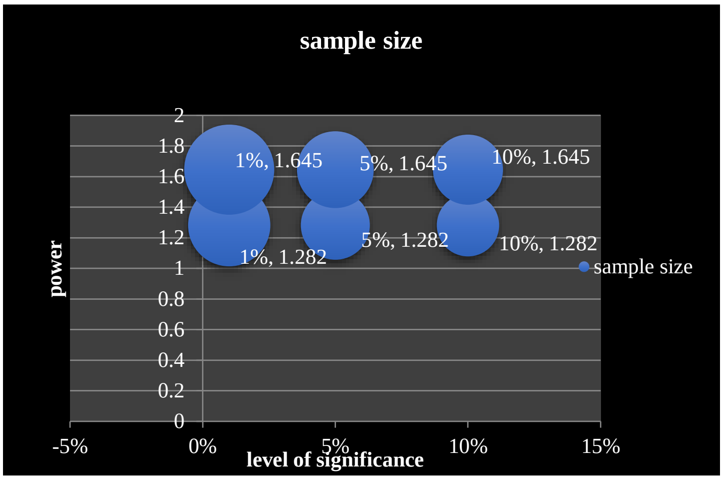
Sample size for two means

n= ((u+v)^2 (σ_1^2+σ_2^2))/(μ_1-μ_2)^2 (5)

where μ_1-μ_2 is difference between means , σ_1^ ,σ_2^ standard deviation ,u one sided percentage point of Normal Distribution corresponding to required power, v two sided percentage point of Normal Distribution corresponding to required significance level.

Table 7 Sample size for mean deference for two groups is 2 unit (assumed) ,standard deviation for both group suppose be same 5 unit, power 95% (u=1.645),significance level is 10%,5% and 1% ,SD 5 (assumed)

Level of significance	v	Mean defference (let)	Standard deviation (let)	U (power)	n(Calculated sample size)	Sample size
10%	1.645	2	5	1.282	107.09	107
10%	1.645	2	5	1.645	135.30	135
5%	1.96	2	5	1.282	131.38	131
5%	1.96	2	5	1.645	162.45	162
1%	2.58	2	5	1.282	186.43	186
1%	2.58	2	5	1.645	223.13	223



Following results are showing that as the level of significance increases, the sample size decreases .similarly when the power increases, the sample size also increases. so maximum level of significance with minimum power require small sample size.

Comparison of two proportion

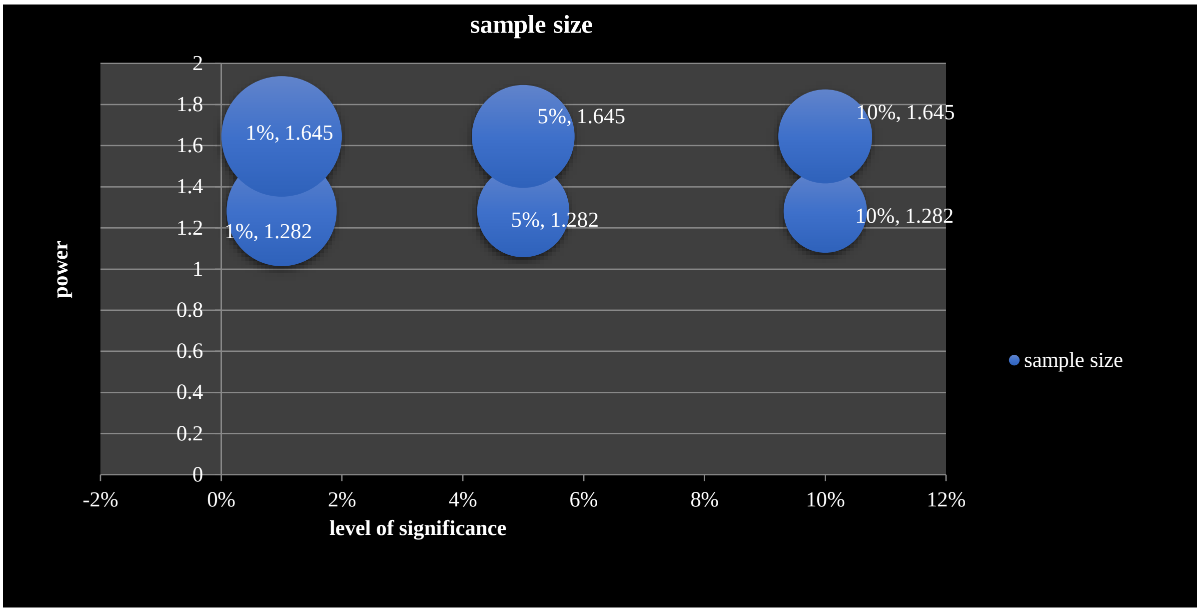
n=(2(u+v)^2 [p·(1-p)])/((P_1-P_2)) (6.1)

p=(p_2+p_1)/2 (6.2)

Where P1, P2 proportions, u one sided percentage point of Normal Distribution corresponding to required power, v two sided percentage point of Normal Distribution corresponding to required Power.

Table 8

Level of significance	v	Proportion (p1)	Proportion (p2)	u (Power)	n(Calculated sample size)	Sample size
10%	1.645	0.25	0.28	1.282	111.24	111
10%	1.645	0.25	0.28	1.645	140.55	141
5%	1.96	0.25	0.28	1.282	136.47	136
5%	1.96	0.25	0.28	1.645	168.75	169
1%	2.58	0.25	0.28	1.282	193.67	194
1%	2.58	0.25	0.28	1.645	231.79	232



Following results are showing that as the level of significance increases, the sample size decreases. similarly, when the power increases, the sample size also increases. so maximum level of significance with minimum power requires small sample size.

There are some other methods for calculating sample size, we will apply the appropriate formula according to the study design in question and determine the sample size.

For qualitative variable

$$n=\frac{r+1}{r}\frac{(p^*)(1-p^*)(z_\beta+z_{\alpha/2})^2}{(p_1-P_2)^2} \qquad (7)$$

r = Ratio of control to cases, 1 for equal number of case and control

p^* = Average proportion exposed = (proportion of exposed cases + proportion of control exposed)/2

z_β = Standard normal variate for power = for 80% power it is 0.84 and for 90% value is 1.28. Researcher has to select power for the study.

$z_{\alpha/2}$ = Standard normal variate for level of significance.

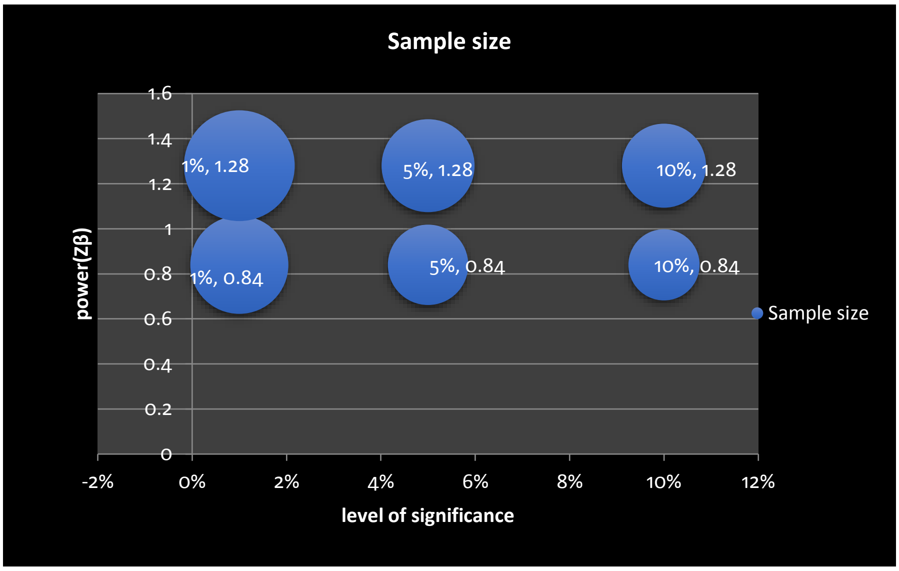
p_1-P_2 = Effect size or different in proportion expected based on previous studies.

p_1 is proportion in cases and

P_2 is proportion in control.

Table 9

Level of significance	r	P1	P2	Zβ	Zα/2	P*	n(Calculated Sample size)	Sample size
10%	1	0.35	0.2	0.84	1.645	0.275	109.43	109
10%	1	0.35	0.2	1.28	1.645	0.275	151.62	152
5%	1	0.35	0.2	0.84	1.96	0.275	138.94	139
5%	1	0.35	0.2	1.28	1.96	0.275	186.04	186
1%	1	0.35	0.2	0.84	2.58	0.275	207.28	207
1%	1	0.35	0.2	1.28	2.58	0.275	264.05	264



Following results are showing that as the level of significance increases, the sample size decreases .similarly when the power increases, the sample size also increases. so maximum level of significance with minimum power require small sample size.

For quantitative variable (association between two variables)

n=(r+1)/r (sD^2 (z_β+z_(α/2))^2)/d^2 (8)

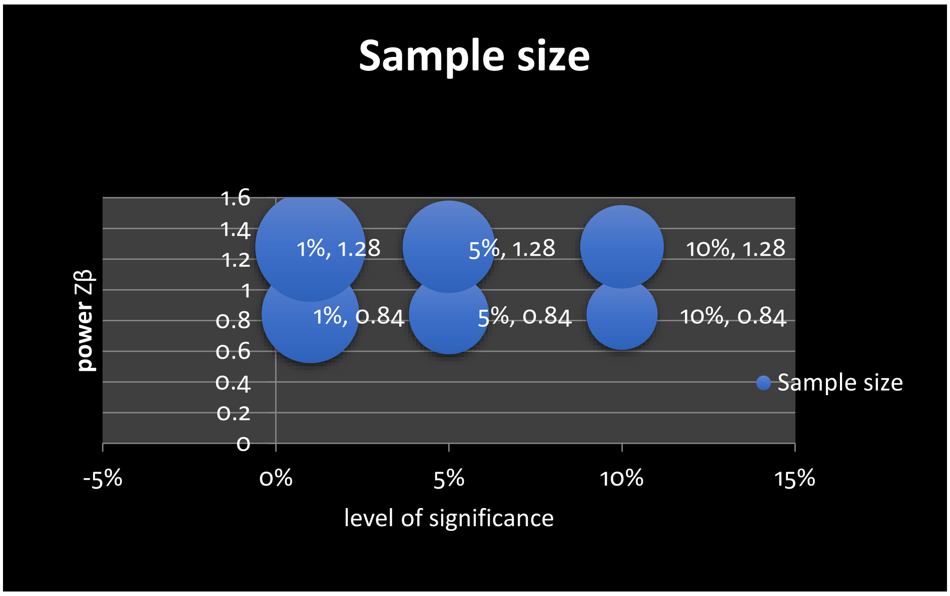
SD = Standard deviation = researcher can take value from previously published studies

d = Expected mean difference between case and control (may be based on previously published studies.)

r, z_β , z_(α/2) are already explained in previous sections.

Table 10

level of significance	r	SD	d	Zβ	Zα/2	n(Calculated Sample size)	Sample size
10%	1	1	0.25	0.84	1.645	197.60	198
10%	1	1	0.25	1.28	1.645	273.78	274
5%	1	1	0.25	0.84	1.96	250.88	251
5%	1	1	0.25	1.28	1.96	335.92	336
1%	1	1	0.25	0.84	2.58	374.28	374
1%	1	1	0.25	1.28	2.58	476.78	477



Following results are showing that as the level of significance increases, the sample size decreases. similarly, when the power increases, the sample size also increases. so maximum level of significance with minimum power requires small sample size.

Formula for sample size calculation for comparison between two groups when endpoint is qualitative.

When the endpoint of a clinical intervention study is qualitative like alive/dead, diseased/non diseased, male/ female etc., then the following formula can be used for sample size calculation for comparison between two groups

$$n=(2(P)(\vdash P)\ (z\_\beta+z\_(\alpha/2)\ )\hat{}2)/(p\_1-p\_2\ )\hat{}2 \qquad (9)$$

z_(α⁄2)= Z0.05/2 = Z0.025 = 1.96 (From Z table) at type 1 error of 5%

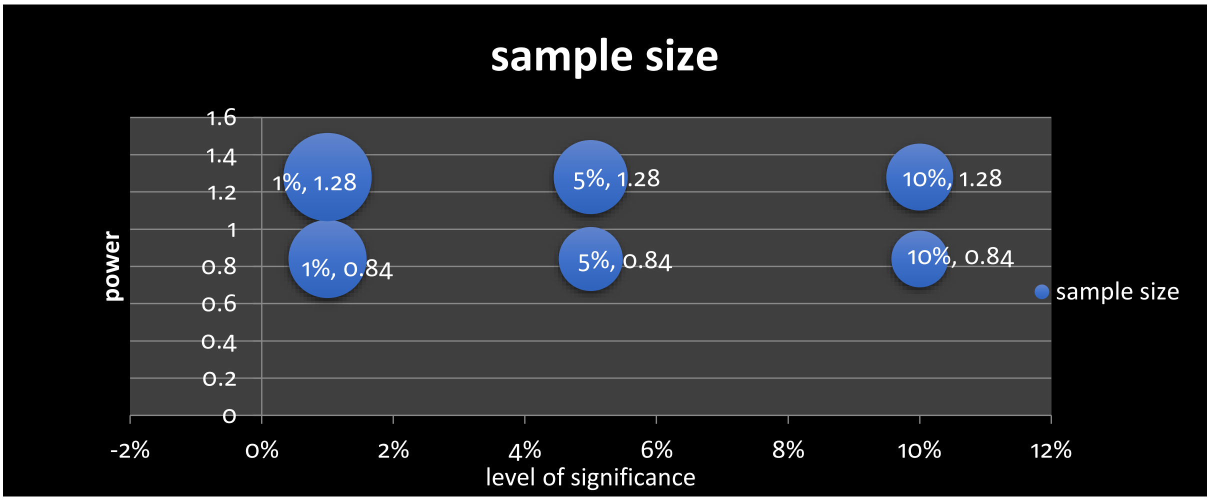
z_β = Z0.20= 0.842 (From Z table) at 80% power

p_1-p_2 = Difference in proportion of events in two groups

P = Pooled prevalence = [prevalence in case group (p_1) + prevalence in control group (p_2)]/2

Table 11

Level of significance	Zα/2	Zβ	P1	P2	p	n(Calculated Sample size)	Sample size
10%	1.645	0.84	0.35	0.2	0.275	109.43	109
10%	1.645	1.28	0.35	0.2	0.275	151.62	152
5%	1.96	0.84	0.35	0.2	0.275	138.94	139
5%	1.96	1.28	0.35	0.2	0.275	186.04	186
1%	2.58	0.84	0.35	0.2	0.275	207.28	207
1%	2.58	1.28	0.35	0.2	0.275	264.05	264



Following results are again showing as similar as previous results that as the level of significance increases, the sample size decreases. similarly, when the power increases, the sample size also increases. so maximum level of significance with minimum power requires small sample size.

Sample size formula for animal studies

Sometimes it is not possible to get information related to the prerequisites needed for sample size calculation by power analysis like standard deviation, effect size etc. In that condition a second method can be used this is called as "resource equation method". In this method a value E is calculated based on decided sample size. The value E should lie within 10 to 20 for optimum sample size. If a value of E is less than 10 then more animal should be included and if it is more than 20 then sample size should be decreased.

$$E = \text{Total number of animals} - \text{Total number of groups} \qquad (10)$$

Suppose in an animal study a researcher formed 4 groups of animals having 8 animals each for different interventions then total animals will be 32 (4 × 8). Hence E will be E = 32 – 4 = 28 This is more than 20 hence animals should be decreased in each group. So if researcher takes 5 animals in each group then E will be E = 20 – 4 = 16

This is a crude method and should be used only if sample size calculation cannot be done by power analysis method.

PRACTICAL ISSUES IN DETERMINING SAMPLE SIZE PARAMETERS

Assumption of Normal Approximation

The above sample size calculation formulae are based on the assumption of normal approximation. It says that nP and n(1-P) must be greater than 5 (Daniel, 1999). Small sample sizes might not fulfill this assumption, and we should check this assumption after calculating the sample size.

$$d = Z \times SE_{(p)} \qquad (11.1)$$

$$SE_{(p)} = \sqrt{\frac{P(1-P)}{n}}, \qquad (11.2)$$

$$d = Z\sqrt{\frac{P(1-P)}{n}} \qquad (11.3)$$

Finite Population Correction

The above mentioned sample size formulae are valid if the calculated sample size is smaller than or equal to 5% of the population size ($n/N \leq 0.05$) (Daniel, 1999). If this proportion is larger than 5% ($n/N > 0.05$), we need to use the formula with finite population correction (Daniel, 1999) as follows.

$$n' = \frac{NZ^2 P(1-P)}{d^2 (N-1) + Z^2 P(1-P)} \quad (12)$$

where n' = sample size with finite population correction,

N = Population size,

Z = Z statistic for a level of confidence,

P = Expected proportion (in proportion of one), and

d = Precision (in proportion of one).

Cluster or Multistage Sampling

The above mentioned all sample size formulae are valid only if we apply the simple random or systematic random sampling methods. Cluster or multistage sampling methods require a larger sample size to achieve the same precision. Therefore, the calculated sample size using the above formulae need to be multiplied by the design effect (deff) (Cochran, 1977).

A Helpful Calculator

In many situations, researcher need to calculate repeatedly, ensure the assumption, and check the necessity of applying the finite population correction. The authors have developed a calculator (Naing et al. 2006) using Microsoft Excel . The calculator is designed to give the sample size for various precisions (error of estimate) with or without finite population correction, and also will suggest the need to apply the finite population correction. It will also determine whether the normal approximation assumption is met or not.

$$n = \frac{p(100-p)z^2}{E^2} \quad (13)$$

n is the required sample size, P is the percentage occurrence of a state or condition, E is the percentage maximum error required, Z is the value corresponding to level of confidence required. There are two key factors to this formula (Bartlett et al., 2001).

First, there are considerations relating to the estimation of the levels of precision and risk that the researcher is willing to accept: E is the margin of error (the level of precision) or the risk the researcher is willing to accept (for example, the plus or minus figure reported in newspaper poll results). In the social research a 5% margin of error is acceptable. So, for example, if in a survey on job satisfaction 40% of respondents indicated they were dissatisfied would lie between 35% and 45%. The smaller the value of E the greater the sample size required as technically speaking sample error is inversely proportional to the square root of n, however, a large sample cannot guarantee precision (Bryman and Bell, 2003). Z concern the level of confidence that the results revealed by the survey findings are accurate. What this means is the degree to which we can be sure the characteristics of the population have been accurately estimated by the sample survey. Z is the statistical value corresponding to level of confidence required. The key idea behind this is that if a population were to be sampled repeatedly the average value of a variable or question obtained would be equal to the true population value. In management research the typical levels of confidence used are 95 percent (0.05: a Z value equal to 1.96) or 99 percent (0.01: Z=2.57). A 95 percent level of confidence implies that 95 out of 100 samples will have the true population value within the margin of error (E) specified. The second key component of a sample size formula concerns the estimation of the variance or heterogeneity of the population (P). Management researchers are commonly concerned with determining sample size for issues involving the estimation of population percentages or proportions (Zikmund, 2002). In the formula, the variance of a proportion or the percentage occurrence of how a particular question, for example, will be answered is P(100-P). Where, P= the percentage of a sample having a characteristic, for example, the 40 % of the respondents who were dissatisfied with pay, and (100-P) is the percentage (60%) who lack the characteristic or belief. The key issue is how to estimate the value of P before conducting the survey? Bartlett et al. (2001) suggest that researchers should use 50% as an estimate of P, as this will result in the maximization of variance and produce the maximum sample size (Bartlett et al., 2001). The formula for determining sample size, of the population has virtually no effect on how well the sample is likely to describe the population and as Fowler (2002) argues, it is most unusual for it (the population fraction) to be an important consideration when deciding on sample size (Fowler, 2002).

The finite population correction factor for the variance of the estimator (Kish, 1965). The variance of a sample mean is which for finite populations is multiplied by the finite population correction factor of the standard error: where N is the size of the population, and n is the size of the sample. When N is much larger than n, the correction factor will be close to 1 (and therefore this correction is typically ignored when populations are very large, even when populations are finite), and will not have a noticeable effect on the variance. When the total population is measured the correction factor is 0, such that the variance becomes 0 as well. For example, when the total population consists of 100 top athletes, and data is collected from a sample of 35 athletes, the finite population correction is = 0.81. The superb R package can compute population corrected confidence intervals (Cousineau & Chiasson, 2019).

CONCLUSIONS

In designing a study, sample size calculation is important for methodological and ethical reasons, as well as for reasons of human and financial resources. When reading an article, the reader should be on the alert to ascertain that the study they are reading was subjected to sample size calculation. In the absence of this calculation, the findings of the study should be interpreted with caution. An appropriate sample renders the research more efficient: Most sample size calculations assume that the population is large (or even infinite). With a finite, small population, the variability of the sample is actually less than expected, and therefore a "finite population correction", FPC, can be applied to account for this greater efficiency in the sampling process. For a large population (greater than 100,000 or so), there's not normally any correction needed to the standard sample size formulae available. For large, finite populations, the FPC will have little effect and the sample size will be similar to that for an infinite population. However, the effect of the FPC will be noticeable if one or both of the population sizes (N's) is small relative to n.

It seems in all sample size calculation methods if the level of significance increases, the sample size decreases. similarly when the power increases, the sample size also increases. so maximum level of significance with minimum power requires small sample size. The use of sample size calculation directly influences research findings. Very small samples undermine the internal and external validity of a study. Very large samples tend to transform small differences into statistically significant differences — even when they are clinically insignificant. As a result, both researchers and clinicians are misguided, which may lead to failure in treatment decisions.

REFERENCES

Bacchetti, Peter, et al. "Ethics and sample size." American journal of epidemiology 161.2 (2005): 105-110.

Charan, Jaykaran, and Tamoghna Biswas. "How to calculate sample size for different study designs in medical research?" Indian journal of psychological medicine 35.2 (2013): 121-126

Cochran, William G. Sampling techniques. John Wiley & Sons, 1977

Daniel, Wayne W., and Chad L. Cross. Biostatistics: a foundation for analysis in the health sciences. Wiley, 2018

Dell, Ralph B., Steve Holleran, and Rajasekhar Ramakrishnan. "Sample size determination." ILAR journal 43.4 (2002): 207-213.

Dodge, Yadolah. The concise encyclopedia of statistics. Springer Science & Business Media, 2008.

Gonick, L., and W. Smith. "The Cartoon Guide to Statistics, HarperPerennial, New York, NY." (1993).

Jones, S., S. Carley, and M. Harrison. "An introduction to power and sample size estimation." Emergency medicine journal: EMJ 20.5 (2003): 453.

Kadam, Prashant, and Supriya Bhalerao. "Sample size calculation." International journal of Ayurveda research 1.1 (2010): 55.

Lakens, Daniel. "Sample size justification." Collabra: Psychology 8.1 (2022): 33267.

Moore, David S., George P. McCabe, and Bruce A. Craig. "Introduction to the practice of statistics. New York: W. H." (1999).

Naing, Nyi Nyi. "Determination of sample size." The Malaysian journal of medical sciences: MJMS 10.2 (2003): 84.

Singh, Ajay S., and Micah B. Masuku. "Sampling techniques & determination of sample size in applied statistics research: An overview." International Journal of economics, commerce and management 2.11 (2014): 1-22.

Taherdoost, Hamed. "Determining sample size; how to calculate survey sample size." International Journal of Economics and Management Systems 2 (2017).

Wang, Hansheng, and Shein- Chung Chow. "Sample size calculation for comparing proportions." Wiley Encyclopedia of clinical trials 10 (2007): 9781118445112

Whitley, Elise, and Jonathan Ball. "Statistics review 4: sample size calculations." Critical care 6.4 (2002): 1-7. *1, No. 3, 2016.*