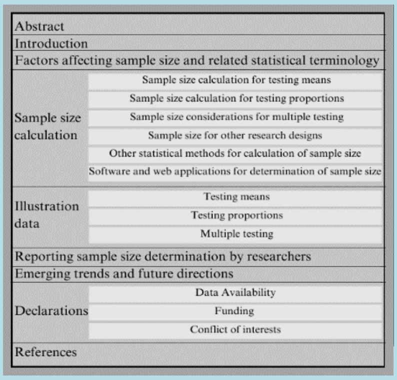
An-Najah University Journal for Research – A
Natural Sciences

A Review of Sample Size Determination for Common Experimental Designs: Further Simplified Equations

Received 27th May. 2023, Accepted 27th Aug. 2023, Published 1st Feb, 2024, DOI: 10.35552/anujr.a.38.1.2120

Jihad M. Abdallah[1]*

Abstract: Determination of the required sample size is an important step in the planning of any research study. A large number of commercial and online resources are readily available for calculation of sample size required for various research designs. However, this abundance of information, the complexity, and the variation in calculation formulations and terminology used make it more confusing to researchers, particularly those with limited statistical knowledge. Therefore, there is a need for more simplified, easy to implement formulas for calculation of sample size. Here, we present a short review of the rules for calculation of sample size for common experimental designs used in research and provide more simplified forms for some of these formulas. Also, data are presented on sample size required for various scenarios with the intent to provide guidelines for researchers. A simple-to-use Excel sheet was developed to perform sample size calculations and is available online as a supplementary file. This Excel sheet was used to generate the data presented in this publication.



Keywords: Effect Size, Multiple Testing, Required Sample Size, Statistical Power, Testing Means, Testing Proportions

Introduction

Sample size calculation is an important early step in any research not only to attain sufficient statistical precision but also for better utilization of available resources. If the sample used is too small, the study will not provide reliable answers to study questions (1). It will lack the statistical power to detect significant differences and effects, the data will not approximate well the underlying statistical distribution (normal or other) and will lack sufficient representation for the results to accurately describe the population (2, 3). Increasing the sample size improves the validity and reliability of results and increases the statistical power to detect significant differences when they truly exist. However, if the sample utilized is too large, it is a poor use of resources and extends the time and effort required to finish the study. Furthermore, a sample larger than the required size may put more individuals at risk in certain interventions (4). It is therefore crucial for researchers to determine the required sample size before conducting research studies to ensure that they have enough sample size to draw meaningful conclusions without wasting available resources.

A large number of commercial and online resources are readily available for calculation of sample size required for various research designs. However, the abundance of information and the variation in the calculation formulations and the terminology used make it more confusing to researchers, particularly those with little statistical knowledge. Furthermore, a review by Meysamie et al., 2014 (5) showed that most of the online sample size calculators are limited to sample size calculation for estimating proportions and considered a fixed value of 0.50, and in certain cases, inaccurate calculations were obtained.

Simplification of formulas for sample size calculation allows the researcher to make a quick determination of sample size while avoiding the overwhelming statistical notations and the mathematical derivations (6). The main objectives of this work are to present a review of the formulas for calculation of sample size for common research designs, to provide these rules in the simplest possible forms, and to explore sample size for various scenarios (i.e., different values of power and effect size, etc.). In addition, an easy-to-use Excel sheet was developed by the author to implement these rules and is available as a supplementary file for interested users.

Factors affecting sample size and related statistical terminology

Many factors affect the calculation of the required sample size including the study design, one-sided or two sided hypothesis testing, the sampling method, the type of population being sampled (homogenous or heterogeneous), the dropout rate (or mortality rate), the nature of the outcome being measured (binary or continuous), the effect size, the statistical power, the significance level, and the variability in the population. The number of

[1] Department of Animal Production & Animal Health, Faculty of Agriculture and Veterinary Medicine, An-Najah National University, Nablus PO Box 7, Palestine
*Corresponding author: jmabdallah@najah.edu

predictors in the regression model, R-square, and effect size are important factors to consider when determining sample size for regression analysis. For a good overview of factors influencing sample size determination, the reader is referred to other reviews in the literature (4, 7-11). Table (1) provides a summary of the main statistical terms related to sample size calculation and their effect on the required sample size.

Table (1): Summary of the statistical terminology and the main factors affecting the required sample size.

Term or factor	Effect on sample size
Significance level, α: the probability of Type 1 error, or the false positive rate. It is the probability to falsely reject the null hypothesis and is set by the researcher. The typical value used is 0.05.	The sample size increases as α decreases and *vice versa*
Statistical power: is the probability to reject a false null hypothesis of no effect or no difference, i.e., the ability of the statistical test to detect a true significant effect. Power = 1-β, where β is the probability of Type 2 error (probability of not rejecting a false null hypothesis). A typical value of 80% is generally used by researchers for statistical power.	The sample size increases as statistical power increases and *vice versa*
Effect size: the magnitude of the effect or difference to be tested; for example, the difference between the treatment and control groups.	The sample size deceases as the effect size increases and *vice versa.*
One-sided vs. two-sided hypothesis testing: a one-sided hypothesis tests if the parameter is larger or smaller than a hypothesized value while a two-sided hypothesis tests if the parameter is different from a specified value.	The required sample size is smaller for one-sided tests compared to two-sided tests (because $Z_{\alpha} < Z_{\alpha/2}$).
Population standard deviation, σ: quantifies the variability among units in the population.	As σ increases, the required sample size increases and *vice versa*
Population proportion, P: the portion of the population having the investigated characteristic (e.g., prevalence of the disease, proportion of smokers, etc.)	The sample size is maximum at P = 0.50 and decreases P gets closer to 0 or 1.
Margin of error: refers to the level of precision required**.** It is half the width of the desired confidence interval. Also called the maximum error of the estimate and is defined as the maximum likely difference between the point estimate of the parameter and the true value of the parameter.	The sample size increases as the desired margin of error decreases and *vice versa*
Nature of the sampled population: refers to how similar are the sampling units in the population (homogenous or heterogeneous).	Homogenous populations require less sample size compared to less homogeneous populations because of lower variability when the population is homogenous.
Dropout rate (or mortality rate): the percentage of the subjects or units in the sample who drop out or die during the course of the study.	The required sample size increases as the expected drop out rate increases and *vice versa*.
R-square: the proportion of the total variation in the dependent variable that is explained by the set of predictors in the regression model.	The required sample size increases when a higher R-square is deemed acceptable and *vice versa*.
Number of predictors (in the regression model)	The required sample size increases as the number of predictors increase and *vice versa*.
Type of study	The sample size required for descriptive studies (such as those based on surveys and questionnaires) is larger than that required for analytical studies. Observational studies need larger samples than experimental studies
Qualitative vs. quantitative research	Quantitative research is generally based on larger samples than qualitative research
Binary vs. continuous outcomes: binary outcomes involve outcomes with two categories (for example, yes/no or presence/absence responses)	Binary outcomes require larger sample size than continuous outcomes

Sample size calculation

Some of the early approaches to deal with sample size determination in experiments include Cochran and Cox (1957), Harris et al. (1948), Harter (1957), Tang (1938), and Tukey (1953) (12-16). Most approaches are based on detecting differences of a specified size or obtaining confidence intervals not larger than a stated width. The first approach will be illustrated herein for determining sample size to test means, and the second approach is illustrated in determining sample size to test proportions. Other available approaches will be also discussed.

Sample size calculation for testing means

Based on the first approach, a general formula to determine the minimum sample size required for testing means with a stated effect size is given by Steel et al. ,1997 (3) as follows (with slightly modified notation and arrangement):

$$n = \frac{\left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right)^2}{\frac{\Delta^2}{\sigma_D^2}} \quad (1)$$

where, n is the sample size per group, α is the desired significance level (probability of Type 1 error, that is the probability to falsely reject H0), β is probability of Type 2 error (probability of not rejecting a false H0, with the power of the test defined as 1- β), Δ is the true difference or effect size to be tested (e.g., $\mu_1 - \mu_0$, $\mu_1 - \mu_2$, for single group and two groups, respectively) and σ_D^2 depends on the research design, and hence, the statistical test used. For pre-test/post-test design (before-after design), σ_D^2 is the variance of the differences. For other designs, it is defined in terms of the error variance, σ^2 ($\sigma_D^2 = \sigma^2$ for single-group design and $\sigma_D^2 = 2\sigma^2$ for designs involving two or more groups including independent-groups designs and the randomized complete block design). The values $Z_{\alpha/2}$ and Z_β are critical values obtained from the standard normal distribution such that $P\left(Z \geq Z_{\frac{\alpha}{2}}\right) = \alpha/2$ and $P(Z \geq Z_\beta) = \beta$. The typical values used by most researchers are 0.05 for α and 0.20 for β (i.e., power = 80%). If one-tailed test is desired instead of a two-tailed test, then $Z_{\frac{\alpha}{2}}$ is replaced by Z_α (in this case the required sample size will be smaller).

The main problem with this approach is that σ^2 is usually not known and an estimate is needed. If σ^2 is underestimated, n is too small and if σ^2 is overestimated, n is too large (3). The problem is cleverly solved by defining Δ in terms of σ, i.e., using a standardized effect size. Therefore, if we define the standardized effect size $\delta = \frac{\Delta}{\sigma}$ (or $\delta = \frac{\Delta}{\sigma_D}$ for the pre-test/post-test design), then Equation (1) becomes:

$$n = \frac{k\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2}{\delta^2} \qquad (2)$$

where k =1 for single-group and pre-test/post-test designs, and k =2 for the independent groups (e.g., the completely randomized design) and the randomized complete block designs. If we apply the typical values of 0.05 for α ($Z_\alpha = 1.64$, $Z_{\alpha/2} = 1.96$) and 0.20 for β ($Z_\beta = 0.84$), then Equation (2) is further simplified to:

$$n = \frac{7.85\,k}{\delta^2} \quad \text{, for two-tailed tests} \qquad (3) \quad \text{and}$$

$$n = \frac{6.15\,k}{\delta^2} \quad \text{, for one- tailed tests} \qquad (4)$$

Therefore, for two-tailed tests, $n = \frac{7.85}{\delta^2}$ for single samples and before-after designs, and $n = \frac{15.70}{\delta^2}$ for independent-groups and randomized complete block designs. Allen, 2011 (6) suggested using $n = \frac{16}{\delta^2}$ as a rule of thumb for calculating sample size for two-independent groups (two-tailed t-test).

Because the critical values are obtained from the standard normal distribution, a correction must be made to the sample size to account for t-distribution as follows (3):

$$n^* = n\frac{(df+3)}{(df+1)} \qquad (5)$$

where *df* is the error degrees of freedom for the specified design. Both *n* and *n** are rounded to the largest integer value.

Sample size calculation for testing proportions

The sample size for testing proportions is usually determined based on obtaining confidence intervals not larger than a stated width. For testing population proportion (e.g., disease prevalence) using a single sample, the following formula is used (17):

$$n = \frac{\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2 [P(1-P)]}{d^2} \qquad (6)$$

where P is the assumed population proportion and *d* is the margin of error which is equal to half-width of the confidence interval with a desired $(1-\alpha)100\%$ confidence cofficient. The problem here is that an estimate of *P* is required. If no previous information is available on *P*, the researcher can use *P* = 0.50 which results in the maximum sample size ($[P(1-P)]$ is maximum when *P* = 0.50).

Equation (6) assumes that sampling is from an infinite population. The sample size is corrected for finite population size as follows (17):

$$n^* = \frac{nN}{n+(N-1)} \qquad (7)$$

where *N* is the population size. Note that *n** is smaller than *n*, that is, correction results in smaller sample size as sampling from a finite population is more efficient than sampling from an infinite population. Correction can be ignored when the sampling fraction (*n/N*) is small.

For testing the difference between two proportions (two-independent samples), the following formula is used (18-20):

$$n = \frac{\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2 [P_1(1-P_1)+P_2(1-P_2)]}{(P_1-P_2)^2} \qquad (8)$$

where n is the sample size per group, P_1 and P_2 are the assumed proportions in group 1 and group 2, respectively. A correction for finite population sizes is performed as follows:

$$n^* = \frac{\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2 [f_1P_1(1-P_1)+f_2P_2(1-P_2)]}{(P_1-P_2)^2} \qquad (9)$$

where $f_1 = \frac{N_1-n}{(N_1-1)}$, and $f_2 = \frac{N_2-n}{(N_2-1)}$ with N_1 and N_2 the sizes of the populations being sampled. As was pointed out for single proportion, the correction for finite population sizes can be ignored when the ratio of the sample size to the population size is small (e.g., less than 0.02).

Sample size considerations for multiple testing

In many experiments, the researcher is interested in making pairwise comparisons among several treatment groups. Multiple tests end up being performed in a single experiment. This raises the issue of Type 1 error rate in multiple testing. In the previous sections, the significance level, α, was defined as the probability of Type 1 error for a single comparison (single test). If *m* independent comparisons are performed, then the probability that at least one null hypothesis is falsely rejected is $1-(1-\alpha)^m$ which is larger than α. This is usually called the family-wise error rate, FWER (3, 21). For example, if the researcher wishes to perform five independent pairwise comparisons, then FWER = 0.226 for $\alpha = 0.05$. Several approaches have been proposed to correct the significance level for multiple testing. The most common are the Sidak correction (also called the Sidak-Dunn correction) and the Bonferroni correction. Based on Sidak correction (22, 23), if a family-wise error rate = α` is desired (typically $\alpha$`= 0.05), then the required significance level for each individual test is calculated as $\alpha = 1-(1-\alpha`)^{(1/m)}$, while the Bonferroni correction (22, 24) uses $\alpha = \alpha`/m$, and both give a value of 0.01 for m = 5 and $\alpha` = 0.05$. In this particular example, the researcher will use a significance level of 0.01 instead of 0.05 in calculation of the sample size required to control the family-wise error rate.

The Sidak and the Bonferroni corrections are considered very conservative when the number of tests is large and when the tests are not independent (22). Therefore, alternative approaches have been proposed to make the correction less stringent (e.g., 25-27). The procedure by Benjamini and Hochberg, 1995 (25) is based on reducing the false discovery rate (FDR) instead of the family-wise error rate (22). If the pairwise comparisons are not independent as in most experiments, one can use the correction suggested by John W. Tukey (28): $\alpha = 1 - (1 - \alpha`)^{(1/\sqrt{m})}$. This results in a larger α (0.02 compared to 0.01 for the example above), and hence smaller sample size is required than when independence is assumed (21).

Sample size for other research designs

The focus in this review was on some widely used experimental designs particularly in agricultural, environmental, social and life sciences. Other designs like case-control and cohort studies are very common in medical studies. Interested readers may consult other reviews (e.g., 29, 30) for further details on sample size determination for such designs. Furthermore, several online tools are available to determine sample size requirements for such designs (e.g., Epitools-Epidemiological Calculators site available at: http://epitools.ausvet.com.au/samplesize). Allen, 2011 (6) extended the formula used for two-independent samples to accommodate the repeated measures design. Crossover designs (within-subject design where the same subject receives different treatments during different periods in random order) are also popular in medical trials. Siyasinghe and Sooriyarachchi, 2011 (31) provided guidelines for calculating sample size in 2x2 crossover trials while Moxely, 2021 (32) provided a review of sample size and efficacy of these designs. Dharmarajan et al., 2019 (33) proposed two new sample size estimation methods that provide a more accurate estimate of the true required sample size for the case-crossover studies than the traditionally used Dupont formula which was originally developed for matched case-control studies (34).

Regression analysis is very important in many research fields, particularly, agricultural, environmental and social sciences. Several methods have been proposed to deal with sample size determination for both linear and logistic regression analyses (35-39).

The equations presented herein for testing two proportions assume independent samples. Conor, 1987 (40) presented sample size calculations for testing differences in proportions for the paired-sample design. Divine et al., 2013 (41) reviewed sample size calculations for different Wilcoxon tests (nonparametric tests). In addition, the present review assumed equal group sizes, the readers are referred to the review by Whitley and Ball, 2002 (11) where unequal group sizes were also described.

Other statistical methods for calculation of sample size

In the previous sections, the focus was on formula-based techniques for calculating the required sample size (closed-form solutions) These widely used methods for sample size determination are based on the frequentist approach where prior point estimates need to be specified. One problem is that we are generally uncertain about these prior estimates and this uncertainty is not accounted for by the frequentist methods. In contrast, Bayesian methods (e.g., 42-45) can deal with the uncertainty associated with prior information by replacing the prior point estimate by a prior distribution which is then updated to a posterior distribution using Bayes rule. Brus et al., 2022 (46) provided an excellent overview of both approaches with application to determine the number of sampling locations required for soil survey. Another approach for estimating sample size is by simulation. The basis for simulation methods is the general approach for estimating power presented by Feiveson, 2002 (47). Simulation-based methods are iterative techniques which start by generating a data set with an initial size from a given distribution and then calculating the statistical power (or any other criteria like minimum error, R-square, etc). The sample size is then iteratively modified, repeating power calculation, until a sample size with a certain desired value of power is reached (48, 49). Simulation-based determination of sample size can be implemented with codes using existing standard statistical software (50). However, when complex models are involved, specialized complex algorithms are required. The major disadvantage of Bayesian methods and simulation-based methods is that they are computationally intensive which has restricted their application.

Software and web applications for determination of sample size

Many software programs and web applications are freely available for calculation of sample size. Researchers should be careful to which resources to use as some online calculators may give erroneous calculations as outlined by Meysamie et al., 2014 (5). Researchers also need to choose the most relevant resource for the type and design of the study they intend to carryout. Some widely used software and online calculators (commercial and free) are listed in Table (2) with information on calculation methods and where to access them.

Table (2): Examples on available software and web applications for calculation of sample size

Software or web application	Determination Method	Where to access
Epitools-Epidemiological calculators (Free web application)	Formula-based (51)	http://epitools.ausvet.com.au
Select Statistical Services Calculators (Free web-application)	Formula-based	https://select-statistics.co.uk/calculators/
G*Power (Free software)	Simulation (52, 53)	https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower
Scalex and ScalaR calculators (Free software)	Simulation (54)	https://sites.google.com/view/sr-ln/ssc
Psychometroscar (Free software)	Simulation (39)	https://psychometroscar.com/2018/07/31/power-analysis-for-multilevel-logistic-regression/
nQuery (Commercial software)	Multiple (Formula-based, Bayesian, and adaptive design)	www.statsols.com/nquery
Power and Precision (Commercial software)	Power analysis software	https://www.power-analysis.com/

Illustration data

Testing means

Figure (1) shows the required sample size for pre-test/post-test designs and for various scenarios of power and standardized effect size for one-tailed and two-tailed tests. A significance level of 0.05 was used in the calculations. The graph illustrates that the required sample size is larger for two-tailed tests compared to one-tailed tests and increases with increased power and for smaller effect size. However, power becomes less important as the effect size exceeds 1.5 standard deviations.

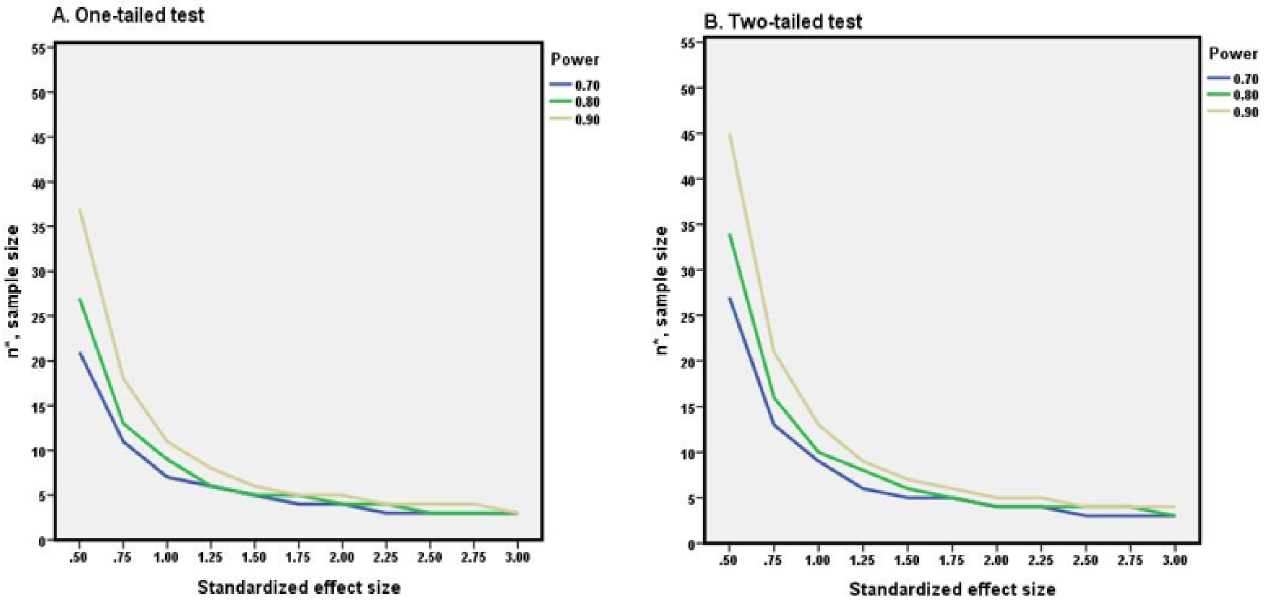


Figure (1): Sample size calculations for pre-test/post-test designs. Calculations were based on a significance level of 0.05**.**

Table (3) shows sample size calculations for independent-groups designs (e.g., the completely randomized design) and the randomized complete block design with different number of treatments or groups and size effects. Calculations were based on a two-tailed test, a significance level of 0.05 and power of 80% ($\beta = 0.20$). The number of treatments has almost no effect on sample size requirement. Furthermore, both types of designs require closely similar sample sizes for the same number of treatments and effect size. However, because randomized block designs are more efficient than independent-groups designs (less error variance is expected due to blocking), researchers can assume larger effect size when determining sample size for randomized block designs. For a two-tailed test and typical power of 0.80, the group size required to detect a size effect of 0.5 to 3 standard deviations varies from 3 to 65 replicates per treatment.

Table (3): Sample size calculations for the independent-groups design and the randomized complete block design for various scenarios of effect size and number of treatments.

Effect size	**Number of treatments or groups**											
	2	**3**	**4**	**5**	**6**	**7**	**2**	**3**	**4**	**5**	**6**	**7**
	n ^{*,a}											
	Independent-Groups Design						**Randomized Complete Block Design**					
0.50	64	64	64	64	64	64	65	64	64	64	64	64
0.75	29	29	29	29	29	29	30	29	29	29	29	29
1.00	17	17	17	17	17	16	18	17	17	17	17	17
1.25	12	11	11	11	11	11	12	12	11	11	11	11
1.50	9	8	8	8	8	8	9	9	8	8	8	8
1.75	7	6	6	6	6	6	7	7	6	6	6	6
2.00	6	5	5	5	5	5	6	6	5	5	5	5
2.25	4	4	4	4	4	4	5	4	4	4	4	4
2.50	4	4	4	3	3	3	5	4	4	4	3	3
2.75	3	3	3	3	3	3	4	3	3	3	3	3
3.00	3	3	3	3	3	3	4	3	3	3	3	3

[a] n^* = minimum number of replicates per treatment assuming α = 0.05, β = 0.20 (power = 80%), and two-tailed tests.

Testing proportions

As shown in Figure (2), the required sample size is largely affected by the assumed value of P. It is maximum at *P* = 0.50 and decreases symmetrically as *P* gets closer to 0 or 1. Therefore, if no prior information is available on P, then the researcher can safely assume P = 0.50. Power has large impact on the required sample size when P is intermediate but its effect diminishes as P approaches 0 or 1.

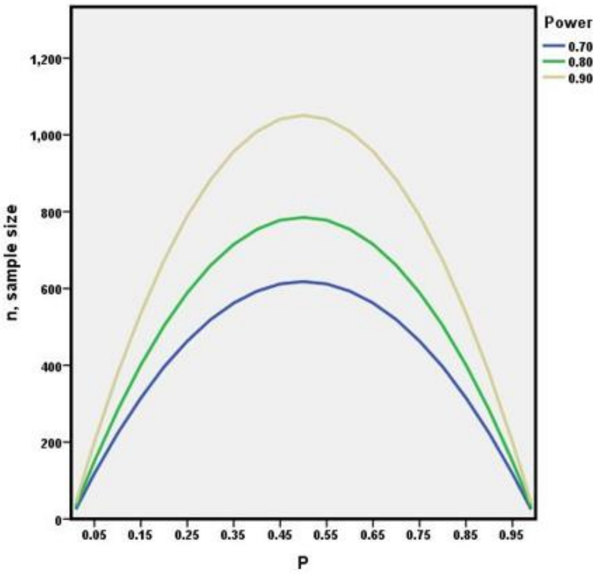


Figure (2): Sample size required for testing single population proportion for various values of power and P. Calculations were made using $\alpha = 0.05$ (95% CI), *d* = 0.05, and assuming two-tailed test and infinite population size.

Figure (3) shows corrected sample size for *n* = 100 and various values of *N*. Correction for finite population size can be neglected only when the sample size is very small compared to the population size, i.e., when the sampling fraction, *n/N*, is small (less than 0.02 in the simulated graph).

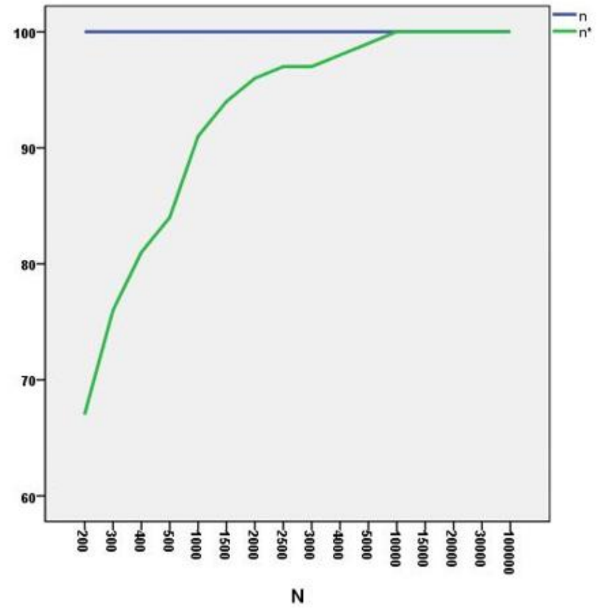


Figure (3): Corrected sample size (n*) in relation to population size (N). Calculations were made for *n* = 100 with *n**= nN/[n+(N-1)]*.

The sample size required per group depends largely on $P_1 - P_2$ (the required sample size decreases as the difference increases and vice versa), while power has little impact when $P_1 - P_2$ exceeds 0.25 (Figure 4). These sample sizes presented in Figure (4) were calculated assuming infinite population sizes.

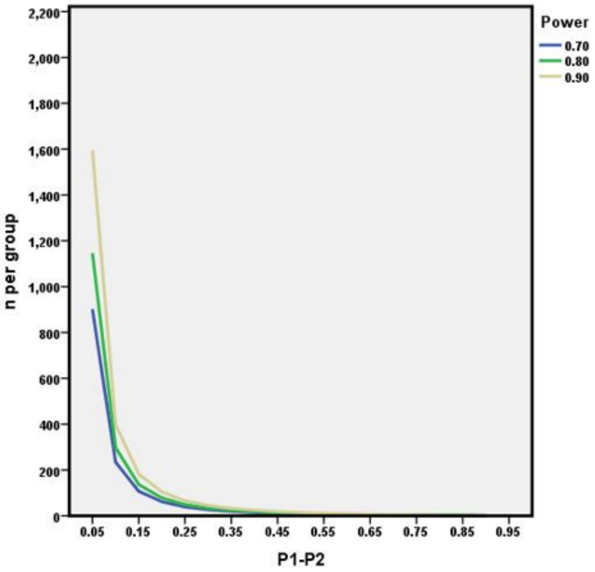


Figure (4): Sample size required per group for testing two proportions for various values of power and $P_1 - P_2$. Calculations were made using $\alpha = 0.05$ (95% CI), and assuming two-tailed test and infinite population sizes.

Multiple testing

Table (4) shows the values of the family-wise error rate when no adjustment is made for multiple testing and the corrected significance level, α, that needs to be used for pairwise comparisons when an overall error rate of 5% (α`= 0.05) is desired. The data are shown for up to twenty multiple comparisons based on Sidak and Bonferroni corrections for independent tests and Tukey correction for dependent tests. The data demonstrates that the required significance level gets smaller as the number of comparisons increases but the adjustment is less stringent (the required significance level is larger) if we assume dependent tests compared to independent tests. Note also that the corrected significance level values are very similar for both Sidak and Bonferroni corrections for independent tests.

Table (4): Required significance level in calculation of sample size to control for the family-wise error rate in multiple testing.

Number of comparisons	**Family-wise Error Rate** [a]	**Corrected significance level (α)** [b]		
		Sidak correction	**Bonferroni correction**	**Tukey correction**
1	0.0500	0.0500	0.0500	0.0500
2	0.0975	0.0253	0.0250	0.0356
3	0.1426	0.0170	0.0167	0.0292
4	0.1855	0.0127	0.0125	0.0253
5	0.2262	0.0102	0.0100	0.0227
6	0.2649	0.0085	0.0083	0.0207
7	0.3017	0.0073	0.0071	0.0192
8	0.3366	0.0064	0.0063	0.0180
9	0.3698	0.0057	0.0056	0.0170
10	0.4013	0.0051	0.0050	0.0161
11	0.4312	0.0047	0.0045	0.0153
12	0.4596	0.0043	0.0042	0.0147
13	0.4867	0.0039	0.0038	0.0141

Number of comparisons	Family-wise Error Rate [a]	Corrected significance level (α) [b]		
		Sidak correction	Bonferroni correction	Tukey correction
14	0.5123	0.0037	0.0036	0.0136
15	0.5367	0.0034	0.0033	0.0132
16	0.5599	0.0032	0.0031	0.0127
17	0.5819	0.0030	0.0029	0.0124
18	0.6028	0.0028	0.0028	0.0120
19	0.6226	0.0027	0.0026	0.0117
20	0.6415	0.0026	0.0025	0.0114

[a] Family-wise error rate if no correction is made for multiple tests $= 1 - (1-\alpha)^m$ for m independent pairwise test.

[b] Calculations of α were made to control the familywise error rate at 0.05 level ($\alpha` = 0.05$).

Reporting sample size determination by researchers

It is very important in this review to highlight the need for transparent reporting of sample size calculations in research articles. Some studies (55) surveyed published research and concluded that sample size calculation is still inadequately reported and often erroneous or based on assumptions that are frequently inaccurate. Guidelines and recommendations have been developed for reporting the outcomes of scientific studies including sample size calculations, such as the Consolidated Standards of Reporting Trials (CONSORT) 2010 statement developed for randomized controlled trials (56, 57) and its 2022 extension (58) and the SPIRIT 2013 statement for clinical trial protocols (59, 60). Researchers need to adhere to these guidelines and ensure accurate and complete reporting of sample size determination and resources used in calculations, which will contribute to improving the quality of scientific publications.

Emerging trends and future directions

Sample size calculation will continue to be an important issue due to its major impact on the results of research. Emerging trends in sample size calculation include adaptive trial designs (which involve re-estimation of sample size based on accumulating interim data to achieve the desired power) and implementation of the “promising zone’ design in clinical trials. The promising zone design was first described by Mehta and Pocock, 2011(61) based on earlier work by Chen et al., 2004 (62). For further details on adaptive designs and the promising zone design, the reader is referred to the reviews by Pallmann et al., 2018 (63) and Edwards et al., 2020 (64), respectively. The extension of these designs to other types of studies should be investigated in the future. Developing easy-to-use software programs for implementing Bayesian and simulation-based methods while reducing their computational burden will facilitate their practical application by researchers. A comprehensive study of available online sample size calculators (their features, advantages and limitations) to filter out unreliable resources would help researchers to avoid inaccurate calculations.

Conclusion

This work presented an overview of the methods for the calculation of the required sample size for common experimental designs and presented the calculation equations in the most possible simple forms. In addition, the calculations were illustrated with simulated data. This review, along with the developed Excel calculation sheet, will make it easier for researchers to understand, choose, apply, and correctly report the appropriate sample size calculations for their studies.

Declarations

Ethics approval and consent to participate Not applicable

Consent for publication Not applicable

Author’s contribution Not applicable

Availability of data and materials All data generated during this study are included in this published article. The Excel sheet used to generate the data is available online as a supplementary file.

Funding No funding has been received for this work

Conflict of interests The author declares that that there is no conflict of interests regarding the publication of this article

Acknowledgments: The author would like to acknowledge the logistic support provided by An-najah National University (www.najah.edu).

References

1) Fitzner C, Heckinger E. Sample size calculation and power analysis: A quick review. The Diabetes Educator. 2010; 36(5):701-707. https://doi.org/10.1177/0145721710380791

2) Di Lorio CK. Review of statistical concepts. In: Measurement in health behavior: Methods for research and evaluation. San Francisco: Jossey-Bass; 2005:153-154.

3) Steel RGD, Torrie JH, Dickey DA. Principles and Procedures of Statistics: a biometrical approach. 3rd edition, NY: McGraw Hill Inc.; 1997.

4) Gumpili SP, Das AV. Sample size and its evolution in research. IHOPE J Ophthalmol. 2022;1(1):9-13. https://doi.org/10.25259/IHOPEJO_3_2021

5) Meysamie A, Taee F, Mohammadi-Vajari M-A, Yoosefi-Khanghah S, Emamzadeh-Fard S, Abbassi M. Sample size calculation on web, can we rely on the results? J Med Stat Inform. 2014;2(3):1-8. http://dx.doi.org/10.7243/2053-7662-2-3

6) Allen JC. Sample size calculation for two independent groups: A useful rule of thumb. Proceedings of Singapore Healthcare. 2011;20(2):138-140.

7) Al-Subaihi A. Sample size determination: Influencing factors and calculation strategies for survey research. Saudi Med J. 2003;24(4):323-330.

8) Althubaiti A. Sample size determination: A practical guide for health researchers. J Gen Fam Med. 2023;24:72–78. https://doi.org/10.1002/jgf2.600

9) Chander NG. Sample size estimation. J Indian Prosthodont Soc. 2017; 17(3):217-218. https://doi.org/10.4103/jips.jips_169_17

10) Das S, Mitra K, Mandal M. Sample size calculation: Basic principles. Indian J Anaesth. 2016;60:652-656. DOI: 10.4103/0019-5049.190621

11) Whitley E, Ball J. Statistics review 4: Sample size calculations. Critical Care. 2002;6:335-341.

12) Cochran WG, Cox GM. Experimental designs. 2nd ed. New York: Wiley; 1957.

13) Harris M, Harvits DG, Mood AM. On the determination of sample sizes in designing experiments. J Amer Statist Ass. 1948;43:391-402.

14) Harter HL. Error rates and sample sizes for range tests in multiple comparisons. Biometrics. 1957;13:511-536.

15) Tang PC. The power function of the analysis of variance tests with tables and illustrations of their use. Statist Res Mem. 1938;2

16) Tukey JW. The problem of multiple comparisons. Mimeograph, Princeton University, NJ; 1953.

17) Daniel WW. Biostatistics: A foundation for analysis in the health sciences. 7th edition. New York: John Wiley & Sons; 1999.

18) Chow SC, Shao J, Wang H. Sample Size Calculations in Clinical Research, Second Edition. Boca Raton: Chapman & Hall/CRC; 2008.

19) Miot HA. Sample size in clinical and experimental trials. J Vasc Bras. 2011;10(4):275-278.

20) Wang H, Chow SC. Sample Size Calculation for Comparing Proportions. In: Wiley Encyclopedia of Clinical Trials, 10. John Wiley & Sons, Inc.; 2007. Available from: https://doi.org/10.1002/9780471462422.eoct005.

21) McConnell B, Vera-Hernandez M. Going Beyond Simple Sample Size Calculations: a Practitioner's Guide. IFS Working Paper W15/17.

22) Abdi H. The Bonferroni and Sidak corrections for multiple comparisons. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Thousand Oaks: Sage; 2007.

23) Šidák ZK. Rectangular confidence regions for the means of multivariate normal distributions. J Amer Statist Ass. 1967;62(318):626–633. https://doi.org/10.1080/01621459.1967.10482935

24) Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilita`. Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936;8:3–62.

25) Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6:65–70.

26) Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75:800–803.

27) Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Statist Soc B. 1995;57(1):289–300.

28) Braun HIE. The collected works of John W Tukey Vol. VIII. Multiple comparisons: 1948-1983. New York: Chapman & Hall; 1994.

29) Charan J, Biswas T. How to calculate sample size for different study designs in medical research? Indian J Psychol Med. 2013;35(2):121-126. https://doi.org/10.4103/0253-7176.116232

30) Sharma SK, Mudgal SK, Thakur K, Gaur R. How to calculate sample size for observational and experimental nursing research studies? Natl J Physiol Pharm Pharmacol. 2019;10(1):1-8. https://doi.org/10.5455/njppp.2020.10.0930717102019

31) Siyasinghe NM, Sooriyarachchi MR. Guidelines for calculating sample size in 2x2 crossover trials: a simulation study. J Natn Sci Foundation Sri Lanka. 2011;39(1):77-89.

32) Moxley KC. A review of sample size and design efficacy in crossover design in peer-reviewed psychology research. Wayne State University Dissertations. 2021;3548. https://digitalcommons.wayne.edu/oa_dissertations/3548

33) Dharmarajan S, Lee J-Y, Izem R. Sample size estimation for case-crossover studies. Stat Med. 2019;38(2):956-968. DOI: https://doi.org/10.1002/sim.8030

34) Dupont WD. Power calculations for matched case-control studies. Biometrics. 1988;44(4):1157-1168.

35) Dupont WD, Plummer WD. Power and sample size calculations for studies involving linear regression. Control Clin Trials. 1998;19(6):589–601. https://doi.org/10.1016/S0197-2456(98)00037-3

36) Green SB. How many subjects does it take to do a regression analysis. Multivariate Behavioral Research. 1991;26(3):499-510. https://doi.org/10.1207/s15327906mbr2603_7

37) Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med. 1998;17:1623–1634. https://doi.org/10.1002/(SICI)1097-0258(19980730)17:14

38) Maxwell SE. Sample size and multiple regression analysis. Psychol Methods. 2000;5(4):434–458. https://doi.org/10.1037/ 1082-989X.5.4.434

39) Olvera Astivia OL, Gadermann A, Guhn M. The relationship between statistical power and predictor distribution in multilevel logistic regression: a simulation-based approach. BMC Med Res Methodol. 2019;19:97, 1-20. https://doi.org/10.1186/s12874-019-0742-8

40) Connor RJ. Sample size for testing differences in proportions for the paired-sample design. Biometrics. 1987;43(1):207–211. https://doi.org/10.2307/2531961

41) Divine GH, Norton J, Hunt R, Dienemann J. A review of analysis and sample size calculation considerations for Wilcoxon tests. Anesthesia & Analgesia. 2013;117(3):699-710. https://doi.org/10.1213/ANE.0b013e31827f53d7

42) Adcock CJ. The Bayesian approach to determination of sample sizes: Some comments on the paper by Joseph, Wolfson and du Berger. The Statistician. 1995;44:155–161.

43) Adcock CJ. Sample size determination: a review. J R Statist Soc D (The Statistician). 1997;46(2):261-283.

44) Joseph L, Belisle P. Bayesian sample size determination for normal means and differences between normal means. The Statistician. 1997;46:209–226.

45) Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Statistical Science. 2002;17(2):193–208. Available from: http://www.jstor.org/stable/3182824

46) Brus DJ, Kempen B, Rossiter D, Balwinder-Singh, McDonald AJ. Bayesian approach for sample size determination, illustrated with Soil Health Card data of Andhra Pradesh (India). Geoderma. 2022;405:115396:1-10. https://doi.org/10.1016/j.geoderma.2021.115396

47) Feiveson AH. Power by simulation. Stata Journal. 2002; 2:107–124.

48) Sutton AJ, Donegan S, Takwoingi Y, et al. An encouraging assessment of methods to inform priorities for updating systematic reviews. J Clin Epidemiol. 2009;62:241–251. https://doi.org/10.1016/j.jclinepi.2008.04.005

49) Growther MJ, Hinchliffe SR, Donald A, Sutton AJ. Simulation-based sample-size calculation for designing new clinical trials and diagnostic test accuracy studies to update an existing meta-analysis. The Stata Journal. 2013;13 (3):451–473.

50) Zhao W, Li AX. Generalized approach to estimating sample sizes. SAS Global Forum 2012 (Statistical Data Analysis), Paper 336:1-7.

51) Sergeant ESG. Epitools Epidemiological Calculators. Ausvet. Available at: http://epitools.ausvet.com.au

52) Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods. 2007;39(2):175-191. https://doi.org/10.3758/bf03193146

53) Faul F, Erdfelder E, Buchner A, Lang A. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods. 2009;41(4):1149-1160. https://doi.org/10.3758/brm.41.4.1149

54) Naing L, Bin Nordin R, Abdul Rahman H, Naing YT. Sample size calculation for prevalence studies using Scalex and ScalaR calculators. BMC Medical Research Methodology. 2022;22:209. https://doi.org/10.1186/s12874-022-01694-7

55) Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ. 2009;338:1-6. https://doi.org/10.1136/bmj.b1732

56) Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869. https://doi.org/doi:10.1136/bmj.c869

57) Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. https://doi.org/10.1136/bmj.c332

58) Butcher NJ, Monsour A, Mew EJ, et al. Guidelines for Reporting Outcomes in Trial Reports: The CONSORT-Outcomes 2022 Extension. JAMA. 2022;328(22):2252–2264. https://doi.org/10.1001/jama.2022.21022

59) Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 Statement: Defining standard protocol items for clinical trials. Ann Intern Med. 2013;158:200-207.

60) Chan AW, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 Explanation and Elaboration: Guidance for protocols of clinical trials. BMJ. 2013;346:e7586.

61) Mehta C, Pocock S. Adaptive increase in sample size when interim results are promising: a practical guide with examples. Stat Med. 2011;30:3267–84.

62) Chen J, DeMets DL, Lan G. Increasing the sample size when the unblinded interim result is promising. Stat Med. 2004;23(7):1023-1038. https://doi.org/10.1002/sim.1688

63) Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16(1):29. https://doi.org/10.1186/s12916-018-1017-7

64) Edwards JM, Walters SJ, Kunz C, Steven A. A systematic review of the “promising zone” design. Trials. 2020;21:1000. https://doi.org/10.1186/s13063-020-04931-w