Review Article

Simplified Sample Size Formulas for Detecting a Medically Important Effect

Abhaya Indrayan, Aman Mishra, Binukumar Bhaskarapillai[1]

Department of Clinical Research, Max Healthcare, Saket, Delhi, [1]Department of Biostatistics, NIMHANS, Bengaluru, Karnataka, India

Abstract

The sample size is just about the most common question in the minds of many medical researchers. This size determines the reliability of the results and helps to detect a medically important effect when present. Some studies miss an important effect due to inappropriate sample size. Many postgraduate students and established researchers often contact a statistician to help them determine an appropriate sample size for their study. More than 80 formulas are available to calculate sample size for different settings and the choice requires some expertise. Their use is even more difficult because most exact formulas are quite complex. An added difficulty is that different books, software, and websites use different formulas for the same problem. Such discrepancy in the published formulas confounds a biostatistician also. The objective of this communication is to present uniformly looking formulas for many situations together at one place in their simple but correct form, along with the setting where they are applicable. This will help in choosing an appropriate formula for the kind of research one is proposing to do and use it with confidence. This communication is restricted to the sample size required to detect a medically important effect when present – known to the statisticians as the test of hypothesis situation. Such a collection is not available anywhere, not even in any book. The sample size formulas for estimation are different and not discussed here.

Keywords: Detecting an effect, medical significance, sample size, simplified formulas, test of hypothesis

Introduction

'How many cases should I study for my research?' is the question faced by almost all medical researchers. Sample size has two distinct roles depending upon the setup. First, serving as a primary determinant of the reliability of the results. Not many realize that reliability is different from validity.[1] Biased samples, however large, will not produce valid results – in fact, a large sample can aggravate bias and give a false sense of security. The larger the sample size, the more is the reliability, when appropriately chosen, although it has diminishing returns. High reliability is required for the replicability of the results – repeated studies in similar populations with the same methodology should give nearly the same result. A reliable result can be biased, giving nearly the same bias every time. Second, and more importantly, is the seminal role of sample size in detecting an effect of a factor when that effect is present. Sample size should be sufficient so that a medically important effect is not missed if present. It is not considered missed when it is statistically significant.

Statistically, the first setup is called estimation, and the second testing of the hypothesis. The sample size formulas are different for these two setups. Estimation and testing of hypotheses are well-known statistical inference methods, but we explain them in brief in the next section so that the researchers can correctly identify the setup and able to choose the right formula.

In the context of medical research, it is easy to understand the testing as 'detecting a medically important effect' setup because that is what it actually is, although the books do not explain this setup in this manner. Different books and different online resources use different formulas for the same situation. This confounds statisticians and medical researchers alike. We reviewed several reputed sources and developed relatively simple versions of sample size formulas for commonly occurring

Address for correspondence: Dr. Abhaya Indrayan, A-037 Telecom City, B-9/6 Sector 62, Noida - 201 309, Uttar Pradesh, India.
E-mail: a.indrayan@gmail.com

Access this article online

Quick Response Code:

Website: www.ijcm.org.in

DOI: 10.4103/ijcm.ijcm_787_23

How to cite this article: Indrayan A, Mishra A, Bhaskarapillai B. Simplified sample size formulas for detecting a medically important effect. Indian J Community Med 2024;49:464-71.

Received: 22-11-23, **Accepted:** 26-02-24, **Published:** 24-05-24

setups so that they can be easily used without much error. Most of the actual formulas are quite complex and this hinders their use in practice. The formulas in this communication have nearly the same uniform pattern. So many formulas in uniform format are not available at any single source – neither a website nor a book, nor a software.

Estimation vs. Detecting an Effect

Estimation is generally descriptive and tells us the magnitude of the problem or the magnitude of an effect. Finding the mortality rate in moderate CoViD cases is an estimation exercise, and so is finding the percentage of obese people ending up with at least one cardiac event in life. In the case of continuous variables, this could be the estimation of the mean of measurements such as the duration of hospitalization of bariatric surgery cases or the hemoglobin level in women facing the first pregnancy. The required sample size in this case depends on the tolerable margin of error in the estimate, called precision, and the confidence level we wish to ensure in our estimate. To keep our focus, and to keep this article within limits, we are not discussing the estimation setup in this communication, instead discussing the more commonly occurring but more complex setup of detecting a medically important effect. As mentioned earlier, this is the same as the testing of the hypothesis setup. 'Detecting' means getting statistical significance at the specified level. If the effect is not statistically significant, the study will miss the effect even if present.

In detecting an effect situation, the interest could be to find whether any effect is present or not or, mostly, whether the effect size reaches a specified threshold – generally called a minimum medically important effect or medically significant effect. The mean systolic blood pressure (BP) level in the first-time electrocardiogram (ECG) positive male patients may be 141 mmHg and in female patients 144 mmHg, and these means may be statistically significantly different, the important medical question is whether the difference of 3 mmHg in their mean systolic level is enough to prescribe different management for the two sexes. In the case of intervention, this is the targeted improvement in the outcome for it to be clinically useful. Medical significance is different from statistical significance. In our example on BP, this is the difference in the mean levels between two groups, but the size of the effect could be measured in terms of any other parameter. The required sample size is large if a small effect is to be detected. This is like finding a needle from a haystack that requires much more effort than finding a brick. The ability of a study to detect the specified effect when present is called the power of the study. In this case, the effect size would be statistically significant. Thus, power is necessarily related to the effect size to be detected and should be stated, for example, as "the sample size has been calculated to detect an effect of at least 7% with a power of 80%". The power is hardly ever described in this manner in the literature. A power of 80% implies that there is an 80% chance that the study will be able to detect the stated medically important effect, when present. This would imply that there is a 20% chance of missing the required effect when present.

No study can detect a medically important effect if not present except by way of (Type-I) error of false positive result. Thus, the threshold of Type-I error also needs to be specified – called the level of significance (generally fixed at 5%). Eighty percent is the conventional power in most medical studies, but one can choose 90% power that will increase the required sample size. A full 100% power is not possible due to omnipresent medical uncertainties. Power is denoted by $(1 - \beta)$, where β is the probability of Type-II error of false negative result.

There are many examples where a study was not able to detect a medically important effect, though present, because of the limited size of the sample. Back in 1978, Freiman *et al*.[2] re-examined 71 negative trials and found that 50 of these had more than a 10% chance of missing a therapeutic improvement because of the inadequate size of the sample. Dimick *et al*.[3] reported similar findings for surgical trials. These are the examples of false negative results.

Basic Requirement for Calculating the Sample Size

More than 80 formulas for sample size are available with us but most of them are derived from nearly 10 basic formulas. The choice of the right formula depends on the objectives of the study, particularly the primary objectives. These objectives must be stated in a measurable format so that the indicator measuring the outcome is clear. Thus, in place of saying that the objective of a study is to assess which treatment is better, specify the outcome of this assessment. This could be mortality, duration of hospitalization, the importance of a factor for prediction, need for rescue analgesia, pain score, the predictive value of a test, or any other. The effect of interest could be measured by the mean, proportion, correlation coefficient, odds ratio (OR), relative risk (RR), hazard ratio (HR), the difference between two groups, or any other such parameter. This specification will determine the formula to be used and will come from the measurable objectives.

The basic format of the simplified sample size formula for detecting an effect is the following for a two-tailed test. This is valid for Gaussian distribution of the estimate of the effect size and requires sample random sampling.

$$n \geq \left[\frac{z_{1-\alpha/2} + z_{1-\beta}}{\delta}\right]^2 * (\text{Numerator of the variance of the estimator of the effect size})$$

where z_t is the value of the standard normal deviate at cumulative probability t. The numerator of the variance (square of the standard error, SE) of the estimator of the effect size changes from situation to situation. Thus, this requires some caution. For example, for the sample mean, this variance is σ^2 / n but only is used in this formula and can be understood as the variance of the original values used to derive the estimator. This article follows this pattern uniformly in all the formulas.

Once the right formula is chosen, the next step is to identify the values of the parameters to be plugged into the formula. For example, the variance in the target population is required but the population values are rarely available. Thus, the reported values in the literature or the ones computed from the records are generally used and called the anticipated values. This is an approximation but is accepted around the world for the calculation of sample size. In case no previous study is available, a pilot study on a small sample is done and the results of this pilot study are used for estimating the values of such parameters. Sometimes one's own clinical experience can be used to make a guess. If the objective is to compare the average blood loss per patient in two methods of surgery, the information required for the calculation of sample size is the variance of blood loss reported in a previous document with each method of surgery. Variance measures the variation – the larger the variation, the more difficult is to spot an effect, and bigger is the requirement of the sample size. In case the variance is not known, and the range is known, range/4 can be used as an approximate estimate of the standard deviation (SD), which is the square root of the variance. In case the median and IQR are given, an approximation of SD is IQR/1.35, where 1.35 comes from the Gaussian distribution covering the middle 50% of the subjects. This though will give larger than actual SD and a larger sample size. To err on this side is considered tolerable. The sample size calculated from the previously reported study or pilot study values should be inflated (we advise, at least 10%) to account for variation in the reported values from study to study. In the case of two or more primary objectives, calculate the sample size for each of them and use the largest. In a large-scale study, this is calculated for each primary and secondary objective, and the maximum is used. If the objective is composite involving several outcomes to be considered together, the level of significance will be split accordingly. For example, for four objectives together, the level of significance for each will be 1.25% to make a total of 5%.

The third requirement for the calculation of sample size for the setup we are discussing in this communication is the minimum medically important effect proposed to be detected if present. This is denoted by δ. A large sample, even if exceedingly large, cannot detect a predetermined medically important effect if that kind of effect is not present, except by way of error. Note that the medically important effect a researcher wishes to detect with sufficient power is not the same as the effect reported in a previous document. Most textbooks, software, and online sample size calculators make this error. For example, G*Power[4] calculates δ from the previously reported mean and SD of the two groups (which they call the effect size) whereas this should be the other way around. The δ should be specified first based on clinical considerations and then the mean of the second group obtained as $\pm\delta$ from the mean of the first group depending upon the second group is anticipated to have a lower or higher mean. The first mean is for the baseline with which the comparison is planned. For example, this may be a placebo or the existing treatment regimen.

We have already mentioned the power and the level of significance as a prerequisite for calculating the sample size. The level of significance will require you to decide the parameter of interest that the situation (alternative hypothesis to statisticians) is one-tailed or two-tailed. If you are not familiar with these concepts, consult any elementary statistics book, such as by Indrayan and Malhotra.[1]

In summary, the sample size calculations for detecting an effect can be done only when (i) the objectives are stated in a measurable format – this will decide the parameter of interest and the right formula, (ii) anticipated values of the parameters in the sample size formula from a relevant previous document (or a pilot study) are available, (iii) the minimum medically important effect size proposed to be detected by the study and the associated power are specified, (iv) and the level of significance is stated, including one- or two-tailed. If you intend to consult a biostatistician, have all this information ready. Note that the confidence level is needed for estimation and not for testing a hypothesis. Similarly, the level of significance is needed for testing of hypothesis and not for estimation. There are some other statistical requirements if somebody wishes to be more accurate. These are mentioned under the Limitations section of this article.

Sample Size Formulas for Detecting a Specified Effect

The minimum medically important effect, δ, defines the alternative hypothesis. For example, a clinician may say that a new treatment regimen must be at least 3% more effective than the previous (established) regimen for discarding the old and adopting the new. Then, $\delta = 3\%$ and a difference of less than 3% is clinically immaterial. The sample size formulas for various situations for detecting δ are as follows. As explained earlier, detecting in this case means getting statistical significance at a specified level. If the effect is less than δ, this will be missed as non-significant by the sample size calculated from these formulas. These formulas have been chosen after a review of several reliable sources[5–16] and, more importantly, modified as needed to make them simple, on uniform pattern, and easily adaptable. These are approximations and subject to the limitations mentioned later but work well. Actual formulas in most cases are quite complex. The commonly used notations in the formulas are explained in Table 1. These formulas are given for a two-tailed setup. For a one-tailed situation, replace $z_{1-\alpha/2}$ with $z_{1-\alpha}$.

Comparison of two independent groups – Test for equality

This is the setup in most studies including clinical trials and case-control observational studies. These studies will have two parallel groups and the objective is to find whether one group has a clinically important different outcome from the other. The null hypothesis under test is that they are equal.

1. For the difference between means:

The actual formula is given in the Supplementary Material. The simplified formula for $n_A = kn_B$ (such as k controls per case) is

Table 1: Notations for sample size formulas

Notations	Explanation
α	The maximum tolerable probability of Type-I error (also called the level of significance in testing of hypothesis setup)
$z_{1-\alpha/2}$	Standard Gaussian (normal distribution) value with α/2 probability in the right tail (1.96 for 5% level of significance – two-tail)
$z_{1-\alpha}$	1.645 for 5% level of significance – one-tail
β	Probability of Type–II error (1 – β is the power)
$z_{1-\beta}$	0.84 for 80% power and 1.28 for 90% power (from the Gaussian distribution)
$z_{1-\beta/2}$	1.28 for 80% power
σ^2	Anticipated variance of a continuous variable (from a previous document)
π	Anticipated proportion of persons with the condition (disease or exposure) under consideration (from a previous document)
k	Number of controls per case – generally group sizes are equal and k=1
K	Number of times a subject is repeatedly assessed such as on day 0, day 7, day 30 and day 60 (K=4 in this case) – K is required only for repeated measure (follow-up) studies. (Note the difference between upper K and lower k.)
τ	Number of pairwise comparisons in case of 3 or more groups in ANOVA setup – for 4 groups, the number of pairwise comparisons is τ=6 (A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, and C vs. D)
r	Anticipated correlation coefficient (from a previous document)
θ	Anticipated hazard rate or ratio (from a previous document)
δ	Minimum medically important effect to be detected (determined by clinical considerations) – could be the difference in means, difference in proportions, odds ratio, correlation coefficient, or any other effect of interest (or testing margin in the case of equivalence or superiority/non-inferiority trials)

Subscripts A and B have been used to identify the groups

$$n_B \geq \left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{\delta}\right]^2\left[\sigma_B^2+\frac{\sigma_A^2}{k}\right]$$

This is the minimum for the smaller group. In most situations, $\sigma_A^2=\sigma_B^2$ and $k=1$. These imply that the variance in group A is the same as in group B and the number of subjects in group A is proposed the same as in group B. The common variance σ^2 can be the pooled estimate. Under these conditions, the formula further simplifies to

$$n \geq \left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{\delta}\right]^2\left(2\sigma^2\right) \quad \text{per group.}$$

Example: A study is planned to compare the reduction in BMI in obese subjects by regular exercise (such as sports) and Yoga in adolescents, and it is considered that Yoga can be recommended when the average reduction in BMI by Yoga is at least 1.5 kg/m^2 more than by regular exercise after 6 months. Thus, δ =1.5. If a previous study indicates that the variance of reduction in BMI by regular exercise is σ_A^2 = 1.29 and σ_B^2 =2.61 by Yoga, the minimum sample size required to detect a difference of at least 1.5 kg/m^2 with a power 80% at significance level 5% is 30 per group by the first formula for the equal number of subjects in Yoga and regular exercise groups (k = 1). This sample size is 80% likely to not miss a difference of at least 1.5 kg/m^2 in the mean BMI if present but may miss if the difference is less. 'Not missing' means that statistical significance will be obtained with $P < 0.05$.

Note: To keep this communication simple, we provide subsequent formulas for equal variances (continuous variable) and equal group sizes. Where the variances are not equal and/or the intention is to recruit k controls for each case (as in some case-control setups), the variance of the larger group is divided by k as mentioned for the first formula stated above. A similar change is required in all the following formulas from 2 to 7 whenever $k \neq 1$.

2. *For difference between proportions:*

The actual formula is complex and is given in the Supplementary Material.

For k = 1, and further simplification gives

$$n \geq \left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{\delta}\right]^2\left[\pi_A\left(1-\pi_A\right)+\pi_B\left(1-\pi_B\right)\right] \text{ per group.}$$

Note that $\pi_B=\pi_A\pm\delta$ so that $\left|\pi_A-\pi_B\right|=\delta$. Thus, only π_A and δ are required and π_B comes from these two, where δ is the minimum medically important effect proposed to be detected. This correct method of sample size calculation rarely appears in the literature.

3. *For difference between risks (attributable risk or absolute risk):*

Since this is the difference between the risk in the exposed and the risk in the unexposed group, the formula is the same as for proportions mentioned above because risk also is a proportion.

4. *For difference between sensitivities and specificities:*

Each of these is also a proportion and hence the formulas are the same as for proportions.[13]

5. *For difference between (Pearsonian) correlations:*

$$n \geq 2\left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{c_A-c_B}\right]^2+3 \quad \text{per group,}$$

where

c_A = Fisher z-transformation of the anticipated correlation coefficient of group A

$$c_A = \frac{1}{2} ln\left[\frac{1+r_A}{1-r_A}\right]$$

c_B = Fisher *z*-transformation of the anticipated correlation coefficient of group B

$$c_B = \frac{1}{2} ln\left[\frac{1+r_B}{1-r_B}\right]$$

Calculate c_B for r_B, which has the minimum medically important difference with r_A. ($r_B = r_A + \delta$ or $r_A - \delta$ as needed, where δ is the minimum medically important difference between the correlations in the two groups).

Example: Consider living donor liver transplant cases, some of whom require tracheostomy and others do not. The objective is to find whether the correlation between donor age and length of ICU stay is the same or different in the tracheostomy group than in the non-tracheostomy group. A previous article reported that the correlation in the non-tracheostomy group is $r_A = 0.65$. This is the anticipated correlation. Clinicians decide that the correlation in the tracheostomy group should be higher (one-tail) by at least 0.10 ($r_B = 0.75$) for it to be clinically different from the correlation in the non-tracheostomy group. The possibility of a lower correlation in this group is ruled out. Thus, in this case,

$$c_A = \frac{1}{2} ln\frac{1+0.65}{1-0.65} = 0.775 \text{ and } c_B = \frac{1}{2} ln\frac{1+0.75}{1-0.75} = 0.973.$$

To not miss a difference of 0.10 (if present) with a power of 80% ($z_{1-\beta} = 0.84$) at significance level 5% ($z_{1-\alpha} = 1.645$- one-tail), we get

$$n \geq 2\left[\frac{1.645+0.84}{0.775-0.973}\right]^2 + 3 = 2\times 157.52 + 3 = 318.04.$$

Inflate it by 10%, for possible variation in the correlation from study to study that we have taken 0.65 in this example from a previous study. Thus, a sample of size 350 is required to not miss a difference of 0.10 in the correlation between the two groups.

6. *For difference between hazard rates (survival studies):*

$$n \geq \left[\frac{z_{1-\alpha/2} + z_{1-\beta}}{ln\theta_A - ln\theta_B}\right]^2 \frac{1}{p_A p_B \pi} \quad \text{per group,}$$

where,

p_A = proportion of subjects in group A (generally $p_A = p_B = \frac{1}{2}$, i.e., equal groups)

p_B = proportion of subjects in group B

θ_A and θ_B are the anticipated hazard rates over time in groups A and B, respectively ($ln\theta_A = ln\theta_B + \delta_1$), where δ_1 corresponds to the minimum medically important effect after taking logarithm.

π = anticipated overall probability of event occurring during the study period

Example: Patients in cardiac rehabilitation with ejection fraction (EF) <55% and EF ≥55% were advised to walk at least 7500 steps a day. The hazard rate of cardiac hospitalization in a previous study in such cases was 1.43 per year in patients with EF ≥55%. Thus, $\theta_A = 1.43$. For different management of patients with EF <55%, the clinicians decide that the hazard rate in them should be at least 3.00 ($\theta_B = 3.00$). Thus, the clinically important difference is δ =3.00 - 1.43 = 1.57. The experience suggests that cardiac hospitalization occurs in about 10% (π =0.10) of the patients in cardiac rehabilitation. It is expected that one out of every four such patients on average has EF <55%. This gives $p_A = 0.75$ and $p_B = 0.25$. Substitution of these values in the formula just mentioned gives n ≥763. Of these, one-fourth are expected to have EF <55%. A study on at least 763 patients is required for getting statistically significance of the difference of 1.57% in the hazard rate between the two groups. This should be inflated by 10% for random errors in the estimates we are using for calculation of the sample size. Note how many parameters are needed to calculate the sample size in this case.

7. *For difference between area under the ROC curve when obtained by two methods on the same group of patients*[14] *(simplified)*

$$n \geq \left[\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)}{\delta}\right]^2 \left[V\left(AUC\right)_A + V\left(AUC\right)_B\right] \quad \text{per group,}$$

where

$(AUC)_A$ = anticipated area under ROC curve with method A

$(AUC)_B$ = anticipated area under ROC curve with method $B = (AUC)_A + \delta$ or $(AUC)_A - \delta$

δ = minimum medically important difference in AUCs between the two groups

$V(AUC)_A$ = anticipated variance of $(AUC)_A$

$V(AUC)_B$ = anticipated variance of $(AUC)_B$

In case any of the $V(AUC)$ s is not available, use the following method:

$a = \varphi^{-1}(AUC) * 1.414$; where φ^{-1} is the cumulative probability in the standard normal distribution, and $V(AUC) = (0.0099 * e^{-\alpha^2/2}) * (6\alpha^2 + 16)$

8. *Each group with K repeated measures – Difference in means in two groups:*

In many medical studies, each subject is followed up and assessed for the condition at several points in time such as at the time of admission, time of discharge, 3 months after discharge, at 6, 9, and 12 months. The sample size calculation in this case is complex and is based on variances of the repeated measures and correlation pattern.[16] We are providing a simplified approximate formula. If *K* is the number of equally spaced time points each subject is measured, the sample size required per group is obtained as

$$\left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{\delta}\right]^2\left(\sigma^2\right)\frac{1+(K-1)\rho}{K}$$

where ρ is the anticipated correlation coefficient between values obtained in different repetitions and when the variance at baseline is the same as at the endpoint. In this case, δ is the minimum medically important difference between group means at the endpoint as the mean difference at baseline between them is expected to be zero because both the groups are supposed to have started with the same baseline. The value of ρ would be rarely available in previous documents and may have to be your best guess based on the clinical experience. Note in this case that σ^2 is the common variance of the repeated measures. This is estimated by the mean error sum of squares (MSE) in the ANOVA table.

The aforesaid procedure requires that the levels of within-subjects factors are same in the two groups under comparison and is valid when the correlation is the same between measurements at time1 and time2 as between time1 and time3, or between any pair of time points (known as sphericity). Thus, ρ is the common correlation coefficient. Experience suggests that the correlation coefficient between measurements at different points in time remains nearly the same because the measurements at different time points are likely to differ by almost a constant margin. It is also seen that this correlation is high because the repeated measurements belong to the same subjects.

Historical controls

In the case of historical controls, only one group is studied. The total sample requirement for investigations becomes half of what is needed for a parallel group setup because the other half is available in records. Such studies have less reliability because historical controls are rarely comparable to the current group under study due to the improvement in the diagnostic, assessment, and treatment approaches. The formulas remain the same despite only one group under study. In this case, the second group in the formula refers to the historical controls. The variance, the proportion, or any other value, as required for these formulas, should be available for the historical controls as well.

Comparison of one group (including paired values) with a pre-fixed value

This arises in at least three situations:

(i) Paired observations (e.g., before–after) where the difference indicates an effect, such as fasting blood glucose level before and after a treatment. In this case, the null hypothesis generally is that the mean difference = 0 and δ is the average medically significant difference we wish to detect.

(ii) Comparison of a new regimen with a predefined standard, such as efficacy of a vaccine is at least pre-fixed 70%. The value of δ in this case will be the difference we wish to detect from the pre-fixed value.

(iii) Comparison of a correlation coefficient with a pre-fixed value such as to find whether the correlation is at least 0.6 or not.

The requirement of the total sample in all these situations is drastically less than required for two groups setup because now only one group is studied (although such studies have much less validity due to the role of the uncontrolled factors), and a fixed value is used in place of the value in the other group.

The sample size formulas for these situations are given in the Supplementary Material. Of special interest are the difference of odds ratio and relative risk from the null value of OR = 1 and RR = 1, respectively, or any other fixed value. For these parameters, two groups (cases and controls) are required but generally the comparison is with a pre-fixed value of OR or RR. Thus, they also come in the one-group setup for the purpose of calculation of sample size.

Equivalence, superiority, and non-inferiority studies

These studies require that a margin is set for assessing equivalence, superiority, or non-inferiority, as the case may be. Call it testing margin and denote it by δ. Whereas the equivalence is tested by two one-sided tests, the superiority and non-inferiority are one-sided tests. The sample size for these setups requires the following two changes:

(i) Switch α with β and β with α since, in these setups, the null hypothesis becomes the alternative, and the alternative becomes the null hypothesis. Thus, $z_{1-\alpha/2}$ becomes $z_{1-\beta/2}$ for equivalence but remains $z_{1-\beta}$ for superiority and non-inferiority and $z_{1-\beta}$ becomes $z_{1-\alpha}$ for all the 3 setups.

(ii) The denominator, in place of δ, becomes the (observed effect in a previous study – δ), where δ is the testing margin.

The sample size formulas thus obtained are in the Supplementary Material for some parameters.

Comparison of three or more groups

The setup of three or more groups requires analysis of variance (ANOVA) for comparison of means and chi-square for comparison of proportions. Both these yield complex formulas for sample size. Instead, a generally accepted simple procedure is to count the number of pairwise comparisons. If there are K groups, the number of pairwise comparisons is $\tau = K(K-1)/2$. With this, the only change required in the sample size formulas for the two groups is to replace $z_{1-\alpha/2}$ by $z_{1-\alpha/(2\tau)}$, where τ is the number of pairwise comparisons. Everything else remains the same as for two groups, but now pooled estimate of the variance σ^2 is used (ANOVA in any case requires equal variances in the groups under comparison) in the case of means, and pooled estimate in the case of proportions. This method is applicable only when the number of subjects in each group is the same ($k = 1$). In case δ is different for different pairs, the sample size is calculated for each pairwise difference with specified δ and the required sample size will be the largest n per group.

Logistic and Linear (Multivariable) Regression

1. *For logistic with one binary predictor*[15] *(equal number of cases and controls):*

$$n \geq \left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{\pi}\right]^2\left[\frac{\pi(1-\pi)}{B(1-B)}\right] \quad \text{per group,}$$

where

π_1 = anticipated event rate at $x = 0$, where x is the binary predictor (such as death rate in cases with no hypertension (no HTN))

π_2 = anticipated event rate at $x = 1$, where x is the binary predictor (such as death rate in cases with hypertension (HTN))

B = proportion of subjects with $x = 1$ (prevalence of the event) (such as percent of subjects with HTN)

$\pi = B*\pi_2 + (1 - B)*\pi_1$ (overall event rate)

δ = minimum medically important difference between the hypothesized and the anticipated logistic coefficient to be detected (mostly, the hypothesis is logistic coefficient $\beta = 0$ – this β is different from β used in $z_{1-\beta}$)

2. *For logistic with multiple predictors (equal number of cases and controls):*

A simulation study indicated that 10 events (outcomes) per predictor may be adequate for testing significance of the logistic coefficient in a multivariable setup.[17] This will have several βs (logistic coefficients). To test any one $\beta = 0$, where $\beta = \ln(OR)$,

$$n \geq \left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{\delta}\right]^2\left[\frac{\pi(1-\pi)}{B(1-B)(1-R^2)}\right] \quad \text{per group,}$$

where all the notations are the same as in the preceding formula

R^2 is the square of the anticipated multiple correlation coefficient, such as Nagelkerke R^2. $(1 - R^2)$ is called the variance inflation factor.

3. *For logistic with one continuous predictor*[15] *(equal number of cases and controls):*

$$n \geq \left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{\delta}\right]^2\left[\frac{1}{\pi(1-\pi)}\right] \quad \text{per group,}$$

where

π = event rate at mean of x, where x is the continuous predictor

δ = the value of the logistic coefficient β to be detected for this predictor (the null is $\beta = 0$).

4. *For linear regression with one predictor:*

Same as for single correlation (Supplementary Material).

5. *For linear regression with multiple predictors:*

Several βs but to test any one $\beta = 0$,

$$n \geq \frac{\left[\left[\frac{z_{1-\alpha/2}+z_{1-\beta}}{c}\right]^2+3\right]}{(1-R^2)}$$

where

R = expected multiple correlation coefficient (the denominator $(1 - R^2)$ is the variance inflation factor)

c = Fisher z-transformation of the correlation coefficient (r) to be detected for the variable under consideration, i.e., $c = \frac{1}{2}ln\left[\frac{1+r}{1-r}\right]$

Other Situations

See Supplementary Material for many other formulas. Despite more than 30 formulas, including those in Supplementary Material, the list in this communication is not exhaustive. For two-stage and adaptive designs, the sample size requires adjustment as given by Indrayan and Holt.[18] For a simple comparison of more than two groups, each with repeated measures, we could not locate any simple sample size formula. Separate formulas are available for bioequivalence and dose-response studies[9] and multilevel data.[5]

Limitations

Exact formulas for sample size are complex. The formulas in this communication are approximations and valid for large samples so that the Gaussian (Normal) distribution can be used. They apply to only those small-sample studies where the distribution is Gaussian. But these are generally proposed formulas and seem to work fairly well. The second important requirement is that the subjects are selected by simple random sampling. The same formulas can be used for systematic random sampling also and for random allocation in the case of clinical trials. In case stratified, cluster, or any other method of sampling is used, an adjustment would be required. For example, in the case of stratified sampling, such as for separate results for males and females, a separate sample size calculation is required for each stratum. For cluster sampling, the adjustment is the multiplication of each variance by the design effect, $D = [1 + \rho_I * (m - 1)]$, where ρ_I is the anticipated intra-cluster correlation coefficient and m is the average cluster size.

As mentioned earlier, all sample sizes arrived at by these formulas should be inflated by at least 10% to account for variation in the values reported in the literature, which are used as an approximation for values in the target population. Second, further inflation is required in case a non-response is expected. If the nonresponse is a%, the new n is (calculated n)/$(1 - a/100)$. This adjustment is widely recommended to account for the attrition in the follow-up studies. Ignoring this adjustment can be a threat to the reliability of the findings.

How to Reduce the Sample Size Requirement

Sample size formulas invariably have variance component in the numerator. Thus, an easy method to reduce the requirement of samples is to control the variation and reduce the variance. In a laboratory study on animals or on biological specimens,

the standardized conditions minimize variability and a small sample of 5 may be enough to provide sufficient evidence. In this case, any effect seen in the subjects can be safely ascribed to the intervention because almost no other factor is operating. In clinical trials, an attempt is made to make the two groups under study as identical as possible at baseline by strict inclusion-exclusion criteria, randomization/matching, blinding, and concealment of allocation. However, even with such strategies, it is seldom possible to choose subjects of the same heredity, same diet, same physiology, same anatomy, and same behavior. Thus, the variance can rarely be so small to give a small size of the sample. When all such factors are under control, such as in a laboratory study on animals, a trial can indeed be done on small size and that would provide equally reliable results. However, this can affect the generalizability because the results would be valid only for the type of cases included in the study.

The other way to reduce the sample size is to relax the size of medically important effect. As mentioned earlier, a bigger effect, if present, is easier to detect with a smaller sample size than a smaller effect. If the minimum medically important effect is 20% improvement in place of 10%, the sample size requirement will steeply decline but any improvement of less than 20% can be missed by this sample size and can give a false negative result. All the resources spent in the study would be wasted. Many researchers do not realize this limitation of small samples.

A large sample is required to control the effect of various aleatory and epistemic uncertainties that affect most medical research. However, if the objective is just to disprove an existing hypothesis, a single contradictory event is enough to demonstrate the possibility. Many breakthroughs occurred in medicine with $n = 1$ such as by Alexander Fleming for penicillin and by Barry Marshall for *Helicobacter pylori.* For important discoveries with small *n*, see Indrayan and Mishra.[19]

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Indrayan A, Malhotra RK. Medical Biostatistics. CRC Press; 2018.
2. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med 1978;299:690-4.
3. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: Equivalency or error? Arch Surg 2001;136:796-800.
4. G*Power. Statistical Power Analysis for Mac and Windows. Heinrich Heine Universitat, Dusseldorf. Available from: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower. [Last accessed 2023 Nov 10].
5. Moerbeek M, Teerenstra S. Power Analysis of Trials with Multilevel Data. CRC Press; 2016.
6. Casagrande J, Pike M, Smith P. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. Biometrics 1978;34:483-6.
7. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 1981;2:93-113.
8. NCSS Statistical Software. PASS 2024. PASS Documentation. Available from: https://www.ncss.com/software/pass/pass-documentation/. [Last accessed on 2024 Apr 11].
9. Ryan TP. Sample Size Determination and Power. Wiley; 2013.
10. Chow S-C, Shao J, Wang H. Sample Size Calculations in Clinical Research. Chapman and Hall/CRC; 2008
11. Lwanga SK, Lemeshow S. Sample Size Determination in Health Studies: A Practical Manual. World Health Organization; 1991.
12. Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case-control designs. Biostatistics 2009;10:94-105.
13. Bujang MA, Baharum N. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. Arch Orofac Sci 2017;12:1-11.
14. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform 2014;48:193-204.
15. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med 1998;17:1623-34.
16. Overall JE, Doyle SR. Estimating sample sizes for repeated measurement designs. Control Clin Trials 1994;15:100-23.
17. Peduzzi P, Concato J, Kemper EM, Theodore R, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373-9.
18. Indrayan A, Holt M. Concise Encyclopedia of Biostatistics for Medical Professionals. CRC Press; 2016.
19. Indrayan A, Mishra A. The importance of small samples in medical research. J Postgrad Med 2021;67:219-23.