Special Article

Sample Size Estimation and Power Calculation - A Guide to Biomedical Researchers

Biju George

Asst professor in Community Medicine, Nodal Officer, Institutional Research Committee
Medical College, Kozhikode, bijugeorge1@gmail.com, Ph- 9846100093

Correspondence: Dr. Biju George, Asst professor in Community Medicine, Nodal Officer, Institutional Research Committee
Medical College, Kozhikode, bijugeorge1@gmail.com, Ph- 9846100093

Two of the most important questions related to all research studies are the way of selecting subjects and the number of subjects required for the study. Why are these two issues given so much importance? Let us take a case of a randomized controlled trial (RCT) for the treatment of hypertension and try to understand this. In an RCT, to show a difference between two drugs used for the treatment of hypertension, the researchers randomized hypertensive patients into two groups. Both the groups were given treatment and were evaluated at the end of the study to compare the desired outcome of reduction in blood pressure below a particular level. Suppose they get an inconclusive result in the study, they would have advocated against the use of this new drug.

One issue relating to the extrapolation of this result is the size of the sample from which the results have been generated. If the result is generated from a large sample then often the results will be close to the truth provided the studied sample is representative of the population. If the sample size is small, often people will not believe the results and say that such results might have been obtained by chance only and the study was not powered enough to detect the difference in outcome. Thus many researchers worry about the sample size before the start of study. They cannot take a very large number of subjects because of administrative, economic, ethical and some scientific issues. Large studies mean more work days to collect the data and thus leads to more expenditure. It also means more chance of inter observer errors. Results obtained from a small sample size are most often unreliable in the sense that, if the study is repeated many times most often we may get different results. Thus chance of obtaining a difference between the two groups will be missed even if the two groups are different in reality. Most of the researchers want to have a sample size that is neither too small (so that they can avoid the chance variation), nor too big (so that they can avoid all kinds of administrative and economic difficulty). So they need to start with an optimum sample size to suit their study objective.

How do we arrive at this optimum sample size?. Sample size is dependent on many factors[1-3] as listed in table 1, some of which are discussed below.

One tailed or a two tailed test - Most of the research situations in bio-medical field assumes a two tailed situations. In the two tailed hypothesis, the researcher uses a non directional alternate hypothesis. In a drug trial situation involving drug A & B, it means that the drug A and B differ in effect. In a non directional alternative, direction of effect is not specified. In one tailed test a directional alternative is specified. For eg- Drug A is superior to drug B. One tailed tests are used only in specific situation where possibility of change is in one direction only as in the case of the study by Dore MP et al[4].

Alpha error (α error or Type I error) - It is the chance of declaring a non effective drug as effective or it is the chance of declaring that there is an association when none is present. It is the probability of rejecting a null hypothesis when it should have been accepted. It indicates the chance of a false positive discovery of an effect /

Table 1

Some common factors affecting sample size

Factors affecting sample size	**Effect on sample size**
Design of the study	Comparative design generally require a larger sample size than descriptive designs
Type of Alternative hypothesis	One tailed tests will require 30% less sample size than two tailed test
Alpha error (α) / level of confidence (1- α)	Small α error means more sample size
Beta error (β) / Power (1-β)	Small β error / high power means more sample size
Type of the variable	Quantitative variable generally require lesser sample size than qualitative variable
Variability of the quantitative variable	Increased variability increases the sample size
Prevalence of the qualitative variable	Rarer the variable larger the sample size
Precision	More precise means an increase in sample size
Minimum clinically relevant effect size	To detect small effect size large sample size is required
Response rate, lost to follow up, missed data	All will increase sample size
Type of statistical test used	Non parametric tests generally require 10-20% more sample size than Parametric tests
Paired / Matched nature of the data	Require less sample size (approximately 50% reduction) than unpaired / unmatched data

association. In all researches,our aim is to reduce the chance of false positive results and so a low alpha level should be chosen. In most studies it is preset at a level of 5% (probability=0.05) or at 1% (probability=0.01) level. Sample size is inversely related with chance of α error. In sample size calculations the value of the standard normal variate (Z value) corresponding to the presetα error is used. Values of Z for commonly used α error (Zα value) is given in table 2

Table 2

Z values for different α and beta (β) error

Preset α error	Two tailed Zα value	Presetβ error	Zβ value
0.001(0.1%)	3.29	0.01 (1%)	2.33
0.01(1%)	2.58	0.05 (5%)	1.64
0.05 (5%)	1.96	0.10 (10%)	1.28
0.1 (10%)	1.64	0.20 (20%)	0.84

Beta error (ß error or Type II error) - It is the chance of declaring an effective drug as ineffective. It is the probability of missing an association when it is present. It is the chance of obtaining false negative results from the study or it is the chance of accepting null hypothesis when it should have been rejected. Alpha and beta errors are depicted in table 3. Similar to α error, ß error should be low. In most studies it is preset at a level of 20% (probability=0.20) or at 10% (probability=0.10) level. Sample size is also inversely related with chance of ß error. In sample size calculations the value of the standard normal variate (Z value) corresponding to the preset ß error is used. Values of Z for commonly used ß error (Zß value) is given in table 2. Power (1- ß) is the inverse of the ß error and represents the probability of true positive result. It is the chance of detecting an effect / association if it exists. Usually it is preset at 80% or at 90%. Alpha and beta error are important for the comparative designs where there is a hypothesis testing.

Level of confidence (1- α) it is commonly used in descriptive designs. Usually a 95% confidence interval (95% CI) is used. A 95% CI indicate that there is 95% probability that the true population parameter will be within the calculated interval. Z value for α error is used in the sample size calculation for descriptive design.

Type of the variable Sample size calculation depends on the type of variable. Variable is the attribute/ character that we measure from each observations. Generally variables are classified as quantitative (eg- age, height, weight, fasting blood sugar etc) or as qualitative (eg- gender, mortality, cure etc). Some variables like Glasgow coma scale (GCS) score, depression score and other similar scores are usually treated as quantitative for the analysis purpose (they are sometime called as semi quantitative variables). In general, sample size is low for quantitative variables and higher for qualitative variables.

Variability of the quantitative variable It is usually measured in terms of standard deviations (SD). SD generally remains same for a variable in different study settings. Take the case of systolic blood pressure (Systolic BP), its variability remains constant in different studies of general population at a value of around 15 to 20 mm of Hg. Sample size requirement will be more for a highly variable character.

Prevalence of qualitative variable- To study a rare character like polydactyly we will require huge sample size. As the prevalence of a condition increases upto 50% in the general population, the sample size requirement will decrease. As the prevalence of a character increase above 50%, prevalence of the absence of the character in

Table 3

Alpha and beta error in terms of Null hypotheses

Study	Truth	Truth (Reality) in the population	
		Null Hypothesis False (Drug Effective / Association Present / Difference present)	Null Hypothesis True (Drug Not effective / No association / No difference)
Study Results from the sample	Null Rejected (Drug Effective / Association Present / Difference present)	Correct decision **(Power)**	**Alpha Error**
	Null not Rejected (Drug not Effective / No association / No difference)	**Beta Error**	Correct decision

population will decrease. Presence and absence of a character being two sides of same coin will require same precision for sample size calculation. A prevalence of 50% for a character will require the lowest sample size.

Precision It is the closeness to the truth that we want to obtain in a study. Highly precise results means that the result obtained in the study is very close to the truth. To get precise results high sample sizes are required. Precision is an important parameter in deciding the sample size in descriptive designs. Size of the precision should be decided based on clinical relevance.

Minimum clinically relevant effect size In all comparative studies, where we need to find a difference between two groups, it should not only be statistically significant but also clinically relevant. In sample size calculation we would like to use this clinically relevant minimum difference (effect size) to estimate the sample size. What constitutes a clinically relevant difference should be decided based on the clinical experience.

Now let us see the different sample size formula for different situations and their worked out examples.

Sample size for Descriptive designs

In descriptive designs, researchers study either the proportion of people with a character (for qualitative variable) or study the average value of a character (for quantitative variable). Every researcher would like to estimate the population parameters with high precision from the sample. If we are taking a large sample then the sample estimate will be very close to the population parameter and thus the precision will be high. If the sample size is small then there is a chance that the sample estimate will be deviated away from the population parameters ie- precision is low. Thus precision with which we have to estimate the population parameter will be the most important factor in the calculation of the sample size in descriptive designs. Formula[5] and a worked out example for the sample size calculation for descriptive designs is given in table 4.

Table 4

Formulas used for descriptive designs

	Quantitative variable	Qualitative variable
Formula	$n = \frac{4SD^2}{d^2}$ SD = Standard deviation of the variable d = precision we want (decided by common sense)	$n = \frac{4pq}{d^2}$ p=expected proportion of subjects with the character (from published literature) q=100-p d=precision we want (often we take 10% or 20% of the expected Proportion)
Worked out example	Question - we want to estimate the diastolic BP of a group of diabetics within ± 2 mm of Hg of the true population mean Expected average diastolic BP is 88 mm of Hg with a SD of 12 mm of Hg. Answer - SD= 12 mm of Hg, d= we want to estimate the population parameter within ± 2 mm of Hg So n= 144, and we need to study 144 subjects to get the results within a ± 2 mm of Hg precision. If we needed higher precision of ± 1mm Hg, then the sample size will be 576	Question - we want to estimate the prevalence of diabetic retinopathy in a group of diabetic patients coming to medicine OP. Expected prevalence of diabetic retinopathy among them is 20% Answer - p= 20% ; q= 100-p = 80%; d= 20% of p = 20 % of 20% = 4% n=400, So we need to study 400 diabetic person to estimate a 20% prevalent character with ±4% precision. If we need to estimate within ±2% precision then the sample size will be 1600.

Finite population correction	Most of the sampling assumes that sample is taken from a large population (infinite). A correction formula has to be applied when the sampling is done from a finite population such as selecting a sample of Tuberculosis patients from a Register kept at the Tuberculosis unit at sub district level. The number of tuberculosis patients will be affixed, finite one. The corrected sample size (n*) can be found out by the formula $$n^* = \frac{n}{1+\frac{n-1}{N}}$$ n*= sample size corrected for finite population n = sample size calculated using original formula N= Size of the finite population from which we are taking the sample

Sample size for Comparative designs

In a comparative study the researcher wants to compare two groups of subjects and make an inference about the difference, in the population represented by them. Most important factor in sample size calculation is the minimum clinically relevant effect size. One of the challenges in sample size calculation is the estimation of this clinically relevant effect size. Ideally this should be based on the consensus of different experts in the same field. For example a difference of 1mg% in fasting blood sugar between two drug groups may not be considered as clinically relevant by most researchers while a difference of 5 or 10 mg% will be considered as relevant. A change in cure rate from 85% to 90 % may not be as relevant compared to change in cure rate from 1% to 6%. Thus the minimum clinically relevant effect size should be decided based on consensus for each situation. Formula[5] and a worked out example for the sample size calculation for comparative designs are given in table 5.

Table 5

Formulas used for comparative designs

	Quantitative variable	Qualitative variable
Formula	$$n = \frac{\left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right)^2 \; SD * 2}{d^2}$$ Z_{α_2}= is the Z value at an α error Z_{β}= is the Z value at an β error SD= average standard deviation of the character = $(SD_1+SD_2)/2$ SD_1,SD_2 are the standard deviation of the character in each group d= clinically relevant effect	$$formula\ 1 \rightarrow n = \frac{\left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right) pq * 2}{d^2}$$ $$formula\ 2 \rightarrow n = \frac{\left(\left(Z_{\frac{\alpha}{2}}\sqrt{2pq}\right) + (Z_{\beta})\sqrt{p_1q_1 + p_2q_2}\right)^2}{d^2}$$ Z_{α_2} = is the Z value at an α error Z_{β} = is the Z value at an β error p= average percentage of the character = $(p_1+p2)/2$ p_1,p_2 are the % of the character in each group d= clinically relevant effect size

	Quantitative variable	Qualitative variable
Worked out example	Question - Estimate the sample size to study reduction of diastolic BP between a new drug and a standard treatment Average reduction (SD) of DBP from base line in standard Rx = 8.6 mm Hg (14.1) (obtained from previous publications) It is expected that the new drug will be marketed only if it is better than this drug in reducing DBP at least by 4 mm Hg (this is assumption / from previous studies) Answers - SD= 14.1, d= 4 And n = 195 So 195 in each group = 195 + 195 = 390 total	Question - Estimate the sample size for a study to compare the effect of a new drug to cure an infectious disease compared to a standard treatment. Cure rate with standard treatment = 70% (found from other published studies) The researchers assumed that / expects that the new drug should be at least 10% better than standard treatment (thus clinically relevant minimum effect size =10%), Answer - So expected cure rate with the new treatment = 80% So p1= 70%; d= 10%, so p2=80% thus p = (70+80)/ 2 = 75% and q = 100-p = 25% And n= 294 Thus 294 subjects should be studied in each group = 294 + 294 = 588 total

Table 6
Change in sample size depending on ratio of subjects in study group

Ratio ->	1:1	1:2	1:3	1:4	1:5	1:6
Size of the Group1	294	221	196	184	176	172
(% decline in size compared to 1:1 ratio)	(0)	(25)	(33)	(37)	(40)	(42)
Size of the Group2	294	442	588	736	882	1032
(% increase in size compared to 1:1 ratio)	(0)	(50)	(100)	(150)	(200)	(251)
Total sample size	588	663	784	920	1058	1204
(%increase in size compared to 1:1 ratio)	(0)	(12)	(33)	(57)	(80)	(105)

Unequal sample size for each group - Formula given in table no 5 is applied if the two study groups are equal in size. Practically ,getting a comparison group may be much easier than obtaining the study group. Thus we may think of taking double/ triple control subjects for each study subject and reduce the number of subjects in the first group without inflating the total sample size much.

If the group size is not 1:1 then the above two formula has to be modified by replacing the 2 in the numerator by (r+1)/r, where the r = ratio of subjects in sample 1 to sample 2. If the ratio is 1: 1 then (r+1)/r = (1+1)/1=2. Thus if 1:1 ratio is not used then the sample size in the worked out example in table 4 will be modified as given in table no 6.

It can be seen from the table 6 that a change in ratio from 1:1 to 1:2 will reduce the sample size requirement in group 1 by 25% without increasing the total sample size requirement much (12% increase). A change in ratio to 1:3 reduces the sample size requirement for group 1 by 33%, a small further reduction from 1:2 ratio. But at the same time it increases the total sample size requirement by 33%. Any further increase in ratio doesn't reduce group 1 size much compared to the increase in total sample size. A sample size calculation with 1:1 ratio is the optimum with highest power. To achieve the same power we will require disproportionately more subject in group 2 as the ratio increase from 1:1. In situation where obtaining a group is difficult / costly then the ratio of group can be increased up to 1:3 without losing power and increasing total sample size much. Any further increase will not be cost efficient.

Power analysis - Power (100-) is the ability of the study to conclude that a relation / effect exist if it was present (ability to reject null hypothesis when it should be rejected). In many comparative studies, researchers often ask what will be the power of the study. Most sample size calculations are based on a power of 80% or 90% (ß error of 20% or 10%). These sample size calculations are based on a set of assumptions regarding the effect size, variability in terms of standard deviation and assumptions regarding the proportion of character (prevalence). Most of these values are taken from other published studies or is based on the best guess by the researcher. Estimates obtained from published studies from other settings may not be the reality in our setting and the assumed values may differ from the truth considerably. So the assumption made during the sample size calculation may not be a good estimate at all time and we may need to get an idea about the reduction in power produced due to these breakages of assumptions. Thus power analysis is very much important in all comparative studies especially in Randomized Controlled Trials. All funding agencies and institutional review board will scrutinize this aspect to know what change will happen to the power if the researchers fail to get the adequate sample size / expected effect size. If the power is less than 60% in these situations, most of the funders will request the researchers to rework the sample size calculations. Power is calculated as the probability density of getting the calculated Zß value. Formula for Zß value is derived from the sample size formula by rearranging the components. Power analysis should be done in the protocol stage itself and not as a post hoc procedure. Post hoc power calculation is not considered as a good practice in bio-medical research[3,6].

To see how power analysis is important let us take the case of a worked out example in table 5 (where sample size for comparing the effect of a new drug to cure an infectious disease compared to a standard treatment was calculated). The sample size for detecting a 10% difference in cure rate with a power of 80% was found to be 294 in each of the group. Suppose the actual difference between cure rates was only 8% then the calculated power will be 61% only. Thus even if the effect size is only 8%, this RCT will have reasonable power. Table 7 shows the power associated with different effect size and sample size. By doing a power analysis we can find out whether the RCT will still have sufficient power to detect a difference / effect even if the assumptions used for calculating sample size deviate from truth to which extent.

Design effect - It is the value of the inflation factor in sample size when sampling was done using a method other than simple random sampling (SRS). It is the ratio of variability of a character found out by a random method other than simple random sampling to variability of the character if sampling is done by SRS. The formula for design effect is given in table 8. For details regarding intraclass correlation see the articles by bennett et al[7].

If we want to estimate a character with a precision similar to that estimated by SRS then we have to have more sample size in the other random sampling method by a fraction of design effect (DE). Design effect need to be applied in all sampling procedure that vary from SRS like systematic random sampling, sampling in stages, stratified random sampling, cluster sampling and also in cluster randomized trials.

Effect of lost to follow up or response rate in sample size - Power will be reduced if all the proposed participants didn't turnout for the study or if the patients dropout during the trial which effectively reduce the sample size. In order to account for this loss in power, sample size should be increased from calculated value. The formula is given in table 9.

Table 7

Power analysis results based on varying effect size and sample size

Actual sample size = 294 in each		Effect size =10%	
Expected effect size	Power	Actual sample size reached	Power
10%	80%	294	80%
9%	71%	275	77%
8%	61%	250	73%
7%	50%	225	69%
6%	39%	200	64%
5%	28%	175	58%

Table 8

Formula for design effect

$$Design\ effect = \frac{Variance_{\ Methods\ other\ than\ SRS}}{Variance_{\ by\ SRS}}$$

$$Design\ effect = 1+(b\text{-}1)\ p$$

b = number of observations in a cluster

p = intraclass correlation coefficient.

Table 9

Formula for sample size corrected for the response rate and lost to follow up

$$n_* = \frac{n}{(1\text{-}r)}$$

n* = corrected sample size

n = original sample size

r = is the response rate/ non participation rate as fractions

Other issues - All the above calculations are for the two tailed, parametric and unmatched/ unpaired studies. If a one tailed test is used, Z values should be modified to suit the one tailed situation. Nonparametric tests in generally have a power of 70-90 of the parametric test. In order to account for this loss in power, sample size have to be inflated by 10-30% from the calculated values. There are specific sample size formulas for nonparametric tests[8], which can also be used for the calculation of the study group size. Paired studies/ Matched studies require lower sample size as in case of crossover trail. Specialized clinical trials like non inferiority designs and equivalence design will require specific formulas for study size calculations[9].

Standalone software's and websites with applets for sample size calculation make the life of an average researcher very easy. They provide a easy alternative to the manual calculation of the sample size. Some of the commonly used software's used for calculating sample size and their abilities are listed in table no 10 and discussed below.

StatCalc - It comes as part of the bundled software Epiinfo 7 created by Centers for Disease Control and prevention (CDC). It has options for sample size calculation for case control studies, cohort studies, RCT, Cross sectional comparative studies and descriptive studies for qualitative variable with or without finite sample size correction. Epiinfo is available free of cost and is useful for data entry and analysis of epidemiological data and is available from the CDC website http://wwwn.cdc.gov/epiinfo/7/index.htm . It works only on a Windows platform.

WinPepi - It is a free of cost software developed by Joe Abramson and available at http:// www.brixtonhealth.com/pepi4windows.html . It is useful for sample size calculation of descriptive studies and can deal with both qualitative variables and quantitative variables. It can also calculate sample size for cluster sampling. It works only on a Windows platform. In addition to sample size it has other useful programs for random sampling and randomization.

Table10

Statistical software used for sample size calculation

software	Free / Paid	Sample size calculation for									
		Estimation of proportion	Estimation of Mean	Comparison of proportion	Comparison of mean	Correlation coefficient	Diagnostic test evaluation	Linear Regression	Logistic Regression	Survival analysis	Others
StatCalc	Free	√		√							
WinPEPI	Free	√	√								
OpenEpi	Free	√		√	√						
G*Power	Free			√	√	√	√	√	√		
Stata	Paid	√	√	√	√	√	√	√	√	√	1, 2
PASS	Paid	√	√	√	√	√		√	√	√	1, 2
nMaster	Paid	√	√	√	√	√	√	√	√	√	1, 2
nQuery	Paid	√	√	√	√	√	√	√	√	√	2

1-Sample size Reliability measurements, 2-Sample size for different types of RCT

OpenEpi -A free software available at http://www.openepi.com/v37/Menu/OE_Menu.htm . It can be accessed online at the above given URL or it can be downloaded from the above URL and will work in Windows, Mac and Linux platforms.

G*Power - It is a free of cost software developed by Franz Faul and available at http://www.gpower.hhu.de/en.html . It works in Windows and Mac platforms.

Stata - it is commercial software by StatCorp for data analysis and works in Windows, Mac and Linux platforms.

PASS - Power analysis & sample size software is commercial software by NCSS statistical software. It works only on a Windows platform.

nMaster - it is a commercial software entirely dedicated for sample size and power calculation and created by the Biostatistics Resources and Training Centre, Christian Medical College, Vellore, South India. It is extremely versatile and cheap software with good support files and references. It works only on a Windows platform.

nQuery Advisor - is a commercial software by Statistical solutions. It works only on a Windows platform.

Sample size calculation is one of the most important steps in the research conduct. Adequate sample size is the pre requisite for the validity of the results obtained in any study. Inadequately powered samples cannot answer the research question in a definite way. Conducting a study with higher than required sample size is also ethically and financially wrong. Most of the mistakes in sample size are due to the error in judgment about the assumptions. Awareness and skill regarding sample size and power calculation is essential for all biomedical researchers. This will help them in addressing the research question in a more scientific, economic and ethical manner. It is essential that all researchers give the assumptions regarding sample size calculation in their protocols and articles so that the reviewers and audience can evaluate this, as this is one of the key issues identified by multiple researchers[10-12]. An appropriate sample size along with reliable data collection methods, appropriate analysis and a clear description of these in the manuscript are the key for high quality evidence generation in science.

References

1. Gogtay NJ. Principles of sample size calculation. Indian J. Ophthalmol. 2010;58:517-8.
2. Suresh K, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. J. Hum. Reprod. Sci. 2012;5:7-13.
3. Noordzij M, Dekker FW, Zoccali C, Jager KJ. Sample size calculations. Nephron Clin. Pract. 2011;118:c319-323.
4. Dore MP, Cuccu M, Pes GM, Manca A, Graham DY. Lactobacillus reuteri in the treatment of Helicobacter pylori infection. Intern. Emerg. Med. 2013;
5. Lwanga SK, Lemeshow S, Organization WH. Sample size determination in health studies?: a practical manual [Internet]. 1991 [cited 2014 Feb 2]. Available from: http://apps.who.int/iris/handle/10665/40062
6. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol. Dial. Transplant. 2010;25:1388-93.
7. Bennett S, Woods T, Liyanage WM, Smith DL. A Simplified general method for cluster-sample surveys of health in developing countries [Internet]. 1991 [cited 2014 Feb 2]. Available from: http://apps.who.int/iris/handle/10665/47585
8. Chow S-C, Wang H, Shao J. Sample Size Calculations in Clinical Research. New York: CRC Press; 2003.
9. Julious SA. Sample Sizes for Clinical Trials. Boca Raton, FL: CRC Press; 2009.
10. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research—a review of common pitfalls. Swiss Med. Wkly. 2007;137:44-9.
11. Nyirongo VB, Mukaka MM, Kalilani-Phiri LV. Statistical pitfalls in medical research. Malawi Med. J. J. Med. Assoc. Malawi. 2008;20:15-8.
12. Carley S, Lecky F. Statistical consideration for research. Emerg. Med. J. EMJ. 2003;20:258-62.