ISSN: 2278-778X
Review Article
www.ijbio.com

THE WHAT, WHY AND HOW OF SAMPLE SIZE ESTIMATION IN CLINICAL TRIALS

Rashmi Pant[1*] and Raghupathy Anchala[1, 2]

[1]Public Health Foundation of India, Indian Institute of Public Health- Hyderabad, India

[2]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, United Kingdom

***Corresponding Author:** Dr. Rashmi Pant, PhD (Statistics), Indian Institute of Public Health, Public Health Foundation of India, Plot # 1, A N V Arcade, Amar Co-operative Society, Kavuri Hills, Madhapur, Hyderabad- 500081

Received for publication: November 21, 2012; **Accepted**: January 17, 2013.

Abstract: The choice of adequate number of subjects that would ensure accurate estimates is an important one in a clinical trial setting. Although complexity of a formula increases with complexity in study design, the basic structure of a sample size calculation formula is very simple. This paper attempts to review the basic requirements for computation of the magic number required to estimate unknown target population quantities and to test treatment differences for continuous as well as binary outcomes. Concepts are demonstrated through simple examples. Using online calculators for sample size is also discussed.

Keywords: Sample Size, Clinical Trials, Power, Treatment Effect

INTRODUCTION

The field of biostatistics is ever evolving and receives much attention in the clinical trial settings. Owing to the huge research and development spent by pharmaceutical and medical device industries, it becomes imperative that a correct sample size is estimated which will answer the research question or test out a hypothesis. During the design stages of a clinical trial, considerable effort is invested in the determination of sample size for two primary reasons: (a). Efficacy (outcome for controlled settings) and (b). Effectiveness (field or community settings) of a study to detect statistically significant results are functioning of sample size. Review studies in the past have shown that inadequate sample size can result in failure to detect therapeutic improvement by as much as 50 percent.[1] Moreover, the number of patients should be adequate enough to detect important difference (between the groups) and not be too large so as to prolong the trial.

We begin with the basic requirements for sample size computations. The first part of the discussion focuses on important considerations that determine what formula will be used. The second part focuses on trials designed to estimate certain quantities in the target population or to detect differences between treatment groups. A generic formula is given for readers to appreciate the key elements required for sample size calculations.

Basic sample size formulae required during formalization stages of a clinical trial are available in all major statistical texts. But randomized control trials (RCTs) involve experimentations on humans and are thus affected by problems typical to studies involving human subjects; such as loss to follow-up, incomplete records etc. This aspect is also touched upon in this paper.

Important considerations for estimating sample size in clinical trials:

Objective of study: Does the study deal with estimation or testing problem? The objective can be equality, non-inferiority, superiority or equivalence.

Study Design: There exist many statistical designs to achieve the set out objectives. The most common designs used are parallel group design and crossover design. Other designs include randomized control trials (RCT), cluster randomized trials (CRT), observational studies, equivalence trials, non-randomized trials, prevalence studies etc. For calculating sample size, the study design should be explicitly defined in the objective of the trial. Each design will have different approach and formula for estimating sample size [2,3].

Outcomes or end-points: The description of primary and secondary outcomes should explicitly state whether they are discrete or continuous or time-to-event variables, as sample size is estimated differently for each of these end points. Further, sample size will have to be adjusted if an outcome involves multiple comparisons. A key point that needs a careful thought at this stage of calculation involves seeking answers for will the outcome be summarized as a mean, proportion, odds ratio or relative risk. For

example, disease outcomes like dead/alive, present/absent, cured/not cured etc. are summarized as proportions or odds. Continuous outcomes like blood pressure, body mass index, test scores etc. are summarized as means.

Unit of randomization and unit of analysis: A clinical trial may be designed so as to randomize hospitals/districts to interventions rather than individuals. Hence, finding out the unit of analysis is the individual or a group or cluster (eg. communities, practices, hospitals) becomes imperative. The structure of data plays a role here; as while randomizing clusters, we may be interested in individuals as our unit of analysis and the effect of clustering on the outcomes. Say for example, we have a community intervention for malnutrition to be delivered to 'x' clusters in a district A and 'y' clusters in same district A. The effectiveness of the intervention for malnutrition may depend on many other factors such as difference in acceptability of the intervention between the communities, presence of high, middle or low socioeconomic group more in number in one community, presence or absence of motivated healthcare professionals. In such a scenario, we would have to keep in mind the effect of clustering (of other variables) on the malnutrition (outcome) per se.

Variation in expected response of treatment: One has to specify the standard deviation of the clinical parameter under consideration a priori, or at least a possible range for it. The more variation in the data, the less "precise" the study results will turn out. Accordingly, large variation may prevent studies from detecting existing clinical differences between the treatment groups. Whereas effect size also called as difference between the group results from clinical consideration, it would be best to derive information on the standard deviation either from similar studies in literature or from an internal pilot study. The information about expected response is usually obtained from previous trials done on the test drug. If this information is not available, it could be obtained from previous published literature or could be hypothesized by clinical experts (a scientific educated guess).

Basic terminology for estimation of sample size:

Often we are interested in estimating an unknown quantity, called a parameter, (X) of the target population. We would require a sample size (n), which would best estimate this unknown parameter. A generic problem would be to obtain estimates of X based on sample quantities, x, called statistics. Table 1 gives a list of parameters and corresponding statistics. Particular values of x are called estimates. These are of two types: point and interval estimates. A point estimate is a single value obtained from the sample values whereas an interval estimate is a range of values. We will concern ourselves with the latter. Hence we would be interested to answer: How many subjects should one recruit to be p% sure that estimate obtained will lie between a and b.

An interval estimate may or may not contain the value of the parameter being estimated. In an interval estimate, the parameter is specified as being between two values. For example, an interval estimate for the average age (μ) of all patients might be 26.9<μ<27.7, or 27.3±0.4 years. Either the interval contains the parameter or it does not. Hence a degree of confidence (usually a percent) needs to be assigned before an interval estimate is made. For instance, one may wish to be 95% confident that the interval contains the true population mean age. This percentage is called a *confidence level*. The confidence level of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter. The interval within which the true value of the population parameter lies is called a *confidence interval*. A confidence interval is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate. Hence in the preceding situation, if the 95% confidence interval for the mean age of patients is stated as (26.9.27.7), this means that there is a 95% chance that the true target population means age will lie between these two values.

A brief intuitive explanation is given below:

The sample of size *n* that we choose to obtain estimate of our unknown parameter is one among many possible samples that can be drawn from the target population. If we could replicate our study with a sample size of *n* such that we get all possible samples of this size from our target population, an empirical rule states that, when the *n* is large, approximately p% of the estimated values (e.g means or proportions) will fall within *k* standard errors (s.e) of the population value, that is,

$$Prob[a < parameter < b] \leq p/100$$

Where

$$a = estimate - k \times s.e(statistic) \text{ and}$$

$$b = estimate + k \times s.e(statistic) \quad \ldots(1)$$

The s. e is a measure of *variation of sample values from the population value*. The standard error formulae for various statistics are provided in table 1. Table 2 provides values of *k* corresponding to different values of *p*.

Here the interpretation of the *p*% confidence interval then is, in approximately *p* out of 100 samples, one is likely to get estimates of parameter between *a* and *b*.

A simple mathematical manipulation of (1) tells us that the minimum sample size required estimating a mean with *p%* confidence level is

$$n = \left(k * \frac{\sigma}{E}\right)^2 \quad \ldots(2)$$

E is the maximum error of estimate and is defined as the maximum likely difference between the point estimate of a parameter and the actual value of the parameter.

Examples:

1. A health care professional wishes to estimate the birth weights of infants. How large a sample must she select if she desires to be 90% confident that the true mean is within 6 ounces of the sample mean? The standard deviation of the birth weights is known to be 8 ounces.
2. *Solution:* From table 2, *k*=1.28 and required sample size from (2) is at least

$$n = \left(1.28 * \frac{8}{6}\right)^2 = 4$$

3. A researcher is trying to estimate the average number of sick days that full-time food service workers use per year. A pilot study found the standard deviation to be 2.5 days. How large a sample must be selected if the researcher wants to be 95% confident of getting an interval that contains the true mean with a maximum error of 1 day?

Solution: from table 2, k= *k*=1.96 and required sample size from (2) is at least

$$n = \left(1.96 * \frac{2.5}{1}\right)^2 = 24$$ (rounding off the decimals)

Withdrawals, missing data and losses to follow up:

Any sample size calculation is based on the total number of subjects who are needed in the final study. In practice, eligible subjects will not always be willing to take part and it will be necessary to approach more subjects than are needed in the first instance. Subjects may fail or refuse to give valid responses to particular questions, physical measurements may suffer from technical problems, and in studies involving follow up (e.g. trials or cohort studies) there will always be some degree of attrition. It may therefore be necessary to calculate the number of subjects that need to be approached in order to achieve the final desired sample size. For example to adjust the sample size for the anticipated loss to follow-up rate: Suppose *n* is the total number of subjects in each group of a treatment/placebo trial not accounting for loss to follow-up, and L is the loss to follow-up rate, then the adjusted sample size is given by

$$n^* = \frac{n}{1 - L}$$

Table.1: Standard error formulae

Parameter	Statistic	Standard error
Population proportion (P)	Sample proportion (*p*)	$\sqrt{p * (1 - p)/n}$
Difference of proportions (P_1-P_2)	p_1-p_2	$\sqrt{\frac{p_1*(1-p_1)}{n_1} + \frac{p_2*(1}{n_}}$
Population mean (μ)	Sample mean $\bar{x}$	$\frac{\sigma}{\sqrt{n}}$, σ-*population standard deviation*
Difference of means (μ_1- μ_1)	$\bar{x}_1 - \bar{x}_2$ $\bar{x}_1$-*mean of group 1* $\bar{x}_2 - mean\ of\ group\ 2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, $s_1^2, s_1^2 - sample$
Difference of paired sample means (D)	d=difference of paired observations (pre-post test)	$\frac{\Sigma(d_i - \bar{d})^2}{(n-1)}$, d-mean of paired differences

Table.2: Confidence level (p%) and values of k used in equation (1)

Confidence level (p%)	K (s.e)
99.90%	3.29
99.50%	2.61
99%	2.58
97.50%	2.24
95%	1.96
90%	1.28
80%	1.04
70%	1.04

Sample size determination for hypothesis testing:

Another situation which we encounter in clinical trials is reflected in the following claim statements (called hypotheses):

Statement 1: Intervention A is as effective as Intervention B. (Null Hypothesis -H_0)

Statement 2: Intervention A is not equal to Intervention B i.e the effect of intervention ((Alternate Hypothesis-H_1 – two sided)

OR

A can be inferior to effect of intervention B or vice versa. (Alternate Hypothesis-H_1 – one sided)

This is an important parameter needed for sample size estimation, which translates the objective of a study into statistical terminology. Statement 2 can be made two-sided also in the case of equality and equivalence trials where as non-inferiority and superiority trials have one-sided statements. Sample

size required to demonstrate equivalence is highest and to demonstrate equality is lowest. One generally starts with the assumption that statement 1 or our null hypothesis is true. The evidence obtained from the sample is then evaluated to determine how likely it is to observe a result similar to or greater than the one observed if the null hypothesis is actually true (in the population). Since inferences about a target group will be based on a sample, there is always the chance of making an erroneous inference. (Table.3)

Table.3: Two by two table indicating error in inference (when we infer results for a whole population based on a sample from that population)

		Truth (population)	
		Treatment benefit	**No treatment benefit**
Clinical Trial Result (in our samples)	Treatment benefit	Correct inference	Type I error (False positive)
	No treatment benefit	Type II error (False negative)	Correct inference

The type I error (α) happens when sample evidence indicates that a treatment is beneficial when it may actually not be true in the population. The type II error (β) occurs when sample evidence does not detect treatment benefit when it is actually there. Ideally, one requires a sample size which generates enough evidence to detect a treatment effect when there is one. Hence, the number of patients enrolled in a study has a large bearing on the ability of the study to reliably detect the size of the effect of the study intervention. This ability is known as the power (1-β) of the trial. The larger the sample size or number of participants in the trial, the greater the statistical power.

The power of a trial is not a single, unique value; it estimates the ability of a trial to detect a difference of a particular size (or larger) between the treated (tested drug/device) and control (placebo or standard treatment) groups. For example, a trial of a weight increasing drug versus placebo with 500 patients in each group might have a power of 0.90 to detect a difference of 6 kilograms or more between patients receiving study drug and patients receiving placebo, but only have a power of 0.70 to detect a difference of 3 kilograms (Note that when large differences between groups exist, power and sample size required will be less to detect this large intergroup difference. Keeping in mind an analogy that huge effort is required to find a needle in haystack, remember that to detect a significantly less difference between the groups, a higher power and a higher sample size would be required.

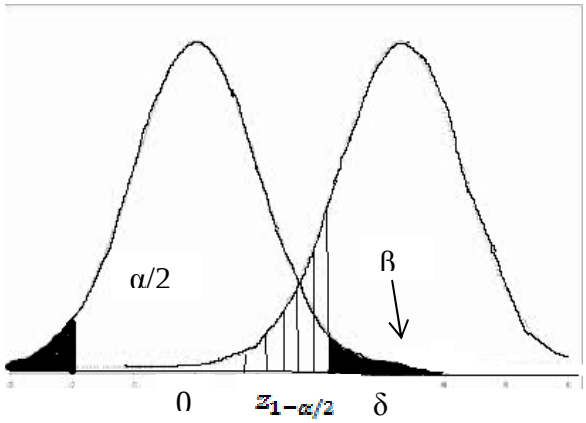

Figure.1 illustrates graphically the ingredients in sample size calculation required to detect treatment effect.

The left hand curve in the above figure shows the distribution of treatment effect if the null hypothesis is true. The right hand curve shows the distribution of treatment effect under the alternative hypothesis. For example, two drugs with different means are overlapping – some portion of A on its right tail and some portion of B on its left tail. Effect size = [(mean of group 1 – mean of group 2) / standard deviation]

Empirical rule:

The basic sample size formula for testing the difference between two treatment groups, with 80% power with a two-sided statement 2 (alternative), gaussian (bell shaped or normal) distribution of mean differences in treatments with equal variances (σ^2) and equal sample sizes (n) in both groups is:

$$n = \frac{16}{\Delta^2} \qquad \ldots(3)$$

Where

$$\Delta = \frac{\delta}{\sigma}$$

is the treatment difference to be detected in units of the standard deviation, the standardized difference (because we are calculating 'difference between the groups in any units' divided by a standard deviation)

Table 4 gives the values of the numerator in equation (3) for different power values.

Table.4: Numerator for Sample Size Formula, Equation (3); Two sided Alternative Hypothesis, Type I Error, α= 0:05, $z_{1-\alpha/2}$=1.96

Type II error	Power=1-Type II error	Numerator in equation (3)		$z_{1-\beta}$
		One sample	Two sample	
0.50	0.50	4	8	0
0.20	0.80	8	16	0.84
0.10	0.90	11	22	1.28
0.05	0.95	13	26	1.64
0.025	0.975	16	32	1.92

Example: Sample size required to detect a standardized treatment difference of 0.5 with 80% power for a two sided test is $n = \frac{16}{0.5^2} = 64$ in each group.

1.1. Exact Rule

Difference in treatment means:

Under the assumption that the difference in treatment means has a normal distribution in the population, the sample size is given by:

$$n = \frac{K(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{\Delta^2} \quad \ldots..(4)$$

Where the values of K can be obtained from table 2 and $z_{1-\frac{\alpha}{2}}$, $z_{1-\beta}$ are available in table 4.

5.2.1 *Difference in sample proportions*

$$n = \frac{2\left[z_{1-\frac{\alpha}{2}}\sqrt{\bar{p}(1-\bar{p})} + z_{1-\beta}\sqrt{p_1(1-p_1)+p_1(1-p_1)}\right]^2}{D^2} \ldots.(5)$$

Where

D=$p_{1,} - p_2$; $\bar{p}$ is pooled proportion

p_1, p_2-prevalence of disease or proportion of subjects possessing particular characteristic in two treatment groups

Examples:

i. Suppose a standard diagnostic procedure has an accuracy of 80% for the diagnosis of a certain disease. A study is proposed to evaluate a new diagnostic procedure that may have greater accuracy. On the basis of their experience, the investigators decide that the new procedure would have to be at least 90% accurate to be considered significantly better than the standard procedure. A significance criterion of 0.05 and a power of 0.90 are chosen. With these assumptions, p_1 = 0.80, p_2 = 0.90, D = 0.10, $\bar{p}$ = 0.85, $z_{1-\frac{\alpha}{2}}$= 1.960, and $z_{1-\beta}$ = 0.842. Equation (5) yields a sample size of N = 398. Therefore, a total of 398 patients should be enrolled: 199 to undergo the standard diagnostic procedure and 199 to undergo the new one.

ii. A randomized controlled trial has been planned to evaluate a brief psychological intervention in comparison to usual treatment in the reduction of suicidal ideation amongst patients presenting at hospital. Suicidal ideation will be measured on the Beck scale[5] (i.e primary outcome is continuous); the standard deviation (σ) of this scale in a previous study was 7.7, and a difference of Δ=5 points is considered to be of clinical importance. It is anticipated that around one third of patients may drop out of treatment.[4] Using equation (4) and relevant values from table 2 and 4, would give a sample size of 38 subjects per group. To allow for the predicted dropout rate of around one third, the sample size is increased to 60 in each group, a total sample of 120.

Online Calculators:

A plethora of online calculators are available for power and sample size calculation. In using these calculators, one must have a fair understanding of the various input values required. Open Epi (www.openepi.com/) is well-known open source software used by trialists, epidemiologists and other researchers in life sciences for power and sample size calculations.

CONCLUSION

The sample size is one of the critical steps in planning a clinical trials and any negligence in its estimation may lead to rejection of an efficacious drug or a device, and an approval of an ineffective drug or device. . A valid calculation of a sample size depends on clarity and interpretation of statistical concepts such as effect size, standard error, standard deviation, standardized mean difference, expected difference in outcomes, null and alternate hypothesis, random error (type 1 error) and type II errors and power. This paper is an attempt to simplify the theory and bring in a critical analytic ability on the 'how, why and what of sample size calculations in clinical trials' for clinicians, research staff, non-clinicians, people working in the clinical trial industry, principal investigators and sponsors. We encourage an early consultation (during the design phase itself) with an experienced statistician in estimation of sample sizes for different scenarios, as it has a strong bearing on the methodological rigor and scientific quality of a study result.

8. Salient features

1. The outcome variables (primary end points) should determine a sample size calculation
2. Power and type I error should be specified a priori. It is a good practice to show sample size calculations for different power (0.8, 0.9, 0.95) and type I error (alpha of 0.05 or 0.10) scenarios.
3. The expected treatment differences must be known or hypothesized (based on clinical judgment) for a correct sample size estimation
4. Careful thought must go in framing a null and alternate hypothesis (equal and not equal are two sided, less than or more than are one sided)
5. The type I error (α) happens when sample evidence indicates that a treatment is beneficial when it may actually not be true in the population.

6. The type II error (β) occurs when sample evidence does not detect treatment benefit when it is actually there.
7. The power of a trial estimates the ability of a trial to detect a difference of a particular size between the treated (tested drug/device) and control (placebo or standard treatment) groups

REFERENCES

1. Sakpal TV, Sample Size Estimation in Clinical Trials. Perspectives in Clinical Research 2010; 1(2): 67-69.

2. Shein-Chung Chow, Shao J & Wang H, Sample size calculation in clinical trials. Chapman & Hall/CRC; 2008.

3. Marcello RP, Gauvreau K, Principles of Biostatistics. Belmont, CA: Duxbury; 1993. P.214-220.

4. Guthrie E, Kapur N, Mackway-Jones K, Chew-Graham C, Moorey J, and Mendel E et al. Randomised controlled trial of brief psychological intervention after deliberate self-poisoning. BMJ 2001; 323(7305):135-138.

5. Beck AT, Steer RA, Kovacs M, Garrison B. Hopelessness and eventual suicide: a 10-year prospective study of patients hospitalized with suicidal ideation. Am J Psychiatry 1985; 142 (5): 559–563.

Source of support: Nil
Conflict of interest: None Declared