Print ISSN: 2319-2003 | Online ISSN: 2279-0780

IJBCP International Journal of Basic & Clinical Pharmacology

doi: 10.5455/2319-2003.ijbcp005112

Educational Forum

Sample size: from formulae to concepts

Rakesh R. Pathak*

Department of Pharmacology,
C.U. Shah Medical College,
Surendranagar 363001, India

Received: 21 November 2012
Accepted: 24 November 2012

***Correspondence to:**
Dr. Rakesh R. Pathak,
Email: rr_pathak@yahoo.com

ABSTRACT

Sample size can be calculated from many online calculators or tables. But the use of these instruments is rational only when we understand our input data and the concept behind them completely. Terminologies like confidence interval, confidence limit, standard error of mean, margin of error, standard normal variate, power, significance level etc. and extent to which population size or chances of occurrence of an outcome can affect our sample size remain to be well understood before using these software solutions.

Keywords: Confidence interval, Confidence limit, Standard error of mean, Margin of error, Standard normal variate, Power

In last two issues of this journal we saw how we can find sample size using readymade tables and sample size calculators available online.[1,2] We also saw that the permitted level and ‘emphasized type’ of error also decide the sample size.[2] Importance of the type of error also changes as per experimental setting.[2] Thus we can decide small or large sample sizes as per our requirements and resources.[2]

Had the sample size calculation been as easy, why we get alarmingly complex formulae for calculations? And why the calculation of sample size is such a “sizeable” problem? Mostly because these formulae are the basis of the aforementioned tables and sample-size-calculators. It remains to be understood that how various parameters can affect our sample size before using these software solutions.

Now if we have no access online or our starting information doesn’t exactly fit into the table or calculator protocol, we need to understand and use these formulae manually. Some PG/ Ph. D students need to understand these formulae to satisfy their guide’s methodist demands or as a finer preparation for the viva on their thesis/ dissertation.

Let’s have a look. If we wish to be 95% certain in an experiment (i.e. there is only 5% chance remaining that “the finding is wrong” – this is called confidence level) with confidence interval of 1% around the true value, then what should be the sample size? This type of estimate is used in sampling for quality assurance (say, measuring thickness of gauze wire, for which margin of error is pre-decided to maintain the quality).

In other words, we can say we wish to find the values to the nearest 1 per 100 (= 1%) – this 1% or 1 per 100 is called confidence interval which is often confused with confidence level which is 95% as said above. Here the standard error is $[p*(1-p)/n]^{1/2}$ and multiplied by normal standard variate (which is 1.96 for 95% confidence level that we can get from any statistical table) the same becomes margin of error.

In this problem, percentage chances of positive (or negative) outcome is not given and would be supposed to be 50% (i.e. equal chances of an outcome being positive or negative, or saying it otherwise, being in or beyond the range of standard wire thickness). Value of p*(100 - p) or its square root value $[p*(1-p)/n]^{1/2}$ is maximal when chances of getting or not getting the expected outcome is 50-50 and would be 50*(100 - 50) = 2500.

If the ‘chances’ are less or more, the value would be < 2500. For example if chances of getting a value within range are 25%, 25*(100 – 25) = 25*75 = 1875 would be the sample size (when failures to comply by quality are more expected).

The same value we get if the chances of getting a value within range are 75%, viz 75 (100 – 75) = 75*25 = 1875 (when successes to comply by quality are more expected). In the given case the chances are supposed to be 50% and thus p*(100 - p) is 2500, the sample size ‘n’ would be decided by putting the pre-decided value of margin of error which is 1% is the given case.

Allowed maximum margin of error is pre-decided as a measure of quality control and if samples fall out of this error range, retrospective improvement is done. This method is used by Deming cycle of PDCA (plan-do-check-act) in Kaizen concept of TQM (total quality management).[3] In the given case, margin of error being 1, normal standard deviate being 1.96, p being 50, 'n'= sample size would be 9600 - any smaller sample can't "convincingly (at 95% confidence level)" ascertain the quality.

Population size, sample size and percentage chances of positive finding – all these three factors affect the confidence interval.[4] It has been statistically shown that effect of "population size increase beyond 20,000" is very less and hence in online sample size calculators, if total population size is unknown, we fit this "20,000" as population size.

Maximum margin of error is a half of the "confidence interval" which is symmetrically distributed on either side of the true value. Thus 1% confidence interval indicates a maximum margin of error = ± 0.5%. Extreme values at the start or end of confidence interval are called confidence limits ("true value + 0.5%" and "true value – 0.5%" is the case above).

A similar question arises when we wish to estimate a disease prevalence in a given population for which we have previously estimated prevalence in other setting like different time period or different locality (for example, roughly 2.5% is the prevalence of G6PD deficiency in Indian population[5]).

Margin of error in such cases is given by [standard normal deviate for the given confidence level of 95%]*[p*(100 – p)/n]$^{1/2}$. Notably [p*(100 – p)/n]$^{1/2}$ is the square root of [p*(100 – p)/n] where 'p' = the chances of getting a value positive or negative. If 'p' is not given, as in the first case, consider it as 50 and we get the maximum size of required sample.

In this case of disease prevalence, where we have a previously suspected 'p' of 2% (prevalence of the disease is 2%), we can use it and get a much lower value of [p*(100 – p)]. The same value of [p*(100 – p)] we would get if we calculate sample size for estimating disease free population i.e. people with normal level of G6PD (i.e. when chances of a person turning out negative for the disease is 98%).

We can get standard normal deviate from any statistical table for the given confidence level (here the confidence level is 95% and this is the most commonly used one) – for two sided P(x), we look for "100% – 95% = 5%" or for one sided P(x) we have to look for 2.5 % – as 5% is symmetrically divided in two halves of 2.5% (in either case, it is 1.96 from the table).[6]

When we compare two means to show that their respective samples and thus respective populations to which these sample are representing differ significantly, we use a different formula: $(\mu_1 - \mu_2) = f(\alpha, P)* \sigma^2*[(1/n_1) +(1/n_2)]$ where μ_1 and μ_2 are the means from two samples and n_1 and n_2 are their respective sample sizes.

The factor f (α, P) is taken from table for a given significance level (most commonly used is $\alpha = 0.05$) and given power (most commonly used power is 0.90 or 0.95) and the value of f ($\alpha = 0.05$, $P = 0.95$) is 15.2 from the table.[7] Standard deviation σ is mostly based on a prior study or a pilot study done at minor level before proper large scale study.

The probability that a test will produce a significant difference at a given significance level is called the power of the test. The power is not zero even if the population difference is zero because there is always a possibility of significant difference even when null hypothesis is true.[8] Conventionally, significant difference is there if the probability of similarity is < 0.05 – thus probability of similarity (P value) is low when significance of difference is high.[9]

If we plan equal size samples and the samples are designed to detect a difference of 0.1 units and the standard deviation is 2.5, the number required in each sample would be 19,000 and a large multicentric clinical trial can safely plan 20,000 subjects in each group. For a smaller size sample, either disease prevalence rate (or probability of occurrence) must be lower still or a higher standard error should be permitted.

REFERENCES

1. Pathak RR. Small size sampling. Int J Basic Clin Pharmacol 2012;1(1):43-4.
2. Pathak RR. Small size sampling? Int J Basic Clin Pharmacol 2012;1(2):118-9.
3. Charantimath PM. Total Quality Management, 3rd Edition. Pearson Education, 2009:180.
4. Bland M. An introduction to medical statistics, Second Edition. Oxford University Press, Great Britain, 1995:110.
5. Glader BE. Glucose-6-phosphate dehydrogenase deficiency and related disorders of hexose monophosphate shunt and glutathione metabolism. In: Wintrobe's Clinical Hematology, 10th ed, Lee GR, Foerster J, Lukens J, et al. (Eds), Baltimore, Williams & Wilkins; 1999:1178.
6. Bland M. An introduction to medical statistics, Second Edition. Oxford University Press, Great Britain, 1995:109.
7. Bland M. An introduction to medical statistics, Second Edition. Oxford University Press, Great Britain, 1995:334.
8. Bland M. An introduction to medical statistics, Second Edition. Oxford University Press, Great Britain, 1995:143-4.
9. Bland M. An introduction to medical statistics, Second Edition. Oxford University Press, Great Britain, 1995:137.

doi: 10.5455/2319-2003.ijbcp005112
Cite this article as: Pathak RR. Sample size: from formulae to concepts. Int J Basic Clin Pharmacol 2012;1:223-4.