Sample size calculation – continuous outcome variable

Jianrong Wu PhD

We are planning to conduct a phase III randomized trial to evaluate the difference between daily short infusions and continuous infusions of a chemotherapy drug on the concentration of an active compound in the blood. We understand randomization is important in conducting such a trial. How do we determine the number of patients to be recruited?

Sample size/power calculations are a very important aspect of randomized clinical trials and should be performed during the design phase of a study. Incorrect sample size calculation alone sometimes can cause the failure of a trial, including making incorrect conclusions and providing false information for clinical practice.

In general, sample size of a trial should be appropriate for answering the research question. If a sample size of a trial is too small, then it might not be able to detect a difference of interest; on the other hand, if the sample size is too large, then the study will take longer and incur greater costs, and it might detect a difference that has no clinical significance. An extreme situation of the latter is to recruit all patients with a certain disease in a trial, so that the entire patient population can be studied. However, such an idea is quite problematic in many aspects, including 1) difficulty in patient recruitment; 2) a longer time to complete the trial; 3) increased costs; and 4) challenges in data analysis.

In this article, we will discuss some issues associated with sample size calculation. Formulas and considerations for sample size calculation differ for different outcome variable types. In this issue, we will focus on situations in which the outcome is a continuous variable, e.g. *drug concentration in the blood.*

Corresponding author: **Jianrong Wu**
Contact Information: Jianrong.wu@uky.edu
DOI: 10.12746/swrccc.v6i25.487

1. Statistical power and effect size

The majority of clinical trials are designed to answer a specific research question. For example, a study might compare two groups and have both null and alternative hypotheses. In general, the null hypothesis states that there is no difference between groups, and the alternative states that there is a significant difference. Based on such a hypothesis, we can define type I and type II errors as follows:

A type I error is the probability of rejecting the null hypothesis (no difference) when the null hypothesis is true, while a type II error is the probability of not rejecting the null hypothesis when the null hypothesis is not true. The statistical power is the complement of type II error, i.e., rejection of the null hypothesis when the null hypothesis is not true. In other words, statistical power is the probability of a test identifying a difference when such a difference truly exists.

Due to random sampling, type I and type II errors are unavoidable in statistical tests, and the error rates that are acceptable need to be pre-specified in sample size calculation. In general, the type I error rate is often set at 0.05, meaning that 1 out 20 trials will potentially make an incorrect conclusion that a difference exists when the truth is that there is no difference. The statistical power is often set at 80%, meaning that, if there is a true difference, then 80% of the time this difference will be detected by such a trial.

Effect size, which is the standardized mean difference between two groups, is another important piece of information needed for sample size calculation:

$$Effect\ size = \frac{Difference\ in\ mean}{Standard\ deviation}.$$

The numerator of effect size is the difference in the mean between the two groups to be compared. This value is usually provided by the investigator based on clinical significance. The denominator is the pooled standard deviation of the two groups calculated using preliminary data. This standard deviation can also be obtained from literature if preliminary data are not available. It is intuitive that effect size is negatively associated with sample size, i.e., if the difference between two groups is small, then a large sample size is needed to detect the difference. Otherwise, a small sample size is needed.

2. One or two-sided test

The choice of using a one-sided or a two-sided test for hypothesis testing depends on the objective of a trial. For example, if the goal is to demonstrate that a treatment is better than placebo, then a one-sided test is appropriate; if two treatments are to be compared, then a two-sided test is more appropriate. Given that all other conditions remain the same, the sample size required for a one-sided test is smaller than that required for a two-sided test. The decision to use a one-sided test is usually based on additional known information and thus compared with using a two-sided test, less new information (smaller sample size) is needed to reach a conclusion.

3. Sample size calculation and assumptions of the corresponding statistical tests

Assumptions made for different statistical tests are often different, and sample sizes calculated for different tests often differ, even with the same data. The derivation of the calculation formula is available in most statistical textbooks, and we will provide only the formula to demonstrate the differences.

a) Sample size calculation for a two sample Z test

Two common assumptions for the sample size calculation of a Z test are that the outcome variable follows a normal distribution and has same variance (known) of the two groups. Then, given a type I error of α and a type II error of β, the sample size of the first group for a two-sided Z test is given as follows:

$$n_1 = \frac{(r+1)(Z_{1-\alpha/2} + z_{1-\beta})^2}{r\Delta^2},$$

where z_γ is the γ^{th} percentile of the standard normal distribution, r is the ratio of sample size of group 1 vs. group 2, and $\Delta = \frac{\mu_1 - \mu_2}{\sigma}$ is the effect size. When the common standard deviation σ is unknown and has to be estimated from data, a two-sample t test can be used to derive the sample size calculation. This is more complicated and will not be discussed in this article.

b) Sample size calculation for a non-parametric test

Comparisons between two groups can also be made by using a Mann-Whitney U test (also called Wilcoxon test), if the distribution of the outcome variable is known to differ from normal. The total sample size can be calculated by,

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{12c(1-c)(p-0.5)^2},$$

where $p = P(Y > X)/P(Y < X)$ is the odds ratio, and $c = \frac{n_1}{N}$ is the proportion of subjects in group 1.

4. Sample size calculation software

Since sample size calculation is critical for a trial, and some formulae are not intuitive, so to avoid mistakes, hand calculations are not recommended. Therefore, a number of software packages have been developed.

a) Software packages

There are several software packages dedicated for sample size calculation, such as PASS and EAST. Since most of these packages are not free, we recommend users to carefully review the user's manual and get necessary support from the developer or the user community, if needed, to ensure correct calculation. There are also generic computation software packages that can be used for sample size calculation, such as SAS, Stata, and R.

b) Online sample size calculation

A few websites can be used for certain simple sample size calculation, such as, https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html (last accessed 3/11/2018), and https://select-statistics.co.uk/calculators/sample-size-calculator-two-means/ (last accessed 3/11/2018).

5. The proposed phase III trial as an example

In the proposed Phase III trial, assume that the difference in mean concentration between the daily short infusion and continuous infusion is 3μg/l and that the standard deviations are the same for both groups (5μg/l). Also set the type I error rate at 0.05, statistical power at 80%, and allow the same numbers of subjects in both groups (r=1). Then sample size can be calculated as, $\Delta = \frac{3}{5}$

$$n_1 = \frac{(r+1)(z_{1-\alpha/2} + z_{1-\beta})^2}{r\Delta^2}$$

$$= \frac{(1+1)(1.96+0.84)^2}{\left(\frac{3}{5}\right)^2} = 43.5 \approx 44$$

Therefore, a total of 88 (44 in each group) subjects are required to achieve 80% power, at a 0.05 type I error rate, to reject the null hypothesis of equal means when the difference in mean concentration is 3μg/l with a standard deviation for both groups of 5μg/l, using a two-sided two-sample equal-variance Z-test. We also did the calculation using the EAST software; the result was the same.

There are other considerations in sample size calculation. For example, trials with long duration may require an interim analysis. The attrition rate of subjects during the long trial may change the sample size. All these considerations should be taken into account. In summary, correct sample size calculation means effective resource utilization and valid study design and thus is an inseparable component of a successful clinical study.

Keywords: sample size, study power, type I error, type II error

From: Department of Biostatistics, University of Kentucky, Lexington, KY 40514
Submitted: 4/24/2018
Accepted: 6/9/2018
Conflicts of interest: none

References

1. Noether GE. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* 1987;82:645–47.
2. Cohen J. Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York, 2nd edition, 1988.