KJA
Korean Journal of Anesthesiology

Statistical Round
pISSN 2005-6419 • eISSN 2005-7563

Considerations when calculating the sample size for an inequality test

Junyong In

Department of Anesthesiology and Pain Medicine, Dongguk University Ilsan Hospital, Goyang, Korea

Calculating the sample size is a vital step during the planning of a study in order to ensure the desired power for detecting clinically meaningful differences. However, estimating the sample size is not always straightforward. A number of key components should be considered to calculate a suitable sample size. In this paper, general considerations for conducting sample size calculations for inequality tests are summarized.

Key Words: Clinical study, Inequality test, Power, Sample size calculation.

Introduction

The proper number of subjects required to accomplish the purpose of a study should be considered during the planning stage of a clinical study. Incorrect calculation of the sample size not only wastes time and resources but also raises ethical problems. Too few subjects can fail to provide appropriate scientific values for the study results, and they may expose the subjects to potential risks without any benefit. If the number of subjects is unnecessarily large, the purpose of the study will have been accomplished even before the end of the study, and the participation of some of the subjects may have been meaningless. In other words, an appropriate sample size is very important for the validity, reliability, accuracy, integrity, and ethicality of a study. However, calculating an appropriate sample size is not always easy.

Sample size calculations are divided into the following four parts depending on the methods and procedures used. These parts are termed sample size estimation/determination, justification, adjustment, and re-estimation or interim analysis [1]. First, sample size estimation/determination is done to estimate the minimum power of a test required by a clinical study (for example, 80% power) and to calculate the sample size necessary to secure the accuracy and reliability of the target statistical values. Second, justification of a sample size is done to verify the statistical evidence or justification of the sample size that has already been determined, as the sample size can be affected by limited research funds, subjects with rare disorders, ethical issues, and other limitations of the study. Third, the sample size is adjusted for several factors, such dropouts or covariates, in order to allocate a proper number of subjects. Finally, re-estimating the sample size enables is done for greater accuracy because the sample size estimated when planning the study includes uncertainty. Generally a re-estimation is done using data collected up to a certain point in time during the study process such that sufficient power of the test can be maintained through an adjustment of Type I error. However, to perform a re-estimation, the clinical study protocol should describe in detail the method used to re-estimate the sample size [1].

The stage of study planning is initially reviewed given that the calculation of an appropriate sample size starts from that stage.

Corresponding author: Junyong In, M.D., Ph.D.
Department of Anesthesiology and Pain Medicine, Dongguk University Ilsan Hospital, 27, Dongguk-ro, Ilsandong-gu, Goyang 10326, Korea
Tel: 82-31-961-7875, Fax: 82-31-961-7864
E-mail: dragona1@dumc.or.kr
ORCID: http://orcid.org/0000-0001-7403-4287

Received: May 16, 2016.
Revised: 1st, June 13, 2016; 2nd, June 20, 2016.
Accepted: June 21, 2016.

Korean J Anesthesiol 2016 August 69(4): 327-331
http://dx.doi.org/10.4097/kjae.2016.69.4.327

Considerations during Sample Size Calculations

Considering that this article focuses on clinical studies, the following factors are assumed. The study design is critical during sample size estimations, and this article focuses on the most frequently used parallel design. The cross-over design, which is also frequently used, differs considerably from the parallel design with respect to the statistical hypotheses and assumptions used. Therefore, it is excluded from the scope of this article. Random allocation as performed during the parallel design process means that subjects are randomly allocated to each group in an independent manner; thus, the response variables may be assumed to be independent.

On the basis of the assumptions described above, the following items are needed when calculating the sample size. In addition to the present article, an article published by Lee and Kang [2] and an article entitled §314.126 Adequate and Well-Controlled Studies[1] published on the website of the US Government Publishing Office may help to clarify each item.

Study objectives

First, the objectives of the study should be clear. Study objectives undergo statistical testing after the establishment of hypotheses. In determining the study objectives, the type of hypotheses test – for example, an inequality test, non-inferiority test, equivalence test, superiority test, and bioequivalence test – should be determined simultaneously. Study objectives to which an inequality test is applied are established in most pilot studies or clinical studies. Such a study objective can be exemplified in the following sentence: 'The anesthesia induction time of the new intravenous anesthetic is different from that of the conventional intravenous anesthetic in healthy volunteers.'

Hypotheses

Hypotheses are assumptions with respect to a population to which a treatment (for example, a study drug, device, or procedure) is applied, and they are established as null hypotheses (H_0) and alternative hypotheses (H_a). A null hypothesis is 'what the researcher wants to investigate,' while an alternative hypothesis is 'what the researcher wants to show.' If the previous example is used again, the null hypothesis is 'The (average) anesthesia induction time of the new intravenous anesthetic (μ_{new}) is equal to the (average) anesthesia induction time of the conventional intravenous anesthetic (μ_{old}),' and the alternative hypothesis is 'The (average) anesthesia induction time of the new intravenous anesthetic (μ_{new}) is shorter than the (average) anesthesia induction time of the conventional intravenous anesthetic (μ_{old}).' In fact, an accurate description of the alternative hypothesis in an inequality test is 'The (average) anesthesia induction time of the new intravenous anesthetic is not equal to the (average) anesthesia induction time of the conventional intravenous anesthetic.' However, in clinical studies, even when the hypothesis is established as if in a one-sided test, a two-sided test is performed in most cases in order to maintain a more conservative point of view, because the probability that a null hypothesis is rejected is higher in a one-sided test where the significance level is set to be high or low on one side. In other words, when a new treatment which does not cause an actual difference is regarded as if it did in fact cause a difference, the result may be used clinically.

Hypotheses in a two-sided test are as follows:

$$H_0: \mu_{new} = \mu_{old} \quad \text{versus} \quad H_a: \mu_{new} \neq \mu_{old}$$

Study design

During the study design process, the allocation and treatment of the subjects are determined. There are many study design methods, but this article focuses on the parallel design, as noted above. Using the previous example again, in a parallel design, subjects are randomly allocated into different groups to which either a conventional intravenous anesthetic or a new intravenous anesthetic is injected. As a result, only one type of intravenous anesthetic is injected into a subject.

Primary endpoint

Among the variety of variables to be measured to verify the study objectives, the response variable, which is the most appropriate for the study objectives and most meaningful from a clinical perspective, is chosen as the primary endpoint, as determined by the study objectives and hypotheses. A primary endpoint uses the normal, binary, ordinal, and time-to-event data types. In the example above, if the effects of a new intravenous anesthetic and a conventional intravenous anesthetic are compared, such variables as the anesthetic induction time, the postanesthetic recovery time, the variation of the hemodynamic parameters during anesthesia, and the existence of pain during the intravenous injection may be selected as a primary endpoint. The choice of an endpoint as a primary endpoint depends on the study objectives. In the previous example, the anesthetic induction time, a normal data type, was chosen as the primary endpoint.

The number of primary endpoints may be two or more,

[1] http://www.ecfr.gov/cgi-bin/text-idx?SID=7868ff31266e298be4f23984a60b6771&mc=true&node=se21.5.314_1126&rgn=div8

though only one is used in most studies. However, given that the risk of false-positive and false-negative results for an evaluation of a treatment may increase with an increase in the number of primary endpoints, it is recommended to choose one primary endpoint appropriate to study the objectives and to calculate the sample size on the basis of that endpoint. Other endpoints apart from the primary endpoint are established as secondary or tertiary endpoints.

Type I error, type II error, and power

Rejection of a true null hypothesis is known as a Type I error (α), while not rejecting a false null hypothesis is a Type II error (β). Therefore, power refers to the probability of avoiding a Type II error (1-β), which is the probability that rejection of a null hypothesis is right when an alternative hypothesis is true.

$$\text{Power} = 1 - \beta = \text{P (reject } H_0 \text{ when } H_a \text{ is true)}$$

Consider the previous example of intravenous anesthetics. The test may give the erroneous result that the anesthetic induction time of a new intravenous anesthetic is not equal to that of a conventional intravenous anesthetic, although they are in fact equal, which is a Type I error. Another erroneous result, the finding that they are equal when they actually are not, is a Type II error. Therefore, power here refers to the probability that the null hypothesis that the anesthetic induction times of the two intravenous anesthetics are equal is tested and found to be false when the anesthetic induction times are actually not equal. Neither Type I error nor Type II error is desired, but the probability of committing one of the two types of error is increased when that of the other is decreased if the sample size is fixed. Therefore, in order to decrease the occurrences of both types of errors, the sample size should be increased.

Power analysis

In a power analysis, the sample size is affected by the level of Type II error. In addition to a power analysis, also available are a precision analysis, where the Type I error or confidence level is calculated; a probability assessment, which is applicable when the occurrence is lower; and sample size re-estimation. However, this article focuses on the power analysis.

Because Type I error is generally considered as more severe in clinical studies, an upper limit of Type I error is established as an appropriate significance level for hypothesis testing, and Type II error is minimized (or the power is maximized) through an appropriate sample size. In other words, a hypothesis test is performed on the basis of a predetermined significance level (generally 5%) and power level (generally 80 to 90%). In addition to the previously mentioned significance level and power, a power analysis requires information about the clinically significant difference (δ) and the standard deviation (s) of the primary endpoint.

Effect size[2)]

In studies to test differences, a null hypothesis is established as 'There is not a difference between the two treatment groups.' Here, the calculated difference in the means (or ratios) is an unstandardized effect size. A standardized effect size is obtained by adjusting (dividing) the difference in the means by the standard deviation. The effect size noted during the calculation of the sample size is always the standardized effect size. When prior knowledge for the calculation of the standardized effect size is not sufficient, a commonly applied effect size is 0.25–0.50, which was initially suggested by Cohen [3] and which is still important.

Clinically significant difference

The effect size is also referred to as the clinically significant difference or the minimum value of the significant difference. Unfortunately, there is not an absolute rule regarding an appropriate effect size. However, determination of the effect size should be based on statistical validity and clinical judgment [4]. An effect size may be expressed as the variation of the absolute value (e.g., a decrease in the anesthetic induction time by one minute) or the variation of the ratio (e.g., a decrease in the anesthetic induction time by 20%) of the primary endpoint.

---

2)# Scripts in R (R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/)

```
# Data from the Korean Statistical Information Service (2013, high school seniors)
# Men have a mean height (ht) of 174 cm with a standard deviation (sd) of 8 cm (n = 100)
# Women have a mean height of 161 cm with a standard deviation of 8 cm (n = 100)
# Standardized effect size (ES) is calculated using the codes below.

ht.men <- 174; sd.men <- 8; n.men <- 100
ht.women <- 161; sd.women <- 8; n.women <- 100
pooled.sd <- sqrt(((n.men - 1) * sd.men^2 + (n.women - 1) * sd.women^2) / (n.men + n.women - 2))
standardized.ES <- ((ht.men - ht.women) / pooled.sd)
round(standardized.ES, digits = 2)
```

Population variance

An estimate of the population variance[3] is another component of a sample size calculation. Because it is obvious that the quality of the variance has a significant effect on the sample size calculation [5], the following considerations should be considered when selecting the references for calculating the estimates of the population variances:

A. Study design: Is the study design of the reference similar to that of the present study? The variance of an observation study may be greater than that of a randomized controlled study. If a multi-center clinical study is planned, questions such as 'Is the reference study designed similarly to the present study?' and 'Is the time interval from the treatment to outcome measurement similar?' should be taken into account.

B. Study subjects: Are the subjects of the reference similar to the subjects of the present study? Demographical similarity is necessary. In the case of a multi-center study, it should be verified as to whether the races or nationalities of the subjects are similar. In addition, questions such as 'Do the subjects have similar diseases or severity levels? and 'Was the study conducted during the same season or period (asthma, influenza, etc.)?' should be considered.

C. Analysis: Are the analytical methods and summary statistical methods applied to the references identical to those of the present study? In addition to the application of the same analytical methods, was a covariate (e.g., a reference value of a response variable) also analyzed if there was one? Including a covariate may reduce the variance estimate and the sample size [5].

In planning a clinical study, reference information is often unavailable. However, even when the references that provide sufficient information are available, it is recommended that the study protocol should include in advance a re-estimation of the variance in order to obtain a more accurate variance estimate.

Other considerations

Potential dropouts

The dropout rate may be determined with reference to previous studies or on the basis of the experiences of the researchers. A common mistake when applying a dropout ratio is to add as many samples as the number of dropouts to the calculated sample size. For example, if the calculated sample size is 100 and the potential dropout rate is 20%, the final sample size should not be 120, but 125, because a dropout rate of 20% is expected with reference to the finally determined sample size (If the dropout rate is 20% out of 120 subjects, only 96 subjects remain, which decreases the power).

Sufficient number of available participants satisfying inclusion criteria

A sample size which is greater than the number of available participants to satisfy inclusion criteria is not useful. If necessary, the criteria for the population to which available participants belong may be changed or the sampling method or sampling location is modified.

Lasagna's Law

It is an empirical rule that the number of available participants is drastically decreased once a study begins, whereas the number recovers after the study ends. The process of collecting participants is difficult to predict even in a well-planned study; thus, the success ratio during the participant collection phase should be anticipated very carefully.

Summary

An appropriate sample size may not be calculated simply by quoting the estimated values of the mean and standard deviation from previously published reference data. When calculating an appropriate sample size, clearly defined study objectives, the study design, the hypotheses and primary endpoints adequate for the study objectives, and proper understanding and determination of the significance level and power are all important. A more accurate sample size may be calculated by deriving useful estimation values from proper data. A sample size may be calculated using a commercially available or free software program or website or even manually. However, an accurate sample size is hardly expected to arise if incorrectly selected estimation values are employed.

An approach which is more conservative than that presented here is asserted in some cases, which is that the difference between the upper confidence limit of a null hypothesis and the lower confidence limit of an alternative hypothesis (or the difference of the reverse case) should be clinically significant [6]. Another conservative approach is the recommendation that the power should set at 90%, which is higher than the 80% mark commonly applied in clinical studies. Such an increased power level reduces the incidence of Type II error by half, and a higher power level is retained even if a sufficient number of subjects are not collected. In any case, the sample size is increased and the time and cost factors are increased as well, but the results are

[3] Pooled variance $(s^2) = \dfrac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2}$

n_1: the number of subjects in group 1. n_2: the number of subjects in group 2. s_1: standard deviation of group 1. s_2: standard deviation of group 2.

more likely to become reliable.

The previously discussed considerations are also important during sample size calculations in other study designs and when using analytical methods that have not been discussed in this article, including the cross-over design, dose-response studies, nonparametric studies, and in analyses of variance with repeated measures.

References

1. Chow SC, Shao J, Wang H. Sample size calculations in clinical research. 2nd ed. Boca Raton, Champman & Hall/CRC. 2008, pp 1-23.
2. Lee S, Kang H. Statistical and methodological considerations for reporting RCTs in medical literature. Korean J Anesthesiol 2015; 68: 106-15.
3. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Abingdon-on-Thames, Routledge. 1988, pp 531-42.
4. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. Stat Med 1999; 18: 1903-42.
5. Julious SA. Sample sizes for clinical trials with normal data. Stat Med 2004; 23: 1921-86.
6. Jia B, Lynn HS. A sample size planning approach that considers both statistical significance and clinical significance. Trials 2015; 16: 213.