

Paper SP05

Power Analysis and Sample Size Estimation using Bootstrap

Xiaomei Peng, Guangbin Peng, Celedon Gonzales

Eli Lilly and Company, Indianapolis, IN

Abstract

Power analysis and sample size estimation are critical steps in the design of clinical trials. Usually, these tasks can be accomplished by a statistician by using estimates of the treatment effect and sample variance from past trials or expert opinion. However, when exact power computations or reasonable approximations are not possible, or when there is no method to estimate the effect size or variability of clinical data, we have to adopt the simulation-based approach. Bootstrap provides a powerful tool to perform the task by directly sampling from existing data. It is especially useful when the study to be designed employs co-primary outcome measurements, or applies special analysis such as stratified Wilcoxon test where available software and traditional approaches are not applicable. In this paper, the bootstrap program was used to perform the power analysis and sample size estimation, and illustrate their application in two clinical trial designs.

Key Words: Simulation, Statistical power, Sample size, Bootstrap

Introduction

Power analysis and sample size estimation are critical step at the design phase of the clinical trial development. Generally statisticians choose a sample size large enough to enhance chances of conclusive results while small enough to lower the study cost, constrained by limited budget and/or some medical considerations.

Current available methods for power analysis include paired and pooled t-test, fixed-effect ANOVA and regression models, binomial proportion comparisons, bioequivalence, correlation, and simple survival analysis models. Numerous mathematical formulas have been developed to calculate sample size for various scenarios in clinical research based on different research objectives, designs, data analysis methods, power, type I and type II errors, variability and effect size (Chow, Shao and Wang, 2003). Some methods have also been incorporated into SAS applications.

One of the most important prerequisite to decide before calculating a sample size is to define a clinically important treatment effect or effects (δ). δ is usually estimated based on knowledge gained from the past trials. Although the estimations of δ may be preliminary, those estimates can be entered into SAS to compute the sample sizes for the subsequent trials.

While traditional sample size computation methods may be appropriate to control the probabilities of type I and type II errors for the evaluation of each individual outcome variable, these methods are not appropriate for the simultaneous evaluation of co-primary outcome measurements without some type of correction. Simulation-based power

analysis provides a viable solution for the before mentioned problems. It can also be used when neither exact power computations nor reasonable approximation are possible, or when there is no means to estimate the effect size or variability of clinical data.

One of the widely used simulation techniques is bootstrap. This permutation-based approach has been gaining popularity with the arrival of high-speed modern computer. It directly re-samples the existing data without any assumption about the underlying distribution of the sampled population. It is applicable to a wide range of analysis scenarios. Also, we can easily incorporate complex designs and special statistical tests as well as co-primary outcome measurements into the bootstrap procedure to get more accurate estimation of power and sample size.

In this paper, we present two cases of power analysis using bootstrap technique. In the first case, we specify values of model parameters and use them to randomly generate a large number of hypothetical datasets and then conduct bootstrap re-sampling. In the second case we directly use bootstrap technique to re-sample the existing data without taking any assumption about the underlying distribution of the sampled population. In both cases, we apply the statistical test to each data set generated, and estimate power as the percentage of times the null hypothesis is rejected.

Sample Size Calculation for Multiple Outcomes in Two Group Comparisons

In clinical studies, it is not uncommon to utilize multiple effects or co-primary outcome variables that are associated with a single treatment. For example, a drug for Alzheimer's disease may improve both cognitive abilities, as assessed by the Alzheimer's disease Assessment Scale-Cognitive Subscale (ADAS-Cog), and global behavior function, as assessed by the Clinician's Interview-Based Impression of Change with Caregiver Input (CIBIC-Plus). The sample size needs to be determined to provide 80% dual outcome power, where dual outcome power refers to the probability of observing a significant drug versus placebo comparison with respect to both primary efficacy variables (Cummings, 1998).

A. Sample Data Generation

The first step is to generate a large number of sample data sets for a given study. The second step is to estimate the power by analyzing each data set of the prescribed same size and computing the percentage of times the null hypothesis is rejected. The minimum size with an estimated power that satisfies the given requirements will then be used as the sample size in the study plan.

To compare multiple outcomes between treatment group and placebo group, we need to generate sample data for each outcome variable of each group. Null hypothesis will not be rejected if p-value of either outcome variable is equal to or larger than the given alpha value.

SAS provides a function (ranuni) to generate pseudo-random decimal numbers between 0 and 1. It also provides a function (normal) to generate random data with a normal

distribution based on given population mean and standard deviation. The second function can be used directly to generate sample data set for continuous variables. For categorical variables, we use the following procedure to generate sample data based the SAS function ranuni:

1. Given a probability distribution of a categorical variable, calculate its cumulative distribution.
2. Generate a given number of psuedo-random numbers using SAS function.
3. Map generated pseudo-random numbers to categorical values by associating the quantile to that random number.
4. Repeat step 2 to 3 for the each outcome variable with and without treatment, and compute p-value based on generated sample data.
5. Increase sample size and repeat step 2 to 4 if needed.

For example,

Categorical Values	C1	C2	C3	C4	C5
Probability	0.1	0.2	0.15	0.25	0.3
Cumulative Probability	0.1	0.3	0.45	0.7	1

Generated Random Numbers	0.12	0.46	0.34	0.23	0.61	0.78
Associated Categorical Values	C2	C4	C3	C2	C4	C5

B. Bootstrap Simulation

SAS 7.0 has a procedure called PROC SURVEYSELECT that can be used directly to generate sample data from a given set of data with replacement, which means a unit of sample data can be selected for the sample more than once. We can also use the following procedure to generate our own samples:

1. Generate a given number (sample size) of uniformly distributed pseudo-random numbers between 0 and 1.
2. Map each random number to one value of the original data by associating the random number to the quantile of the original value.
3. Calculate the p-value from the “bootstrapped” sample.
4. Increase the sample size and repeat step 1 and 2 if needed.

For example (assume original sample size = 5):

Original Value	C1	C2	C3	C4	C5
Associated Quantile	0.2	0.4	0.6	0.8	1
Generated Random Number	0.12	0.34	0.23	0.61	0.78

Associated Categorical Value	C1	C2	C2	C4	C5
------------------------------------	----	----	----	----	----

We have applied this program in an efficacy study of a drug for Alzheimer’s disease. Two patient groups, one with our drug treatment, the other with placebo, are compared for both cognitive (ADASCOG) and behavioral (CIBIC+) improvements. ADASCOG has a continuous measurement ranging from 1 to 100, and CIBIC+ has a categorical measurement ranging from 1 to 7. As an example, we have done an analysis based on the following input:

ADASCOG for Treatment 1:

CIBIC+ for Treatment 1:

ADASCOG for Treatment 2:

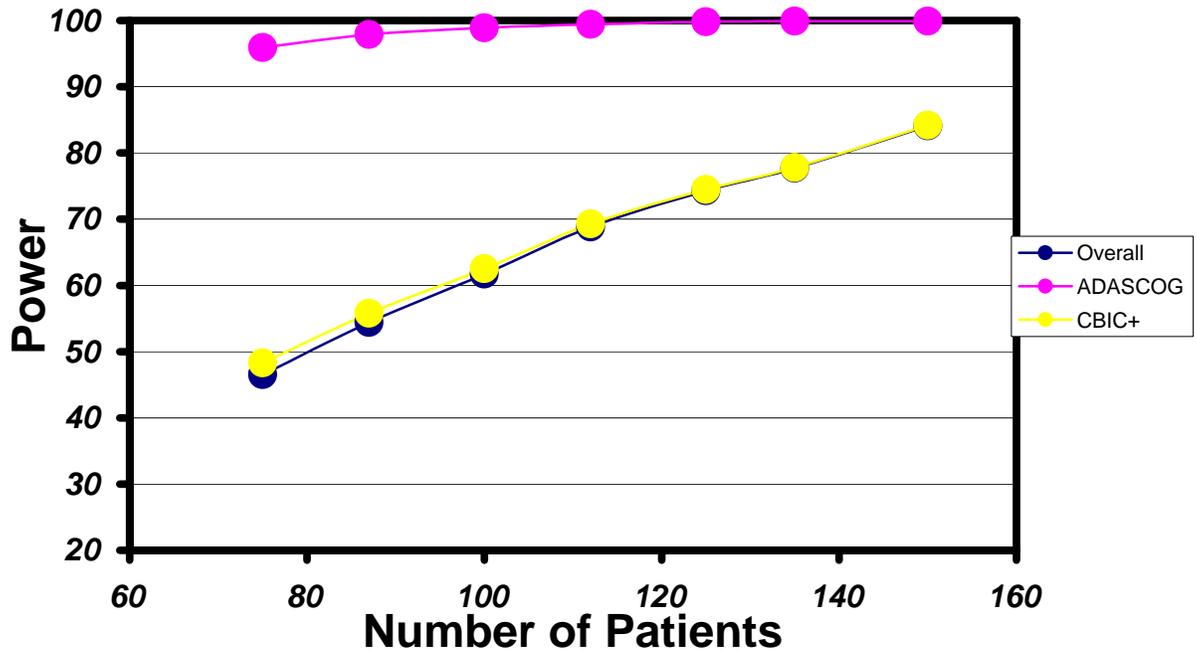
CIBIC+ for Treatment 2:

Type-I error for both ADASCOG and CIBIC+ improvements:

Type-II error for both ADASCOG and CIBIC+ improvements:

By varying the sample size from 75 to 125 with an increment of 10 each time, we obtain a curve as follows showing the relationship between the count of rejecting null hypothesis for both ADASCOG and CIBIC+ improvements and the number of patients involved. For comparison, the curves for either ADASCOG or CIBIC+ improvements are also displayed (Figure 1).

Figure 1. Power of Co-Primary Outcome



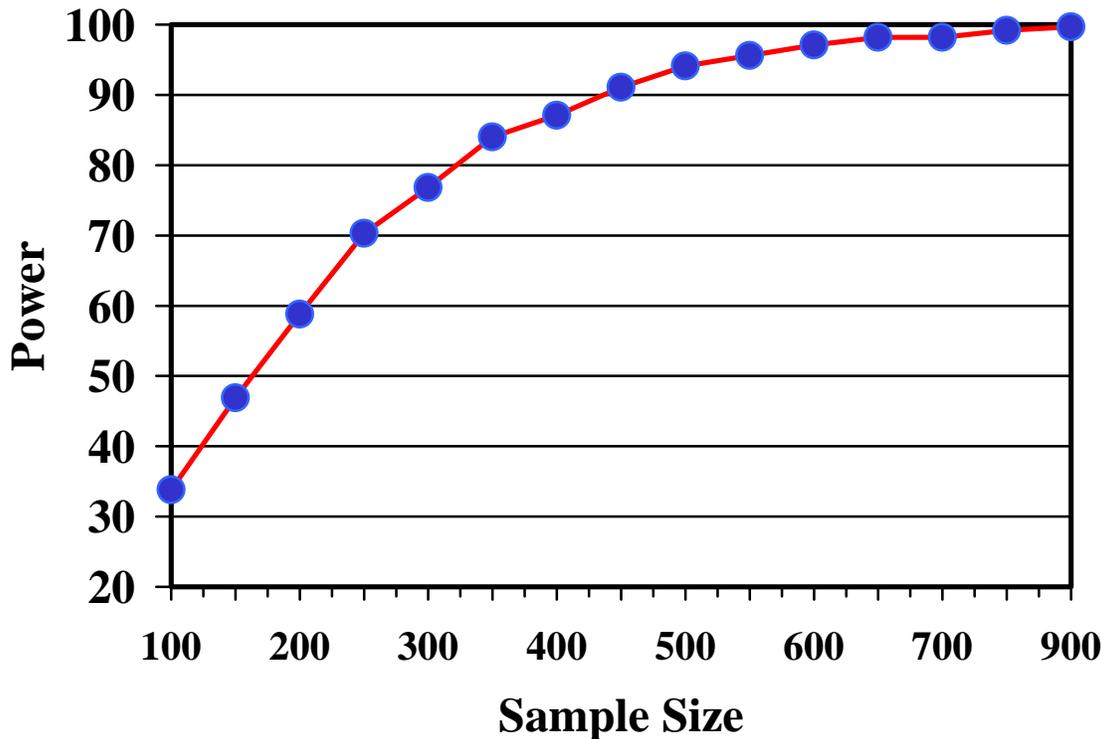
From the picture above, we see the blue curve that stands for ADASCOG improvement is far away from the black curve that stands for CIBIC+ improvements. The red curve which stands for the improvements of ADASCOG and CIBIC+ simultaneously is very close to the black one representing CIBIC+ improvement. In this example, if we use Bonferroni correction instead, we are going to end up with a much bigger (larger) sample size to achieve the same significance level for both ADASCOG and CIBIC+ improvements, which is, of course, too conservative and would result in a much higher monetary cost due to the larger sample size.

Bootstrap Power Analysis with Stratified Wilcoxon Test

In one of the phase III registration trial, we employed the stratified Wilcoxon test as our primary analysis due to the fact that the primary endpoint is obviously deviated from normality assumption, and a baseline covariate (baseline severity) has to be incorporated into the analysis through stratification. To estimate the power, we resample the data from a previous completed phase II b trial containing patients with similar characteristics.

At the first, we assume there is a treatment difference (this assumption is valid based on the data from the previous trial). Therefore, we are operating under the alternative hypothesis. The input data include baseline stratification, endpoint measurements and therapy assigned. The missing endpoint will be excluded from re-sampling, and will be added to the estimated sample size as dropout later on. In the program, we first separate input dataset into two groups according to subjects' therapies assigned and resample the two groups independently. Here we can specify how many subjects we want to get from each group. Then the two samples are combined to form an analysis dataset. The test now is applied to this dataset. We declare the trial is positive when p-value is less than 0.05. We follow the same step to resample the input data again to form a new analysis dataset and check if the p-value is significant. The new dataset is independent from the previous one. The whole process is inside a do-loop. So, if we fix the sample size and repeat 1000 times of re-sampling process, we will obtain 1000 independent datasets and be able to calculate the percentage of positive trials in those 1000 simulations. This percentage is essentially the power at the specified sample size. Using the same idea, we can get powers at different sample sizes then plot the power against the corresponding sample size. In the end, we can pick the proper sample size by examining the power curve (Figure 2)

Figure 2. Power curve



The most attractive advantage of this approach is the ability to incorporate different tests easily, such as stratified Wilcoxon test in this case. The flexibility also can allow more complicated designs. We use the balanced randomized design with two treatments in this example. Obviously you can have unbalanced designs by set different number of subjects to samples in different groups, you also can add more groups if the input data allowed. Finally, you can replace the test with other appropriate tests. However, not all tests can be used in this approach. The size of input data had strong influence on the final result. In this case, if the input data only had a few observations, the samples will have many ties, this in turn leads to the loss of power in the test itself. Also, if one stratum had only few observations in the input data, the test will lose power too. The magnitude of the impact can be investigated through this simulation.

Conclusion and Discussion

It has been demonstrated that simulation-based power analysis is suitable to compute the sample size for a study where precise or estimation formula cannot or is very difficult to be compute by classical sample size calculations. As an example of this, we have conducted an analysis using our simulation program for a study involving two co-primary outcome variables, one with a continuous measurement, and the other with a categorical measurement. It is shown that simulation-based power analysis gives more reasonable sample size estimation than Bonferroni correction, which is especially true in a situation where the actual significance levels of the two outcome variables are far apart from each other. We also have shown the flexibility of bootstrap technique by estimating power curve with stratified Wilcoxon test. Since the bootstrap directly re-sampling the data in

hand and does not require that distributions be normal or that sample sizes be large. A common misunderstanding is that bootstrap builds “something from nothing”. In fact, the technique has a sound theoretical foundation (Efron, 1994). However, there are assumptions made with bootstrap. We assume the sample data are independent and representative. We can obtain independent sampling by carefully programming. But it may be hard to determine whether the data is representative, since the original data has to represent the population studied accurately.

References

Shein-Chung Chow, Jun Shao and Hansheng Wang. Sample Size Calculation in Clinical Research. Marcel Dekker Inc, 2003

Cummings JL, Cyrus PA, Bieber F, Mas J, Orazem J, Gulanski B. Metrifonate Treatment of the Cognitive Deficits of Alzheimer’s Disease. Neurology. 1998 May;50(5):1214-21.

Efron, B., Tibshirani, R.J. (1994) An introduction to bootstrap. Chapman & Hall/CRC

SAS Institute Inc. SAS Procedure Guide, Version 8, and Cary, NC: SAS Institute Inc., 1999.