한국전문물리치료학회지 2016년 23권 2호 93-99
Phys Ther Korea 2016;23(2):93-99

ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
http://dx.doi.org/10.12674/ptk.2016.23.2.093

An Overview of Bootstrapping Method Applicable to Survey Researches in Rehabilitation Science

Bong-sam Choi, PhD, MPH, PT
Dept. of Physical Therapy, College of Health and Welfare, Woosong University

Abstract

Background: Parametric statistical procedures are typically conducted under the condition in which a sample distribution is statistically identical with its population. In reality, investigators use inferential statistics to estimate parameters based on the sample drawn because population distributions are unknown. The uncertainty of limited data from the sample such as lack of sample size may be a challenge in most rehabilitation studies.

Objects: The purpose of this study is to review the bootstrapping method to overcome shortcomings of limited sample size in rehabilitation studies.

Methods: Articles were reviewed.

Results: Bootstrapping method is a statistical procedure that permits the iterative re-sampling with replacement from a sample when the population distribution is unknown. This statistical procedure is to enhance the representativeness of the population being studied and to determine estimates of the parameters when sample size are too limited to generalize the study outcome to target population. The bootstrapping method would overcome limitations such as type II error resulting from small sample sizes. An application on a typical data of a study represented how to deal with challenges of estimating a parameter from small sample size and enhance the uncertainty with optimal confidence intervals and levels.

Conclusion: Bootstrapping method may be an effective statistical procedure reducing the standard error of population parameters under the condition requiring both acceptable confidence intervals and confidence level (i.e., p=.05).

Key Words: Bootstrap; Error; Measurement; Population; Sample size.

Introduction

Bootstrapping, the word itself saying, is pulling oneself up by one's bootstraps over a fence. It is somewhat an analogous term to a self-sustaining process that performs an impossible task without external helps. From a statistical standpoint, bootstrapping is a procedure that permits the iterative re-sampling with replacement from a sample when the population distribution is unknown (Efron, 1979). The method eventually allows setting confidence intervals and estimating significance levels from the re-sampled distribution. The method appears to be unrealistically promising to estimate population parameters using the re-sampling method over time, however it actually allows optimal estimates of population distribution (Kulesa et al, 2015). After the method was first introduced to statistical sciences by Efron (1979) and computer technologies was updated, the procedure has become widespread because it provides methodological reasoning for inferential statistics.

In inferential statistics, sample statistics such as mean and standard deviation are to estimate population parameters with some acceptable variations, which would later be used in evaluating the margin of errors. These statistics often vary from sample to

Corresponding author: Bong-sam Choi bchoi@wsu.ac.kr

sample. More specifically, one would like to investigate the magnitude of these variations around the corresponding population parameter under assumptions with which the variation of sample would be similar to that of population. The overall sense of all possible values of a sample statistic may be presented with respect to a possible distribution which may closely be matched to the population studied. This is called a sampling distribution. Based on the sampling distribution, sample statistics may later be determined to infer population parameters. However, in many cases, one is unable to determine how the population would be distributed, thus its parameters are commonly estimated by sample data only. This uncertainty resulting from the lack of representativeness of the population being studied in relation to small sample size may lead to a shortcoming of inferential statistics. Therefore obtaining optimal sample size may be critical to determine stable estimates of population parameters and to select realistic statistical procedures. (Pedhazur and Schmelkin, 1991; Tabachnick and Fidell, 1996; Wolf et al, 2013).

Efron (1979) introduced and developed the bootstrap method drawing repeated samples from the population studied and obtaining the overall sense of idea about the sampling distribution. The primary concept of the method is based on which a simulated distribution of population estimators obtained by bootstrap method is able to provide the closest approximation to the parameter distribution (Efron, 2012; Efron and Tibshirani, 1993). The stochastic method applicable to literally any type of sample statistic leads to a surrogate population approximating the sampling distribution of a statistic. The sample summary statistic is then computed on each of the bootstrap samples. This values may be transferred to a histogram and referred to as the bootstrap distribution of the sample statistic. Bootstrapping is the most widely accepted method for overcoming the limitations resulting from small sample sizes and the unreality of parametric statistical procedures.

Most, if not all, of researches in rehabilitation fields have been focus on meaningful treatment effects anchoring the empirical results to evidence-based practice. These studies are often carried out in healthy or patient cohorts using the different phases of clinical trials. To be determined if any changes in the treatment effects are meaningful “clinically-important changes” or “minimal detectable changes”, which can be determined by standard error of measurement (SEM) (Page, 2014; Wyrwich, 2004). In addition the changes, effect size indicating clinical significance in measurements between groups and statistical power indicating the likelihood of detecting an effect if the effect actually exists are interrelated with the issues of sample size (Cohen, 1988). Due to the reason, the most challenging aspect of many rehabilitation researches is conveniently drawing a small sample from the target patient population to which the outcomes drawn would be generalized. These small samples may typically be biased or not representative to the target patient population. This may also lead to type II error on a greater risk of small sample being unusual just by chance. That is, the possibility of getting type II error increases because statistical hypothesis testing with small samples may results in accepting the null hypothesis when it is false (Banerjee et al, 2009; Banerjee and Chaudhury, 2010).

The purpose of this study is to review the bootstrapping method to overcome shortcomings of limited sample size in rehabilitation researches.

Bootstrapping

Theoretical Basis

In the bootstrapping method, one may draw a large number of repeated samples, in other word “ghost samples”, from the corresponding population and postulate a sampling distribution of a specific statistic from the repeated samples to obtain a Monte Carlo distribution describing for translating uncertainties in model inputs into uncertainties in model outputs (Wolf et al, 2013). The newly obtained distribution is referred to as the bootstrap distribution of the statistic. The method may also be used only when the distributions is unable to be estimated analytically. Theoretically, the method is based on central limit theorem, which assumes a bell shaped normal distribution curve with μ for mean and σ/ for standard deviation. Hence, the sampling distribution of ($-\mu$)/SE (standard error of the mean) with SE=$\sigma/\sqrt{n}$ will be approximated by the bootstrap distribution of $(X - \overline{X}\ )/SE$ with $\overline{X_i}$=bootstrap sample mean and $\widehat{SE}$=$s/\sqrt{n}$.

For example, we typically accept samples when the sampling distribution of the estimates is able to estimate its population parameters. That is, the distribution representing a symmetrically bell-shaped with μ at the center and $\sigma/\sqrt{n}$ for standard deviation can approximate the population distribution. The estimates of sample mean or median may randomly be bootstrapped by which the repeated samples represents the same statistic from the population. The bootstrap distribution now better approximates to the sampling distribution with which the statistical function is of the form of $(\overline{X_i} - \overline{X})/\widehat{SE}$ where $\widehat{SE}$ is the estimate of the SE of $\overline{X}$ and $\overline{X_i}$ is the mean of bootstrap sample based on bootstrap central limit theorem (Singh, 1981). The method includes; 1) treating a sample of size n from a population as a virtual population, 2) re-sampling k samples of size n by permitting replacement (i.e., bootstrapping sample), 3) creating a simulated distribution for the parameters such as mean, SEs and confidence intervals (CIs). It should be noted that identical measurements may be selected over time in the bootstrapping sample due to allowing replacement. At extreme cases, all measurements selected can be identical (Figure 1).

Bias correction by bootstrap

The mean value of sampling distribution of $\overline{X}$ often differs from its true mean because the estimator of $\overline{X}$ is a function of individual data (i.e., X_1, X_2, X_3, ..., X_n). That is, the difference between the estimator and true mean is determined by large n, which replaces the population by the empirical population of the sample. This is so called bias correction by bootstrap. In addition, the SE of the estimator can be computed by the simple bootstrap sample as the sample varies on all possible samples (Efron, 2012; Singh and Xie, 2003).

Bootstrap confidence interval (CI)

The CI, a sample based range, are often provided

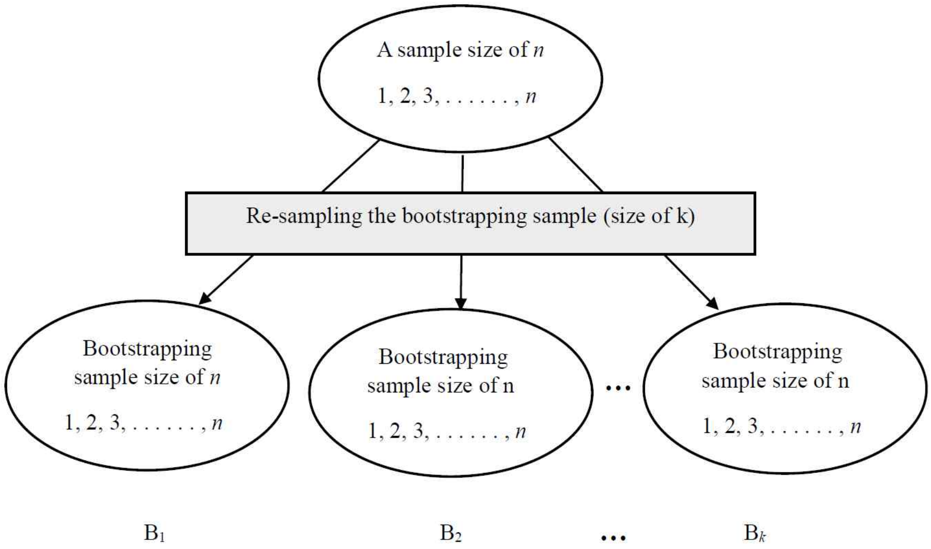
A sample size of n
1, 2, 3, ... , n
Re-sampling the bootstrapping sample (size of k)
Bootstrapping sample size of n
1, 2, 3, ... , n
Bootstrapping sample size of n
1, 2, 3, ... , n
...
Bootstrapping sample size of n
1, 2, 3, ... , n
B1
B2
...
Bk

Figure 1. Framework of the bootstrapping sample.

for the unknown mean of   as an estimator of population parameter. This is a property that true mean values would fall into within the range with a specified probability with respect to all possible samples depending on how the samples are drawn. That is, the CI is asymptotically determined by sampling distribution of the estimator of true mean as the sample size is approximate to infinity (i.e., n→∞). In bootstrap method, one may draw 1,000 bootstrap replicated estimators for true mean, rank from the lowest to the highest, and determine the CI at 95% when choosing 95% CI. For example, when the replicated estimators of true means ($\overline{X}$) are denoted by $\widehat{X}$, Individual bootstrap values are denoted by $\widehat{X}_1$, $\widehat{X}_2$, ...,$\widehat{X}_{1000}$. The CI for the bootstrap samples at 95% would be $\widehat{X}_{25}$, $\widehat{X}_{975}$. Since it should be noted that the sampling distribution of $\widehat{X}-\overline{X}$ is symmetrically distributed. Hence, the sampling distribution of $\widehat{X}-\overline{X}$ is approximated by the bootstrap distribution of $\widehat{X}-\widehat{X}_B$, which is contrary to the bootstrap concept and could now be approximated by the bootstrap distribution of $\widehat{X}_B-\widehat{X}$.

Bootstrap-t method

For better accuracy, it is possible to bootstrap a statistical function of the form t=$(X-\overline{X})/SE$, where $\widehat{SE}$ is a sample estimate of the standard error of $\widehat{X}$. In other words, Edgeworth correction by the bootstrap (Hall, 1988; Hall, 1992). The bootstrap-t is analogous to t-statistic where the SE of population mean is unknown, in most cases, and the standard deviation of the sample is replaceable for the unknown SE (i.e., $\widehat{SE}$=s/   ). Likewise the bootstrap-t is denoted by $t_B=\widehat{X}_B-\widehat{X}/\widehat{SE}_B$ where $\widehat{SE}_B$ is exactly like the SE as previously discussed under the bootstrap CI. Additionally the tB statistic can be obtained from the bootstrap sample and considered within the bootstrap CI. For instance, the $\overline{X}$ would lies between $\widehat{X}-t_{B\ .975}SE$ and $\widehat{X}-t_{B\ .025}SE$ when tB statistic from 1,000 bootstrap replicated estimators is denoted by $(\overline{X}_i-\overline{X}\ )/\widehat{SE}$. This range for $\overline{X}$ is the bootstrap-t CI obtained by bootstrap-t method at the 95% probability. Such an interval is now known to believe better accuracy in comparison to the traditional methods. It should be noted that the bootstrap-t method.

Sample size issues in healthcare research

Of the factors affecting quality sampling, selecting an optimal sample size is a matter of falsely implying no significant difference. In other words, the probability of accepting the null hypothesis when it is false (i.e., type Ⅱ error) depends on the sample size. In addition, the sample size is often too small to give a reliable test. Accepting the null hypothesis leads to further considerations such as whether the null is true or false in the real situation. Needless to say, investigators may consider accepted methods to determine how large a sample size should be such as power analyses or differently design the study to compare parameters between different populations. Most, if not all, researches in health care settings are typically conducted to demonstrate treatment effects in the forms of treated versus control group or survey design in which the numbers of subject are always limited by realities. The reason for that would be limited time and cost, or high drop-out rate. Consequently, it is difficult to recruit an ample sample size in health care researches, despite determining the optimal sample size with accepted methods (Anderson and Vingrys, 2001).

An application of bootstrap method on a real data

Several simulation studies emphasized that sample statistics are sensitive to sample size (Claudy, 1972; Pernet et al, 2015) and the ratio of sample size versus the number of variable around 5:1～10:1 was recommended as for stable results (Green, 1991). In bootstrapping method, an alternative of stably estimating the parameters is recommended as empirical evidences are unable to provide a standard but merely a compendious guideline for the determination of sample size. The method practically regards a virtual population as a sample of n measurements from population and draws samples with same size of n measurements allowing replacement to create bootstrapping sample. Under given conditions, an application to an original sample of n observations is as follows; 1) conducting a bootstrapping sample by random sampling method allowing replacement from the original sample, 2) estimating the mean and CI of the bootstrapping sample, 3) obtaining simulated distributions for two conditions with replicated the previous two procedures (i.e., 500 and 1,000 replications) and comparing those 3 estimates (i.e., the original sample, bootstrapping samples with 500 and 1,000 replications).

For a bootstrapping example with CI changes, a real data of studies (Choi and Park, 2012; Velozo et al, 2006) was used. The instrument of the study comprises functional capacity scales for measuring 10 functional activities from dictionary of occupational title at admission and discharge of worker's compensation clients. The rating scales were rated four categories: 1) severely impaired, 2) moderately impaired, 3) mildly impaired and 4) not impaired. When applying descriptive statistic procedure with bootstrapping method the total score of the functional capacity scale at discharge, mean value remains the same but standard error slightly changes (Table 1) as well as the confidence interval (Figure 2).

Whether applying the bootstrapping method or not, the mean value provided best estimates. That is, the estimator of remains the same whether the bootstrapping applied or not. In addition, increasing sample size by applying bootstrapping procedure from 123 to 500 and 1,000 cases, the SEs get smaller. This is the case where very similar sample may be drawn from its original sample during the procedure. Thus, bias correction by bootstrapping procedure was not necessary (Button et al, 2013; Efron, 2012). In addition, in a comparison of the two bootstrapping samples with size of 500 and 1,000, the SEs of the estimator allowing replications were more precise relative to that of its original sample as sample sizes increase. This would eventually lower the type II error, also known as a false negative, in which the newly drawn samples may be less biased or more representative to the target population being studied (Ioannidis, 2008; Ioannidis et al, 2011).

In general, both CI and p-value are acceptable to confirm the uncertainty in a point estimate from samples (Masicampo and Lalande, 2012). However CI is well known for its less subjective judgement to misinterpretation and better descriptive statistic to the range of possible value in comparison with p-value. High confidence level is typically recommended for estimating a parameter because there is a smaller chance of including the parameter (e.g., mean) in a particular confidence interval (Tabachnick and Fidell, 1996). Therefore, acceptable CI should be as narrow as possible and confidence level should be as high as possible. A dilemma between these two factors is that CI gets larger when confidence level is high and vice versa. That is, a best strategy increasing a narrow CI would be lowering standard error as much as possible by increasing sample size after setting a particular confidence level. Because, as stated previously, the standard error SE is calculated from s/ in which there is no control for s but for $\sqrt{n}$.

Table 1. The changes of a descriptive statistic following the bootstrapping procedure

Sample size	Mean	Variance	Bootstrapping		
			SE[a]	Upper CI[b]	Lower CI
Original sample (n=123)	38.10	83.04	.86	36.44	39.84
Bootstrapping (n=500 replications)	38.10	83.04	.82	36.47	39.82
Bootstrapping (n=1,000 replications)	38.10	83.04	.81	36.57	39.71

[a]standard error, [b]confidence interval.

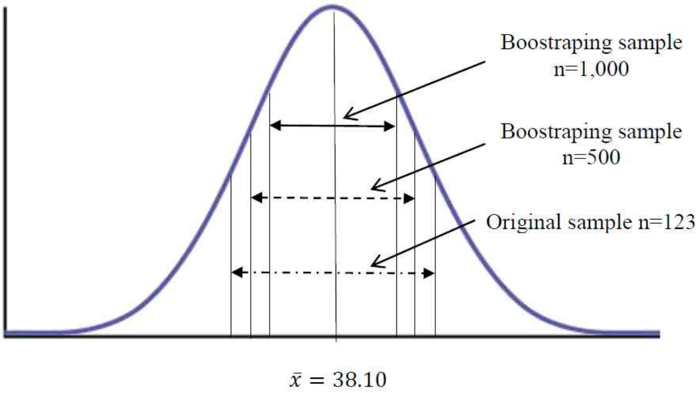

Figure 2. The changes of confidence intervals among the original sample, two bootstrapping samples with 500 and 1,000 replications.

Conclusion

Applying bootstrapping method to a sample with small size of n may be an effective procedure reducing the SE of parameter under the condition requiring both acceptable CI and optimal confidence level (p=.05 in general). Therefore most, if not all, researches in rehabilitation science fields can resolve greater risk of type II error resulting from small sample sizes.

References

Anderson AJ, Vingrys AJ. Small samples: Does size matter? Invest Ophthalmol Vis Sci. 2001;42(7):1411-1413.

Banerjee A, Chaudhury S. Statistics without tears: Populations and samples. Ind Psychiatry J. 2010;19(1):60-65.

Banerjee A, Chitnis UB, Jadhav SL, et al. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127-131. http://dx.doi.org/10.4103/0972-6748.62274

Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-376. http://dx.doi.org/10.1038/nrn3475

Choi BS, Park SY. Responsiveness comparisons of self-report versus therapist-scored functional capacity for workers with low back pain. Phys Ther Korea. 2012;19(3):91-97. http://dx.doi.org/10.12674/ptk.2012.19.3.091

Claudy JG. A comparison of five variable weighting procedures. Educ Psychol Meas. 1972;32:311-322.

Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, N.J, Lawrence Erlbaum Associates Inc., 1988:112-122.

Efron B. Bootstrap methods: Another look at the jackknife. Ann Stat. 1979;7(1):1-26.

Efron B. Bayesian inference and the parametric bootstrap. Ann Appl Stat. 2012;6(4):1971-1997.

Efron B, Tibshirani R. An Introduction to the Bootstrap. 1st ed. London, Chapman & Hall/CRC, 1993:23-29.

Green SB. How many subjects it take to do a regression analysis? Multivariate Behav Res. 1991;26(3):499-510. http://dx.doi.org/10.1207/s15327906mbr2603_7

Hall P. Theoretical comparison of bootstrap confidence intervals. Ann Stat. 1988;16(3):927-953.

Hall P. On bootstrap confidence intervals in nonparametric regression. Ann Stat. 1992;20(2):695-711.

Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology. 2011;22(4):450-456. http://dx.doi.org/10.1097/EDE.0b013e31821b506e

Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008:19(5):640-648. http://dx.doi.org/10.1097/EDE.0b013e31818131e7

Kulesa A, Krzywinski M, Blainey P, et al. Points of significance: Sampling distributions and the bootstrap. Nat Methods. 2015;12(6):477-478.

Masicampo EJ, Lalande DR. A peculiar prevalence of p values just below .05. Q J Exp Psychol (Hove). 2012;65(11):2271-2279. http://dx.doi.org/10.1080/17470218.2012.711335

Page P. Beyond statistical significance: Clinical interpretation of rehabilitation research literature. Int J Sports Phys Ther. 2014;9(5):726-736.

Pedhazur EJ, Schmelkin LP. Measurement, Design and Analysis: An integrated approach. 1st ed. Hillsdale, NJ, Lawrence Erlbaum Associates Inc., 1991:89-97.

Pernet CR, Latinus M, Nicholas TE, et al. Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. J Neurosci Methods. 2015;250:85-93. http://dx.doi.org/10.1016/j.jneumeth.2014.08.003

Singh K. On the asymptotic accuracy of Efron's bootstrap. Ann Stat. 1981;9(6):1187-1195.

Singh K, Xie M. Bootlier-plot: Bootstrap based outlier detection plot. Sankhya Ser A. 2003;65(3):532-559.

Tabachnick BG, Fidell LS. Using Multivariate Statistics. 3rd ed. New York, Haper Collins Publishers, 1996:120-122.

Velozo CA, Choi B, Zylstra SE, et al. Measurement qualities of a self-report and therapist-scored functional capacity instrument based on the Dictionary of Occupational Titles. J Occup Rehabil. 2006;16(1):109-122.

Wolf EJ, Harrington KM, Clark SL, et al. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. Educ Psychol Meas. 2013;76(6):913-934.

Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: Is there a connection? J Biopharm Stat. 2004;14(1):97-110.

This article was received April 11, 2016, was reviewed April 11, 2016, and was accepted May 11, 2016.