International Journal of Innovative Mathematics, Statistics & Energy Policies
9(3):22-28, July-Sept., 2021

 www.seahipaj.org **ISSN: 2467-852X**

Bootstrapping - An Introduction And Its Applications In Statistics

***Chibo Onyesom & **Aboko, S. I.**

Department of Statistics
Captain Elechi Amadi Polytechnic
Rumuola, Port Harcourt, Nigeria
E-mail: *chiboonyesom5@gmail.com./abokoigboye@gmail.com**

Abstract
A completely different way of calculating confidence intervals is called bootstrapping. This is phrased as ''pulling yourself up by your own bootlaces''. It is used in the sense of getting 'something for nothing'. The idea is very simple. You have a single sample of 'n' measurements but you can sample from this in many ways, so long as you allow some values appear more than once, and other samples to be left out (i.e sampling with replacement).All you do is calculate the sample mean lots of times ,once for each sampling from your data ,then obtain the confidence interval by looking at the extreme highs and lows of the estimated means using a function called quartile to extract the interval you want(e.g. a 95% interval is specified using c (0.0275, 0.975) to locate the upper and lower bounds). This paper shows how successfully the bootstrapping technique can be used in regression, estimation, hypothesis testing, confidence interval, prediction and model selection through empirical investigation
Keywords: Replacement, bootlaces, simulate, resample, confidence intervals, symmetrical.

1.0 INTRODUCTION

Bootstrapping is a resampling technique that can be used to study the sampling distribution of estimators, to compute approximate standard errors, and to find appropriate confidence intervals. In statistics, bootstrapping can be described also as any test or metric that uses random sampling with replacement (example in mimicking the sampling process), and falls under the broader class of resampling methods. Bootstrapping assigns measures of accuracy (bias, variance, confidence intervals, prediction errors …) to sample estimators. The bootstrapping technique allows estimation by measuring the properties of an estimate (such as variance) by sampling distribution of almost any statistic using random sampling methods.

Bootstrap was published by Bradley Efron in (1979). Bootstrapping is a computer – intensive procedure that was developed to allow us to determine reliable estimates. One standard choice for an approximating distribution is the empirical distribution function of the observed data.

In the case where a set of observations are assumed to be from an independent and identical distributed population, this can be implemented by constructing a number of resamples with replacement of the observed data set (and of equal size to the observed data set). It may also be used for constructing hypothesis tests.

Bootstrap is often used as an alternative to statistical inference based on the assumption of a parametric model when the assumption is in doubt, or where parametric inference is impossible or requires complicated formulas for the calculation of standard errors.

The basic idea of bootstrapping is that inference about a population from sample data (sample→population) can be modeled by resampling the sample data and performing inference about a sample from resampled data (resampled → sample).

When the population value is unknown, the true error in a sample statistic against its population value is unknown. In bootstrap → resamples, the population is the sample that is known; hence the quality of inference of the true sample from resampled data (resampled → sample), is measurable. More formally, the bootstrap works by treating inference of the true probability distribution θ, given that the original data, as being analogous to inference of the empirical data.

The accuracy of inferences regarding using the resampled data can be assessed because if we know a reasonable approximation to θ, and then the quality of inference on θ this can in turn be inferred.(David Hinkley posited)

For example, assume we are interested in the average (or mean) of the height of people in a nation. We cannot measure all the people in the national population, so instead a sample of only a tiny part of it is taken and measured. Assuming the sample size is N that is; we measure the heights of N individuals. From this single sample, only one estimate of the mean can be obtained. In order to reason about the population, we need some sense of the variability of the mean that we have computed.

The simplest bootstrap method involves taking the original data set of heights by using a computer, sampling from it to form a new sample called a resampled or bootstrap sample that is of size N. The bootstrap sample is taken from the original by sampling with replacement (we might resample 5 times from 1, 2, 3, 4, 5 and get 2, 5, 4, 4, 1 so assuming N is sufficiently large, for all practical purposes, there is virtually zero probability that it will be identical to the original real sample.

This process is repeated a large number of times (typically 1,000 or 10,000 times), and for each of these bootstrap samples, we compute its mean (each of these are called bootstrap estimates). We now can create a histogram of bootstrap means

This histogram provides an estimate of the shape of the distribution of the sample mean from which a question can be answered about how much the mean varies across samples. The method described here for the mean can be applied to almost any other statistic or estimator.

1.1 Types Of Bootstrap Scheme In Statistics

In univariate problems, it is usually acceptable to resample the individual observations with replacement unlike subsampling in which resampling is without replacement and is valid under much weaker conditions compared to the bootstrap. In small samples, a parametric bootstrap approach might be preferred.

The bootstrap method requires us to select a random sample of 'n' with replacement from this original sample. This is called bootstrap sample. Since it is selected with replacement, the bootstrap sample will contain observations from the original sample with some of them duplicated and some of them omitted.

The model is then fit to this bootstrap sample, using the same analysis procedure as the original sample. This produces the first bootstrap estimate say; $\widehat{\beta i*}$. This process is repeated a large number of times on each repetition, a bootstrap sample is selected, the model is fit and an estimate of $\widehat{\beta i*}$ is obtained for *i*= 1,2,…. m bootstrap samples. This is because repeated samples are taken from the original sample.

Bootstrapping is also called resampling procedure. If we denote the estimated standard deviation of the bootstrap $\widehat{\beta i*}$ by S ($\widehat{\beta i*}$). The bootstrap standard deviation S($\widehat{\beta i*}$) is an estimate of the standard deviation of sampling distribution of $\hat{\beta}$ and consequently, it is a measure of the precision of estimation for the regression coefficient β in regression analysis.

1.2. Another Bootstrap Scheme In Statistics Is The Bootstrapping Cases (or Bootstrapping Pairs)

In statistics, sometimes, the adequacy of the regression function or the error variance is not constant or the regressors are not fixed type variables: then bootstrapping is used to remedy the situations such that the 'n' sample pairs (X_i, Y_i) that are considered to be the data that are to be resampled.

The 'n' original sample pairs X_i, Y_i are sampled with replacement 'm' times, yielding a bootstrap sample, say ($X*_i, Y*_i$) for I = 1, 2, …. Then we fit a regression model to this bootstrap sample say, y* = x*β+e resulting in the first bootstrap estimate of the vector m times.

Generally, the choice of m depends on the application. However, 200 - 1000 bootstrap samples are employed. One way to select m is to observe the variability of the bootstrap standard deviation S ($\widehat{\beta i *}$) as m increases when S ($\widehat{\beta i *}$) stabilizes, a bootstrap sample of adequate size has been reached.

1.3 Bootstrap And Sampling Distribution Of Statistic

The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a data set with replacement. It can be used to estimate summary statistics such as the mean or standard deviation.

When using the bootstrap you must choose the size of the sample and the numbers of repeats say m times. Bootstrapping then assigns measures of accuracy (bias, variance, confidence interval, prediction errors etc) to sample estimates. The technique allows estimation of the sampling distribution of almost any statistic using random sampling methods.

The bootstrap method on the other hand takes the original sample data and then resamples it to create many (simulated) samples. This technique also allows for accurate estimates of statistics, which is crucial when using data to make decisions.

Bootstrap is a Monte Carlo simulation approach used to estimate the uncertainty of a statistic or an estimator on a data.

2.1 Bootstrap Estimation of The Mean

The statistical mean is the measure of Central Tendency. It is calculated by summing all the observations in a batch of data and then dividing the total by the number of items involved. While bootstrapping estimation involves an initial sample of the batch of data and then repeated resampling from the initial sample.

What makes bootstrapping estimation useful is that the procedures are based on the initial sample and makes no assumption regarding the shape of the underlying population. In addition, the procedures do not require knowledge of any population.

From the resampling distribution of the statistic of interest (i.e. the distribution of the sample mean obtained from the m resamples) using a stem-and-leaf display or an ordered array.

To form a 100(1- $\propto$) bootstrap confidence interval of the population mean μ_x, we use the stem and leaf display or ordered array for the resampling distribution and find the cuts off the smallest ($\frac{\propto}{2}$)100% and the value that cuts off the largest ($\frac{\propto}{2}$) 100% of the statistic. These values provide the lower and upper limits for the bootstrap confidence interval estimate of the unknown parameter.

2.2 Bootstrapping Confidence Interval In Regression

We can use bootstrapping to obtain approximate confidence intervals for regression coefficients and other quantities of interest such as the mean response at a particular point x- space or an approximate production interval for a future observation on the response.

Simple procedure for obtaining an approximate 100(1- $\propto$) percent confidence interval through bootstrapping is the reflection method (also known as Percentile method). This reflection confidence interval method uses the lower ($\frac{\propto}{2}$) 100% and upper 100(1- $\propto$) percentile of the bootstrap distribution of $\widehat{\beta\imath*}$. If we denote these percentiles by $\widehat{\beta\imath*}(\frac{\propto}{2})$ and $\beta\imath*\widehat{(1-\propto)}$, respectively, where $\widehat{\beta\imath*}$,i=1,2,….m.

Define the distances of these percentiles from $\hat{\beta}$, the estimate of the regression coefficient obtained for the original sample as follows:

$D_1 = \hat{\beta} - \widehat{\beta\imath*}(\frac{\propto}{2})$

$D_2 = \beta\imath*\widehat{(1-\propto)} - \hat{\beta}$

Then the approximate 100 (1- $\frac{\propto}{2}$) percent bootstrap confidence interval for the regression coefficient $\hat{\beta}$ is given by $\hat{\beta} - D_2 \leq \beta \leq \hat{\beta} + D_1$

3.1 The Role Of Bootstrapping In Statistics

3.1.1 Regression Models

Bootstrapping in regression model can be presented in-terms of a linear regression model, but could be applied to a nonlinear regression model or a generalized linear model in essentially the same way. The basic approach for bootstrapping regression estimates are:

In the linear regression model

Y = Xβ +e and obtain the 'n' residuals $e' = [e_1, e_2 - - - e_n]$

Choose a random sample of size n with replacement from these residuals and arrange them in a bootstrap residual vector e*. Attach the bootstrap residuals to the predicted values;

$\hat{y}$ = X$\hat{\beta}$ to form a bootstrap vector of responses y*.That is, calculate $y* = X\hat{\beta}$ + e*

These bootstrapped responses are now regressed on the original regressors by the regression procedure used to fit the original model. These produce the first bootstrap estimate of the vector of regression coefficient. We could similarly obtain bootstrap estimates of any quantity of interest that is a function of the parameter estimates.

3.1.2 Case Resampling

This can be computationally done through different samples with the formula:

$\binom{2n-1}{n} = \frac{(2n-1)!}{n!(n-1)!}$ Where n is the size of the data set .

Thus, for n = 5,10,20,30 these are 126,92378,6.89$\times 10^{10}$ and 5.91 $\times 10^{16}$ different resamples respectively.

3.1.3 Estimating The Distribution Of Sample Mean

Consider a coin flipping experiment, when a coin is flipped a record of either tail or head is recorded.

Let X* = x_1, x_2 - - -,x_{10} be 10 observations from the experiment, xi= 1 if the *i* th flip lands heads and 0 otherwise. From normal theory, we can use the t-test to estimate the distribution of the sample mean $\bar{x}$ = 1/10 [$x_1, + x_2$ +- - -+,x_{10}].

Instead, we use the bootstrap, to derive the distribution of $\bar{x}$. We first resample the data to obtain a bootstrap resample such as: X_1 =$x_2, x_1, x_{10} x_3 \ x_4 x_6 \ x_7 \ x_9$ There are some duplicates since a bootstrap resample comes from sampling with replacement from the data.

Also the number data points in a bootstrap resample is equal to the number of points in the original observations. Then we can compute the mean of this resample and obtain the first bootstrap mean μ*. This process is repeated to obtain the second resample μ_2 * if we repeat this process 100 times, then we have $\mu *_1$,$\mu *_2$ - - - $\mu *_{100}$.

This represents an empirical bootstrap distribution of sample mean. From this empirical distribution, one can derive a bootstrap confidence interval for the purpose of hypothesis testing.

3.1.4 Bayesian Bootstrap

In statistics, bootstrap can be useful in the Bayesian modeling. Bootstrapping can be interpreted in a Bayesian framework using a scheme that creates new data set through reweighting the initial data.

Given a set of N data points, the weighting assigned to data point θ in a new data set $k^\theta$is $\emptyset_i^\theta$ = $x_i^{\theta T}$- x_{i-1}^θ, where x^θ is a low-to-high ordered list of N-1 uniformly distributed random numbers on [0,1], preceded by 0 and succeeded by 1.The distributions of a parameter inferred from considering such data sets k^θ are then interpretable as Posterior distribution on that parameter

3.1.5 Smooth Bootstrap

Under this scheme, a small amount of (usually normally distributed) zero-centered random noise is added on to each resampled observation. This is equivalent to sampling from a Kernel density estimate of the data.

Assume K to be a symmetric Kernel density function with unit variance, the standard Kernel estimator $\widehat{f_h}$ (x) if f(x) is $\hat{f}$(x) =$\frac{1}{nh}\sum_{i=1}^{n} k\left[\frac{x-x_i}{h}\right]$;

Where h is the smoothing parameter and the corresponding distribution function estimator: $\widehat{F_h}$ (x) is: $\widehat{F_h}$ (x) = $\int_{\infty}^{x} \widehat{f_h}$(t) dt.

3.1.6 Parametric Bootstrap

Based on the assumption that the original data set is a realization of a random sample from a distribution of a specific parametric type, in this case a parametric model is fitted by parameterθ, often by maximum likelihood, and samples of random numbers are drawn from this fitted model.

Usually the sample drawn has the same sample size as the original data. Then the estimate of original function F can be written as $\hat{F}$ = F_θ

This sampling process is repeated many times as for other bootstrap methods. The use of a parametric model at the sampling stage of the bootstrap methodology leads to procedures which are different from those obtained by applying basic statistical theory to inference for the same model.

3.1.7 Resampling Residuals

In regression problems we fit the model and retain the fitted values y_i and the residuals

$\widehat{e_i} = y_i - \widehat{y_i}$,i = 1,- - - ,n. For each pair, (x_i,y_i) in which x_i is the (possibly multivariate) explanatory variable, add a random resampled residual e_i to the fitted value y_i

In order words, create synthetic response variables $y^*_i = \widehat{y_i} + e_j$, where j is selected randomly from the list (1, 2 - - -, n) for every *i*.

A repeated step of this model retains the information in the explanatory variables. Resample can resample residuals of standardized residuals or raw residuals.

4.1 Advantages Of Bootstrapping In Statistics

A great advantage of bootstrapping is its simplicity. It is a straight forward way to derive estimates of standard errors and confidence intervals for complex estimators of the distribution, such as percentile points, proportions, odd ratios and correlation coefficients. Bootstrap is also an appropriate way to control and check the stability of the result. Although for most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality. Bootstrapping is also convenient method that avoids the cost of repeating the experiment to get other groups of sample data.

4.2 Disadvantages Of Bootstrapping In Statistics

Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees. The result may depend on the representative sample. The apparent simplicity may conceal the fact that important assumptions are being made when undertaking the bootstrap analysis (eg independence of samples) where these would be more formally stated in other approaches. Also, bootstrapping can be time consuming.

5.1 CONCLUSIONS

The bootstrap is said to yield more rapid convergence to normality of the estimates; and also yields estimates of error where it is difficult or impossible to obtain asymptotic expressions. It serves as a means of reducing bias and of providing a realistic assessment of error, in cases where the estimator is a function of all the data values concerned. The bootstrap appears to behave reliably, both as a means of reducing bias and of providing a realistic measurement of error, in cases where the estimator (for entering and partial estimates) is a function of all the data values concerned. Situations involving variances, maximum likelihood or least squares estimation, regression and ratios fall in this category but those involving order statistics do not.

5.2 RECOMMENDATIONS

(1) Use as many sub-groups (ideally r = n) as is consistent with computational constraints.
(2) Do not bootstrap estimates involving extreme values.
(3) Return one or two most significant figures in the overall and partial estimates than would normally be the case, to allow for the differencing operation.

REFERENCES

Beard, R.E. et al (1977).*Risk Theory. The stochastic Basis of Insurance*.3rd edition 233-239.

Bissell, A.F. and Ferguson, R.A.(1995).The jackknife-Toy, Tool or Two-edged weapon? The statistician. *Journal of the institute of statisticians*.79-100

David, M.L., and Mark, L.B.(1996).Basic business statistics: Concepts and applications.6th edition.356-357.

David,V.H(2000).Bootstrap Methods: Journal of the Royal Statistical Society series B(Methodology)

Douglas,C.M. et al {2006).Introduction to Linear Regression Analysis.4th edition.418,493-494.

Duncan,G.T. and Layard, M.W.J (1993) A Monte Carlo study of asymptotically robust tests for correlation coefficients.*Biometrica*,60.551-8.

Michael,J.C(2015).Statistics: An introduction using R. 2ND edition.