Bootstrapping

Caelan Bryan, Jenna Dufresne, Jamielee Jimenez Perez

Fall 2022

Literature Review

Bootstrap Methods: Another Look at the Jackknife (B. Efron 1979) is credited as being the creation of bootstrapping by author B. Efron. It starts by explaining that bootstrapping is a step above the jackknife method, and that the jackknife is a linear expansion for estimating the bootstrap. Starting in the introduction, it is stated that the bootstrap is shown to estimate the variance of the sample median, which is an area that the jackknife fails at. It is also mentioned that the bootstrap does well at estimating the error rates in certain problems, which outperforms other non-parametric estimation methods. The problem attempting to be solved is estimating the sampling distribution based on the observed data. The idea behind the bootstrap method is listed in three parts. First, construct the sample probability distribution. Second, draw a random sample of size n. Lastly, approximate the sampling distribution by the bootstrap distribution. It is mentioned that that the difficult part of bootstrapping is calculating the bootstrap distribution and three methods are given to accomplish this; direct Monte Carlo approximations, and Taylor series expansion methods. A few applications are listed. These include estimating the median, error rate estimation in discrimination analysis, relationship with the jackknife, Wilcoxon's statistic, and regression models. Finally, a list of some remarks regarding the bootstrap method are listed. Some important ideas listed are that the calculation of the bootstrap distribution using the Monte Carl method is easy to implement on the computer and that the bootstrap and jackknife provide approximate frequency statements and not approximate likelihood statements.

Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Methods (Bradley Efron 1981) starts by giving some background on what the purpose of the paper is. In this case, we want to estimate the standard error based on the data. This is normally done through parametric modeling methods, but in the paper it is being done through nonparametric methods. All methods are being tested using the same scenario which is bivariate normal distribution. This paper has four points. They are to describe the various methods, show how the methods derive from the same idea, relate the methods to each other, and finally show how the methods perform differently even though they are similar. The Monte Carlo experiment that the data was obtained from is then described. The bootstrap method is the first that is looked at. The steps are described to obtain the bootstrap samples. Random samples are created from the original sample and the bootstrap estimate is obtained. This process is than repeated multiple time to achieve enough samples to accurately estimate the standard error. 128 and 512 total samples were used separately, and while the 512 total sample provided a slightly more accurate result, the increase was minimal. on this result, the size of N samples is not overly important past 50-100. A smoothed bootstrap method is than looked at. This produced by compromising the normal theory maximum, likelihood estimate and the nonparametric maximum likelihood estimate. The results of the smoothed bootstrap method were overall better than those in the on-smoothed bootstrap method. Next, the jackknife, infinitesimal jackknife, half-sampling, and random sub sampling methods are discussed. All results are displayed in a table containing th estimated values, standard deviation, and confidence interval for each estimate. Overall, the bootstrap method produced the results that closest matches the theoretical values.

In Bootstrapping - An Introduction And Its Applications In Statistics (Hossain 2000) the authors take a high-level approach to introducing the reader to bootstrapping along with applications to different parts of statistics. The introduction begins to explain what bootstrapping is and how it is accomplished. They make use of an example to help explain what bootstrapping is. In this case, they look at performing analysis on the population of the United States. It would be difficult, costly, and timely to sample the entire population

so instead you might sample a smaller subset of the population and create new bootstrap samples using replacement. In this sense the bootstrap samples might contain duplicates or even omit certain responses from the original sample. Doing so allows interpretation of the entire population based on a subset of the population. One of the first applications looked at is estimation of means and confidence intervals for the mean. Another application is constructing confidence intervals for regression coefficients. The authors also look at regression models, case re-sampling, estimating the distribution of sample mean, Bayesian bootstrap, smooth bootstrap, parametric bootstrap, and re-sampling residuals as applications of bootstrapping. Lastly, some of the advantages and disadvantages are listed. One advantage is the simplicity of deriving estimates of standard errors and confidence intervals where a disadvantage would be that the result still depends on the original sample.

In The Importance of Discussing Assumptions when Teaching Bootstrapping (Totty, Molyneux, and Fuentes 2021) the authors spend time ensuring that the readers understand when it is appropriate to use bootstrapping methods by presenting the math along with experimental results. The paper starts with an introduction to what bootstrapping is. They also mention that bootstrapping has increased in popularity since it's inception with applications in linear regression and neural networks among other fields. The methods focused on during the paper are studentized, basic, and percentile bootstrap intervals and their hypothesis tests. The next section focuses on why it's important to teach statistical computing and bootstrapping. One feature relevant to bootstrapping is that students understandings of confidence intervals and statistical inference relies on their understanding of sampling distributions. Next, some of the assumptions for bootstrapping are discussed. The assumptions were split based on interval estimation and hypothesis testing. For both cases, the largest assumption is that the distribution can be made approximately pivotal through shifting or studentization. Lastly, simulation-based performance was evaluated. This was done by using the metrics coverage proportion, significance level, and power. The most important results were that when the assumptions are broken, there can be differences in the performance of the different bootstrapping methods. There also was not a improvement between bootstrapping and other methods whose assumptions were also broken. The smaller the sample size and non-normalcy also impacted the performance of the methods. Lastly, an R package was created, which encompasses the functions used throughout the paper, that can be used to create intervals using bootstrapping methods.

Bootstrapping with R to make generalized inference for regression model (Sillabutra et al. 2016) looks at a specific application of bootstrapping, validating a generalized regression model to make generalization of statistical inference to different cases outside of the original sample. In the introduction, the authors explain the different types of regression models and explain the different ways to complete model validation. Some of those listed include cross-validation, Jackknife, and bootstrap methods. The idea of the bootstrap method is then also explained. The methodology and design of the experiment is then outlined. In this case, the original observations will be resampled leading to a set of bootstrap samples. The mean estimate and regression coefficient estimates can then be found for each bootstrap sample. Finally, confidence intervals can be created for the mean, regression coefficients, and standard errors for both the mean and regression coefficients. A table is provided to outline the values received from the original sample and bootstrap samples. Finally, the results are discussed in the conclusion. Based on the results of this experiment, the values found are very similar among the original and bootstrap samples, although the confidence intervals for the bootstrap samples are often wider. Some advantages to bootstrapping are also listed.

Data Description

The dataset we will be using to explore Bootstrapping is the "Causes of Death - Our World in Data" dataset from kaggle (Chavez 2022). The causes of death dataset was also expanded with population data from the world bank (2022).

The raw data contains a multitude of death statistics broken down by the continent, region, country, and territory. The statistics for cause of death are given in number of deaths and split up by the cause of death. Data is presented over the span of multiple years. The population dataset from the world bank is a list of countries, regions, and territories but contains the population for each year from 1960 to 2021 where available.

To get more meaningful numbers, the population for each year 1990 through 2019 was populated into the

causes of death dataset. In order to perform a more predictable experiment, the causes of death dataset was cleaned by first removing all non-country entries. The number of executions and terrorism deaths columns were also removed due to a lack of data for the majority of countries. As a final note, the Vatican and Liechtenstein are the two countries that did not have cause of death statistics available in the dataset.

The columns in the cleaned dataset are as follows:

Name	Description	Type
Entity	Name of Country	Nominal
Population	Population of Country at Specific Year	Discrete
Code	Three Letter Country Code	Nominal
Year	Year for Causes of Deaths	Nominal
Deaths_Meningitis	Number of Deaths Caused by Meningitis	Discrete
Deaths_Neoplasms	Number of Deaths Caused by Neoplasms	Discrete
Deaths_FireHeatHotSubstances	Number of Deaths Caused by Fire, Heat, or Hot Substances	Discrete
Deaths_Malaria	Number of Deaths Caused by Malaria	Discrete
Deaths_Drowning	Number of Deaths Caused by Drowning	Discrete
Deaths_InterpersonalViolence	Number of Deaths Caused by Interpersonal Violence	Discrete
Deaths_HIVAIDS	Number of Deaths Caused by HIV/AIDS	Discrete
Deaths_DrugUseDisorders	Number of Deaths Caused by Drug Use Disorders	Discrete
Deaths_Tuberculosis	Number of Deaths Caused by Tuberculosis	Discrete
Deaths_RoadInjuries	Number of Deaths Caused by Road Injuries	Discrete
Deaths_MaternalDisorders	Number of Deaths Caused by Maternal Disorders	Discrete
Deaths_LowerRespiratoryInfections	Number of Deaths Caused by Lower Respiratory Infections	Discrete
Deaths_NeonatalDisorders	Number of Deaths Caused by Neonatal Disorders	Discrete
Deaths_AlcoholUseDisorders	Number of Deaths Caused by Alcohol Use Disorders	Discrete
Deaths_ExposureToForcesOfNature	Number of Deaths Caused by Exposure to Forces of Nature	Discrete
Deaths_DiarrhealDiseases	Number of Deaths Caused by Diarrheal Diseases	Discrete
Deaths_EnvironmentalHeatAndColdExposure	Number of Deaths Caused by Environmental Heat and Cold Exposure	Discrete
Deaths_NutritionalDeficiencies	Number of Deaths Caused by Nutritional Deficiencies	Discrete
Deaths_Selfharm	Number of Deaths Caused by Self-Harm	Discrete
Deaths_ConflictAndTerrorism	Number of Deaths Caused by Conflict and Terrorism	Discrete
Deaths_DiabetesMellitus	Number of Deaths Caused by Diabetes Mellitus	Discrete
Deaths_Poisonings	Number of Deaths Caused by Poisoning	Discrete
Deaths_ProteinEnergyMalnutrition	Number of Deaths Caused by Protein Energy Malnutrition	Discrete
Deaths_CardiovascularDiseases	Number of Deaths Caused by Cardiovascular Diseases	Discrete
Deaths_ChronicKidneyDisease	Number of Deaths Caused by Chronic Kidney Disease	Discrete
Deaths_ChronicRespiratoryDiseases	Number of Deaths Caused by Chronic Respiratory Diseases	Discrete
Deaths_CirrhosisOtherChronicLiverDiseases	Number of Deaths Caused by Cirrhosis or Other Chronic Liver Diseases	Discrete
Deaths_DigestiveDiseases	Number of Deaths Caused by Digestive Diseases	Discrete
Deaths_AcuteHepatitis	Number of Deaths Caused by Acute Hepatitis	Discrete
Deaths_AlzheimersDiseaseOtherDementias	Number of Deaths Caused by Alzheimers Disease or Other Dementias	Discrete
Deaths_ParkinsonsDisease	Number of Deaths Caused by Parkinsons Disease	Discrete

Introduction

Methods

Statistical Modeling

Conlusion

References

2022. https://data.worldbank.org/indicator/SP.POP.TOTL.

Chavez, Ivan. 2022. https://www.kaggle.com/datasets/ivanchvez/causes-of-death-our-world-in-data.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7 (1): 1–26. http://www.jstor.org/stable/2958830.

Efron, Bradley. 1981. "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods." *Biometrika* 68 (3): 589–99. http://www.jstor.org/stable/2335441.

Hossain, Mohammad. 2000. "Bootstrapping – an Introduction and Its Applications in Statistics." *Bangladesh Journal of Scientific Research* 18 (January): 75–88.

Sillabutra, Jutatip, Prasong Kitidamrongsuk, Chukiat Viwatwongkasem, Chareena Ujeh, Siam Sae-tang, and Khanokporn Donjdee. 2016. "Bootstrapping with r to Make Generalized Inference for Regression Model." *Procedia Computer Science* 86: 228–31. https://doi.org/https://doi.org/10.1016/j.procs.2016.05.103.

Totty, Njesa, James Molyneux, and Claudio Fuentes. 2021. "The Importance of Discussing Assumptions When Teaching Bootstrapping." arXiv. https://doi.org/10.48550/ARXIV.2112.07737.