Quantifying uncertainty with the bootstrap: introduction and simulation study

Greta Gašparac[1], **Erik Štrumbelj**[1]

[1] *University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia*
E-mail: greta.gasparac@gmail.com, erik.strumbelj@fri.uni-lj.si

Abstract

We introduce the reader to the bootstrap, a simple and flexible resampling-based alternative for quantifying uncertainty. We describe the basic characteristics of the non-parametric bootstrap and illustrate its practical behaviour with simulations in the context of a typical task in machine learning - estimating and comparing the performance of different prediction models. We compare the standard normal, percentile and BC_a bootstrap confidence intervals. As theory suggests, the BC_a performs the best over a wide range of situations.

1 Introduction

Empirical research is an integral part of science. However, we must be aware that any result that is based on measurements has a certain degree of uncertainty. Not accounting for this uncertainty can lead to incorrect and possibly misleading conclusions. This is why it is very important to quantify uncertainty and why statistics play such an important role in modern science.

Our motivation is the common task in machine learning - estimating a model's performance (predictive accuracy, speed, ...) and/or comparing it with other models on one or more different scenarios. In many cases researchers do not even attempt to quantify the uncertainty in their estimates. When they do, they typically use either confidence intervals (CI) based on the normal distribution or null-hypothesis statistical testing (NHST) [3].

Performance data are typically not distributed normally, so CIs based on the normal distribution make sense only in cases when we reason about the mean and the sample size is not too small. The same applies to most parametric NHST which are based on the normal distribution. Using non-parametric NHST alleviates this, but is more difficult to apply and interpret correctly. Furthermore, most standard approaches apply directly only to the mean/median and cannot be readily used to quantify the uncertainty of more complicated functions (for example, the root-mean-square error or a correlation coefficient).

In this paper we advocate the use of the bootstrap for quantifying the uncertainty in our estimates. The method was introduced by Efron [4] and is a representative of resampling-based methods, which also include, among others, permutation tests and cross-validation. Most of these methods have been known for decades, but have been limited by their computationally-intensive nature. Today, however, computation has become very cheap (in particular, very cheap compared to learning how to correctly perform classical statistical analysis), and the full potential and flexibility of resampling-based methods can be realized. Cross-validation has already been widely adopted by the machine learning community as a means for estimating how the model's performance will generalize to new and unseen data. Permutation tests and the bootstrap are, however, in our opinion, still heavily under-utilized. Both due to their usefulness and their pedagogical advantage over NHST.

Bootstrap is a resampling method used for estimating standard errors, bias, computing CI and for constructing NHST. Theoretically complicated methods are replaced by computer simulation, which is much more intuitive and easier to apply. Furthermore, it is more widely applicable - the process remains the same no matter what our statistic of interest is and number of assumptions is minimized. There are many different variants of the bootstrap, but the content of this paper applies to the most common one - the non-parametric bootstrap. We aim to provide a better understanding of the method, point out the advantages it offers, and demonstrate its weaknesses.

The remainder of the paper is organized as follows. In Section 2 we explain the idea behind the bootstrap. Section 3 offers a brief overview of bootstrap CI. In Section 4 we present empirical evidence of some of the method's characteristics, which are relevant for practical application. We see how changing the number of generated bootstrap samples affects the bootstrap distribution. We also compare coverage of different CI for mean, median and the 95-th percentile when our data come from different distributions and different sample sizes. With 5 we conclude the paper and offer some direction for further work.

2 The bootstrap

Suppose $\mathbf{x} = (x_1, ..., x_n)$ is a sample of size n from population F. Let θ be the population mean, our statistic of interest, and we want to compute the 95% CI.

Assuming normality and relying on the central limit theorem, our CI is $\pm 1.96 \cdot se$, where se is the standard error of the mean. But what happens if our statistic of interest is the median or correlation coefficient? What happens if our data are not distributed normally? We could

get those values from the sampling distribution of θ, but in order to obtain such a distribution, one we would need much more than just one sample, which is in practice typically impossible.

2.1 The basic idea behind the bootstrap

If **x** is a set of identically distributed and independent observations from F, it should be a good estimate of the population. Instead of sampling from the original population F, we will sample from an estimate of the population $\hat{F}$, which we will construct from **x**. That is, we will use the so-called *plug-in principle*, where we substitute an unknown variable for an estimate. Based on what type of bootstrap we choose, $\hat{F}$ can take different forms. For non-parametric bootstrap it is the empirical distribution function (EDF)

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=0}^{n} I(x_i = x), \tag{1}$$

which assigns probability $\frac{1}{n}$ to each data point (sampling with replacement). With sampling from $\hat{F}$ we obtain bootstrap samples. In total, there are n^n different samples. Evaluating all of them is known as the theoretical bootstrap. In practice, however, this is infeasible, so we use the Monte Carlo implementation and draw B (usually $B > 10^4$) bootstrap samples $\mathbf{x}^{*1}, ..., \mathbf{x}^{*B}$. For each of them we calculate the test statistic and get the bootstrap sampling distribution $\theta^{*1}, ..., \theta^{*B}$. In order for our results to reflect the actual data, we normally draw bootstrap samples of size n. On account of the EDF representing the whole population, the bootstrap distribution is usually too narrow by approximately $\sqrt{(n-1)/n}$ [6].

3 Bootstrap confidence intervals

Efron and Tibshirani [5] provide great insight into different ways the bootstrap can be used to compute CI. Below, we describe three: a normal distribution-based approach, a first-order accurate approach, and a second-order accurate approach[1].

3.1 Standard normal interval

We assume normality and construct the CI using the bootstrap standard error:

$$\hat{se}_b = \sqrt{\frac{\sum_{i=0}^{B}(\hat{\theta}^{*i} - \overline{\hat{\theta}^{*}})}{B-1}}. \tag{2}$$

This symmetrical CI does poorly on skewed data.

3.2 The percentile interval

We construct the percentile interval by using percentiles of the bootstrap distribution. If we want to calculate a $1-2\alpha$ percentile interval, we order our bootstrap estimates $\theta^{*1}, ..., \theta^{*B}$ and take the $100 \cdot \alpha$-th and $100 \cdot (1-\alpha)$-th values as limits of our interval:

$$[\hat{\theta}_{lo}, \hat{\theta}_{up}] = [\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}]. \tag{3}$$

The percentile interval is first order accurate and transformation invariant.

3.3 The BC_a interval

BC_a stands for bias-corrected and accelerated. This method also uses percentiles of the bootstrap distribution to calculate interval endpoints, but the percentiles used are based on the shape of bootstrap distribution. Two values determine this interval: the bias-correction factor $\hat{z}_0$ and the acceleration factor $\hat{a}$:

$$[\hat{\theta}_{lo}, \hat{\theta}_{up}] = [\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}], \tag{4}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right) \tag{5}$$

and

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right). \tag{6}$$

The bias correction factor measures the median bias of $\hat{\theta}^*$ and is calculated by plugging the proportion of bootstrap replications less than our observed statistic $\hat{\theta}$ into the inverse function of a standard normal cumulative distribution function

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=0}^{B} I(\hat{\theta}^{*(b)} < \hat{\theta})\right). \tag{7}$$

The acceleration factor measures the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter value θ:

$$\hat{a} = \frac{\sum_{i=0}^{n}\left((\sum_{i=0}^{n}\hat{\theta}_{(i)}/n) - \hat{\theta}_{(i)}\right)^3}{6\left[\sum_{i=0}^{n}\left((\sum_{i=0}^{n}\hat{\theta}_{(i)}/n) - \hat{\theta}_{(i)}\right)^2\right]^{3/2}}, \tag{8}$$

where $\hat{\theta}_{(i)}$ is a jackknife estimate of the statistic[2]. Note that if both factors equal zero, we get the percentile interval.

The method seeks to correct for bias and skew. It is second-order accurate, transformation invariant and range preserving. In the non-parametric setting it usually proves to be the most accurate [5].

4 Simulations

4.1 Number of bootstrap samples

Efron and Tibshirani [5] suggest that 25 bootstrap samples should be enough to get a good approximation of the standard error and 1000 samples for CI. Hesterberg [6], however, argues we should do at least 10^4 bootstrap samples to minimize sampling variability. Figure 1 illustrates how the number of bootstrap samples affect the bootstrap distribution.

[1] A CI is considered first-order accurate if the errors of the non-coverage probabilities differ from the true values by $O(n^{-\frac{1}{2}})$. It is second-order accurate if error size is of order $O(n^{-1})$.

[2] A jackknife estimate $\hat{\theta}_{(i)}$ is calculated from the original sample **x** with the i-th observation removed.

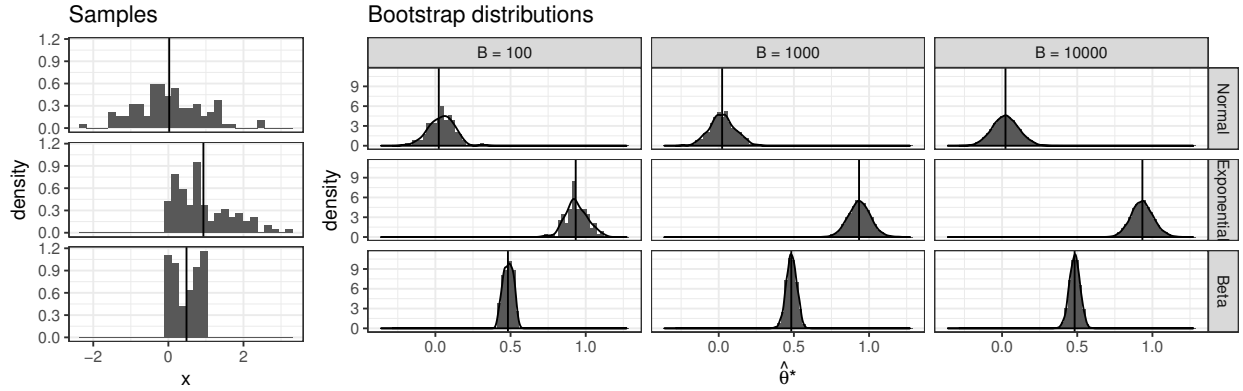

Figure 1: Changes in bootstrap distributions as the number of bootstrap samples B increases. Samples of size $n = 100$ are drawn from N(0,1) [top], Exp(1) [middle], and Beta(0.5, 0.5) [bottom]. Statistic of interest is the mean. Vertical lines are sample means.

4.2 Comparing confidence intervals

We applied 3 different variants of estimating the 95% CI to 3 distributions N(0,1), Exp(1), and Beta(0.5, 0.5) and estimated their coverage probability at different sample sizes n. The number of bootstrap samples was $B = 10^4$. The perfect CI would have coverage probability 0.95. We run the experiments with three statistics: mean, median and the 95-th percentile. To minimize variability our results are based on 10^4 repetitions for each n.

Simulation results in Figure 2 are in line with expectations. For $n \leq 25$ the bootstrap CI underestimate the CI. However, with increasing n, all three intervals converge to the specified coverage probability at their own rate. Overall, bootstrapped CI are biased (too narrow), which confirms what we previously stated and it needs to be taken into account when dealing with small n.

BC_a performs the best, but percentile intervals are almost as good and the advantage of BC_a decreases with increasing n. The normal interval performs best when the bootstrap distribution is roughly normal, so it does not perform well for the median and 95-th percentile. For the Beta distribution it even overestimates the CI, because in that case the booststrap distribution is bounded and concentrated at the upper bound.

4.3 Examples that arise in practice

To illustrate how the bootstrap can be expected to perform in practice, we prepared several examples of distributions that are the result of measuring the predictive performance of a machine learning model on a dataset:

- The distribution of the squared error obtained with leave-one-out cross-validation (LOOCV) for the *random forests* (RF) model on the *ozone* regression dataset [2]. From this distribution we bootstrap the mean and the rooted mean, effectively obtaining CI for the mean squared error (MSE) and root mean squared error (RMSE), respectively.
- The distribution of the difference in squared error obtained for the *lasso* and RF model in the same setting as above. We bootstrap the mean, effectively obtaining CI for how much the *lasso* is better/worse (in terms of mean squared error).

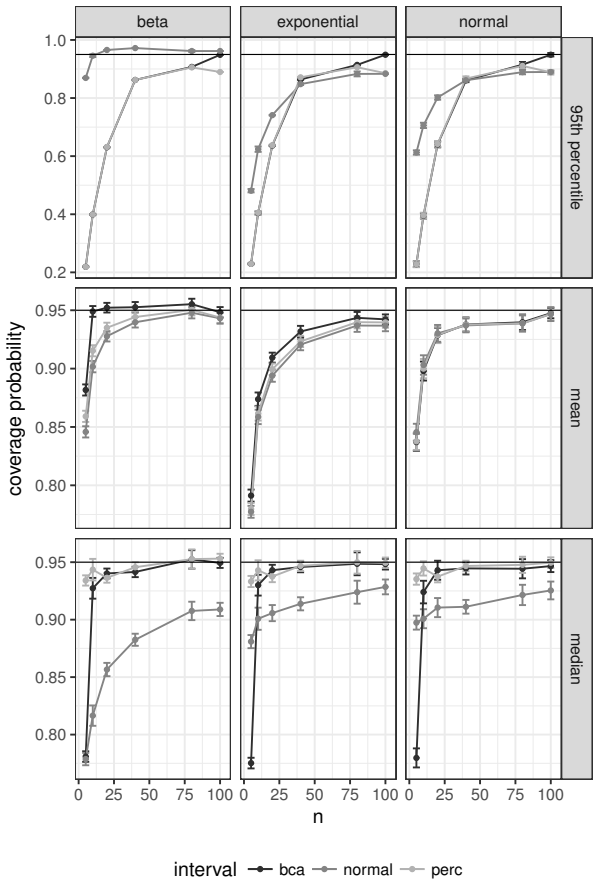

Figure 2: Observing how coverage probability changes when sample size n increases. The error bars denote 95% confidence in the interval coverages.

- The distribution of the 0-1 loss and the log-loss obtained with LOOCV for the RF model on the *Glass* dataset from the UCI Machine learning repository. We bootstrap the mean, effectively obtaining CI for the accuracy (CA) and log-score (LS).

The empirical distributions are shown in Figure 3. We pretend that these empirical distributions are the population distributions and obtain samples by resampling with

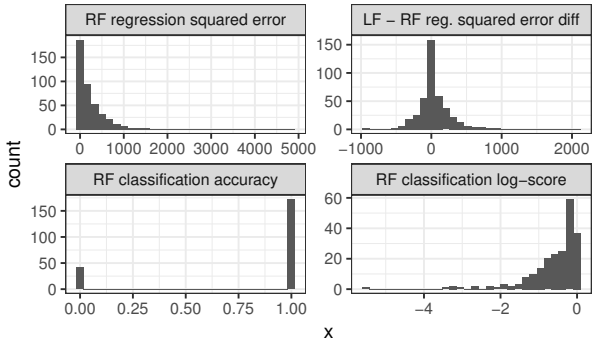

Figure 3: In practice, normally distributed performance data are not very common. 0-1 loss is distributed Bernoulli. Bounded variables, such as squared error and log-loss will often be close to exponential (or gamma) distributed. The difference of two such distributions will be close to a Laplace distribution.

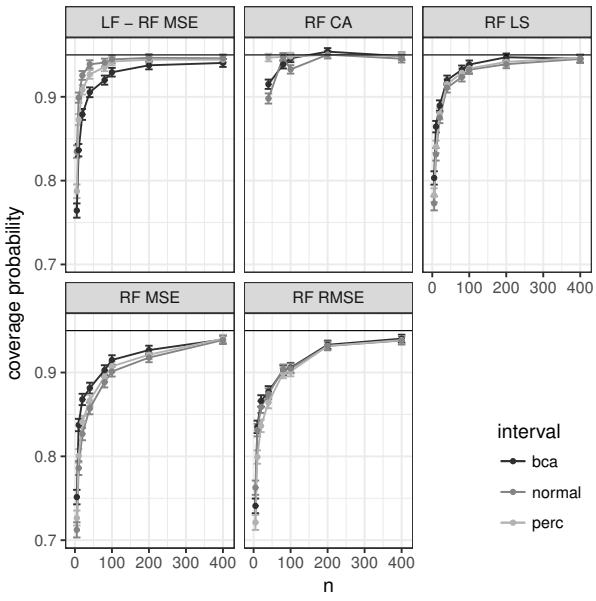

Figure 4: Coverage probabilities of CI when bootstrapping the empirical distributions.

replacement. This enables us to 'know' the true mean of the population and estimate coverage probabilities.

Again, we drew 10^4 bootstrap samples and calculated the CI based on 10^4 repetitions. Figure 4 shows that a larger sample size is needed for accurate CI for heavy-tailed distributions, such as the RF MSE. As expected, the normal interval's performance is relatively poor when the data are discrete and far from normal. The binary sample can also pose a problem when the sample size is too small (in our case $n < 40$), as sampling with replacement can result in a trivial distribution with zero variability. BCa performs the worst for a Laplace-like distribution (LF - RF). Similar results were observed in [8, 9].

5 Conclusion

The bootstrap has several advantages over classical approaches: it is easy to understand, implement, and apply and it can be applied to almost any statistic of interest.

It can fail in practice with very small sample size n and/or problematic statistics, such as extreme quantiles. However, simulations show that this is not a serious problem as long as we have at least $n \geq 50$.

BCa CI perform the best, but percentile intervals also perform well-enough. We should avoid normal intervals unless the bootstrap distribution is expected to be close to normally-distributed. Today, computation is relatively cheap, and the number of bootstrap samples B should be as large as possible, but simulations show that even 1000 is good enough for almost all practical situations.

Recently, there has been a shift towards Bayesian methods for model comparison [1]. Bayesian methods are in many ways more intuitive and therefore easier to interpret than frequentist ones. We did not touch on Bayesian approaches as the problems of frequentist approaches also apply to Bayesian approaches (distributional assumptions, problems when we are interested in more than just the mean). If one prefers the Bayesian view of probability, one can always apply the Bayesian boostrap [7].

References

[1] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

[2] Rok Češnovar and Erik Štrumbelj. Bayesian lasso and multinomial logistic regression on gpu. *PLoS ONE*, 12(6):e0180343, 2017.

[3] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[4] Brad Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979.

[5] Brad Efron and Rob Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, Inc., 1993.

[6] Tim C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, 2015.

[7] Donald B. Rubin. The Bayesian bootstrap. *The annals of statistics*, pages 130–134, 1981.

[8] Klairung Samart, Naratip Jansakul, and Mitchai Chongcheawchamnan. Exact bootstrap confidence intervals for regression coefficients in small samples. *Communications in Statistics - Simulation and Computation*, 2017.

[9] Yaqian Zhu and John Kolassa. Assessing and comparing the accuracy of various bootstrap methods. *Communications in Statistics - Simulation and Computation*, 2017.