Pamukkale Univ Muh Bilim Derg, 26(1), 174-183, 2020

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

An empirical look at the variation associated with bootstrap estimates of location parameters

Konum parametrelerinin bootstrap tahminleri ile ilişkili varyasyona ampirik bir bakış

Levent ERİŞKİN[1*]

[1]Industrial Engineering Department, Turkish Naval Academy, National Defense University, İstanbul, Turkey.
leriskin@dho.edu.tr

Received/Geliş Tarihi: 26.12.2018, Accepted/Kabul Tarihi: 04.07.2019
doi: 10.5505/pajes.2019.39260
* Corresponding author/Yazışılan Yazar
Research Article/Araştırma Makalesi

Abstract

Bootstrap is a technique for estimating standard error and bias of the statistic of interest. The idea behind the bootstrap technique is that bootstrap distribution generated by resampling from the sample at hand mimics the sampling distribution of the statistic. Nevertheless, the effect of sample size and number of bootstrap replications on the accuracy of bootstrap predictions is rarely considered and ignored while applying bootstrap. Although there exist limited studies on this matter in the literature, results obtained in these studies are expressed based on the population distribution. In this paper, we provide results of an empirical study that examines the relationship between sample size and number of bootstrap replications and standard errors of bootstrap estimates of location parameters for different population distributions. To that end, we focus on the representativeness of bootstrap distribution to sampling distribution for different continuous and discrete population distributions and different sample sizes, firstly. According to application results, we observe that sample size has more impact on accuracies of bootstrap estimates as regards to number of bootstrap replications. Additionally, we confirm that bootstrap distributions of median based on small sample sizes are inadequate for representing sampling distribution. Lastly, in order to model relationship between standard errors of bootstrap estimates and sample size and number of bootstrap estimations independently of population distribution, we propose a methodology based on jackknife-after-bootstrap technique and regression modeling.

Keywords: Bootstrap, Bootstrap repetitions, Jackknife-after-bootstrap, Sampling variability, Resampling variability

Öz

Bootstrap bir istatistiğin standart hatasını ve yanlılığını tahmin etmek üzere kullanılan bir tekniktir. Bootstrap tekniği; eldeki örneklemden yeniden örnekleme ile üretilen bootstrap dağılımının, istatistiğin örneklem dağılımını temsil edeceği ana fikri üzerine kuruludur. Buna karşın; bootstrap uygulanırken örneklem büyüklüğünün ve bootstrap yineleme sayısının bootstrap tahminlerinin doğruluğuna olan etkisi genelikle dikkate alınmamakta ve ihmal edilmektedir. Her ne kadar literatürde bu konuyu ele alan sınırlı sayıda çalışma olsa da, bu çalışmalarda elde edilen sonuçlar örneklemin alındığı ana kütle dağılımına bağımlı olarak ifade edilmektedir. Bu makalede, örneklem büyüklüğü ve bootstrap yineleme sayısı ile konum parametrelerinin bootstrap tahminlerinin standart hataları arasındaki ilişkiyi farklı ana kütle dağılımları için inceleyen ampirik bir çalışmanın sonuçları sunulmaktadır. Bu maksatla öncelikle farklı sürekli ve kesikli dağılımlardan çekilmiş farklı büyüklüğe sahip örneklemlere uygulanan bootstrap işlemi sonrası bootstrap dağılımının örneklem dağılımını ne oranda temsil ettiği incelenmektedir. Uygulama sonucunda, bootstrap tahminlerinin doğruluğuna örneklem büyüklüğünün bootstrap yineleme sayısına göre daha fazla etki ettiği görülmüştür. Ayrıca, medyana ilişkin bootstrap dağılımlarının özellikle küçük örneklemler için örnekleme dağılımını temsil etmede oldukça yetersiz olduğu tespit edilmiştir. En son olarak da bootstrap tahminleri standart hataları ile örneklem büyüklüğü ve bootstrap yineleme sayısı arasındaki ilişkinin ana kütle dağılımından bağımsız olarak tahmin edilebilmesi için jackknife-sonrası-bootstrap tekniği ve regresyon modeli tabanlı bir yöntem önerilmektedir.

Anahtar kelimeler: Bootstrap, Bootstrap yinelemesi, Jackknife-sonrası-bootstrap, Örnekleme değişkenliği, Yeniden örnekleme değişkenliği

1 Introduction

A statistical inference is based on the sampling distribution of sample statistics. A sampling distribution of a statistic is the distribution of the statistic for all possible samples of some size n from the same population. If we knew the sampling distribution of a statistic, then we would be able to draw conclusions about the behavior of the statistic under random sampling. When the statistic of interest is the mean, then the Central Limit Theorem states that the sampling distribution follows $N(\mu, \sigma^2/n)$ provided that the population is normal, or sample size is large. However, life is usually not that easy; sometimes the statistic of interest is median, trimmed mean or quantiles. We do not have smooth approximations for these statistics. Additionally, most of the time we have no idea about the population from which the sample comes. In these situations, bootstrap methods help. First introduced by Efron [1], bootstrap method is a kind of Monte Carlo simulation where sample at hand is treated like a pseudo-population. The sample serves as an estimate of the population. Many resamples are drawn from the pseudo-population (i.e. the sample) with replacement and the statistic of interest is calculated each time. The distribution of the statistic calculated from resamples forms the bootstrap distribution of the statistic. What makes bootstrapping special is that we do not have any assumption about the underlying population and it applies to any statistic other than mean. The idea behind the bootstrap technique is that bootstrap distribution generated by resampling from the sample at hand mimics the sampling distribution of the statistic. Hence, we draw conclusions about the statistic by looking at the bootstrap distribution.

Since the main purpose of bootstrap is to estimate standard errors and bias, it has found a wide range of application in building confidence intervals for the statistics of interest. The

works of Efron and Tibshirani [2], Efron [3], DiCicio and Efron [4] provided good examples to application of bootstrap procedure to confidence interval building. In his study Pawitan [5] showed how to construct a bootstrap likelihood from a single bootstrap, without any nested bootstrapping nor any smoothing. Hall [6] provided a theoretical comparison of bootstrap confidence intervals. In another study Hall [7] described the connection between Edgeworth expansions and the bootstrap. Hesterberg et al. [8] provided a complete overview to the bootstrap methods and permutation tests. In their work they posed the question: How accurate is a bootstrap distribution? In order to answer this question, they compared empirical sampling and bootstrap distributions by looking at their shapes. They also examined the effect of sample size on the variation of bootstrap estimates.

Regarding application of the bootstrap to finite populations, Lo [9] proposes Bayesian analogue of finite population bootstrap (FPB). He suggests that finite population Bayesian bootstrap can be defined in terms of Polya's urn scheme. He shows that FPBB and FPB present similar operational characteristics and for a large population size FPNN reduces to the Bayesian bootstrap. Booth et al. [10] present bootstrap applications for finite population sampling problems. They focus on estimating the distribution function of a Studentized estimate of a population mean. They show that second-order accurate bootstrap estimates can be generated for the distributions of stratified sample means, separate ratio estimates, and other estimates of a finite population mean. They also conduct Monte-Carlo simulation to present performance of the proposed method. Shao [11] discusses the impact of the bootstrap on sample surveys. In sample surveys, the original data is generally sampled without replacement from a finite population. In this paper Shao [11] explains why bootstrap is so important in sample surveys and presents developments about bootstrap applications in this field. In his study Aitkin [12] extends the Bayesian bootstrap analysis to regression models for numerically-valued response variables in stratified and clustered samples. He discusses disadvantages of Bayesian bootstrap approach and remarks that these disadvantages are shared with survey sampling analysis, as well. Antal and Tille [13] remark that when sampling design is not considered, classical bootstrap methods tend to produce biased variance estimators. In order to overcome this problem, they propose novel resampling methods where they select subsamples under a completely different sampling scheme from that of original sample. They show that their technique generates unbiased estimators of variance.

Some other studies focused on estimating standard errors of the bootstrap estimations and selection of proper n and B figures in order to reduce variability. In his prominent study, Efron [14] described the jackknife-after-bootstrap procedure where he used Tukey's jackknife technique to attach standard errors to bootstrap estimates. In his paper he also proposed using bootstrap-after-bootstrap to compute standard errors, however he ended up recommending using jackknife-after-bootstrap due to the efficiency of the latter. Hill, Cartwright and Arbaugh [15] considered assessing the reliability of the bootstrap using jackknife-after-bootstrap procedure. They examined the accuracy of the jackknife-after-bootstrap using Monte Carlo experiments. They used jackknife-after-bootstrap in the context of three statistical models: the model of the mean of a normal population, the linear regression model and the seemingly unrelated regression model. They found that jackknife-after-bootstrap overestimates the standard error by a large amount when $B \leq 200$. With B = 10,000 jackknife-after-bootstrap estimates are very much reliable. Andrews and Buchinsky [16] proposed a three-step method for choosing the number of repetitions (B) in order to have reliable bootstrap estimates. Reliability is defined in terms of accuracy which is measured by the percentage deviation of the bootstrap standard deviation estimate based on B bootstrap simulations from the corresponding quantities for which $B = \infty$. Davidson and MacKinnon [17] pointed out that using finite number of bootstrap samples cause a loss of power, hence, they propose a pretest procedure for choosing the number of bootstrap samples in order to minimize experimental randomness. Regarding the number of bootstrap repetitions required, Lunneborg [18] suggested continuing drawing resamples until successive standard error estimates vary by less than 1%. Pattengale et al. [19] proposed a threshold value, which they call as stopping criteria, used to determine if enough bootstrap samples are generated.

Nowadays computation power has improved enormously. Therefore, focusing on the number of bootstrap replications necessary for reliable bootstrap estimates might be pointless. As Chernick [20] emphasized, it may be silly to argue between 100 and 800 iterations while it is easy to bootstrap 5,000 to 10,000 iterations. Even though it is true, we still need to know how variability of bootstrap estimates relies on n and B. In this respect, Efron and Tibshirani [21] give approximate forms for the variance of bootstrap estimate of standard error and percentiles. In their study they provide theoretical closed-form solutions for the relationship between variance of bootstrap estimate of standard error and n and B. The main problem with this form is that unknown constants are distribution dependent and they are not easy to obtain. A distribution-dependent estimate contradicts the idea of bootstrap which is proposed mainly as a remedy for unknown population distributions.

Considering this fact, the main aim of this paper is to provide an empirical study that examines the relationship between n and B and standard errors of bootstrap estimates. Additionally, we propose a simple methodology to build this relationship empirically. The paper is organized as follows: Section 2 provides theoretical background for bootstrap and jackknife-after-bootstrap methods. In section 3 we present an application to examine the representativeness of bootstrap distribution to sampling distribution for different population distributions. Sources of variation of bootstrap estimates are examined in Section 4. Last section concludes our study.

2 Methods

2.1 Bootstrap

The bootstrap is a method to calculate standard error and bias of the statistic of interest where we make no assumptions about the population from which the sample is drawn. Rather, we treat the sample at hand as the pseudo-population and draw samples from it with replacement. In other words, the sample is used to estimate the population. Let's illustrate it with an example. Suppose we are interested in the mean of some unknown population and we have a sample of size n. Say our sample is $\mathbf{x}$ = (6, 8, 1, 7, 9). We resample from this sample with replacement and calculate the statistic of interest. Two of our resamples could be $\mathbf{x}^{*1}$ = (9, 8, 1, 9, 6) and $\mathbf{x}^{*2}$ = (7, 6, 1, 9, 7) then the bootstrap estimates for the mean would be $\hat{\bar{x}}^{*1} = 6.6$ and $\hat{\bar{x}}^{*2} = 6$, respectively. We replicate this procedure B times,

where B is a large number. Calculated values form the bootstrap distribution of the statistic. As mentioned before, the bootstrap distribution mimics the sampling distribution. By using the bootstrap distribution, we calculate the standard error and bias of the statistic. Let θ be the statistic of interest (i.e. mean), the bootstrap estimate of standard error can be formulated as;

$$S\hat{E}_B(\hat{\theta}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*b} - \bar{\hat{\theta}}^{*} \right)^2 \right\}^{\frac{1}{2}} \quad (1)$$

where

$$\bar{\hat{\theta}}^{*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b} \quad (2)$$

Based on this idea, standard errors of other statistics of interests can be easily calculated with this approach. Additionally, bootstrap confidence interval of a statistics can be found by;

$$(\hat{\theta} - z^{\left(1-\frac{\alpha}{2}\right)} SE_{\hat{\theta}}, \hat{\theta} - z^{\left(\frac{\alpha}{2}\right)} SE_{\hat{\theta}}) \quad (3)$$

where $SE_{\hat{\theta}}$ is the standard error of the statistic of interest found by using bootstrap.

2.2 Jackknife-After-Bootstrap

Bootstrap estimates of standard error and bias are also estimates, meaning that they also have errors associated with them. Jackknife-after-bootstrap procedure was introduced by Efron [14] to calculate errors associated with the bootstrap estimates. Say we estimated the standard error of some statistic (θ) using the bootstrap method. Then we would like to measure the uncertainty of the standard error (var($s\hat{e}_B$)). In order to compute jackknife-after-bootstrap estimate of the variability of $s\hat{e}_B$, we leave out one data point at a time and calculate ${s\hat{e}_B}^{(-i)}$ using the bootstrap method on the remaining n-1 points. We repeat this procedure till we have all ${s\hat{e}_B}^{(-i)}$ values [22]. The jackknifes-after-bootstrap estimate of variance can be formulated as;

$$\widehat{var}_{jack}(s\hat{e}_B) = \frac{n-1}{n} \sum_{i=1}^{n} \left(s\hat{e}^{(-i)} - \overline{s\hat{e}} \right)^2 \quad (4)$$

where

$$\overline{s\hat{e}} = \frac{1}{n} \sum_{i=1}^{n} s\hat{e}^{(-i)} \quad (5)$$

In order to improve the efficiency of the procedure, we use the original bootstrap samples to apply jackknife instead of producing a new set of bootstrap samples where we leave out a data point. We find bootstrap samples which do not contain the point x_i. These are the bootstrap samples used to calculate ${s\hat{e}_B}^{(-i)}$.

3 Application

3.1 Sampling and bootstrap distributions

The idea behind the bootstrapping lies in the assumption that bootstrap distribution mimics the sampling distribution of the statistic. In order to verify this assumption, we need to compare these two distributions in terms of spread, shape and center. If these two distributions are similar, then we can use bootstrap distribution as a substitute for sampling distribution to draw conclusions about the statistic. In our application, we will sample from three different population distributions. First population distribution is normal with mean 8 and standard deviation 2. Second population distribution is exponential with mean 5. Another population distribution is a multimodal distribution generated with mixture of two distributions; N(5, 2) and N(10, 2). Apart from these continuous distributions, we also consider Poisson distribution with mean rate 5. The probability density/mass distributions of these population distributions are displayed in Figures 1 to 4.

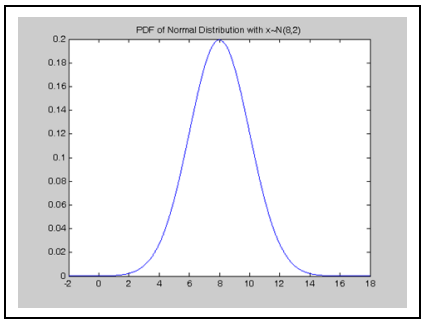

Figure 1: Normal distribution used in the application.

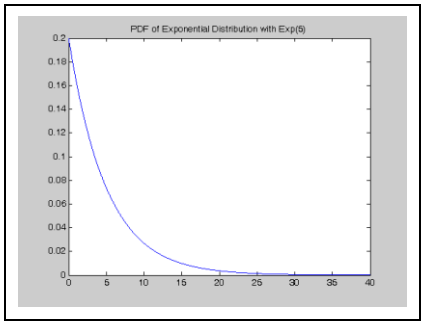

Figure 2: Exponential distribution used in the application.

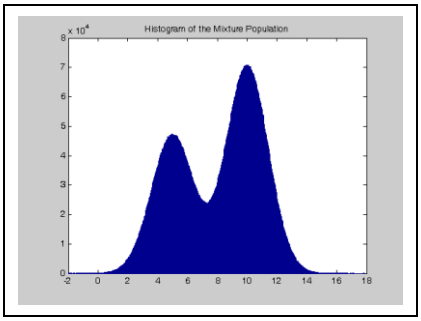

Figure 3: Mixture population used in the application.

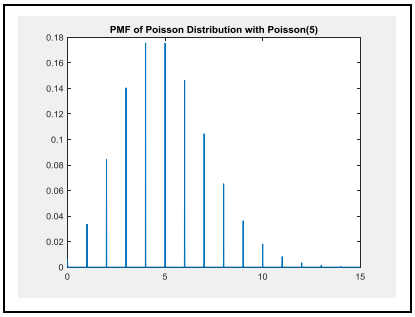

Figure 4: Poisson population used in the application.

Normal distribution has some nice properties regarding sampling distributions of the statistics. For example, if the population is normal, Central Limit Theorem holds even for small sample sizes. Normal population mean and median are the same, therefore, sampling distributions of both statistics are centered at the same common number. In order to see what happens when population is nonnormal, we also considered exponential population and a multimodal mixture-generated population. A Poisson population is included in the study so that we could observe performance of bootstrapping with respect to discrete populations. Additionally, in order not to be distracted by the nice properties of the mean, we also considered median as the statistic of interest. In order to obtain empirical sampling distributions, we conducted a Monte Carlo procedure where we drew large number (i.e. 1,000,000) of samples of size 50 from the populations and calculated the statistics of interest for each sample. The calculated statistic values form the sampling distributions. Table 1 summarizes the values of the statistics of these three distributions, while Figures 5 and 6 display sampling distributions of mean and median with respect to these three populations.

Table 1: Means and medians of the population distributions.

Statistic	Normal	Exponential	Mixture	Poisson
Mean	8.000	5.000	8.000	5.000
Median	8.000	3.466	8.650	5.000

Even though exponential sampling distribution is somewhat positively skewed, we see that four sampling distributions of mean are close to normal regardless of the underlying population. This is in harmony with the Central Limit Theorem. Sample size in the experiment is 50; we expect four sampling distributions to become more normal as sample size increases. Three sampling distributions are centered at the population means, as expected. When we considered sampling distributions of median, we observe that only the first one of three continuous distributions is close to normal while others are skewed. Sampling distribution of the median for Poisson distribution is rather discrete and far from being normal. This stems from the fact that, median of a sample is very much dependent on the limited number of middle values of the sample and sample quantiles for discrete populations are not consistent for the population quantiles, in general [23].

In order to generate bootstrap distributions, we first drew one sample of size 50 from each of the population and then applied bootstrap procedure. In the bootstrap procedure the number of resamples is 1000. Table 2 gives the sample means and medians, while Figures 7 and 8 present histograms of the bootstrap distributions generated from these samples.

Table 2: Means and medians of the samples.

Statistic	Normal	Exponential	Mixture	Poisson
Mean	8.426	5.435	8.280	4.960
Median	8.427	4.772	9.155	5.000

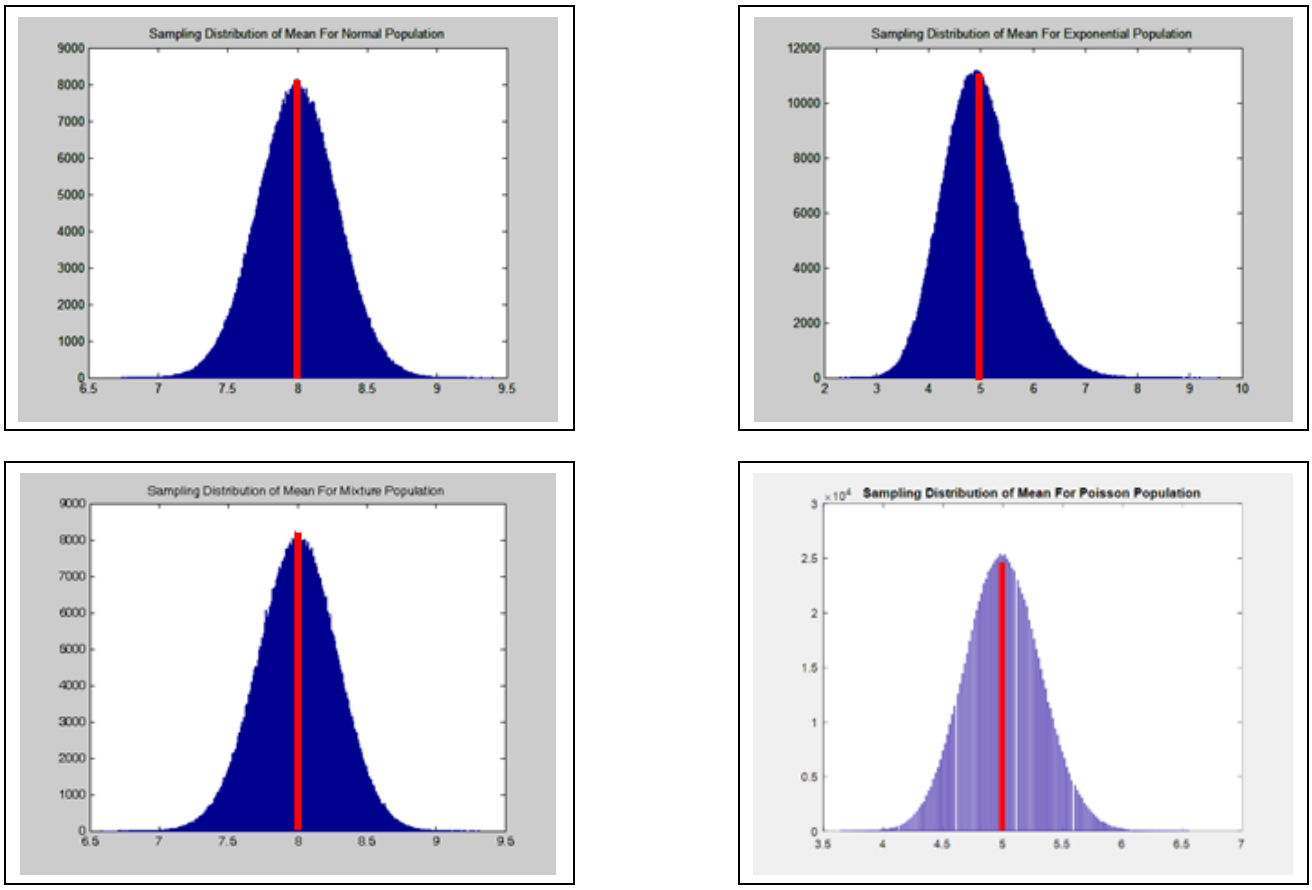

Figure 5: Sampling distributions of mean with respect to 4 populations.

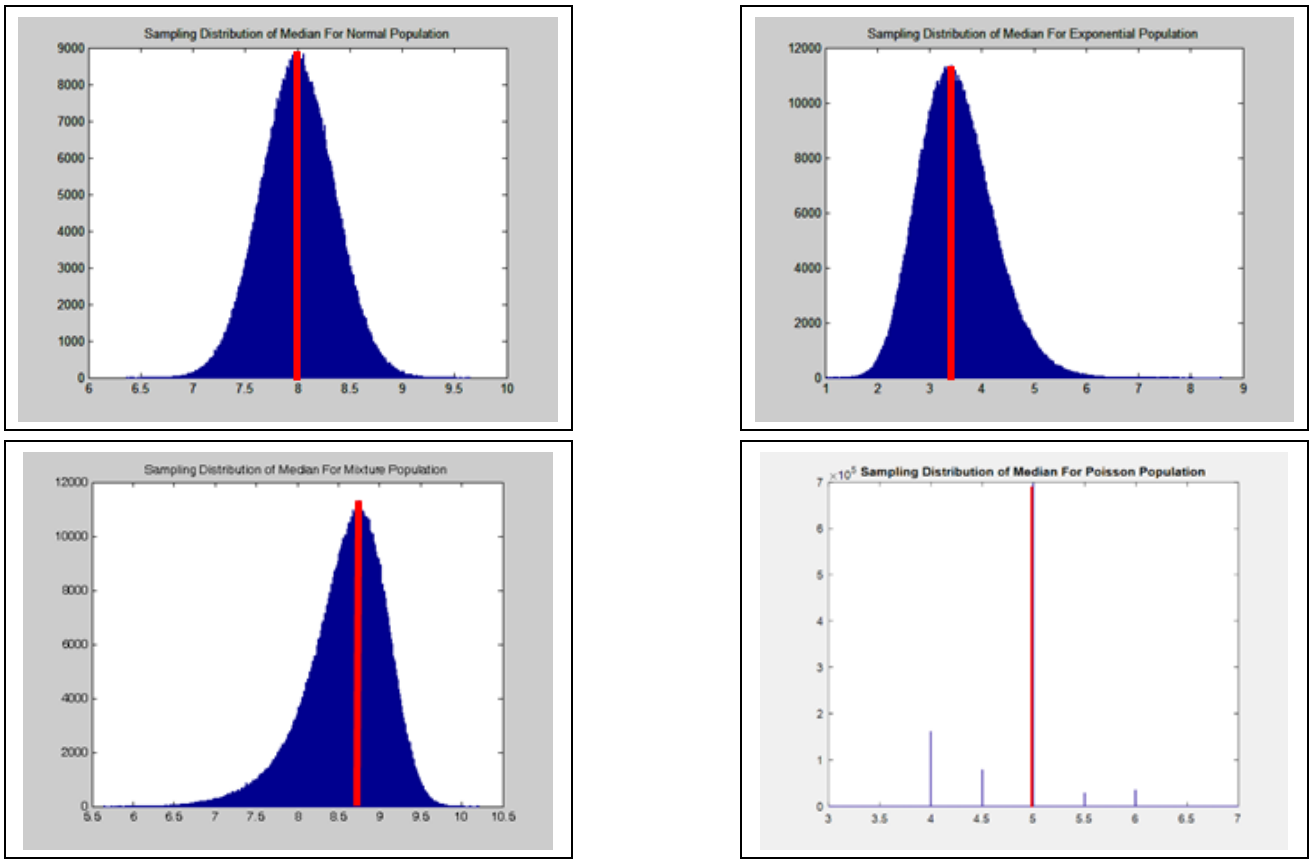

Figure 6: Sampling distributions of median with respect to 4 populations.

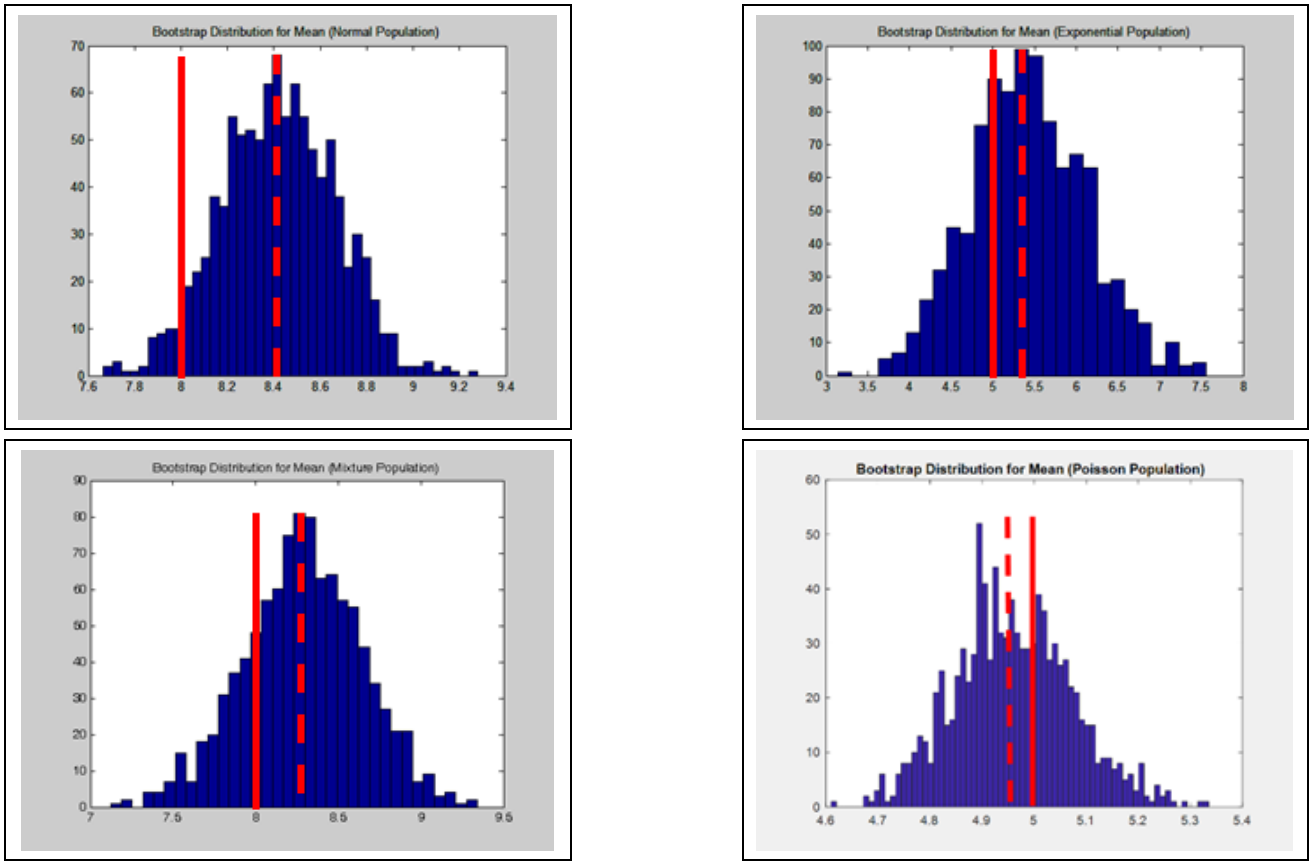

Figure 7: Bootstrap distributions of mean with respect to 4 samples.

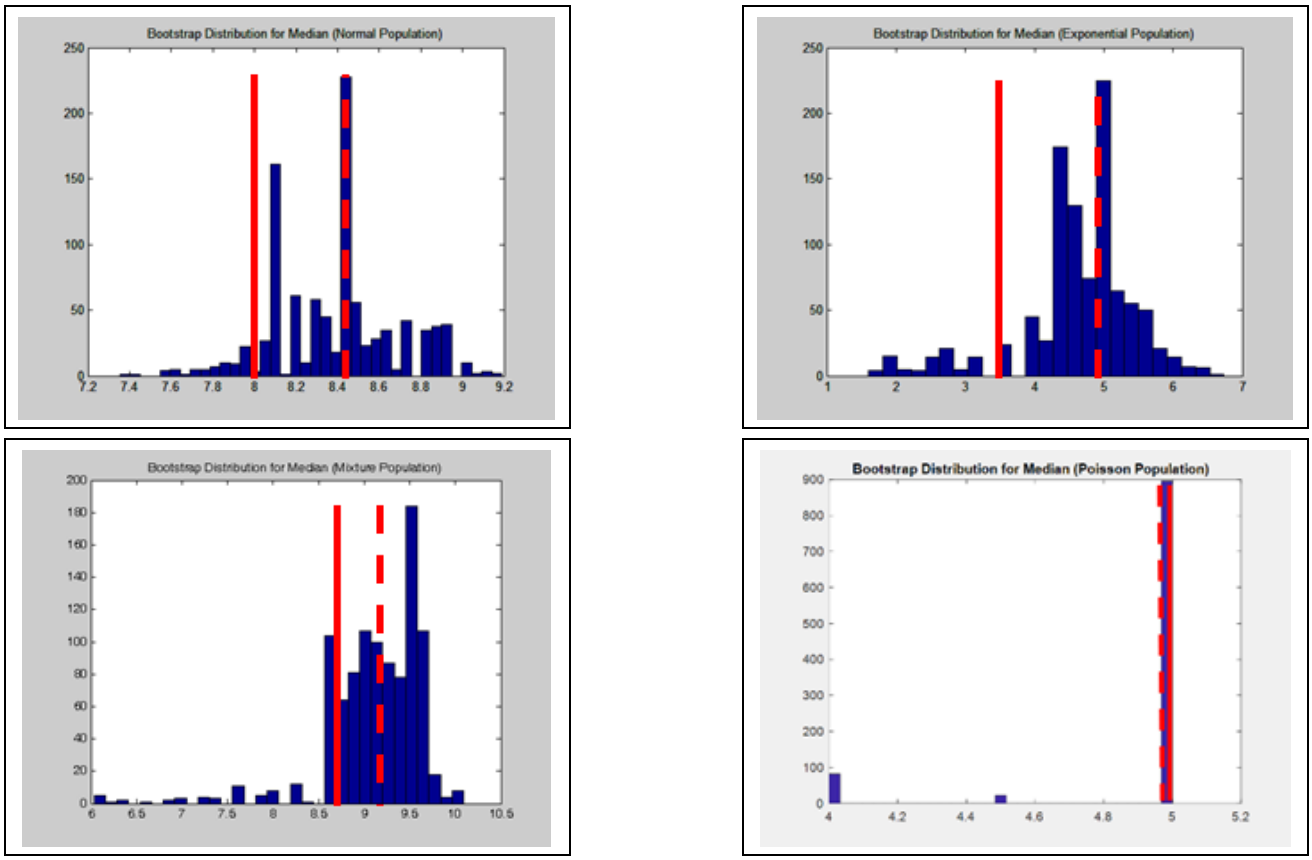

Figure 8: Bootstrap distributions of median with respect to 4 samples.

In the figures solid lines indicate population means and population medians while dashed lines indicate sample means and sample medians. We observe that bootstrap distributions of mean and median are centered at the sample means and sample medians rather than population means and population medians. This is the consequence of bootstrap procedure. Because bootstrap procedure treats the sample like the pseudo-population. Therefore, the bootstrap distribution is centered at the sample statistics. As number of resamples increases and approaches to infinity, the bootstrap distribution center converges to the sample center. The bootstrap distributions of medians are not good in shape in terms of representability of the sampling distributions. This is because of the sample size. The bootstrap procedure for the median requires quite large sample sizes. Value of the median depends on the middle observations of the sample. When we resample with replacement from the same sample, we repeat the same few middle observations. Conversely, sampling distribution contains the medians of all possible samples and is not confined to a few values [8]. Bootstrap distribution of median for the Poisson sample suffers from the same problem as sampling distribution does. Central Limit Theorem does not hold true for median and bootstrap distribution does not consistently mimic distribution of the sample quantiles [23]. To overcome this problem, Jentsch and Leucht [23] propose two different strategies. On the other hand, bootstrap distributions of means are similar to sampling distributions in terms of shape and spread.

In order to compare bootstrap distributions and sampling distributions, we employ QQ plots. If the bootstrap distributions and sampling distributions are similar in shape and spread, we expect the quantiles fall on a straight line in the QQ plot. Figures 9 and 10 are QQ plots of bootstrap and sampling distributions of the statistics.

QQ plots for the bootstrap and sampling distributions of mean propose that bootstrap distribution mimics sampling distribution quite well for all three populations. This means that it is safe to use bootstrap distribution in order to draw conclusions about the statistic. On the other hand, when we consider the QQ plots for the median, we observe that only the QQ plot corresponding to normal population proposes the same inference. QQ plots corresponding to exponential and mixture populations are bad in shape, contradicting the theory. We explained the reason why we expect to have such a picture, previously. We need to have more observations in the sample in order to rely on the bootstrap distributions of the median.

For the discrete population case, on the other hand, QQ plot for the Poisson bootstrap and sampling distributions presents a staircase pattern, as expected. Additionally, this plot proposes that quantiles of these two distributions do not match well enough. This is because bootstrap distribution inherits inconsistency of sampling distribution for the discrete cases. Interested reader may refer to Jentsch and Leucht [23] for details and solution methodologies in this subject.

QQ plots for the bootstrap and sampling distributions of mean propose that bootstrap distribution mimics sampling distribution quite well for all three populations. This means that it is safe to use bootstrap distribution in order to draw conclusions about the statistic. On the other hand, when we considered the QQ plots for the median, we observe that only the QQ plot corresponding to normal population proposes the

same inference. QQ plots corresponding to exponential and mixture populations are bad in shape, contradicting the theory. We explained the reason why we expect to have such a picture, previously. We need to have more observations in the sample in order to rely on the bootstrap distributions of the median.

To see if the bootstrap distribution of median gets better in terms of representability of the sampling distribution as sample size gets larger, we performed the same experiment with larger samples. For the reasons we explained previously, we exclude Poisson population from the further analysis. We replicated the procedure for sample sizes 100, 200 and 500. Figures 11-13 display the results.

We observe that QQ plots get better as sample size increases. When sample size 100 and 200, we cannot claim that bootstrap distribution mimics sampling distribution as desired. However, as sample size increases to 500, representation becomes quite good. When $n = 500$, it's safe to use bootstrap. As a consequence, we need to be careful when applying bootstrap to statistics like median, which relies on one or two observations from the sample.

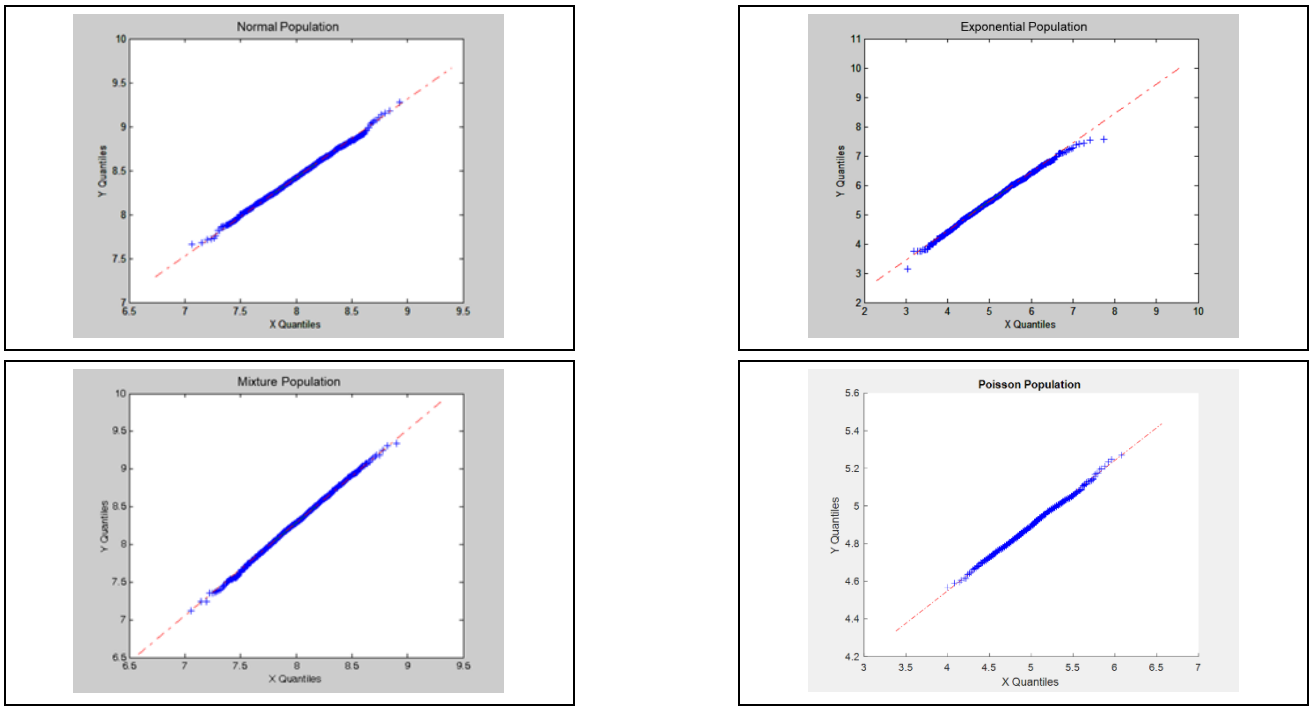

Figure 9: QQ plots of the bootstrap and sampling distributions of the mean.

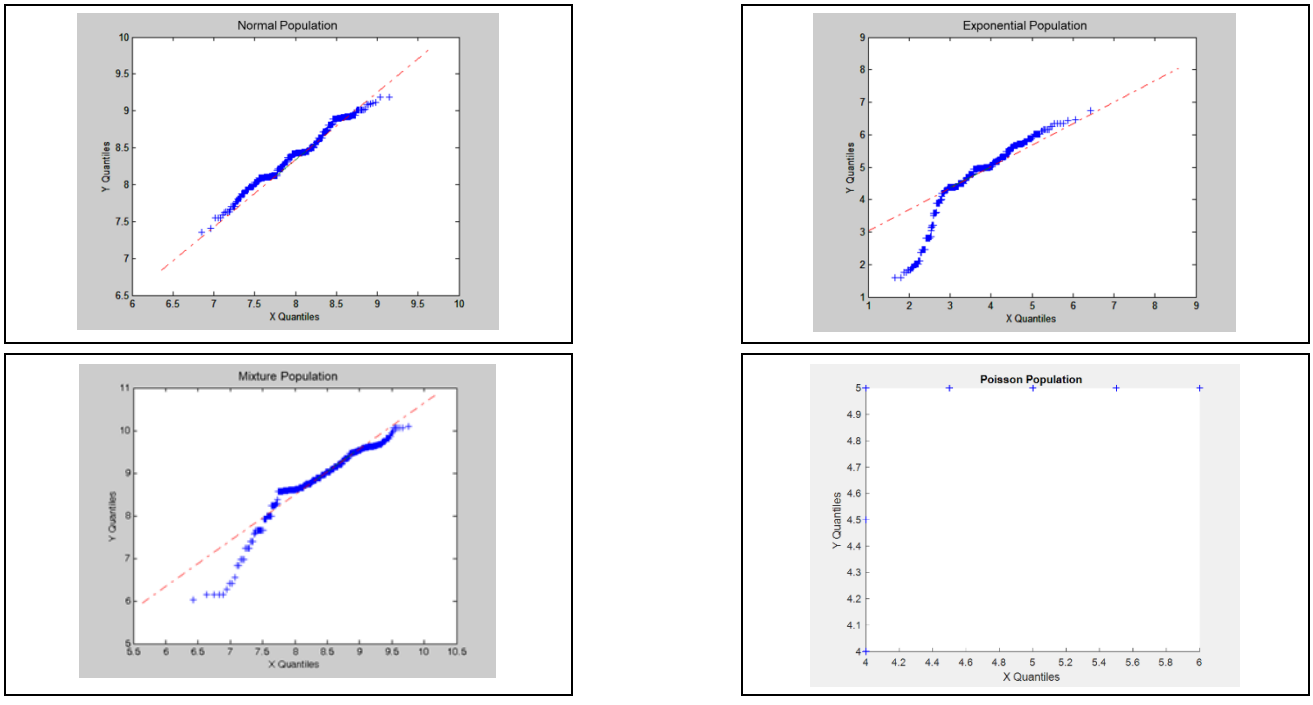

Figure 10: QQ plots of the bootstrap and sampling distributions of the median.

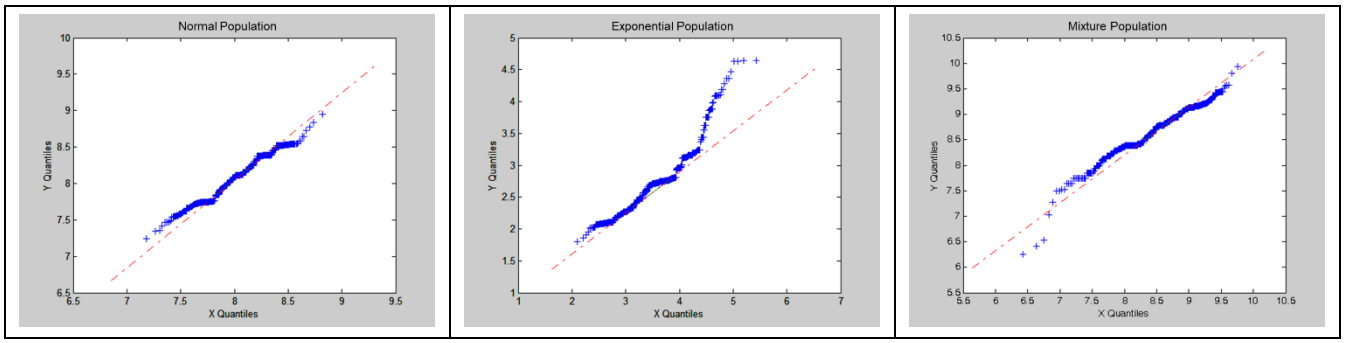

Figure 11: QQ plots of the bootstrap and sampling distributions of the median (Sample Size 100).

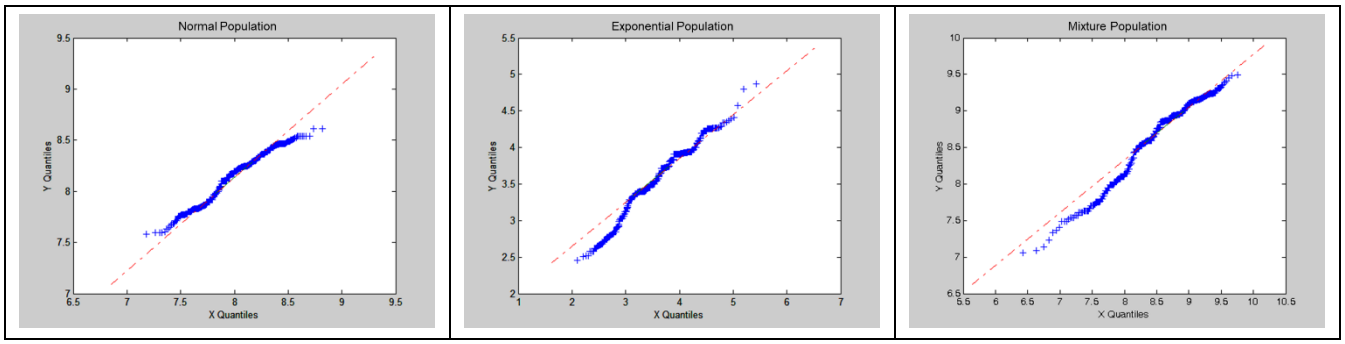

Figure 12: QQ Plots of the Bootstrap and Sampling Distributions of the Median (Sample Size 200).

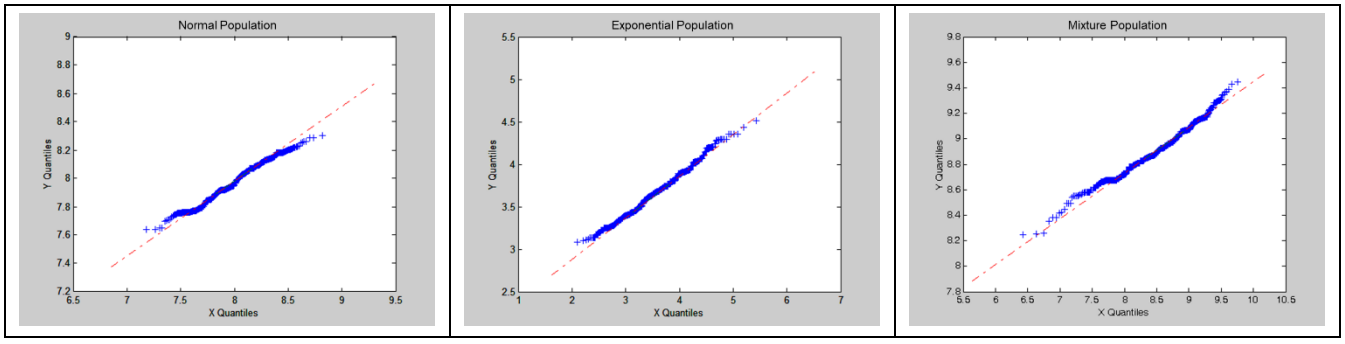

Figure 13: QQ plots of the bootstrap and sampling distributions of the median (Sample Size 500).

3.2 Sources of Variation

The sampling distribution of a statistic contains the variation in the statistic because samples are drawn randomly from the population. Since we treat bootstrap distribution as a substitute for the sampling distribution, this adds a second source of random variation. Therefore, bootstrap estimates have two sources of variation associated with them [21]:

- Sampling variability: Arises since we have a sample of size n rather than the entire population.
- Bootstrap resampling variability: Arises since we take only B bootstrap samples rather than an infinite number.

Figure 14 shows the sampling and resampling components of variance. As a consequence, choice of n and B has impact on the accuracy of the bootstrap estimates.

Efron and Tibshirani [21] gave approximate forms for the variance of bootstrap estimate of standard error and percentiles and showed that jackknife-after-bootstrap procedure can be used to estimate variation associated with bootstrap estimates provided that number of bootstrap replicates (B) is big enough.

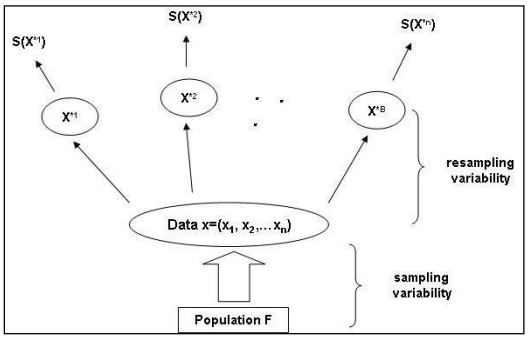

Figure 14: Components of Variance for the Bootstrap Estimates (Source: Efron and Tibshirani [21]).

As an example, for the bootstrap estimate of standard error, variance associated with this estimate has the form

$$var(se_B) = \frac{c_1}{n^2} + \frac{c_2}{nB} \quad (6)$$

Where c_1 and c_2 are constants depending on the underlying population F. First part of the variance corresponds to sampling variation while the second part represents the resampling variation. It is clear that sampling variability is function of the sample size and resampling variability is function of both the number of bootstrap replicates and the sample size. It is intuitive to assert that resampling variability adds little variation to the bootstrap estimates because we are only limited with analyst's time and power of the computer when selecting number of resamples. On the other hand, most of the time we do not have control over the sample size. This can also be shown empirically. Hesterberg et.al [8] drew five random samples from a mixture distribution and drew 1000 resamples from one of these samples. After drawing histograms of these samples and bootstraps distributions, they pointed out that each bootstrap distribution is centered close to mean of the original sample. Additionally, they are very similar to the sample from which they are drawn in terms of shape, center and spread. On the other hand, each of 5 samples vary in terms of center, shape and spread. This analysis shows that bootstrapping adds little variation to the bootstrap estimates.

Can we visualize the variation of bootstrap estimates for a given sample size and number of resamples combination? Figure 15 shows the relationship between variance of the bootstrap estimate of standard error of the mean and these two parameters. The variance figures are computed empirically with jackknife-after-bootstrap procedure.

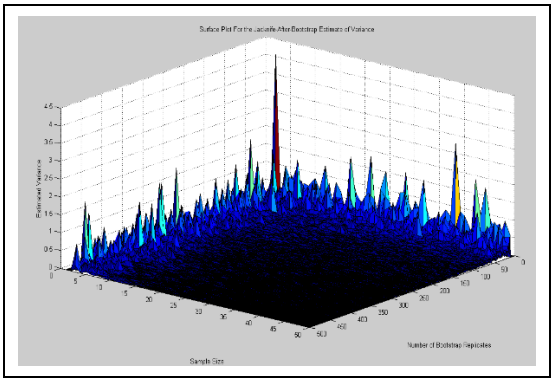

Figure 15: 3D plot for the jackknife-after-bootstrap variance of the bootstrap estimate of standard error of the mean.

The variance of the bootstrap estimate decreases as the sample size and the number of resample size increases. Even though small sample size and small number of resample size increase the variance, we can assert that sample size is more critical between two, since achieving large number of bootstrap replicates is only analyst's limitation. When we implement the same procedure for the median, we obtain Figure 16.

In general, variances of the bootstrap estimate of the standard error of the median is larger than those of the mean for all levels of sample size and number of resamples. This observation is in harmony with the inference we made earlier about the sample size of the bootstrapping procedure for the median. As stated there, the reason for this is that median requires larger sample size than mean in order to achieve a good representation of the sampling distribution.

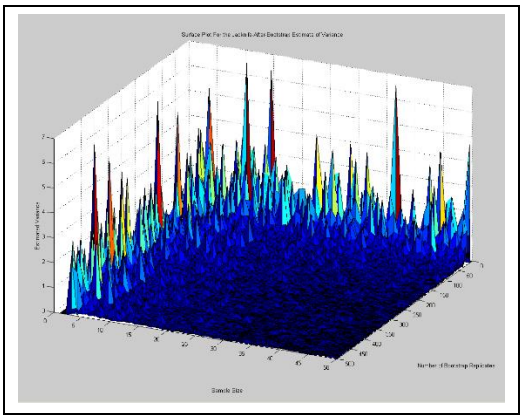

Figure 16: 3D plot for the jackknife-after-bootstrap variance of the bootstrap estimate of standard error of the median.

4 Model for the variance associated with bootstrap estimates

After observing the relationship between the variance and sample and number of resample sizes, the question arises: Could we build an empirical model for this relationship? It seems possible to build a regression model for this relationship. For this purpose, we devised a 3^2 factorial design. We selected levels 20, 50 and 100 for the sample size factor and 500, 1000 and 5000 for the number of resample size factor and replicated the design 5 times. We computed the variances of the standard error for the mean at these levels with jackknife-after-bootstrap procedure. The reason why we chose these levels lies in the limitations of the jackknife-after-bootstrap procedure. Efron and Tibshirani [21] showed that jackknife-after-bootstrap runs into trouble when $n < 10$ and $B < 20$. Because when the sample size and the number of resample size are below these figures, then the probability that every bootstrap sample contains a given point i is high. Also, they showed that jackknife-after-bootstrap overestimates the variance by a large margin when B is as small as 20, but seems to improve as B gets up to 200. When the number of resample size is near 500, estimate becomes reasonable.

After getting responses for the design points, a regression analysis was performed. A quadratic model with interactions terms was considered for the regression analysis. In order to decide which predictors to include, stepwise regression was performed first. Then regression model was built with the suitable predictors. We applied log transformation to response variable based on Box-Cox analysis. Additionally; normality (with Ryan-Joiner test), homoscedasticity (with Bartlett test) and auto-correlation (with Durbin-Watson statistic) of residuals are checked for the regression analysis. Predictors and coefficients in the model are summarized in Table 3.

Table 3: Regression coefficients.

Predictor	Coefficient	SE Coefficient	P
Constant	-1.16	0.107400	0.000
B	-2.3×10^{-4}	0.000032	0.000
n	-3×10^{-2}	0.003920	0.000
n^2	-2.4×10^{-4}	0.000035	0.000
n^2B	-4.6×10^{-7}	0.000000	0.000
nB^2	9.3×10^{-10}	0.000000	0.000

The regression equation to model the relationship between the variance associated with the bootstrap estimate of standard error of the mean and sample and number of resample size is:

$$\log\left(\hat{var}(\hat{se}_B)\right) = -1.16 - 2.3x10^{-4}B - 3x10^{-2}n - 2.4x10^{-4}n^2 - 4.6x10^{-7}n^2B + 9.3x10^{-10}nB^2 \quad (7)$$

This model is obtained empirically, rather than theoretically as done by Efron and Tibshirani [21]. It may not be easy to determine values of c_1 and c_2 for different populations immediately as shown in Equation (6). On the other hand, it is possible to model the relationship between the variance of the bootstrap estimates and sample and number of resample sizes for other statistics in the same fashion for all kinds of population distributions. Consequently, we can use this model to decide sample size and number bootstrap repetitions in order to obtain bootstrap estimates with acceptable variance.

5 Conclusion and further research

Even though bootstrap methods are very popular and find many application areas, their limitations are rarely considered. In this paper, we examined the variation associated with bootstrap estimates empirically. We need to be aware that bootstrap estimates have two sources of variance. One of them is the sampling variability, decreases as sample size increases, and resampling variability, decreases as number of resamples increases. Since number of bootstrap resamples is limited with time of the analyst and power of the computer, sample size needs more attention.

The relationship between variance of bootstrap estimates and n and B is examined theoretically by Efron and Tibshirani [21], however, it may not be easy to find closed form formulas for this relationship all the time. Additionally, constants in this formula are distribution specific, hence, depends on the underlying distribution from which the random sample is drawn. However, we almost never know the true underlying distribution of the sample. Consequently, distribution free techniques to build this relationship is important, which constitutes the main contribution of this study. In this respect, we propose a simple and empirical way to model this relationship by using jackknife-after-bootstrap procedure.

In this paper two statistics are considered: mean and median. An analysis similar to this can be performed for different statistics of interest. Because different statistics have different limitations in terms sample size and number of resamples.

6 References

[1] Efron B. “Bootstrap methods: another look at the jackknife”. *The Annals of Statistics*, 7(1), 1-26, 1979.

[2] Efron B, Tibshirani RJ. “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical measures of statistical accuracy”. *Statistical Science*, 1(1), 54-75, 1986.

[3] Efron B. “Better bootstrap confidence intervals”. *Journal of the American Statistical Association*, 82(397), 171-185, 1987.

[4] DiCiccio TJ, Efron B. “Bootstrap confidence intervals”. *Statistical Science*, 11(3), 189-212, 1996.

[5] Pawitan Y. “Computing empirical likelihood from the bootstrap”. *Statistics & Probability Letters*, 47(4), 337-345, 2000.

[6] Hall P. “Theoretical comparison of bootstrap confidence intervals”. *The Annals of Statistics*, 16(3), 927-953, 1988.

[7] Hall P. *The Bootstrap and Edgeworth Expansion*. New York, USA, Springer-Verlag, 1992.

[8] Hesterberg T, Monaghan S, Moore DS, Clipson A, Epstein R. *Bootstrap Methods and Permutation Tests.* Editors: Moore D, McCabe GP, Duckworth WM, Sclove SL. The Practice of Business Statistics, 18.1-18.78, New York, USA, Freeman, 2003.

[9] Lo AY. "A Bayesian bootstrap for a finite population". *The Annals of Statistic*s, 16(4), 1684-1695, 1988.

[10] Booth JG, Butler RW, Hall P. "Bootstrap methods for finite populations". *Journal of the American Statistical Association*, 89(428), 1282-1289, 1994.

[11] Shao J. "Impact of the bootstrap on sample surveys". *Statistical Science*, 18(2), 191-198, 2003.

[12] Aitkin M. "Applications of the Bayesian bootstrap in finite population inference". *Journal of Official Statistics*, 24(1), 21-51, 2008.

[13] Antal E, Yves T. "A direct bootstrap method for complex sampling designs from a finite population". *Journal of the American Statistical Association*, 106(494), 534-543, 2011.

[14] Efron B. “Jackknife-after-bootstrap standard errors and influence functions”. *Journal of the Royal Statistical Society*, 54(1), 83-127, 1992.

[15] Hill RC, Cartwright PA, Arbaugh JF. “Jackknifing the bootstrap: some monte carlo evidence”. *Communications in Statistics-Simulation and Computation*, 26(1), 125-139, 1997.

[16] Andrews DW, Buchinsky M. “A three-step method for choosing the number of bootstrap repetitions”. *Econometrica*, 68(1), 23-51, 2000.

[17] Davidson R, MacKinnon JG. “Bootstrap tests: How many bootstraps?”. *Econometric Reviews*, 19(1), 55-68, 2000.

[18] Lunneborg CE. *Data Analysis by Resampling: Concepts and Applications*. Pacific Groove, California, USA, Brooks/Cole, 2000.

[19] Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A. “How many bootstrap replicates are necessary?”. *Journal of Computational Biology*, 17(3), 337-354, 2010.

[20] Chernick MR. *Bootstrap Methods: A Guide For Practitioners and Researchers*. New Jersey, USA, Wiley, 2008.

[21] Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York, USA, Chapman and Hall, 1993.

[22] Martinez WL, Martinez AR. *Computational Statistics Handbook with MATLAB*. 2nd ed. New York, USA, Chapman and Hall/CRC, 2007.

[23] Jentsch C, Leucht A. “Bootstrapping sample quantiles of discrete data”. *Annals of Institute of Statistical Mathematics*, 68(3), 491-539, 2016.