November 2017

Investigation on the properties of bootstrap confidence intervals

Bootstrap confidence intervals are an invaluable tool in the statistician's toolbox, allowing them to obtain measures of uncertainty when many of the standard assumptions fail. Due to the varying algorithms that are available to produce them, these bootstrap intervals don't all behave in the same way, and each have their strengths and weaknesses. We conduct a computation heavy simulation in order to examine the behaviours of common choices of bootstraps in certain situations, and attempt to make inference on which methods are better suited to various scenarios.

Introduction

When conducting a study or an analysis, a statistician's work is never over until some robust measure of uncertainty is established. When the distributions of estimators are known, this task becomes far more trivial. Maximum likelihood estimates have convenient properties which aid in this task, making Wald intervals a very popular and effective choice. However, when our estimators are non-linear functions of these estimates, or sample size is far too small to assume the distribution of estimates is close to that of their asymptotic form, we must turn to other methods.

The bootstrap allows us to resample many times from the distribution of our data, either parametrically (assuming a known distribution) or non-parametrically (by resampling from the empirical distribution of our sample). By calculating statistics of interest on each of these samples, we can then obtain a bootstrap distribution of these statistics. This 'free' way of obtaining a measure of uncertainty when traditional methodologies would fail or lead to large amounts of bias is invaluable, and allows us to remain confident in the robustness of our confidence intervals.

We produce a computer simulation to aid us in describing the behaviours of common choices of bootstrap intervals, and to see how these changed in different (standard) situations.

Method

Given finite computing time and an infinite number of interesting cases to investigate, we reach the obvious conclusion that not all of them can be explored. We choose to focus on the humble mean as our statistic of interest; a respected robust and unbiased estimator of the expectation of the distribution of interest.

We use the R programming language to perform a suite of simulations, producing bootstrap confidence intervals in a spread of common situations, varying both the distributions the random samples are generated from, and the methodologies the bootstraps use to calculate their confidence intervals. We fix the value of α to be 0.05, such that all produced confidence intervals are of the (1- α)×100% form. This allows more consistent comparison of the different methods we investigate.

We sample our deviates from Normal, Poisson and Gamma distributions, each hoping to illustrate different properties of the intervals we examine. The Normal deviates allow us to set a 'baseline' behaviour, when there are no obvious difficulties and all assumptions should be being met. The Poisson deviates introduce a small amount of strain on our bootstraps, allowing us to see how they behave when the statistic of interest is dependent on the standard deviation. Gamma distributed deviates allow us to examine what happens when the true distribution is skewed, making it more difficult to produce intervals with accurate coverage, or equal coverage on either side of the true value.

We examine four bootstraps; the non-parametric percentile method, the non-parametric bias-corrected and accelerated (BCa) bootstrap, the non-parametric (percentile) smooth bootstrap, and the parametric percentile bootstrap. In practical situations, it is extremely unlikely the true parameters of the sample distribution is known. Because of this, we choose to implement an MLE parametric bootstrap (referred to as 'par.fit' in the figures) , a parametric bootstrap which instead of using the true parameters of the sample distribution, uses maximum likelihood estimates for these values instead. Hence this bootstrap resamples the data from an estimated distribution instead, rather than a known distribution which is usually unattainable. Early iterations of this simulation did not include this feature, leading to parametric bootstraps over-performing with perfect coverage. This was a reflection of the unrealistic

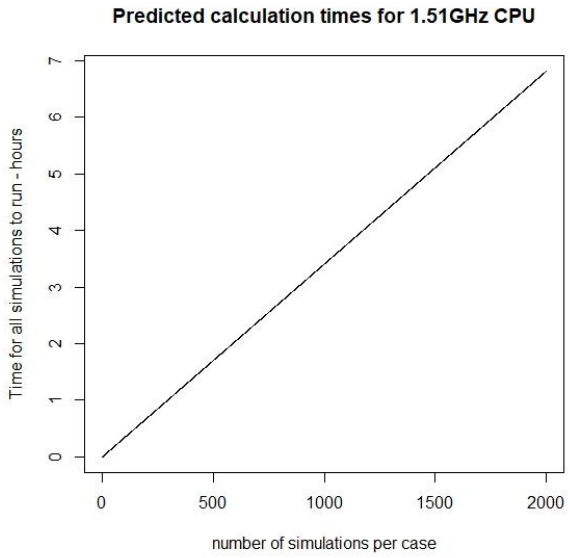

Figure 1 - Expected total calculation time for various simulation numbers per case, for the machine used to produce the simulation study

assumptions the standard parametric bootstrap requires and the impracticality of its use in a real-world context.

For each of the considered methods, we vary sample size of the original sample, we consider sample values of 20, 50, 100, and 500. For each of these in turn, we also vary the number of resamples the bootstrap method uses to construct its interval, using the values 999, 1999 and 4999. In this way we produce simulated intervals for every combination of the above sample sizes and bootstrap resample numbers. Each bootstrap also includes the original sample.

The smooth bootstrap we implement introduces some noise into the resampled data, and then calculates the confidence interval using the percentile method on the bootstrap distribution of the statistic of interest. There is no obvious universal choice for the distribution of this noise or its variance, however, normal noise seems very reasonable due to its symmetry. We set the standard deviation of this noise as a fraction of the observed noise in the data (the sample standard deviation). This takes a default value of 0.1 in all simulations, unless specified otherwise. We conduct a small side analysis which looks at the smooth bootstrap in more detail to ascertain how changing this value affects the bootstrap's behaviour.

We run a small pilot study which estimates computation time given the number of simulated bootstraps we produce for each combination of sample size and number of bootstrap resamples. This pilot study lead to figure 1, which in turn lead to choosing a study of 1000 simulations. This number struck the desired balance between reducing the variance in our results, whilst keeping the required computational resources under a reasonable threshold.

Results

It should be noted that all references to statistics which measure the performance and behaviour of each method of producing confidence intervals refers only to the observed statistic, which we use as a (hopefully accurate) estimate of the true value.

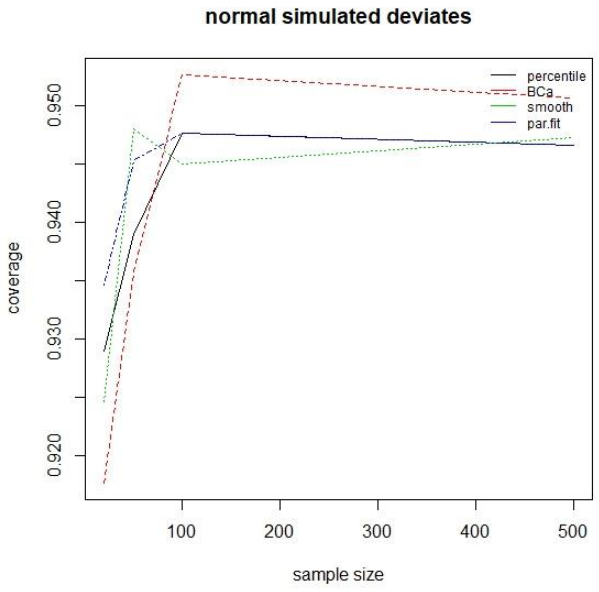

Figure 2 - Coverage for normal samples as sample size increases, averaged over all values of bootstrap resamples

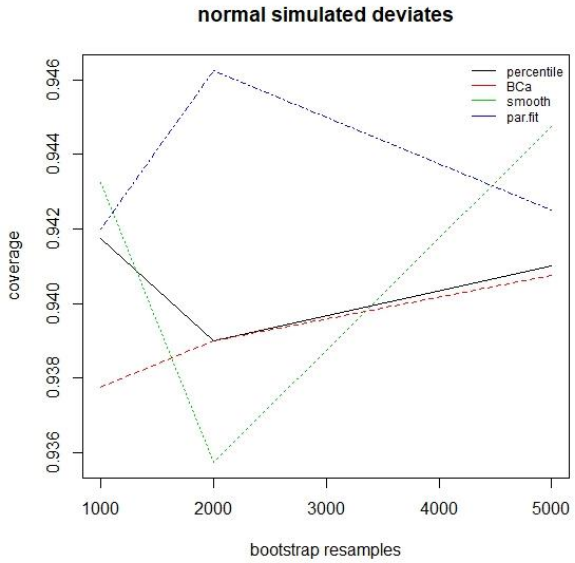

Figure 3 - Coverage for normal samples as bootstrap resamples increases, averaged over all values of sample size

We first consider our simulations from normally distributed samples. Averaging across all sample sizes and number of resamples, coverage for the percentile method, BCa, smooth bootstrap and the MLE parametric bootstraps was 0.941, 0.939, 0.941 and 0.944 respectively. We produce figures 2, and 3 to visualise performance (measured by coverage) as sample size and number of bootstrap resamples increases. Figure 2 shows that small samples sizes (20 and 50) tend to have drastically poorer coverage than greater values. This is accompanied with a significant increase in coverage from samples sizes of 20 to 50, and again from 50 to 100. From 100 onwards we seem to observe a plateau across all bootstrap methods, indicating that after a certain point, increases in samples sizes may follow a law of diminishing returns when it comes to coverage. This reaffirms the bootstrap's ability to produce reasonable intervals when sample size is reasonable, but not large enough to assume the distributions of our estimates are close to their asymptotic form. For most methods, we also notice a general increase in coverage as the number of resamples increases. This is to be expected, as larger resamples give us a better approximation of the bootstrap statistic. For larger values of bootstrap resamples, we see that the BCa and percentile methods actually perform extremely similarly. Indeed, in figure 3, where values are averaged over sample size, their coverage at 1999 and 4999 is almost identical. However, surprisingly, the BCa does not perform better at small sample sizes, having a coverage of 0.927 on average for samples sizes 20 and 50, versus 0.934 for the percentile method. This is unexpected, due to the BCa's theorised better coverage for small samples sizes. On the other hand, the BCa does have higher coverage for larger samples (an average of 0.952 for all bootstrap resamples and samples sizes 100 and 500, and 0.947 for the percentile in the same scenario). Looking specifically at a sample size of 500 and 4999 resamples, the BCa has a coverage of 0.953, versus 0.948 for the

percentile, which is closer to 0.95, the theorised target. Whether a small sacrifice in coverage accuracy is desirable in order to gain in absolute coverage is not immediately clear, and largely depends on the preference of the statistician in question. The smooth and MLE parametric bootstraps tend to perform similarly to the percentile method when averaging over bootstrap resamples, however their coverage seems erratic and unpredictable as we average over sample size, suggesting that perhaps their performance varies more, making them less reliable. Intuitively this would seem reasonable, given the smooth bootstrap's reliance on random noise added to its resamples and the parametric MLE's estimates of the true distribution the samples are generated from varying with each sample.

We define the notion of failure tendency, which when examining interval which have not contained the true value, refers to the proportion of the time the true statistic was to the left of the confidence interval. Thus this value ranges between 0 and 1. A value of 0 indicates the true value was always to the right of the interval, indicating our intervals are too far to the left, conversely, a value of 1 suggest our interval is too far to the right. Ideally any confidence interval should have a failure tendency of 0.5, indicating equal probability of the interval being too far left or right, given the true value is not in the interval.

Figure 4 shows us the behaviour of this failure tendency for our simulations from normal deviates, and demonstrates that the BCa is generally our best option if we require equal coverage on both tails of the distribution of the statistic of interest. It is curious that the smooth bootstrap, which is simply the percentile with noise added to its resamples, has a trend in failure tendency which goes in the other direction. It is unclear why the percentile seems to behave so badly, or why it performs worse as sample sizes increases, especially given our normal deviates are perfectly symmetric. This is

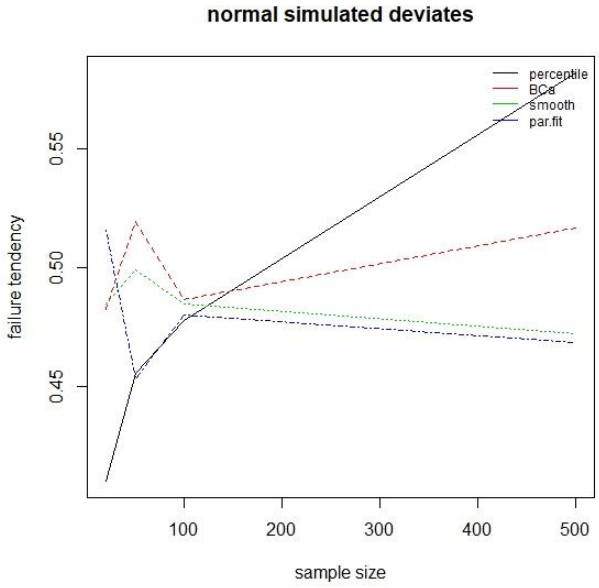

Figure 4 – Failure tendency for normal samples as sample size increases, averaged over all values of bootstrap resamples

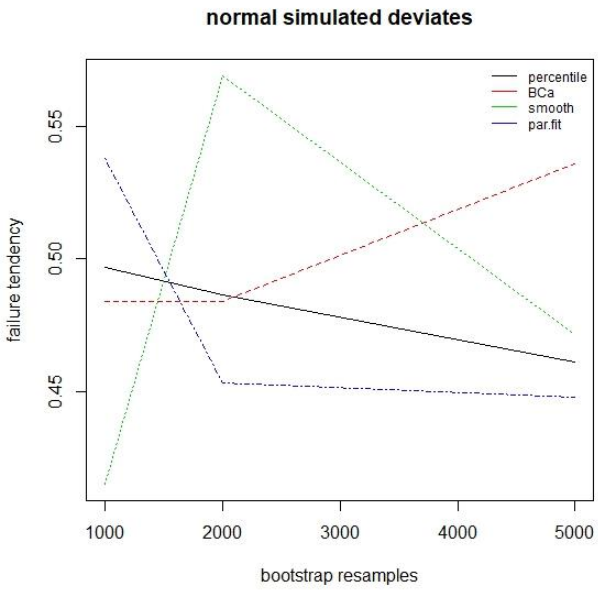

Figure 5 – Failure tendency for normal samples as bootstrap resamples increases, averaged over all values of bootstrap sample size

perhaps an indication that failure tendency varies significantly, and may not follow any particular pattern as sample sizes increases, implying that we should be cautious not to over interpret this plot. However, if we plot this failure tendency over bootstrap resamples (figure 5), we see a clear downward trend for the percentile method, and a steady increase for the BCa.

We move onto our intervals generated when the distribution of our data is Poisson, hoping to illustrate how the behaviours of these confidence intervals varies when the mean is dependent on the standard deviation of the distribution. We simulate from a Poisson with mean 100, which is therefore reasonably symmetric, and so hope to have successfully isolated the aforementioned property.

Figures 6, and 7 illustrate the results we obtained, which are again, surprising. The BCa is again worse than the percentile when sample size is small (20 and 50), and similar to the percentile when sample size is 100 or 500. In fact, the BCa only has significantly better coverage than the percentile when the number of resamples is high (4999). In this scenario it has an average coverage of 0.943, as opposed to 0.941, 0.947 and 0.948 for the percentile, smooth and parametric MLE bootstraps respectively. The smooth and parametric MLE bootstraps perform surprisingly well in this scenario, although given sample size is large enough to accurately estimate the parameters for our sample's distribution, we clearly gain considerable information from the ability to be certain of the shape of its distribution. This would suggest the parametric bootstrap is indeed very valuable when we are extremely confident in the likelihood of our data, but require estimates of its parameters. Plots of failure tendency for these data are available with the source code, but were

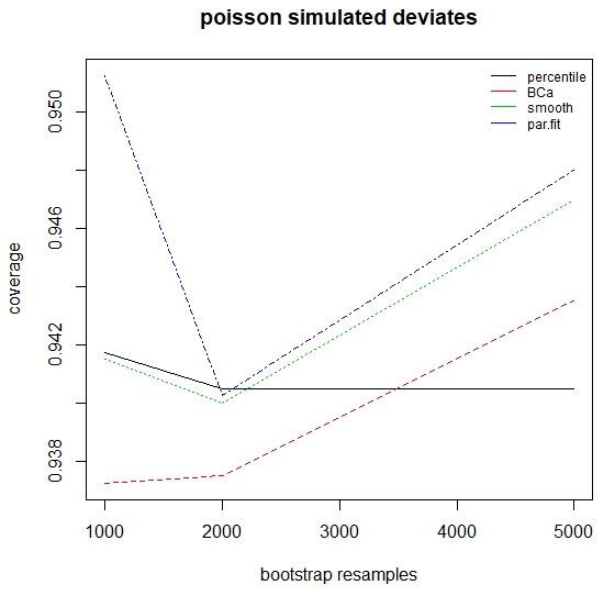

Figure 6 – Coverage for Poisson samples as bootstrap resamples increases, averaged over all values of sample size

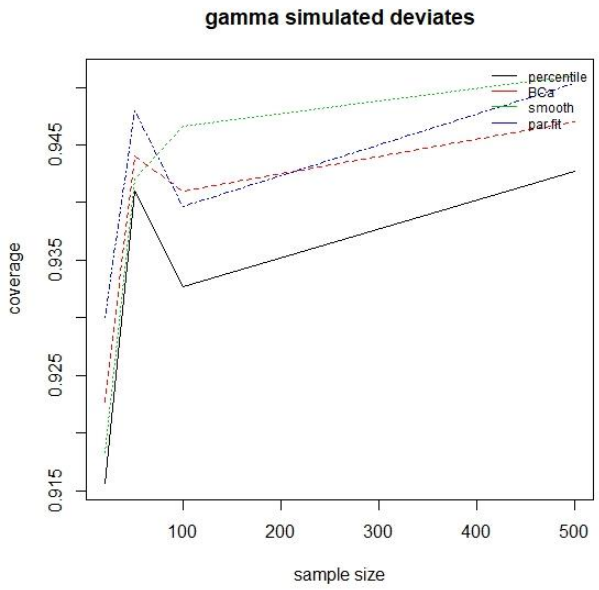

Figure 8 – Coverage for Gamma samples as sample size increases, averaged over all values of bootstrap resamples

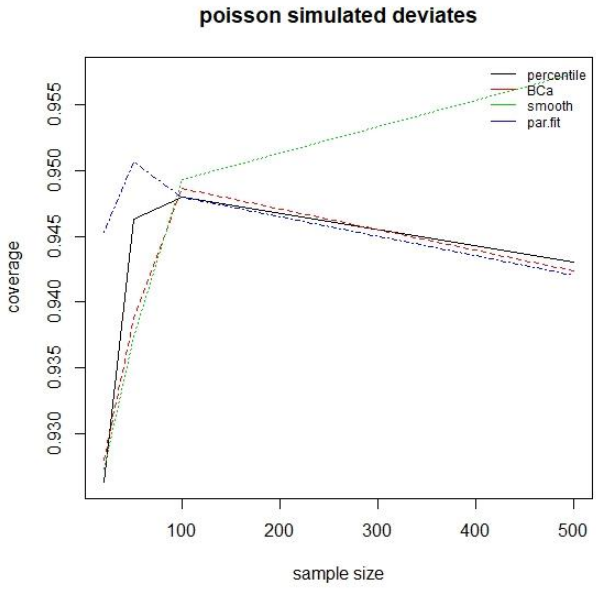

Figure 7 – Coverage for Poisson samples as sample size increases, averaged over all values of bootstrap resamples

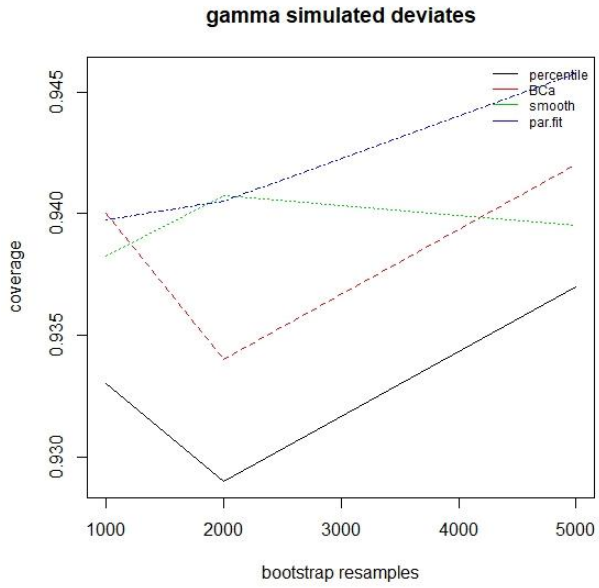

Figure 9 – Coverage for Gamma samples as bootstrap resamples increases, averaged over all values of sample size

omitted here due to erratic behaviour, indicating no real pattern of significant interest which could be explored in detail.

We move on to our simulations with gamma distributed samples. These deviates all had a shape parameter of 3 and a rate of 10, this produces samples with a significant skew and hopefully reveals other interesting properties of the bootstraps being investigated when conditions are less easy to handle. Figures 8 and 9 show the results of these simulations.

We again observe that coverage is poor for all methods when sample size is low (20 or 50), except this time the skew of the distribution allows for the BCa to demonstrate its advantages. It outperforms the percentile at all levels of sample size and bootstrap resamples, and performs relatively similarly to the smooth bootstrap and parametric MLE bootstraps, with the four methods having an average coverage of 0.933, 0.939, 0.940 and 0.942, given in the same order as the legend in figure 9. We again observe that although all the methods are negatively affected by the strong skew of these samples, the parametric MLE bootstrap remains fairly robust, due to its strong assumptions on the shape of these samples. The smooth bootstrap also shows an excellent ability to cope, especially when we consider figure 8, which averages over all values of bootstrap resamples. This makes it an extremely viable option for all sample sizes when we expect the distribution of our

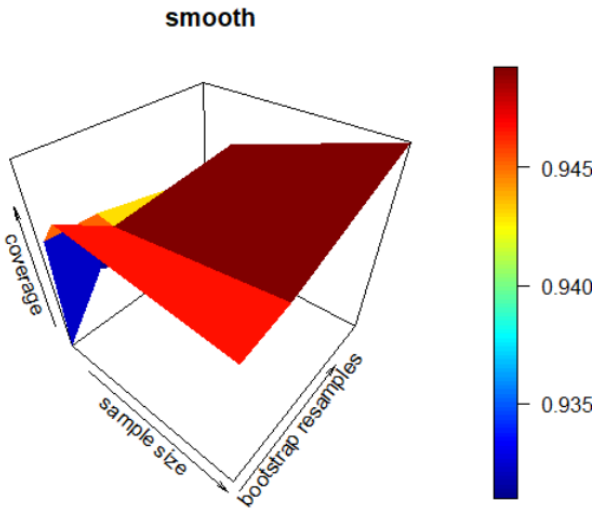

Figure 10 – Gamma samples, smooth bootstrap average coverage surface for all 1000 simulations

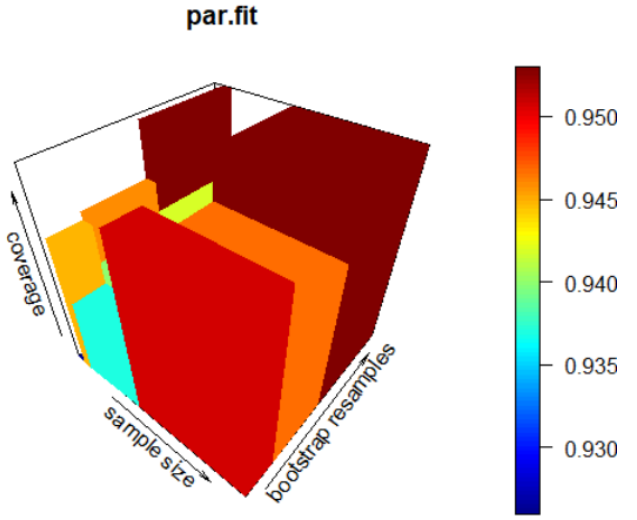

Figure 11 – Gamma samples, parametric MLE bootstrap coverage histogram for all 1000 simulations

data to be skewed, and we are unwilling to make other strong assumptions on its shape. We notice again that all methods generally improve their average coverage as sample size and bootstrap resamples, which is of no surprise. This can also ne illustrated with 3D plots of the surface of coverage or its histogram, as illustrated by figures 10 and 11.

To investigate the smooth nonparametric percentile bootstrap further, we fix the value of sample size to 500, and that of bootstrap resamples to 999. We then set smooth.sd (the standard deviation of the noise added to resampled observations as a fraction of the sample standard deviation) to be 10 evenly spaced values between 0.01 and 0.33. For each of these values we produce 1000 samples from a standard normal distribution and calculate a confidence interval for each one. Figures 14 and 15 illustrate the results of this

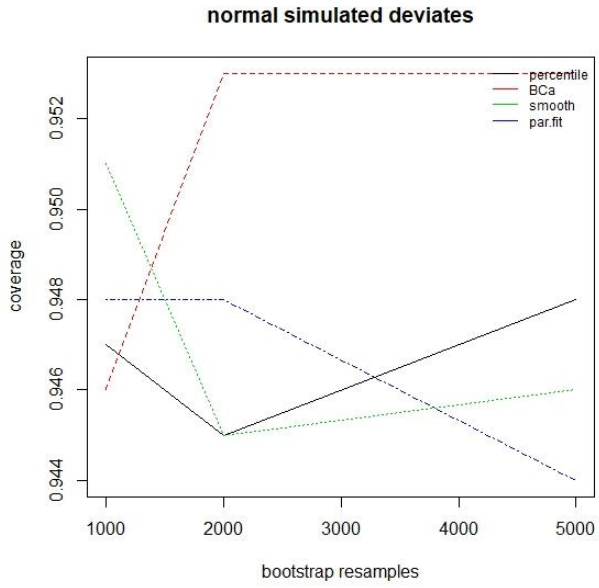

Figure 12 – Coverage for normal samples as bootstrap resamples increases, with sample size fixed at 500

simulation. Coverage generally increases, with some variability as noise increases. This is likely simply a product of the fact that the length of the interval increases along with the noise. It would seem a value of smooth.sd around 0.15 is a good balance for producing 95% confidence intervals. Unfortunately, since this is a rather artificial way of increasing coverage, it may be difficult to calculate or ascertain which values of smooth.sd would be the most appropriate for a given situation, and we may have to run simulations in each of these cases. In our case, where the samples came from a symmetric distribution where the statistic of interest had no dependence on the sample standard deviation, there was no obvious relationship between smooth.sd and the failure tendency of the confidence intervals, as is illustrated by figure 15.

Discussion

We note that in multiple figures, such as (2, 3 and 12), coverage can sometimes decrease as the number of resamples (or sample size) increases. This goes against both our intuition and the theory which underpins these bootstraps. It is likely that this is the result of small simulation numbers, leading to larger than expected variance in our observations. It is also likely due to the small number of cases that were explored for both sample size and bootstrap resamples; with only 3 or 4 of these options, respectively. When we average across a dimension to produce a point on a plot, any particularly bad (or good) result has a high weight in the arithmetic mean, leading to bias in our point estimate for the value of 'true coverage'. However, this second explanation is not necessarily the largest influence, as is illustrated by figure 12, which still displays this counterintuitive decrease in coverage even though we fixed the value of sample size to 500, rather than taking an average over values 20, 50, 100 and 500. Hence, the proposed solution to this problem is simply to significantly increase the number of total simulations for each

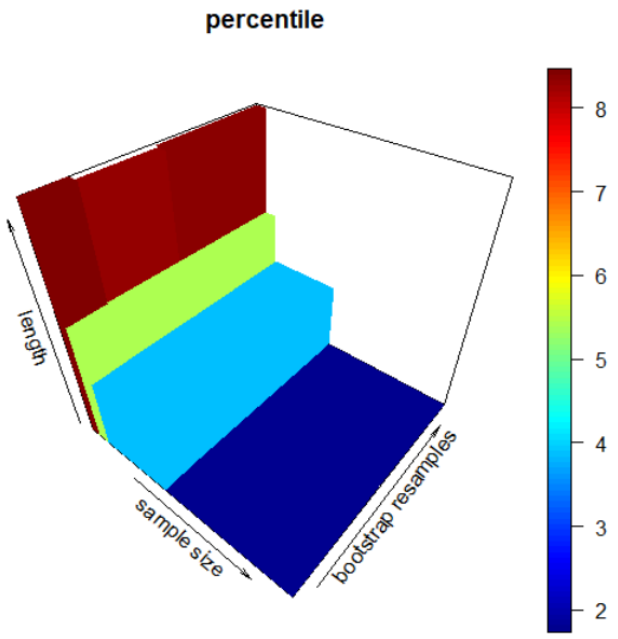

Figure 13 – Histogram of average confidence interval length for all 1000 simulations for Poisson samples, for the nonparametric percentile method

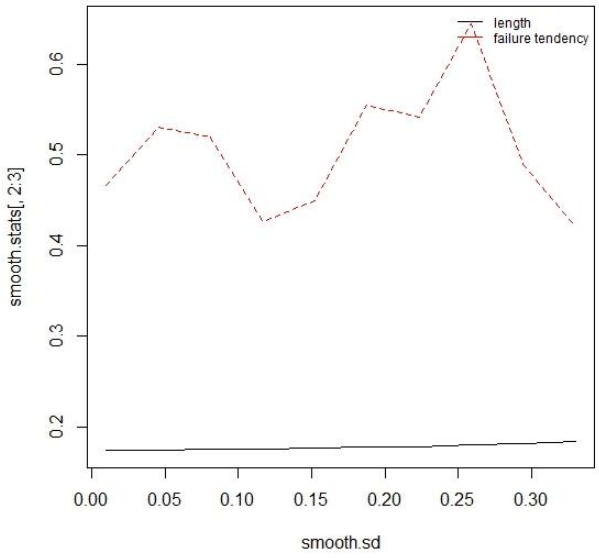

Figure 14 – length and failure tendency for the smooth percentile bootstrap with a sample size of 500, 999 bootstrap resamples and 1000 simulations per case

case (perhaps to 5000), computation time allowing, and if possible, to introduce a few new sample sizes and number of bootstrap resamples to investigate, particularly for small sample sizes, where we expected to see different behaviour from the BCa.

We observe that the length of bootstrap intervals decreases significantly, for all methods, as sample size increases. This trend is not at all as strongly visible when we consider increases in bootstrap resamples. This is illustrated by figure 13. This suggests that as sample size increases, coverage increases primarily because intervals become longer, and are therefore more likely to include the true value of the statistic of interest. This also confirms that confidence intervals obtain better coverage as resamples increase, not because their length changes, but because we obtain a better approximation of the true distribution of the statistic of interest.

It should be noted that short tests, which compared our implementation of the BCa to that of a standard bootstrap library, found that the BCa intervals produced by this library had marginally better coverage, even when the number of simulations was high. This test is available as part of the source code. This result could invalidate the results of our investigation, if the reason the BCa under-performed in certain situations was due to the specific implementation. Further work in understanding why this difference in coverage was present is required.

Conclusion

It would appear that the nonparametric percentile method is, in most scenarios, completely outdated. In situations where strong assumptions can be made on the distribution of the observed data, more robust intervals for the statistic of interest

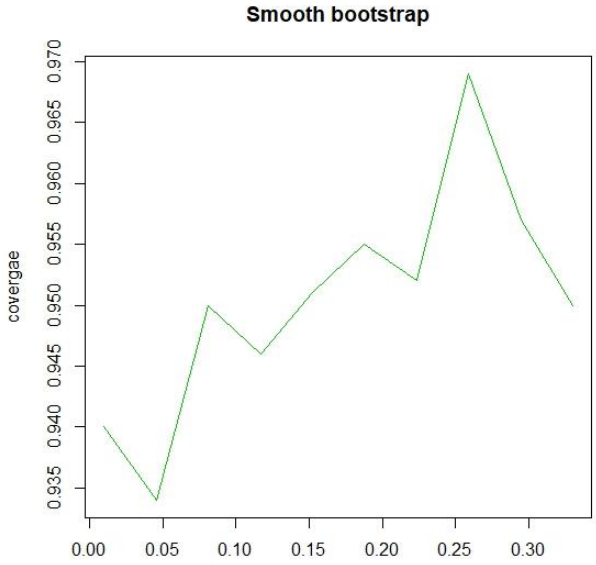

Figure 15 - observed coverage for the smooth percentile bootstrap with a sample size of 500, 999 bootstrap resamples and 1000 simulations per case

can be obtained using a parametric percentile bootstrap, even if the parameters must be obtained through maximum likelihood. In cases where this is not possible, the smooth nonparametric percentile bootstrap and the nonparametric BCa bootstrap still provide better alternatives which remain more robust to skew and a statistic which is dependent on the standard deviation of the sample distribution.

The BCa is not a good option when computational resources are scarce, given its performance is usually significantly better with high numbers of resamples. We require further simulation and investigation into this bootstrap and its properties once the

nature of our implementation has been examined in more detail, and has been corrected, if this is required.