Ilmari Juva
45727R

Mat-2.108 Independent research projects in applied mathematics

Confidence Intervals in Bootstrap
1 Introduction

Bootstrap is a computational tool for statistical analysis. It can be used to estimate the accuracy of statistical estimators in situations where there is a random sample from an unknown distribution. With the bootstrap-\(t\) approach, it is possible to obtain confidence intervals for statistical estimators without making normality assumptions.

In this work, the accuracy of bootstrap-\(t\) confidence intervals with small sample sizes is evaluated. Random samples are drawn from known distributions to test how well the bootstrap works. Both normal distribution and exponential distribution are used, as well as different sample sizes.

The theoretical part of the work in section 2 is based on Efron’s book [1].
2 Bootstrap

2.1 Notation

Let $F$ be a probability distribution. A random sample drawn from the distribution $F$ is denoted by $x$, which is a vector of the random data points $x_i$. That is $x = (x_1, x_2, \ldots, x_n)$.

$\theta = t(F)$ is a parameter of distribution $F$. If the distribution is not known, statistic $s(\cdot)$ can be used to calculate an estimate $\hat{\theta}$ for the parameter $\theta$.

$$\hat{\theta} = s(x)$$ (1)

2.2 The Plug-in principle

If we have a random sample $x = (x_1, x_2, \ldots, x_n)$ from an unknown distribution $F$, the empirical distribution $\hat{F}$ is a discrete distribution that is an estimate of distribution $F$. $\hat{F}$ is obtained by assigning a probability $\frac{1}{n}$ on each $x_i$ [1].

If some values appear several times in sample $x$ the empirical distribution can be expressed by a list of the values of the sample, and the proportion of times each value occurs. So the probability of the $k$th item in the empirical distribution would be

$$\hat{f}_k = \frac{\# \{x_i = k\}}{n},$$

the frequency of the value $k$ in the sample. And the empirical distribution is

$$\hat{F} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_K),$$

where the number of different values in the sample $K \leq n$, the sample size.

Now, to estimate, for instance, the mean or median of the unknown distribution $F$, the corresponding aspect of the empirical distribution can be used.
Using the empirical distribution this way is called the plug-in principle. The \textit{plug-in estimate} of parameter 

\[ \theta = t(F) \]  

is 

\[ \hat{\theta} = t(\hat{F}), \]  

where a parameter \( \theta \) of the probability distribution \( F \) is estimated by the same function of the empirical distribution \( \hat{F} \). In equation (1) the estimate \( \hat{\theta} \) was defined through the sample \( x \), which is the same thing as here, since the empirical distribution is constructed from the sample \( x \).

\subsection*{2.3 Bootstrap sample}

A Bootstrap sample \( x^* = (x^*_1, x^*_2, \ldots, x^*_n) \) is drawn from the random sample \( x = (x_1, x_2, \ldots, x_n) \). It is a randomized version of the original sample, drawn with replacement, meaning that any value \( x_i \) may appear several times in the sample \( x^* \), and some may not appear at all. The bootstrap sample is the same size than the original sample and consists of members of the original sample.

\subsection*{2.4 Bootstrap replication}

Let \( s(\cdot) \) be some statistic, and \( s(x) \) would be the statistic taken from the original data set. Then \( s(x^*) \) is the \textit{bootstrap replication} of the statistic. For example if \( s(\cdot) \) is the sample mean, \( s(x^*) \) is the mean of the bootstrap sample. The bootstrap replication of the estimate \( \hat{\theta} \) is then 

\[ \hat{\theta}^* = s(x^*), \]  

which is an estimate for parameter \( \theta \) based on the bootstrap data set \( x^* \) [1].
2.5 Bootstrap algorithm

The bootstrap computation goes as follows [1]:

1. Select $B$ independent bootstrap samples

2. Evaluate the bootstrap replication for each bootstrap sample

3. Using the $B$ bootstrap replications make the bootstrap estimate of the statistic in question

For example, for estimating the standard error of statistic $\hat{\theta}$, the steps would be as follows.

First draw the bootstrap samples $x^{*b}$, $b = 1, 2, \ldots, B$, as described in section 2.3. the number of bootstrap replications $B$ is usually between 25 and 200 for estimating the standard error.

In the second step, evaluate the bootstrap replications of $\hat{\theta}$ for each of the $B$ independent bootstrap samples. That is, evaluate

$$\hat{\theta}(b)^* = s(x^{*b})$$

for each sample $b$.

Finally, estimate the standard error of $\hat{\theta}$ by the sample standard deviation of the replications.

$$\hat{se}_B = \left\{ \frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B - 1} \right\}^{\frac{1}{2}}, \quad (5)$$

where

$$\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B},$$

and the $B$ in $\hat{se}_B$ refers to the number of replications used to obtain the estimate.
2.6 Confidence intervals

2.6.1 Standard confidence interval

The previous section presented the bootstrap estimate for standard error. Standard errors can be used to approximate confidence intervals. Confidence interval gives the range of plausible values for the statistic $\theta$, having observed the sample $x$ [1]. That is, in which range is the actual value of $\theta$ likely to be, if it produces the sample at hand.

The $1 - 2\alpha$ standard confidence interval for $\theta$ is

$[\hat{\theta} - z^{(1-\alpha)} \cdot \hat{se}, \hat{\theta} - z^{\alpha} \cdot \hat{se}]$  \hspace{1cm} (6)

in which the lower and upper bound can be denoted respectively by:

$\hat{\theta}_{low} = \hat{\theta} - z^{(1-\alpha)} \cdot \hat{se}$

$\hat{\theta}_{up} = \hat{\theta} - z^{\alpha} \cdot \hat{se}$

This is to say that the probability that the actual value of $\theta$ is between the confidence interval bounds is $1 - 2\alpha$:

$P\{\theta \in [\hat{\theta}_{low}, \hat{\theta}_{up}]\} = 1 - 2\alpha$  \hspace{1cm} (7)

Furthermore, since the confidence intervals considered in this work are all equal-tailed, the value of $\theta$ is as likely to be lower than the lower bound as it is to be higher than the upper bound [1]. Hence

$P\{\theta < \hat{\theta}_{low}\} = \alpha$,  \hspace{1cm} (8)

and

$P\{\theta > \hat{\theta}_{up}\} = \alpha$.  \hspace{1cm} (9)

Here, $\hat{\theta}$ is the plug-in estimate for statistic $\theta$, and $\hat{se}$ is some estimate for the standard error of $\hat{\theta}$, the bootstrap estimate for standard error from (5), for instance.
The standard confidence interval holds true for normal distributions, as well as for any other distribution when the sample size \( n \) grows, as the distribution of \( \hat{\theta} \) comes more and more normal. But for small sample sizes the standard confidence intervals are not accurate. The properties discussed above hold true also for the bootstrap confidence interval, the values for the percentiles \( z^\alpha \) just cannot be taken from a standard normal table.

To improve the confidence interval for small samples the percentiles are taken from student’s \( t \)-distribution table, instead of the normal distribution table.

### 2.6.2 Bootstrap-\( t \) interval

With bootstrap it is possible to obtain accurate confidence intervals without making the normality assumptions. The bootstrap-\( t \) approach builds intervals that are appropriate specifically for the data set in hand [1]. The bootstrap-\( t \) interval is constructed as follows.

First, generate \( B \) bootstrap samples and for each sample \( x^{*b} \), \( b = 1, 2, \ldots, B \), compute the approximate pivot:

\[
Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\hat{se}^*(b)},
\]

(10)

where, as in (4), \( \hat{\theta}^*(b) = s(x^{*b}) \), which is the value of estimator \( \hat{\theta} \) for the bootstrap sample \( x^{*b} \). \( \hat{se}^*(b) \) is the estimated standard error of \( \hat{\theta}^* \) for the sample \( x^{*b} \), as in (5).

Next, the \( \alpha \)th percentile of \( Z^*(b) \) is estimated by \( \hat{\ell}^{(\alpha)} \), such that

\[
\frac{\#\{Z^*(b) \leq \hat{\ell}^{(\alpha)} \}}{B} = \alpha.
\]

(11)

That is, the number of observed values lower than \( \hat{\ell}^{(\alpha)} \) is the proportion \( \alpha \) of the sample size \( B \). For example, for \( B = 1000 \), the estimate for the \( \hat{\ell}^{(5\%)} \) is the 50th largest value of the \( Z^*(b) \)s, and similarly the estimate for the \( \hat{\ell}^{(95\%)} \)
is the 950th largest value. If $B\alpha$ is not an integer, the value

$$[(B + 1)\alpha]$$

is used.

The bootstrap-t confidence interval is then

$$(\hat{\theta} - t^{1-\alpha} \cdot \hat{s}e, \hat{\theta} - t^{\alpha} \cdot \hat{s}e)$$

(12)

The sufficient number of bootstrap replications is not considered in this work, and is set large enough, $B = 5000$, so that it will not contribute significant inaccuracy.
3 The simulation

In order to test the bootstrap, known distributions are used, so that it is easy to observe whether the bootstrap can accurately estimate the confidence intervals. The simulation draws \( n \) random samples from a pre-determined distribution. The information about the distribution is not used in the bootstrap calculations, which uses only the random sample to calculate the confidence intervals. Based on the sample, the bootstrap makes an estimate for the mean of the distribution and an confidence interval for the estimate. If the \( 1 - 2\alpha \) confidence interval is accurate, the actual mean of the distribution used to draw the sample should fall within the interval with corresponding probability:

\[
P\{\hat{\theta}_{\text{low}} < \theta < \hat{\theta}_{\text{up}}\} = 1 - 2\alpha
\]

Also, since the bootstrap-t interval is equal-tailed, the true mean \( \theta \) should be higher than the upper bound of the estimated confidence interval with probability \( \alpha \), and lower than the lower bound with the same probability:

\[
P\{\hat{\theta}_{\text{low}} > \theta\} = \alpha
\]

\[
P\{\hat{\theta}_{\text{up}} < \theta\} = \alpha.
\]

3.1 The algorithm

Pseudocode for evaluating the Bootstrap-t interval:

- For \( i = 1 \) to \( I \)
  - Draw a sample of \( n \) random numbers from selected distribution
    - Generate \( B \) bootstrap samples
    - For each sample compute the \( Z^* \)-values of equation (10)
• Estimate the \( \alpha \)th percentile by the \( \lfloor \alpha(B + 1) \rfloor \) largest value of the \( Z^* \)s.

• \((\hat{\theta} - \hat{t}^{(1-\alpha)\hat{s}e}, \hat{\theta} - \hat{t}^{(\alpha)\hat{s}e})\) is the bootstrap-\( t \) confidence interval.

Next i

The outermost loop as well as the sampling from the known distributions are part of the simulation set up, and do not belong to the bootstrap-\( t \) algorithm. A random sample of \( n \) numbers is drawn \( I \) times from the selected distribution. Then the bootstrap estimates confidence interval for \( \hat{\theta} \) based on the sample at hand, after which another sample is drawn from the same distribution. The bootstrap computations in the algorithms are marked with the bullets (●) and follow the presentation of section 2.6.2.

3.2 The distributions

Two distributions are used to draw the samples.

• The normal distribution \( x \sim N(10, 2^2) \)

• Exponential distribution \( x \sim \text{Exp}(10) \)

For each distribution simulations are done with different sample sizes \( n \), to study the effect of small samples on the accuracy of the bootstrap-\( t \) confidence interval.

3.3 Sample size in the simulation

The simulation sample size refers to the number of times a different sample is drawn from the known distributions. In the algorithm of section 3.1 this is denoted by \( I \). The simulation sample size is not to be confused with the
bootstrap sample size $B$, which indicates the number of different bootstrap samples resampled from any sample. The bootstrap sample size $B$ is related to the bootstrap computation, while the simulation sample size $I$ is only for the purpose of testing the accuracy of the bootstrap computations.

Using the 95% confidence interval, any given bootstrap-$t$ interval should cover the actual mean with probability 0.95, be above the actual mean with probability 0.025, and below it with the same probability. This leads to situation where the number of observed such occasions in the simulation study is binomially distributed with expected value 0.025, if the interval indeed is accurate. The simulations sample size $I$ is selected so that the confidence interval is sufficiently small, but the sample size does not become excessively large for the simulation. Note, that the confidence interval considered now is for the results of the simulation, namely the number of occasions the bootstrap-$t$ interval covers the actual mean, not the confidence interval computed by the bootstrap.

The sample size for estimating $p$, so that $d$ is the distance of the confidence intervals bounds from the mean, is

$$I = \frac{z_{\alpha/2}^2 \cdot p(1 - p)}{d^2}$$

since we have a estimate of $p$ available, because it should be 0.95 if the interval is accurate [2]. Now we have selected $d = 0.005$, and $z_{\alpha/2} = 1.96$ for the 95% confidence interval. So (16) yields

$$I = \frac{1.96^2 \cdot 0.95(0.05)}{0.005^2} = 7299$$

(17)

If the bootstrap-$t$ interval would be accurate, then the percentage of times it covers the actual mean in the simulation should be in the range

$$(0.945, 0.955).$$

For the cases where the actual mean is above or below the bootstrap-$t$ inter-
val, the confidence interval on the percentage of observed occasions is

\[ p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} = 0.0250 \pm 0.0035, \]

so if the bootstrap-t is accurate the percentage of times this occurs in the simulation should be between

\[ (0.0215, 0.0285). \]
4 Results

Figure 1 shows the simulation results for different sample sizes. The rectangles represent the samples drawn from the exponential distribution and the crosses represent the samples drawn from normal distribution. The figure displays the occasions observed in the simulation were the bootstrap-t interval covered the actual mean, that is the cases where

\[ \hat{\theta}_{\text{low}} < \theta < \hat{\theta}_{\text{up}}. \]

For the normal distribution, the coverage is well within the acceptable limits for sample size \( n = 10 \), and is even more accurate for sample sizes \( n = 20 \) and \( n = 30 \). Simulations for larger sample sizes were omitted, since there is no

Figure 1: Proportion of occasions the confidence interval covered the actual mean
room for improvement, and the value stayed the same for \( n = 30 \) as it was for \( n = 20 \). In figure 2 the equal-tailed quality is studied, by checking separately the number of occasions were the actual mean is below the lower bound of the interval and the cases were it is above the upper bound of the interval. For each sample size, the percentage of cases where bootstrap overestimated the actual mean so much that \( \hat{\theta}_{\text{low}} > \theta \) is shown on the left, and the percentage of cases where the actual mean was underestimated so much that \( \hat{\theta}_{\text{up}} < \theta \), is shown on the right. It can be seen from the figure that these values too are easily within the acceptable limits, defined in section 3.3, for all the sample sizes.

For the exponential distribution the accuracy is not as good. From figure 1
it is seen that for the smallest sample size \( n = 10 \) the interval covered the actual mean in 94.3% of the cases, which is outside the acceptable limits. Although the situation is much better for \( n = 20 \), it deteriorates again for \( n = 30 \) for which the percentage is barely within the acceptable limits. This may indicate that the seeming accuracy of the \( n = 20 \) case could be by chance. Only for sample sizes \( n = 50 \) and larger, the value is accurate. Figure 3 shows the results separately for \( \hat{\theta}_{\text{low}} > \theta \) and \( \hat{\theta}_{\text{up}} < \theta \), for sample sizes \( n = 10, n = 20 \) and \( n = 30 \). For all of these the observed number of such occasions are drastically outside the acceptable limits, although the observed values clearly get closer to 0.025 as the sample size grows larger. Figure 4 shows the situation for larger sample sizes. The sample size has to be \( n = 100 \), until the equal-tailed quality is satisfied for the exponential distribution.
Figure 4: Proportion of occasions the confidence interval did not cover the actual mean for the exponential distribution, with larger sample sizes $n$. 
5 Conclusion

For the normal distribution the bootstrap-$t$ interval is very accurate. even for sample size as small as $n = 10$ the results are very good. The reason for this is that the normal distribution is a symmetrical distribution.

For the asymmetrical exponential distribution the results are not as good. The 95% confidence intervals covers the actual mean accurately for sample sizes $n = 20$ and larger, but the equal-tailed quality does not hold. That is, the actual mean is much more likely to be above the bootstrap-$t$ interval as it is below it. The reason for this could be the fact that for exponential distribution, median is lower than the mean. So majority of the probability mass is below the mean. Thus, when a small sample is drawn from the distribution, it is possible that majority of the values are below the mean.

To summarize, sample sizes as small as $n = 10$ are sufficient if there is some reason to assume that the underlying distribution is symmetrical. If no such assumptions can be made, the sample size needs to be at least $n = 20$, or for the equal-tailed property to hold, at least $n = 100$. So, for the bootstrap-$t$ to be used on small samples, there has to be a reasonable assumption that the sample was produced by a symmetrical distribution.

References
