Assignment 2 - Bootstrap

Naveen Narayanan Meyyappan

22/10/2019

## Bootstrap Assignment

### Question

Q1. Build a bias, standard deviation, and confidence interval estimator for the mean based on the bootstrap (use 10000 =nboot) and the jackknife

Q2. Build a simulator that draws n samples form a lognormal distribution (rlnorm) and builds both the central limit theorem based confidence interval, and compares it to the coverage rate for the bootstrap and the jacknife normal based confidence interval (with bias correction (how would you do that)) (confidence interval based on the 1st program). (1000 simulation runs minimum)

• Normal approximation

• Pivotal CI

Compare the coverage rates for the bootstrap confidence interval, the jacknife normal approximation confidence interval and the central limit theorem based confidence interval. For sample sizes 10, 30, and 100 alpha=0.05 (95% confidence)

### Answer

Confidence Intervals are methods of statistical inferences wherein we try to infer the statistical parameters of the population using the statistical parameters from a truly random sample of the population. Confidence intervals are generally used to estimate the population parameter. There are 2 types of population parameter estimation:

1. Point Estimate: A single best guess for the population parameter

2. Confidence Interval: A range of values that the population parameter can take.

Here we will write R program to estimate the confidence interval based on the central limit theorem, bootstrap based confidence interval and Jackknife confidence interval.

For this assignment we have to use log normal distribution.

X is log normal if ln(x) is normal, so if y is normal exp(y) is log normal. If m is mean of the normal distribution, then exp(m+((s^2)/2)) is the mean of the lognormal distribution.

Central Limit Theorem based confidence interval: Central Limit theorem states that the mean of the samples from the population will be normally distributed. This is not depended on the distribution of the population. We calculate the confidence interval by taking samples from a population and calculating its mean. This process is repeated and the resulting distribution of the sampling means will be a normal distribution. Suppose if we take 100 random samples from a population and try to calculate the 95% confidence interval for all these samples, then 95 of these 100 samples will contain the true population mean.

Bootstrap confidence interval: the confidence intervals are estimated resampling the samples with replacement. This process makes the process even more random and these resamples will also be normally distributed.

Jackknife confidence Interval: Jackknife method is similar to Bootstrap as in we will be resampling the samples. It’s a leave one out method that is we will leave one sample out each time while resampling the samples. Suppose we have a sample of size n then we will make a resample of size (n-1). This resampling is done n times, leaving out each data point from sample once.

Let us create a simulation function which take 3 arguments: The mean, the number of data points per sample and the number of simulations.

Sim.func<-function(mu.val=3,n=30,nsim=1000)  
{  
 #create coverage indicator vectors for bootstrap and normal  
 cvec.boot<-NULL # Create an empty vector for the bootstrap, normal and Jackknife confidence interval  
 cvec.norm<-NULL  
 cvec.jack<-NULL  
 #calculate real mean  
 mulnorm<-(exp(mu.val+1/2)) #Calculating the mean of the lognormal distribution  
 #run simulation  
 for(i in 1:nsim)  
 {   
 #Running the loop for nsim simulations where i indicates the current simulation nummber  
 #sample the simulation vector  
 vec.sample<-rlnorm(n,mu.val) # Creating a random sample from the lognormal distribution of the given mean  
 #bootstrap it  
 boot.list<-my.bootstrapci.ml(vec.sample) #Now for the obtained sample we call the bootstrapci function which calculates the bootstrap and normal confidence intervals   
 jack.list<-myJackknife(vec.sample) #This calls the myJackknife function to determine the jackknife confidence interval  
 boot.conf<-boot.list$bootstrap.confidence.interval  
 norm.conf<-boot.list$normal.confidence.interval  
 jack.conf<-jack.list$Jackknife.confidence.interval  
 #calculate if confidence intervals include mu  
 #count up the coverage by the bootstrap interval  
 cvec.boot<-c(cvec.boot,(boot.conf[1]<mulnorm)\*(boot.conf[2]>mulnorm))  
 #count up the coverage by the normal theory interval  
 cvec.norm<-c(cvec.norm,(norm.conf[1]<mulnorm)\*(norm.conf[2]>mulnorm))  
 #count up the coverage by the Jackknife interval  
 cvec.jack<-c(cvec.jack,(jack.conf[1]<mulnorm)\*(jack.conf[2]>mulnorm))  
 }  
 #calculate and output coverage probability estimates  
 list(norm.coverage=(sum(cvec.norm)/nsim),jack.coverage=(sum(cvec.jack)/nsim),boot.coverage=(sum(cvec.boot)/nsim))  
}

The above function calls the my.bootstrapci.ml and myJackknife function nsim times and each time a different sample is take from the lognormal distribution. The lower and upper bound for the 3 confidence intervals are obtained. Using the bounds for each sample we determine the coverage by taking the average for nsim samples

Now let us define the my.bootstrapci.ml function to determine the central limit theorem based confidence interval and bootstrap confidence interval for a given sample

my.bootstrapci.ml<-function(vec0,nboot=1000,alpha=0.05)  
{  
 #extract sample size, mean and standard deviation from the original data  
 n0<-length(vec0) #Calculate the length of the samples this will be the value of n  
 mean0<-mean(vec0) # Mean of the sample is calculated  
 sd0<-sqrt(var(vec0)) # Standard deviation of the sample is calculated  
 # create a vector to store the location of the bootstrap studentized deviation vector  
 bootvec<-NULL # Create a vector for storing the values required for calculating the bootstrap confidence interval  
 bootbiasvec<-NULL # A vector which stores the bootbias values in each iteration  
 #creating the bootstrap distribution using a for loop  
 for( i in 1:nboot)  
 {   
 #nboot denotes the number of times the bootstrap sample is to be taken from the given sample  
 vecb<-sample(vec0,replace=T) # creating a sample from the sample with replacement. This is bootstrapping  
 #calculate mean and standard deviation for the bootstrap sample  
 meanb<-mean(vecb)  
 sdb<-sqrt(var(vecb))  
 #Since we are resampling full vector we can use n0 for sample size of vecb  
 bootvec<-c(bootvec,(meanb-mean0)/(sdb/sqrt(n0)))  
 bootbiasvec<-c(bootbiasvec,meanb-mean0) # calculating the bias and adding it to the bootbias vector which tracks the boot bias in each iteration  
 }  
 # Calculate lower and upper quantile of the bootstrap distribution  
 bootbias<-mean(bootbiasvec)  
 lq<-quantile(bootvec,alpha/2)  
 uq<-quantile(bootvec,1-alpha/2)  
 #Add the other two confidence intervals  
 LB<-mean0-(sd0/sqrt(n0))\*uq  
 UB<-mean0-(sd0/sqrt(n0))\*lq  
 #The mean and standard deviation are calculated for the normal confidence interval  
 NLB<-mean0-(sd0/sqrt(n0))\*qnorm(1-alpha/2)  
 NUB<-mean0+(sd0/sqrt(n0))\*qnorm(1-alpha/2)  
 list(bootstrap.confidence.interval=c(LB,UB),normal.confidence.interval=c(NLB,NUB))  
}

From the above function we will get the bootstrap and central limit theorem based confidence intervals.

Let us define the function of myJackknife function to determine the jackknife based confidence interval

myJackknife<-function(v1,alpha = 0.05)  
{  
 n1<-length(v1) #Identify the size of sample  
 jackvec<-NULL #Initialize an empty vector  
 mu0<-sd(v1) #Calculate the sample standard deviation  
 for(i in 1:n1){ #Each sample is omitted once in the calculation  
 mua<-sd(v1[-i]) #Now calculate the standard deviation by leaving one data point  
 jackvec<-c(jackvec, n1\*(mu0)-(n1-1)\*mua) #Calculating the pseudo value and adding it to the jack vector  
 }  
 jackmean<-mean(jackvec) #Compute the mean of the jack vector  
 jackvarvec<-NULL #Initialize a new Jack variance vector  
 for (i in jackvec)  
 {  
 jackvarvec<-c(jackvarvec,(i-jackmean)^2) #This for loop is used to compiute variance/standard error for the jack vec  
 }  
 jackse<-sqrt((sum(jackvarvec))/(n1\*(n1-1))) #Standard error of jack vec is calculated  
 JUB=mean(v1) + qt(0.975,n1-1)\*jackse #Compute the lower and upper bound of jackknife confidence interval using the t distribution  
 JLB=mean(v1) - qt(0.975,n1-1)\*jackse  
 list(Jackknife.confidence.interval=c(JLB,JUB))  
}

The above function gives the Jackknife based confidence intervals for the given sample

Now we run the Sim.Func with nsim to be 1000

Sim.func(3,10,1000)

## $norm.coverage  
## [1] 0.794  
##   
## $jack.coverage  
## [1] 0.789  
##   
## $boot.coverage  
## [1] 0.902

Sim.func(3,30,1000)

## $norm.coverage  
## [1] 0.872  
##   
## $jack.coverage  
## [1] 0.915  
##   
## $boot.coverage  
## [1] 0.92

Sim.func(3,100,1000)

## $norm.coverage  
## [1] 0.926  
##   
## $jack.coverage  
## [1] 0.973  
##   
## $boot.coverage  
## [1] 0.951

The above results show that the bootstrap based confidence interval provides more coverage than the central limit theorem based confidence interval.

This is especially true for small sample sizes. As the sample sizes increases the bootstrap and the normal based confidence interval show almost the same coverage rate. For larger samples the bootstrap based confidence interval requires more time to compute and also requires more computational power as the same size becomes very large. In such cases we can compromise by computing the central limit theorem based confidence interval. But for lower sample sizes, it is better to go with the bootstrap based confidence interval as it provides more coverage and the computations can be handled by normal computers.

Another inference is that the bootstrap confidence interval coverage reduces with decrease in the value of nboot. This is because the number of resamples from the sample is lesser as a result the coverage by bootstrap confidence interval reduces.

The Jackknife based confidence interval is computationally less expensive than the bootstrap based confidence interval calculation. It performs well for medium sample sizes.

The Jackknife provides more coverage than the central limit theorem based confidence interval but less than the bootstrap based confidence interval for smaller sample sizes

As sample size increases, the 3 confidence intervals gave almost the same coverage. For large samples we use central limit theorem based confidence interval as it is computationally less expensive

### References

1. <http://www.math.montana.edu/jobo/thainp/jack.pdf>

2. <https://www.biostat.washington.edu/sites/default/files/modules/2017_sisg_1_9_v3.pdf>

3. <https://rutgers.instructure.com/courses/30723/files?preview=6200512>