Introduction to Bootstrap

Preetham Reddy Armoor
University of Applied Sciences, Mittweida, Germany

28 Nov 2016

Contents

1 Introduction

The Bootstrap was introduced by Bradley Efron in 1979. It is computer based method for assigning measure of accuracy to statistical estimates.The basic idea behind the bootstap is that, in absence of any other information about the distribution, the observed sample contains all the available information about the underlying distribution, and hence resampling the sample is the best guide to what can be expected from resampling from the distribution.

Consider an unknown probability distribution F has a data $x = (x_1, x_2, x_3, x_4, x_5.........x_n)$ and when we take a random sample $x = (x_3, x_5, x_2, x_7, x_{10})$ some bootstrap samples can be:

$$x_1^* = (x_2, x_3, x_{10}, x_5, x_2)$$

$$x_2^* = (x_2, x_{10}, x_3, x_3, x_2)$$

$$x_3^* = (x_5, x_5, x_7, x_3, x_5)$$

program to show that how a bootstrap sample distribution will similar to the original data distribution

```
set.seed(333)
x<-rnorm(30) #generate 30 normal random variable
bootmean<-rep(NA,1000)
samplemean<-rep(NA,1000)
# take a mean from sample with replacement from 30 variables
for(i in 1:1000)
{
  bootmean[i]<-mean(sample(x,replace=TRUE))
}
```

```
# take a mean each time with 30 new normal random variables
for(i in 1:1000)
{
  samplemean[i]<-mean(rnorm(30))
  }
#we plot the two means
plot(density(bootmean))
lines(density(samplemean),col="red")
```

Below plot is generated using Rstudio.

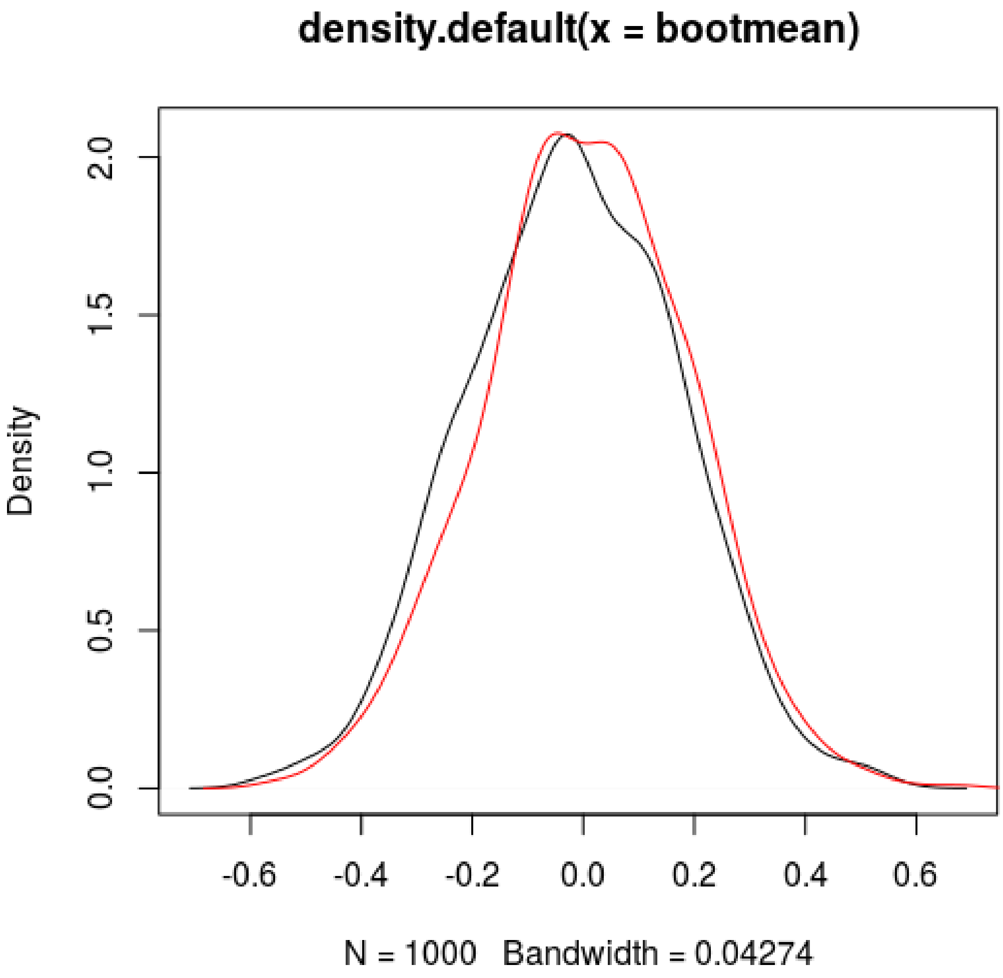

Figure 1:

The plot of both the distributions(samplemean,bootmean) of above program are remarkably similar to each other.so,what we can say is the repeated sampling of original data to approximate the distribution, we would get from repeated sampling from original data.

The empirical distribution : for a sample $x_1, ..., x_n$ of independent real-valued random variables with distribution F, we define a probability distribution $\hat{F}$ by

$$\hat{F}(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(X_i)$$

for $A \subset \mathbb{R}.\hat{F}$ is called the empirical distribution of the sample x. $\hat{F}$ can be thought as the distribution which puts mass $\frac{1}{n}$ on each observation x_i (for values that occurs more than once in the sample the mass will be a multiple of $\frac{1}{n}$). It follows that $\hat{F}$ is a discrete probability distribution with effective sample space $x_1, ..., x_n$. It can be shown that $\hat{F}$ is a nonparametric maximum likelihood estimator of F which justifies to estimate F by $\hat{F}$ if no other information about F is available (such as e.g. F belongs to a parametric family).

2 Preliminaries

2.1 The Plug-in Principle

The plug-in principle is a simple method of estimating parameters from samples. To estimate some parameter of a probability distribution F on the basis of a random sample drawn from F. The empirical distribution function $\hat{F}$ is a simple estimate of the entire distribution F .To estimate some parameter of F,like its mean or median or correlation is to use the corresponding parameter of $\hat{F}$

The Plug-in estimate of a parameter $\theta = t(F)$ is defined to be:

$$\hat{\theta} = t(\hat{F})$$

the function $\theta = t(F)$ of the probability distribution function F is estimated by the same function $t(.)$ of the empirical density $\hat{F}$.

with each bootstrap sample $x^{*(1)}$ to $x^{(B)}$, we can compute a bootstrap replication $\hat{\theta^*}(b) = s(x^{*(b)})$ using the plug-in principle.
The next thing we want to know is the accuracy of $\hat{\theta}$ compared to θ .The bootstrap uses Standard Error and Bias to measure the accuracy between θ and $\hat{\theta}$.

2.2 Standard Error

The standard error is the standard deviation (SD) of sampling distribution of statistic $\hat{\theta}$. As such, it measures the precision of an estimate of the statistic of a population distribution.

$$se(\hat{\theta}) = \sqrt{var_F(\hat{F})}$$

the standard error of mean $\overline{x}$ is the square root of the variance of $\overline{x}$ so, when we calculate the sample mean we are usually interested not in the mean of this particular sample, but in the mean for individuals of this type in statistical terms, of the distribution from which the sample comes. We usually collect data in order to generalise from them and so use the sample mean as an estimate of the mean for the whole distribution. Now the sample mean will vary from sample to sample. The way this variation occurs is described by the sampling distribution of the mean. We can estimate how much sample means will vary from the standard deviation of this sampling distribution, which we call the standard error (SE) of the estimate of the mean. As the standard error is a type of standard deviation.

$$se(\overline{X}) = \sqrt{var_F(\overline{X})} = \frac{\sigma_F}{\sqrt{n}}$$

The standard error of the sample mean depends on both the standard deviation and the sample size, by the simple relation $SE = SD/\sqrt(samplesize)$. The standard error falls as the sample size increases, as the extent of chance variation is reduced this idea underlies the sample size calculation for a controlled trial

$$\hat{se}(\overline{X}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

Estimate the standard error $se_F(\hat{\theta})$ by the standard deviation of the B replication:

$$s\hat{e}_B = [\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2}{B-1}]$$

where $\hat{\theta}^*(.) = \frac{\sum_{b=1}^{B}[\hat{\theta}^*(b)]}{B}$

The bootstrap estimated standard error is an unbaised estimated of standard deviation is the calculation from a statistical sample of an estimated value of the standard deviation of a population of values, in such a way that the expected value of the calculation equals the true value.

We have used the plug-in principle twice first to estimate the expectation $E(x)$ by sample mean $\overline{x}$ and then to estimate the standard error $se_F(\overline{x})$ of $se_{\hat{F}}(\overline{x})$

2.3 Bias

The Bias is the difference between the expectation of an estimator $\hat{\theta}$ and the quantity θ being estimated:

$$Bias_F(\hat{\theta}, \theta) = E_F(\hat{\theta}) - \theta$$

A large bias is usually an undesirable aspect of an estimator's performance.when bias is equal to zero the expectation of an estimator $\hat{\theta}$ is equal to the quantity being estimated θ.since we donot known the probability distribution of given data $x = (x_1, x_2, x_3, x_4, x_5.........x_n)$ we measure the bootstrap estimate of bias using plug-in estimate $\hat{\theta}$

The bootstrap estimate of bias is defined to be the estimate:

$$Bias_{\hat{F}}(\hat{\theta}) = E_{\hat{F}}[S(\mathbf{x}^*)] - t(\hat{F}) = \theta^*(.) - \hat{\theta}$$

here $t(\hat{F})$, the plug-in estimate of θ and $E_{\hat{F}}[S(\mathbf{x}^*)]$ is the expectation of bootstrap estimate.

3 Non-Parametric Bootstrap

Algorithm:

1. Assume a data set $x = (x_1, x_2, ..., x_n)$ is available.
2. Fix the number of bootstrap re-samples B.
3. Sample a new d ata set x^* set of size n from x with replacement.
4. Estimate θ from x^*.call the estimate $\hat{\theta}_i^*$,for $i = 1.....N$.
5. Repeat step 3 and 4 B times.
6. Consider the emperical distribution of $(\hat{\theta}_1^*, ..., \hat{\theta}_N^*)$ as an approximation of the true distribution of $\hat{\theta}$.

Program for statistic mean calculation of non-parametric bootstrap model

```
#package boot
library(boot)

# generate 50 observation between 1 to 365
x <- sample(1:365,50)

# statistic mean
samplemean <- function(x, d) {
  return(mean(x[d]))
}

# boot function call statistic mean R(5000) times.Each time, it generates
# set of random indices, with replacement, from the integers 1:nrow(data).
boot.non <- boot(x,samplemean,5000)
```

R simulated result of above program we can seen that the number of bootstrap sample increases the standard error and bias decreases

```
ORDINARY NONPARAMETRIC BOOTSTRAP
Call:
boot(data = x, statistic = samplemean, R = 5000)

Bootstrap Statistics :
    original    bias     std. error
t1*   172.04 0.304444     14.75959


Call:
boot(data = x, statistic = samplemean, R = 15000)
Bootstrap Statistics :
    original      bias     std. error
t1*   209.76 0.06263867     13.26952
```

4 Parametric Bootstrap

we assume that the distribution F belongs to a parametric family of distributions.we can still use the bootstrapping method to obtain an estimate of the sampling distribution of $\hat{\theta}$

Algorithm:

1. we assume data set $x = (x_1, x_2, ..., x_n)$ has a known distribution F_ψ.
2. Estimate the parameters with F_ψ.
3. Fix the number of bootstrap re-samples B.
4. Sample a new data set x^* set of size n from x with replacement.
5. Estimate θ from x^*.call the estimate $\hat{\theta}_i^*$,for $i = 1.....N$.
6. Repeat step 4 and 5 B times.
7. Consider the emperical distribution of $(\hat{\theta}_1^*, ..., \hat{\theta}_N^*)$ as an approximation of the true distribution of $\hat{\theta}$.

Program for statistic mean calculation of parametric bootstrap model

```
#package boot
library(boot)
# generate 50 observation between 1 to 365
x <- sample(1:365,50)

#Function for simulating from parametric distribution
gen.data <- function(x,mle)
  rnorm(length(x),mle$mu,mle$sd)

# boot function call statistic mean R(5000) times.Each time, it generates
# set of random indices, with replacement, from the integers 1:nrow(data).

boot.mean <- boot(x,mean,5000,sim=”parametric”,
                      ran.gen=gen.data,mle=list(mu=mean(x),sd=sd(x)))
```

R simulated result of above program if the data is not from the normal distribution the standard error and bias does not converge.

PARAMETRIC BOOTSTRAP

```
Call:
boot(data = x, statistic = mean, R = 5000, sim = ”parametric”,
    ran.gen = gen.data, mle = list(mu = mean(x), sd = sd(x)))

Bootstrap Statistics :
    original        bias    std. error
t1*   176.82 -0.01754264      13.79105

Call:
boot(data = x, statistic = mean, R = 17000, sim = ”parametric”,
    ran.gen = gen.data, mle = list(mu = mean(x), sd = sd(x)))

Bootstrap Statistics :
    original      bias    std. error
t1*   176.82 0.1108351      13.59146
```

5 Conclusion

- In Parametric bootstrap,$\hat{F}_{par}$ is not anymore the emperical density function.
- If the prior information on F is accurate,then $\hat{F}_{par}$ estimates better F than the empirical p.d.f .In this case the parametric bootstrap gives better estimation for the standard errors.
- If the parametric model is mis-specified then it rapidly converges to the wrong distribution.

  when might bootstrap fail:

  - Incomplete data.
  - Dependent data.
  - Noisy data.

References

[1] B.Efron,R.J.Tibshirani. *An Introduction to Bootstrap*

[2] Tim Hesterberg,Shaun Monaghan,David S.Moore. *Bootstrap Methods And Permutation Test*

[3] R Documentation `https://www.rdocumentation.org/`