**Practical**

STAT - 402

(Advanced Data Analytics)

Q. Generate a univariate dataset of size 5 from the normal distribution and find the boostrap and jackknife variance estimators for the sample pth quantile for some fixed p. Compare these to large sample approximations based on normality. How do the results vary with p?

***Solution –***

First, we generate a univariate sample of size 5 from a normal distribution.

We take the parameters µ = 0 and σ2 = 1.

The sample we got is,

-0.9792872, 0.1945809, 1.0522872, -1.1825058, 1.0587979.

From the above sample, we get 5 samples for our jackknife estimator, each sample having 1 value missing from the above 5 values, and the 5 samples have 5 different values missing.

These 5 jackknife samples are,

0.1945809, 1.0522872, -1.182506, 1.058798;

-0.9792872, 1.0522872, -1.182506, 1.058798;

-0.9792872, 0.1945809, -1.182506, 1.058798;

-0.9792872, 0.1945809 , 1.052287, 1.058798;

-0.9792872, 0.1945809 , 1.052287 ,-1.182506;

For the bootstrap sample, we take B= 100, i.e. we generate 100 bootstrap samples each of size 5.

Each bootstrap sample has been taken by sampling with replacement from the 5 values we got in the original sample.

The bootstrap samples are,

-0.9792872, -0.9792872 , 1.0587979, 1.0522872, -0.9792872;

-1.1825058, -0.9792872, 0.1945809, 0.1945809, 1.0587979;

-0.9792872, -1.1825058, 1.0522872, 1.0522872, -1.1825058;

0.1945809, 1.0522872 , 1.0522872 , 1.0587979, 0.1945809;

…. 100 rows.

Now, for each p from (0, 0.01, 0.02, 0.03,….… ,0.99, 1) ,

we calculate the pth quantile for each of 5 jackknife samples and 100 Bootstrap samples.

Now, the jackknife variance is,

VJACK, p = (n-1)\*Vp ;

Where, Vp is the variance of the 5 pth quantiles got from the 5 jackknife samples.

i.e. VJACK,p = ((n-1)/n)\* ∑ (Tn-1;i,p – (1/n)\* ∑ Tn-1;i,p)2 , where Tn-1;i,p is the ith Jackknife sample’s pth quantile value.

The bootstrap variance is,

VBOOT,p = (n/(n-1))\*V(p);

Where, V(p) is the variance of the 100 pth quantiles got from the 100 bootstrap samples.

i.e. VBOOT = [n/((n-1)\*B)]\*∑(Tn;b,p - (1/B)\* ∑ Tn;b,p)2 , where Tn;i,p is the pth quantile value of the bth bootstrap sample.

Thus for p in (0, 0.01, 0.02, 0.03,….… ,0.99, 1) , we get 101 jackknife variances and bootstrap variances for each value of p.

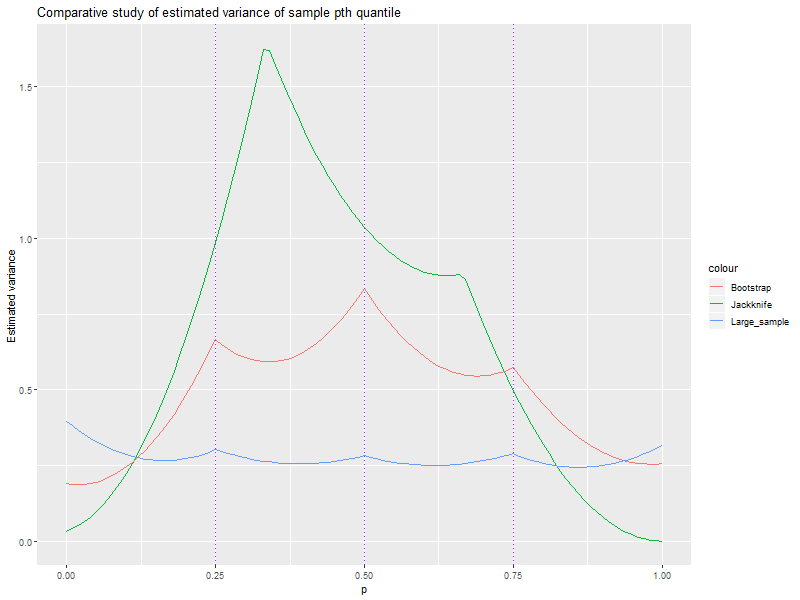
Similarly for large sample procedure,

We take 100 independent samples from the same Normal distribution, and calculate their pth quantiles, and corresponding variances.

Now, for each p, we have the jackknife, bootstrap and large sample variances.

We get the plot –

Fig. – 1.1



We have repeated the simulations 4 times by changing the base samples, and we get the following 4 plots –

Fig.- 1.2

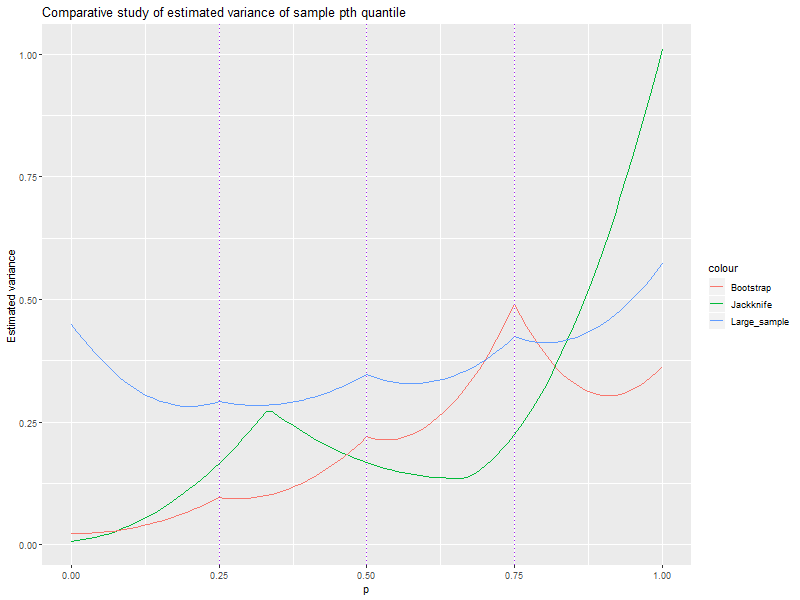


Fig.-1.3

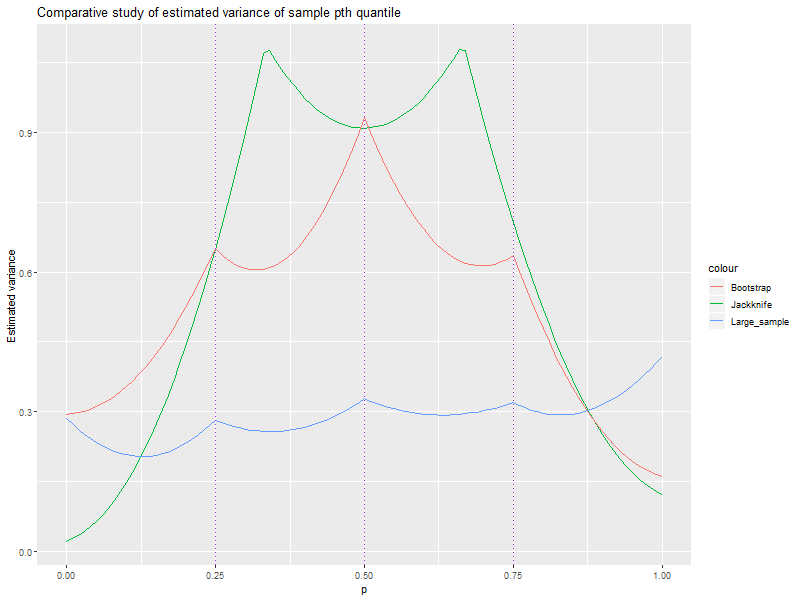
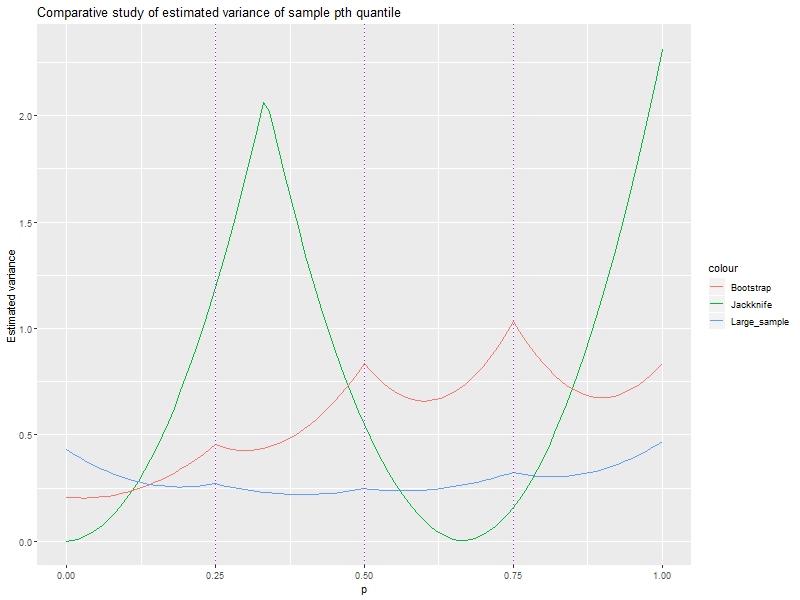
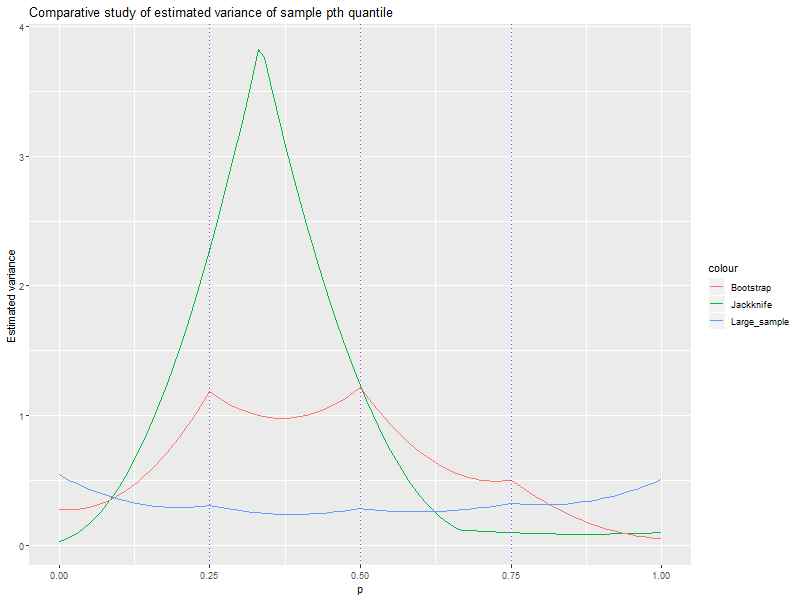


Fig.1.4



Fig.1.5

It is evident that the Jackknife estimator is the least consistent among the three, then the bootstrap, and the large sample method is the most consistent estimator.

Also in all the figures it can be seen that below 25th quantile, the three variances have a crossing point, and the jackknife variance becomes lowest bootstrap comes next to that, and large sample is the highest. Same phenomena happens at the tail regions above 75th quantile. This occurs since, the jackknife and bootstrap samples are originally from one univariate sample of size 5, but the large sample variances are estimated from 100 independent samples. Thus the extreme lower and higher sample observations might have been more varied than the bootstrap and jackknife. Except in Figure 1.3 and 1.4 where the right tail of the Jackknife variance also takes a peak. This may occur if the original sample itself is very dispersed.

It can be seen that the Jackknife variance is peaking around p=0.25 to 0.35 in most cases.

Index :

File 1: Readme.txt

!!!!Instructions!!!!

1. Copy both the files (init.R and source.R) into a folder.

2. Open init.R .

3. Uncomment the first line by removing the hash.

4. inside the inverted commas, write the folder path where the two scripts are copied.

5. Run the script.

!!!Description!!!

A) Description of init.R

Sources the source.R script, and calls the main function.

Try with different parameter values.

B) Description of source.R

Loads the required packages, if not installed, installs it, then loads.

1) generating(n ,B, mu, sigma, seed)

Parameters -----

n <- the size of the initial sample (in our question it is 5)

B <- the number of Bootstrap and large samples

mu <- mean parameter for generating normal samples

sigma <- sd parameter for generating normal samples

seed <- seed for random number generation (reproducibility of results)

Values -----

Returns a list of 4 objects, viz.

[[1]] The initial sample of size n

[[2]] A matrix containing the Jackknife samples, i.e. each row is a sample of size n-1

[[3]] A matrix containing the Bootstrap samples, i.e. each row is a bootstrap sample of size n, there are B rows

[[4]] A matrix containing B rows and n columns, where each row is a independent sample drawn from normal population with mean = mu, sd = sigma

This is used for the large sample approximations

2) est(samp, n, B, p)

Parameters -----

samp <- a list of samples, that is returned from the generating function

n <- same as before

B <- same as before

p <- can be a vector or a scalar containing the quantile values at which to calculate the estimates

Values -----

Returns a tibble of estimated variances,

with columns denoting p, Jakckinfe variance estimates,

Bootstrap variance estimates, Large sample variance estimates.

There are as many rows as there are values of p.

3) main(n, B, mu, sigma, seed)

The main function takes parameters n, B, mu, sigma, seed

saves the following outputs.

The Output.txt file contains the output of the est function.

The Samples.txt file contains the output of the generating function

ComparativePlot.png is the plot for p vs Estimated variance for different mode of estimation

i.e. Jackknife, Bootstrap, or Large sample

File 2: init.R

#setwd(<"insert here the path of the folder where you copy both scripts">)

#calling source functions

source("source.R", echo = T)

main(5, 100, 0, 1, 17)

File 3: source.R

options(scipen=999)

packages<-c("ggplot2", "tibble")

inst<-packages %in% installed.packages()

if(length(packages[!inst])>0) install.packages(.packages[!inst])

lapply(packages, require, character.only = T)

generating<-function(n, B, mu, sigma, seed)

{

set.seed(seed)

#generate a sample of size n

prim.sample<-rnorm(n, mu, sigma)

#Generating jackknife samples from the original

jk.sample<-matrix(NA, nrow = n, ncol=n-1)

for (i in 1:n)

{

jk.sample[i,]<-prim.sample[-i]

}

#Generating B bootstrap samples from the original

boot.sample<-matrix(NA, nrow = B, ncol = n)

for (i in 1:B)

{

boot.sample[i,]<-sample(prim.sample, size = n, replace = T)

}

large.sample<-matrix(rnorm(B\*n, mu, sigma), nrow = B, ncol = n)

samples<-list(prim.sample, jk.sample, boot.sample, large.sample)

return(samples)

}

#estimating function

est<-function(samp, n, B, p)

{

jk.mat<-matrix(NA, nrow = n, ncol = length(p))

boot.mat<-matrix(NA, nrow = B, ncol = length(p))

large.mat<-matrix(NA, nrow = B, ncol = length(p))

jk.var<-c()

boot.var<-c()

large.var<-c()

for (i in 1:n)

{

jk.mat[i,]<-quantile(samp[[2]][i,], p)

}

for (i in 1:B)

{

boot.mat[i,]<-quantile(samp[[3]][i,], p)

large.mat[i,]<-quantile(samp[[4]][i,], p)

}

for ( i in 1:length(p))

{

jk.var[i]<-(n-1)\*var(jk.mat[,i])

boot.var[i]<-(n/(n-1))\*var(boot.mat[,i])

large.var[i]<-var(large.mat[,i])

}

output<-cbind(p, jk.var, boot.var, large.var)

colnames(output)<-c("p", "Jackknife", "Bootstrap", "Large\_sample")

output<-as\_tibble(output)

return(output)

}

main<-function(n, B, mu, sigma, seed)

{

if (n<2) stop("n is less than 2")

p.seq<-seq(0, 1, by = 0.01)

samples<-generating(n, B, mu, sigma, seed)

vars<-est(samp=samples, n, B, p=p.seq)

vars<-as\_tibble(vars)

p.plot<-ggplot(data=vars, aes(x=p))+

geom\_line(aes(y=Jackknife, colour = "Jackknife"))+

geom\_line(aes(y=Bootstrap, colour = "Bootstrap"))+

geom\_line(aes(y=Large\_sample, colour = "Large\_sample"))+

geom\_vline(xintercept = p.seq[c(26,51,76)], linetype = 3, colour = "purple")+

labs(title = "Comparative study of estimated variance of sample pth quantile", x = "p", y = "Estimated variance")

png(filename = "Comparative plot.png", height = 600, width = 800)

plot(p.plot)

dev.off()

capture.output(print(samples), file = "Samples.txt")

capture.output(print(vars, n=length(p.seq)), file = "Output.txt")

return(list(samples, vars))

}