Sagnik Bhattacharyya, Roll – 22, 4th Semester, Non-Parametric Assignment -1

Setup:

For each [[i]], i=1(1)4, the values for MSE and variance for different estimates of location parameter for different distributions, and, the ARE's between different statistics for different distributions, are given.

For each [[i]], from each distribution there are 'm' samples of size 'n' each. For each of the m samples the values of different statistics are computed. The statistics are :

1. Median
2. Mean
3. min plus max mean, i.e., $\frac{(X_{(1)}+X_{(n)})}{2}$
4. max ordered welsh mean, i.e. $\max_{i=1(1)n} \{\frac{(X_{(i)}+X_{(n-i+1)})}{2}\}$
5. Winsorized mean
6. Hodges-Lehmann Estimator

Then their corresponding MSE and Variances are computed. And the ARE between the Median & Mean, the HL estimator & Mean, and the HL estimator and Median are also computed by the empirical variances (say the empirical variance of the HL estimate is σ_{HL} and that of Mean be $\sigma_{\overline{X}}$ then the ARE is $\frac{\sigma_{\overline{X}}}{\sigma_{HL}}$).

Discussion:

The results are obtained for four pairs of sample size.
The first one with 10 samples each of size 10, the second with 10 samples each sample with size 100, the third with 100 samples of size 10 and the last with 100 samples each size 100.

From the data we can confirm our knowledge about the appropriate estimators for location parameter for different distribution. For example for the Normal distribution for any sample size the MSE for mean is lowest among the estimators.

Below the theoritical values for the ARE's are given in a table :

Distribution	ARE ($\widetilde{X}, \overline{X}$)	ARE($\widehat{X_{HL}}, \overline{X}$)
Normal	64%	45%
Logistic	82%	109.6%
Double Exponential	200%	150%
Uniform	33%	100%

The empirically computed values of ARE are almost similar to those of the theoritical values.
For example, take the ARE between Median and Mean for Uniform distribution.
The values are 38.8 % for m = 10, n = 10; 39.5% for m = 10, n =100; 42.7% for m=100, n=10; and also 39% for m=100, n=100, whereas the theoritical value is 33%.

All the other ARE's also confirm similarly, with Normal being around 64% for Median vs. Mean etc.
Interestingly the Hodges-Lehmann estimator serves as a better estimate in comparison to both Median and Mean in respect to ARE, for all the distributions.

Considering the **Cauchy distribution**, the ARE between Median and mean, and HL estimate and mean seem to plunge suddenly to higher values, this is because **the theoritical expression goes to infinity since for Cauchy the Expectation does not exist.** But behaves regularly for the ARE between HL estimator and the Median. In computation there is nothing like infinity, so we observe a sudden peak in those ARE value.

Hodges-Lehmann estimator doesnot always give the least MSE, as we can see from the data that in general for **Normal – Mean has the least MSE** as we know, for **Logistic the Winsorised Mean has the least MSE** (here we have taken the location parameter as 13 and the scale parameter as 1; note that the tails of logistic are a little bit more bulged than that of normal under similar parameter setting. As a result the higher tail values would interfere in the estimation of location parameter with the usual mean. Thus the Winsorized mean (which truncates the extreme ends) gives a better MSE), for **Uniform ($\theta-0.5, \theta+0.5$)** $\frac{(X_{(1)}+X_{(n)})}{2}$ **has the least MSE** (which we have also seen during parametric estimation), and for **Cauchy and Double Exponential the Hodges-Lehmann estimate has the least MSE.** This is because **Cauchy distribution has thick tails, and the Expectation doesn't exist**. In case of Double Exponential though the tail is usual, it has a sharp peak at the central value. Note that DE is a symmetric distribution thus the mean , median and mode should converge. Since there is a sharp peak, **it is wiser to take some measure based on Median**. It can be seen that the MSE for Median and HL estimate is very close for these two distributions.

Also another interesting observation is that **for any distribution the MSE of HL is lesser than that of the Median**, so it can be looked upon as **an improved measure of estimating location parameter (over median)** when other options like mean are not useful.

Note/Query :- After all the observation the question that is posed is whether calculation of ARE in this way is justified? That is we computed the values of ARE from a purely theoritical perspective, and now the empirically computed values are close by but not exactly same. Why is this discrepancy? If we are able to find some value precisely by theoritical calculations, why should the computed values not be far more accurate and close?

Outputs

```
[[1]]
[[1]]$Detail
     m     n Theta  Seed
    10    10    13    23

[[1]]$MSE
            Median     Mean min plus max mean max ordered welsh mean Winsorised HodgesLehmann
Normal      0.2043   0.1078            0.1315                23.175     0.1188        0.1107
Logistic    0.3599   0.3284            0.9041                24.375     0.3122        0.2915
Cauchy      0.2498 564.2309        14074.8192                26.525     6.4019        0.3436
Double Expo 0.0805   0.0872            0.3772                28.575     0.1007        0.1106
Uniform     0.0186   0.0074            0.0040                24.375     0.0084        0.0085

[[1]]$Variance
            Median     Mean min plus max mean max ordered welsh mean Winsorised HodgesLehmann
Normal      0.2085   0.1135            0.1441                0.6806     0.1255        0.1170
Logistic    0.3310   0.2842            0.9018                0.9472     0.2655        0.2714
Cauchy      0.2775 571.1082        14203.7031                1.1361     6.7519        0.3798
Double Expo 0.0892   0.0861            0.3951                1.1250     0.1009        0.1131
Uniform     0.0175   0.0068            0.0043                0.9472     0.0077        0.0081

[[1]]$ARE
              Median/Mean HL/Mean  HL/Median
Normal             0.5444    0.9701    1.7821
Logistic           0.8586    1.0472    1.2196
Cauchy          2058.0476 1503.7077    0.7306
Double Expo        0.9652    0.7613    0.7887
Uniform            0.3886    0.8395    2.1605

[[2]]
[[2]]$Detail
     m     n Theta  Seed
    10   100    13    23

[[2]]$MSE
            Median    Mean min plus max mean max ordered welsh mean Winsorised HodgesLehmann
Normal      0.0117  0.0105            0.0765               6067.775     0.0121        0.0110
Logistic    0.0680  0.0629            1.1307               6542.650     0.0549        0.0577
Cauchy      0.0221 56.2319        80922.9935               6159.975     0.0347        0.0172
Double Expo 0.0131  0.0202            1.2172               6271.525     0.0166        0.0127
Uniform     0.0042  0.0016            0.0000               6542.650     0.0017        0.0016

[[2]]$Variance
            Median    Mean min plus max mean max ordered welsh mean Winsorised HodgesLehmann
Normal      0.0119  0.0101            0.0828                7.9472     0.0105        0.0106
Logistic    0.0712  0.0690            1.2048               15.5667     0.0599        0.0618
Cauchy      0.0225 52.0124        74011.2924               23.6139     0.0342        0.0176
Double Expo 0.0140  0.0222            1.2625                7.5583     0.0184        0.0141
Uniform     0.0043  0.0017            0.0001               15.5667     0.0018        0.0017

[[2]]$ARE
              Median/Mean HL/Mean  HL/Median
Normal             0.8487    0.9528    1.1226
Logistic           0.9691    1.1165    1.1521
Cauchy          2311.6622 2955.2500    1.2784
Double Expo        1.5857    1.5745    0.9929
Uniform            0.3953    1.0000    2.5294
```

```
[[3]]
[[3]]$Detail
     m     n Theta  Seed
   100    10    13    23

[[3]]$MSE
            Median     Mean min plus max mean max ordered welsh mean Winsorised HodgesLehmann
Normal      0.1452   0.0919            0.1535                26.6250     0.1008        0.0943
Logistic    0.3991   0.3799            1.0092                25.9075     0.3578        0.3481
Cauchy      0.3233 396.9880         9797.0246                25.2200     6.3141        0.4773
Double Expo 0.1440   0.2009            0.8450                25.4450     0.1851        0.1602
Uniform     0.0202   0.0086            0.0036                25.9075     0.0097        0.0109

[[3]]$Variance
            Median     Mean min plus max mean max ordered welsh mean Winsorised HodgesLehmann
Normal      0.1453   0.0915            0.1528                 0.7241     0.1005        0.0944
Logistic    0.3978   0.3829            1.0138                 0.8661     0.3588        0.3486
Cauchy      0.3237 391.4821         9654.4573                 0.8246     6.2877        0.4777
Double Expo 0.1445   0.2028            0.8432                 0.8519     0.1869        0.1617
Uniform     0.0201   0.0086            0.0036                 0.8661     0.0097        0.0109

[[3]]$ARE
             Median/Mean HL/Mean  HL/Median
Normal           0.6297   0.9693    1.5392
Logistic         0.9625   1.0984    1.1411
Cauchy        1209.3979 819.5145    0.6776
Double Expo      1.4035   1.2542    0.8936
Uniform          0.4279   0.7890    1.8440

[[4]]
[[4]]$Detail
     m     n Theta  Seed
   100   100    13    23

[[4]]$MSE
            Median    Mean min plus max mean max ordered welsh mean Winsorised HodgesLehmann
Normal      0.0164  0.0103            0.0976                6208.880     0.0117        0.0111
Logistic    0.0387  0.0347            0.7167                6299.810     0.0326        0.0321
Cauchy      0.0249 45.8115       103455.6161                6271.203     0.0802        0.0323
Double Expo 0.0122  0.0244            0.9634                6288.477     0.0212        0.0163
Uniform     0.0024  0.0009            0.0001                6299.810     0.0010        0.0010

[[4]]$Variance
            Median    Mean min plus max mean max ordered welsh mean Winsorised HodgesLehmann
Normal      0.0164  0.0104            0.0983                16.9332     0.0117        0.0111
Logistic    0.0379  0.0350            0.6966                21.0428     0.0328        0.0320
Cauchy      0.0246 44.7227       101824.7388                18.5322     0.0810        0.0323
Double Expo 0.0121  0.0246            0.9700                15.1974     0.0214        0.0164
Uniform     0.0023  0.0009            0.0000                21.0428     0.0010        0.0010

[[4]]$ARE
             Median/Mean  HL/Mean  HL/Median
Normal           0.6341    0.9369    1.4775
Logistic         0.9235    1.0938    1.1844
Cauchy        1817.9959 1384.6037    0.7616
Double Expo      2.0331    1.5000    0.7378
Uniform          0.3913    0.9000    2.3000
```

Scripts

```
#File: Source.R

library(psych)
library(nimble)

#function for calculating the max ordered welsh stat
or_mean=function(x)
{
  l=length(x)
  x=order(x,decreasing = F)
  b=rep(0,l)
  for (i in 1:l)
  {
    b[i]=(x[i]+x[l-i+1])/2
  }
  return(max(b))
}



#function for calculating the stats for different distributions
#Choices are described below
#0 for normal
#1 for logistic
#2 for cauchy
#3 for double expo
#4 for unif

gen_func=function(m,n,choice,theta,seed)
{
  set.seed(seed)

  #Creating the observation matrix
  X=matrix(NA, nrow = m, ncol = n)

  #Creating the matrix of the statist-ics
  stat=matrix(NA, nrow = m, ncol = 6)
  colnames(stat)=c("Median", "Mean", "min plus max mean", "max ordered welsh mean", "Winsorised",
"HodgesLehmann")

  #taking the choice of the distribution
  if (choice == 0)
  {
    for (i in 1:m)
    {
      X[i,]=rnorm(n, theta, 1)
    }
  }
  else if (choice == 1)
  {
    for (i in 1:m)
    {
      X[i,]=rlogis(n, theta, 1)
    }
  }

  else if (choice == 2)
  {
    for (i in 1:m)
    {
```

```
      X[i,]=rcauchy(n, theta, 1)
    }
  }

  else if (choice == 3)
  {
    for (i in 1:m)
    {
      X[i,]=rdexp(n, theta, 1)
    }
  }

  else if (choice == 4)
  {
    for (i in 1:m)
    {
      X[i,]=runif(n, theta-0.5,theta+0.5)
    }
  }

  #Computing the stats
  for (i in 1:m)
  {
    stat[i,1]=median(X[i,])
    stat[i,2]=mean(X[i,])
    stat[i,3]=(min(X[i,])+max(X[i,]))/2
    stat[i,4]=or_mean(X[i,])
    stat[i,5]=winsor.means(X[i,], trim = 0.1)
    stat[i,6]=wilcox.test(X[i,], conf.int = T)$estimate
  }

  return(stat)
}

#Creating the Mean-squared error matrix for all the estimators and distributions
MSE=matrix(NA, nrow=5, ncol=6)
colnames(MSE)=c("Median", "Mean", "min plus max mean", "max ordered welsh mean", "Winsorised",
"HodgesLehmann")
rownames(MSE)=c("Normal","Logistic","Cauchy","Double Expo","Uniform")

#Creating the variance matrix for all the estimators and distributions
V=matrix(NA, nrow=5, ncol=6)
colnames(V)=c("Median", "Mean", "min plus max mean", "max ordered welsh mean", "Winsorised",
"HodgesLehmann")
rownames(V)=c("Normal","Logistic","Cauchy","Double Expo","Uniform")

#Creating the ARE matrix for all distributions b/w Median-mean, HL-mean, HL-median
ARE=matrix(NA, nrow=5, ncol=3)
rownames(ARE)=c("Normal","Logistic","Cauchy","Double Expo","Uniform")
colnames(ARE)=c("Median/Mean", "HL/Mean", "HL/Median")


#File: Output.R

#Setting the folder containing all the files are set as working directory
setwd("~/Documents/Non para/MSE compare of estimate of locations")

#run the Source.R script to load the required functions
source("Source.R", echo=T)
```

```
#disabling exponential notation
options(scipen = 999)

#parameter values to be specified
theta=13
seed=23
m=10 #number of sample sets
n=10 #number of observation in each sample set

for (k in 1:5)
{
  temp=gen_func(m,n,k-1,theta,seed)
  for (l in 1:6)
  {
    MSE[k,l]=mean((temp[,l]-theta)^2)
    V[k,l]=var(temp[,l])
  }
}
detail=c(m,n,theta,seed)
names(detail)=c("m","n","Theta","Seed")
MSE=round(MSE, 4)
V=round(V, 4)
output=list(detail,MSE,V)
names(output)=c("Detail","MSE","Variance")

ARE[,1]=output$Variance[,2]/output$Variance[,1]
ARE[,2]=output$Variance[,2]/output$Variance[,6]
ARE[,3]=output$Variance[,1]/output$Variance[,6]

output[["ARE"]]=round(ARE, 4)

output
output.10.10=output #storing outputs for a specific m, n.
```