Question 2

Description and Context of population:

The data set is of Billboard Top 30 songs. The following data has different variates of every song in the Billboard Top 30 for the years 2010 to 2019. We will compare and contrast the mean and median attributes.

The ratio of the $\frac{popularity\ of\ song}{duration\ of\ song}$ will help us understand how the duration of the song and the popularity depend on each other. The dataset of interest is

```
spotify <- read.csv("./spotify.csv", header=TRUE)
Spotify = na.omit(spotify)
```

Calculating the ratio $\frac{Popularity\ of\ Song}{Duration\ of\ Song}$

```
ratio.variable = Spotify$popularity / Spotify$duration
```

Calculating α which makes the skewness equal to 0.

```
ratioSkew = createSkewFunction( ratio.variable )
ratio.alpha= uniroot(ratioSkew, interval=c(-1,1))$root;
ratio.alpha
```

```
## [1] 0.3349566
```

Generating transformed variable

```
trans.ratio = powerfun(ratio.variable, ratio.alpha)
```

Standardizing the variable i.e performing a location and scale shift

```
st.ratio = (trans.ratio - mean(trans.ratio))/sdn(trans.ratio)
```

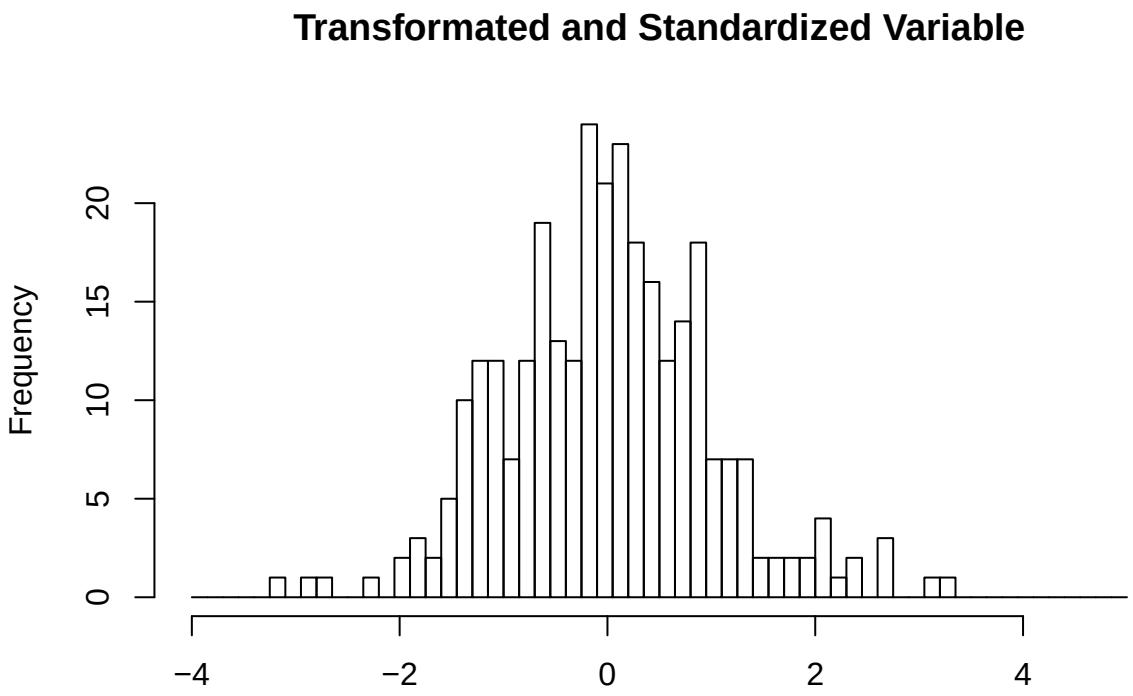
Verifying the mean, standard deviation and skewness

```
c(mean(st.ratio), sd(st.ratio), skew(st.ratio))
```

```
## [1] -7.900704e-16  1.001671e+00 -4.516119e-06
```

Plotting the histogram of the transformed and standardized variable

```
hist( st.ratio, breaks=seq(-4, 5, 0.15), main="Transformated and Standardized Variable", xlab="")
```




```
sim.avg.mean <- function(pop=NULL, n=NULL, m=10^4) {
N = length(pop);
set.seed(341)
temp = unlist(Map(function(rep) {
  sam.values = pop[sample(N, n, replace=FALSE)]
  c(mean(sam.values), median(sam.values))}, 1:m))
  temp = matrix( temp, nrow=m, ncol=2, byrow=TRUE)
  temp = c(apply(temp, 2, mean)- c(mean(pop) , median(pop)) , apply(temp, 2, sd) )
  result=c(temp,temp[1]^2+temp[3]^2,temp[2]^2+temp[4]^2)
  return(result)
}
```

Numerical Comparision

We will now find

- Average Standard Bias and Meadian Standard Bias
- Average Standard Deviation and Median Standard Deviation
- Average Mean Squared Error and Median Mean Squared Error

For the sample sizes n=30,60,...,300 to find the effects of changes in some parameters such as the sample size.

```r
n.set = seq(30,300, by=30)
result = matrix(nrow=length(n.set), ncol=6,
dimnames = list(n.set, c("Avg. SB", "Median SB", "Avg. SD", "Median SD", "Avg. MSE","Median MSE" )))
for (i in 1:length(n.set)){
  result[i,] = sim.avg.mean(st.ratio, n=n.set[i])
}
round(result,4)
```

```
##     Avg. SB Median SB Avg. SD Median SD Avg. MSE Median MSE
## 30   0.0002   -0.0093  0.1725    0.1800   0.0298     0.0325
## 60  -0.0011   -0.0041  0.1166    0.1171   0.0136     0.0137
## 90  -0.0004   -0.0014  0.0883    0.0902   0.0078     0.0081
## 120 -0.0002   -0.0003  0.0707    0.0730   0.0050     0.0053
## 150  0.0001    0.0025  0.0577    0.0606   0.0033     0.0037
## 180  0.0000    0.0036  0.0470    0.0496   0.0022     0.0025
## 210 -0.0002    0.0046  0.0380    0.0401   0.0014     0.0016
## 240  0.0000    0.0049  0.0290    0.0308   0.0008     0.0010
## 270  0.0000    0.0042  0.0190    0.0222   0.0004     0.0005
## 300  0.0000    0.0000  0.0000    0.0000   0.0000     0.0000
```

Graphical Comparision

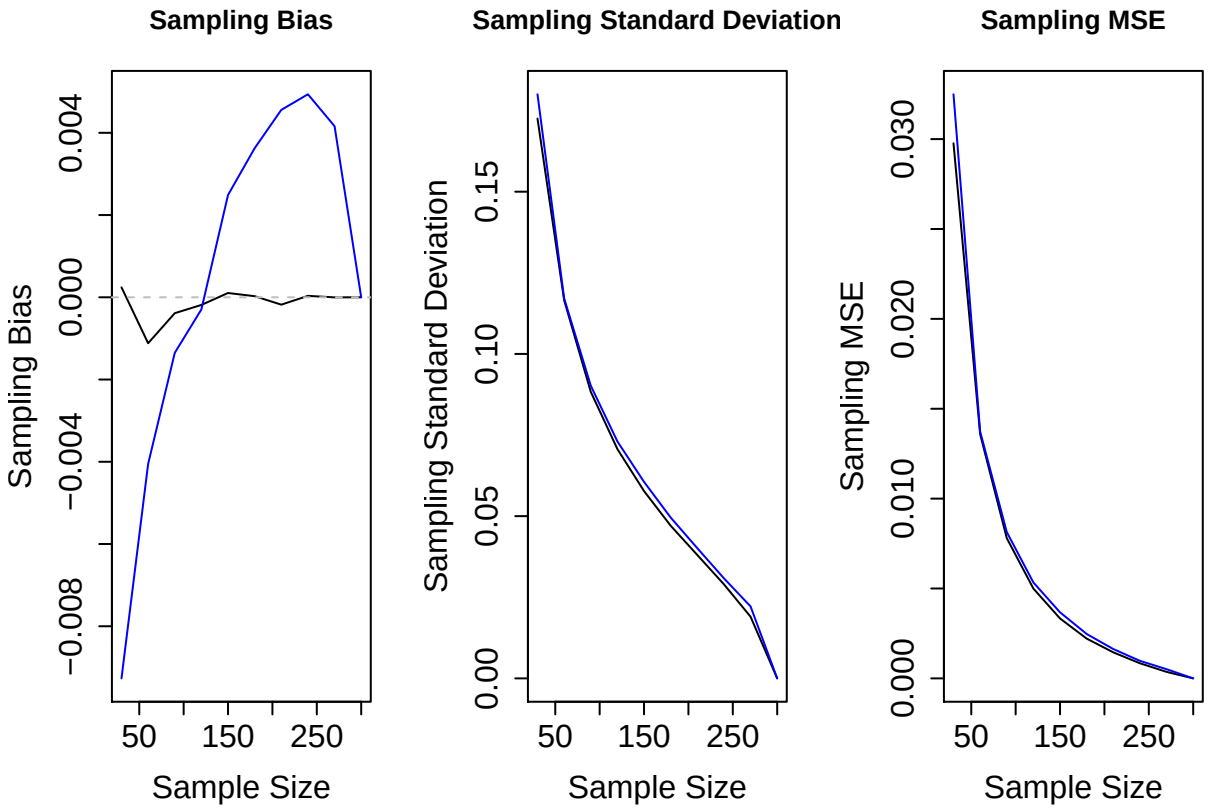
We will now look at the above estimators graphically. The black line will correspond to the average estimator and the blue line corresponds to the median estimator.

```r
par(mfrow=c(1,3),oma=c(0,0,0,0))

plot( result[,1]~n.set, main="Sampling Bias", type='l', ylim=range(result[,1:2]), ylab="Sampling Bias",
lines( result[,2]~n.set, col=4)
abline(h=0, lty=2, col="grey")

plot(result[,3]~n.set, main="Sampling Standard Deviation", type='l', ylim=c(0,max(result[,3:4])), xlab=
lines( result[,4]~n.set, col=4)

plot( result[,5]~n.set, main="Sampling MSE", type='l',
ylim=c(0,max(result[,5:6]) ), xlab="Sample Size", ylab="Sampling MSE",
cex.lab=1.5 , cex.axis=1.5)
lines( result[,6]~n.set, col=4)
```



Conclusion

We would like to choose the estimator whose MSE is the least.

- SB denotes the estimate of the sampling bias and SD denotes the estimate of the sampling standard deviation. The last two columons in the table above show the MSE calculated from the sample.
- The Average performs better then the median in this situation as its MSE is consistently smaller than that of the median.