HW#6

Haichuan Du

7/8/2020

A close up of a piece of paper

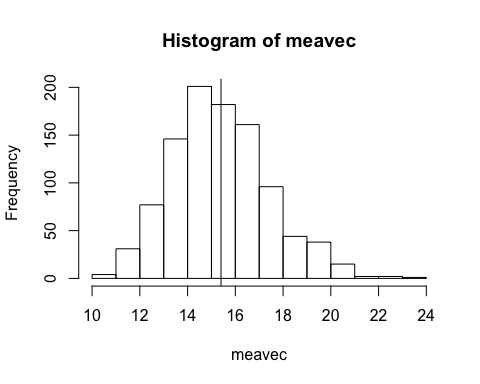
Description automatically generated

setwd("~/Desktop/statistics/Nonparametric/project")  
churn\_data <- read.csv("churn.csv")  
head(churn\_data)

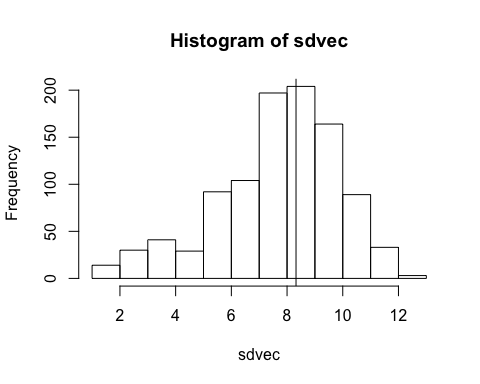
## customerID gender SeniorCitizen Partner Dependents tenure PhoneService  
## 1 7590-VHVEG Female 0 Yes No 1 No  
## 2 5575-GNVDE Male 0 No No 34 Yes  
## 3 3668-QPYBK Male 0 No No 2 Yes  
## 4 7795-CFOCW Male 0 No No 45 No  
## 5 9237-HQITU Female 0 No No 2 Yes  
## 6 9305-CDSKC Female 0 No No 8 Yes  
## MultipleLines InternetService OnlineSecurity OnlineBackup  
## 1 No phone service DSL No Yes  
## 2 No DSL Yes No  
## 3 No DSL Yes Yes  
## 4 No phone service DSL Yes No  
## 5 No Fiber optic No No  
## 6 Yes Fiber optic No No  
## DeviceProtection TechSupport StreamingTV StreamingMovies Contract  
## 1 No No No No Month-to-month  
## 2 Yes No No No One year  
## 3 No No No No Month-to-month  
## 4 Yes Yes No No One year  
## 5 No No No No Month-to-month  
## 6 Yes No Yes Yes Month-to-month  
## PaperlessBilling PaymentMethod MonthlyCharges TotalCharges  
## 1 Yes Electronic check 29.85 29.85  
## 2 No Mailed check 56.95 1889.50  
## 3 Yes Mailed check 53.85 108.15  
## 4 No Bank transfer (automatic) 42.30 1840.75  
## 5 Yes Electronic check 70.70 151.65  
## 6 Yes Electronic check 99.65 820.50  
## Churn  
## 1 No  
## 2 No  
## 3 Yes  
## 4 No  
## 5 Yes  
## 6 Yes

attach(churn\_data)

set.seed(0)  
samp <- c(7, 10, 10, 10, 11, 12, 12, 13, 13, 14, 14, 15, 22, 31, 37)  
n=length(samp)  
B=1000 #Number of bootstrap samples to draw.  
sdvec=double(B) #Saving the bootstrap standard deviations.  
meavec=double(B) #Saving the bootstrap sample means.  
coevar=double(B)  
sd=sd(samp) #Observed standard deviation.  
mea=mean(samp) #Observed sample mean.  
cov=sd(samp)/mean(samp)  
for (brun in 1:B){  
 bsamp=sample(samp,n,replace=T) #Sample with replacement from the sample., 从sample里面取10次  
 meavec[brun]=mean(bsamp)  
 sdvec[brun]=sd(bsamp)  
 coevar[brun]=sd(bsamp)/mean(bsamp)}  
hist(meavec) #Plotting the bootstrap means.  
abline(v=mea)



hist(sdvec) #Plotting the bootstrap standard deviations.  
abline(v=sd)



#Now studying the mean.  
  
ehat=mean(meavec) #Mean of the bootstrap values. -------point estimate  
msehat=mean((meavec-mea)^2) #Estimated MSE.  
se=sqrt(msehat) #Bootstrap SE---------> standard error.  
bhat=ehat-mea #Estimated ----------> bootstrap bias.  
CI=quantile(meavec,c(0.025,0.975))  
ehat

## [1] 15.39653

msehat

## [1] 4.140818

se

## [1] 2.0349

bhat

## [1] -0.003466667

CI

## 2.5% 97.5%   
## 11.73333 19.80167

#Now studying the standard deviation.  
  
ehat=mean(sdvec) #standard deviation. of the bootstrap values.----> point estimate  
msehat=mean((sdvec-sd)^2) #Estimated MSE.  
se=sqrt(msehat) #Bootstrap SE.  
bhat=ehat-sd #Estimated bias.  
CI=quantile(sdvec,c(0.025,0.975))  
ehat

## [1] 7.659678

msehat

## [1] 5.348207

se

## [1] 2.312619

bhat

## [1] -0.6624092

CI

## 2.5% 97.5%   
## 2.260737 11.342461

#Now studying the population coefficient of variation.  
  
ehat=mean(coevar) #coefficient of variation of the bootstrap values.----> point estimate  
msehat=mean((coevar-cov)^2) #Estimated MSE.  
se=sqrt(msehat) #Bootstrap SE.  
bhat=ehat-cov #Estimated bias.  
CI=quantile(coevar,c(0.025,0.975))  
ehat

## [1] 0.4897533

msehat

## [1] 0.01428481

se

## [1] 0.1195191

bhat

## [1] -0.050642

CI

## 2.5% 97.5%   
## 0.2021840 0.6418488

1. For the population mean point estimates = 15.39653; bootstrap standard errors = 2.0349 bootstrap bias estimates = -0.003466667 95% bootstrap percentile confidence intervals = (11.73333, 19.80167)
2. For the population standard deviation point estimates = 7.659678 bootstrap standard errors = 2.312619 bootstrap bias estimates = -0.6624092 95% bootstrap percentile confidence intervals = (2.260737, 11.342461)
3. For the population coefficient of variation point estimates = 0.4897533 bootstrap standard errors = 0.1195191 bootstrap bias estimates = -0.050642 95% bootstrap percentile confidence intervals = (0.2021840, 0.6418488)

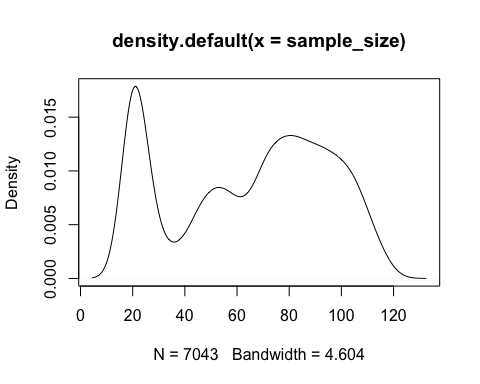
The population standard deviation is seriously biased (bootstrap bias estimates = -0.6624092)

#3

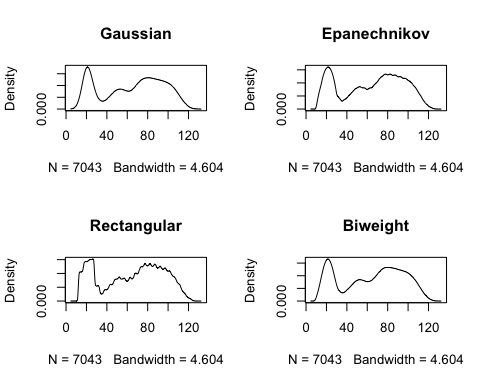
sample\_size <- churn\_data$MonthlyCharges  
length(sample\_size)

## [1] 7043

plot(density(sample\_size))



par(mfrow=c(2,2))  
fit1=density(sample\_size,kernel='gaussian') #Using default bandwidth.  
fit2=density(sample\_size,kernel='epanechnikov')  
fit3=density(sample\_size,kernel='rectangular')  
fit4=density(sample\_size,kernel='biweight')  
plot(fit1,main='Gaussian')  
plot(fit2,main='Epanechnikov')  
plot(fit3,main='Rectangular')  
plot(fit4,main='Biweight')



#Different bandwidths for gaussian kernal  
# par(mfrow=c(4,4))  
# plot(density(sample\_size,kernel='gaussian',bw=0.1),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=1),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=3),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=5),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=10),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=20),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=25),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=30),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=40),main='')  
# plot(density(sample\_size,kernel='gaussian',bw=50),main='')  
#   
# #Different bandwidths for epanechnikov kernal  
# par(mfrow=c(4,4))  
# plot(density(sample\_size,kernel='epanechnikov',bw=0.1),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=1),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=3),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=5),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=10),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=20),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=25),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=30),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=40),main='')  
# plot(density(sample\_size,kernel='epanechnikov',bw=50),main='')  
#   
# #Different bandwidths for rectangular kernal  
# par(mfrow=c(4,4))  
# plot(density(sample\_size,kernel='rectangular',bw=0.1),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=1),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=3),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=5),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=10),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=20),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=25),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=30),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=40),main='')  
# plot(density(sample\_size,kernel='rectangular',bw=50),main='')  
#   
# #Different bandwidths for biweight kernal  
# par(mfrow=c(4,4))  
# plot(density(sample\_size,kernel='biweight',bw=0.1),main='')  
# plot(density(sample\_size,kernel='biweight',bw=1),main='')  
# plot(density(sample\_size,kernel='biweight',bw=3),main='')  
# plot(density(sample\_size,kernel='biweight',bw=5),main='')  
# plot(density(sample\_size,kernel='biweight',bw=10),main='')  
# plot(density(sample\_size,kernel='biweight',bw=20),main='')  
# plot(density(sample\_size,kernel='biweight',bw=25),main='')  
# plot(density(sample\_size,kernel='biweight',bw=30),main='')  
# plot(density(sample\_size,kernel='biweight',bw=40),main='')  
# plot(density(sample\_size,kernel='biweight',bw=50),main='')

For Gaussian:

*A screenshot of a cell phone

Description automatically generated*

For epanechnikov:

*A close up of a piece of paper

Description automatically generated*

For rectangular:

*A close up of a piece of paper

Description automatically generated*

For biweight:

*A close up of a piece of paper

Description automatically generated*

1. The sample size: Sample size for MonthlyCharges is 7043
2. The kernel that I used: Gaussian, Epanechnikov, Rectangular, Biweight
3. The bandwidth: The bandwidth I used including 0.1, 1, 3, 5, 10, 20, 25, 30, 40, 50

Comment on what the plot of the estimated density suggests about the shape of the distribution. Please indicate which bandwidths you tried (Do try some!) in addition to the default choice.

In general, with the constant increase of the bandwidth of the four kernels, the shape of the density gradually presents unimodal shape, approaching their original kernel function.

For gaussian kernel: When the bandwidth is less than 3, the density shape appears jagged; when the bandwidth is greater than 3 and less than 10, the density shape changes from 3 humps to 2; when the bandwidth is greater than 20, the density shape initially appears unimodal and tends to the original kernel function

For epanechnikov kernel: similar to gaussian kernel, when the bandwidth <= 3, the density shape appears jagged; when the bandwidth is greater than 3 and less than 10, the density shape changes from 3 humps to 2 but has a smoother curve; when the bandwidth is greater than 20, the density shape initially appears unimodal and tends to the original kernel function。

For rectangular kernel: When the bandwidth is less than 10, the density shape is jagged, when the bandwidth is greater than 3 and less than 20, the hump gradually changes from 3 to 1 and the curve tends to be flat. When the bandwidth is greater than 25, the top shape of the hump tends to be close to the original kernel function like straight line.

For biweight kernel: very similar to the gaussian kernel, gaussian kernel is little bit smoother than biweight.