load("C:/Users/ggard/stat 135/labs/KaiserBabies.rda")

#NUMBER !  
#part a)  
  
set.seed(7)  
mysample = sample(na.omit(infants$wt), 10)  
  
s1 = infants[mysample,]  
  
#For this project, I will measure variance of Xbar using the formula s^2\*(N-n)/(N\*n) ~= the sd of Xbar. The book gave the impression this is the recommended estimate to use.  
unbvar = function(x, N){  
 return(var(x)\*(N-length(x))/((length(x)-1)\*N))  
}  
  
unbvar(s1$wt, length(infants$wt))^.5

## [1] 13.93359

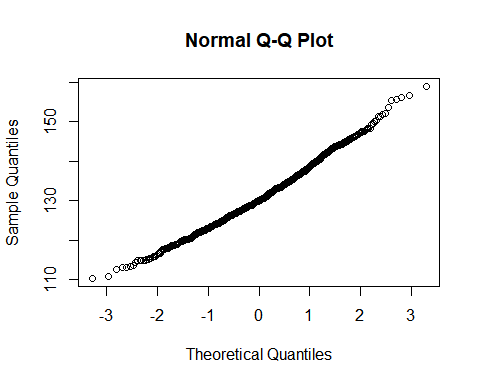
#sd ~= 13.9  
  
mean(s1$wt)

## [1] 144

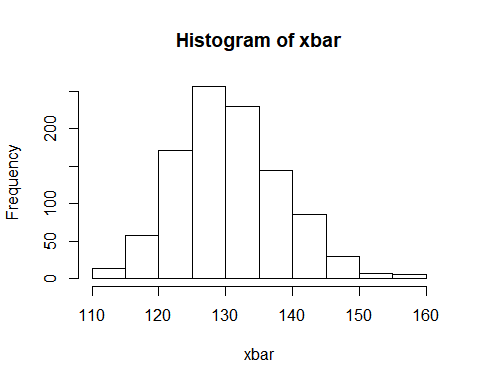
# = 144  
  
#So the 95% condifence interval will be 144-13.9\*1.96=116.75 to 144+13.9\*1.96=171.244 or (116.75, 171.244)

#part b)  
  
# Deleting data and trimming NA values  
infants = infants[!is.na(infants$wt),]  
  
trum = mean(infants$wt)  
# ~=128.6258  
  
lower = 0.0[-1]  
upper = 0.0[-1]  
xbar = 0.0[-1]  
staddev = 0.0[-1]  
  
for (i in 1:1000){  
 sam = infants[sample(na.omit(infants$wt), 10),]  
 usd = sqrt(unbvar(sam$wt, length(infants$wt)))  
 staddev = c(staddev, usd)  
 lower = c(lower, mean(sam$wt) - 1.96\*usd)  
 xbar = c(xbar,mean(sam$wt))  
 upper = c(upper, mean(sam$wt) + 1.96\*usd)  
}  
  
fish = (lower < trum) & (trum < upper)  
  
confidence = sum(fish, na.rm=TRUE)/length(fish)  
#It appears as if only around 94% of our confidence intervals include our true value (though we woud expect 95% to cover it)!

#part c)  
  
#Plots of WEIGHT  
#qqnorm(infants$wt-mean(infants$wt)/sd(infants$wt))  
#hist(infants$wt-mean(infants$wt)/sd(infants$wt))  
  
#Plots of XBAR for sample size 10  
  
qqnorm(xbar)



hist(xbar)



# Looking at the distribution of Xbar, it appears to be almost normal, and the qq plot is very close to linear. The distribution seems to still be slighty skew right however.  
  
sd(xbar)

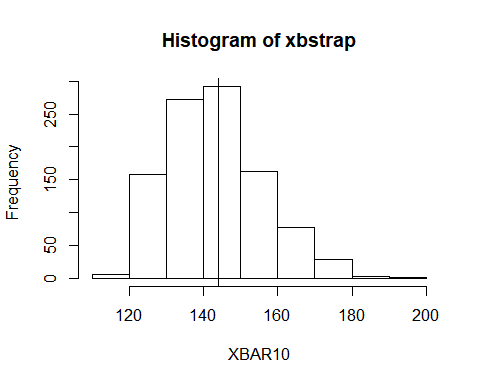
## [1] 7.794421

#~ 7.87  
  
#the estimate for the SD of xbar using our sample is larger than the SD of Xbar claculated when repeated samples of xbar are teaken. I looked at the SD's of some of the 1000 xbars calculated and the <13 SD of orginial sample seemed to be on the upper end of SD's estimated for our thousand samples. Our original sample had an abnormally high SD, possibly leading to an inflated s^2   
  
mean(staddev)

## [1] 8.022787

# sd ~= 8  
  
#when I use the formula to calculate the expected SD of XBAR for each sample individually using our estimatation formala, i get an avergae of around 8 for the SD of XBar. This is extremely close to the SD found by looking at the distributions of Xbars themselves.

#after we removed our NA's, the original sample is around 1,200 entries, so we will replicate our 10 point smaple 120 times to bootstrap.  
  
bs = 0.0[-1]  
for(i in 1:120)  
 bs = c(s1$wt,bs)  
  
xbstrap = 0.0[-1]  
sdbstrap = 0.0[-1]  
  
for (i in 1:1000){  
 sam = bs[sample(na.omit(bs), 10)]  
 usd = sqrt(unbvar(sam, length(bs)))  
 sdbstrap=c(sdbstrap, usd)  
 xbstrap = c(xbstrap, mean(sam))  
}  
  
hist(xbstrap, xlab = "XBAR10")  
abline(v=mean(bs))



sd(xbstrap)

## [1] 12.4412

#~=12.14  
mean(sdbstrap)

## [1] 12.46442

#Mean is right of mode, so the distribution is again right skew. The mean of the boostrap population should be the expected value the mean of a random sample, so it shoud also be the mean of XBAR (in expectation).  
# the SD of xbar caluculated from the distribution of XBAR is around 12, farily close to that estimated in part 1a) than that found from the actual distribution. When I used to the formula to estimate the sd of XBAR from each sample, I found their avergae was around 12.47, also fairly close.

quantile(xbstrap,c(.025,.975))

## 2.5% 97.5%   
## 124.2575 170.9000

# ~= (122.5, 170.5)   
#This looks similar to the confidence interval I found in 1a). however, in 1a the CI extended slightly further in the negative direction (the left endpoint). The original CI ended went to 116.75. In my opion, this discprepcy is caused by the skew distribution of XBAR, while our estimate in 1a assumed a normal distribution. Due to the actua distribution being "bunched up"" on the left side of the mean, the normal distribution over estimated how far left many of these values actually are.

#PART 3  
  
set.seed(7)  
mysample = sample(na.omit(infants$wt), 100)  
  
s2 = infants[mysample,]  
  
unbvar(s2$wt, length(infants$wt))^.5

## [1] 2.601528

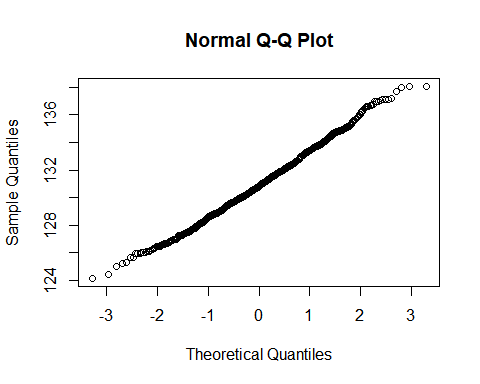
#sd ~= 2.602  
  
mean(s2$wt)

## [1] 130.97

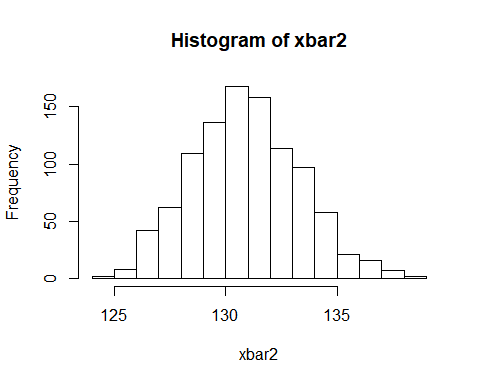
# = 130.97, this is much close to our true mean of 128.6 than our estimate in part a  
  
#So the 95% condifence interval will be 130.97-2.602\*1.96=125.8701 to 130.97+2.602\*1.96=136.0699 or (125.87, 136.0699)

lower2 = 0.0[-1]  
upper2 = 0.0[-1]  
xbar2 = 0.0[-1]  
staddev2 = 0.0[-1]  
  
for (i in 1:1000){  
 sam = infants[sample(na.omit(infants$wt), 100),]  
 usd = sqrt(unbvar(sam$wt, length(infants$wt)))  
 staddev2 = c(staddev2, usd)  
 lower2 = c(lower2, mean(sam$wt) - 1.96\*usd)  
 xbar2 = c(xbar2,mean(sam$wt))  
 upper2 = c(upper2, mean(sam$wt) + 1.96\*usd)  
}  
  
fish2 = (lower2 < mean(xbar2)) & (mean(xbar2) < upper2)  
  
confidence2 = sum(fish2, na.rm=TRUE)/length(fish2) # ~= 87.5%  
  
#only about 87.5% of confidence intervals contain the true mean. again, our seed seems to have yeilded somewhat of an outlayer as our sample mean is about 1 SD (of XBAR) above the population mean or in the 85th percentile. I found a peculairity in that the average of the samples means was signifigantly higher than the actual population value. (about and SD of XBAR). A much higher proportion (around 94%) of CI's included the mean of XBAR.

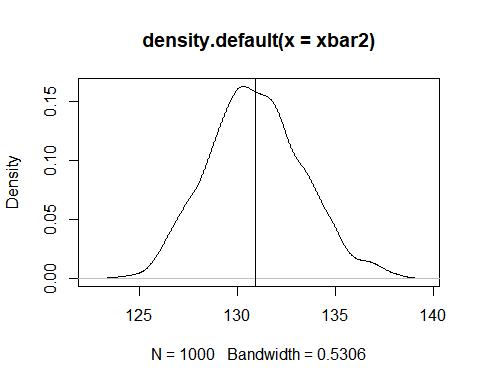
#Plots of XBAR for sample size 100  
  
qqnorm(xbar2)



hist(xbar2)



plot(density(xbar2))  
abline(v=mean(xbar2))



# As with the smaller samples, the distribution of XBar looks very close to normal, even more so than in part 1. The distribution again looks slighty right skew (more bunched up towards left), but this may be due to the nature of bar plots in R. Looking at the desnity plot, this appear to be the case  
  
sd(xbar2)

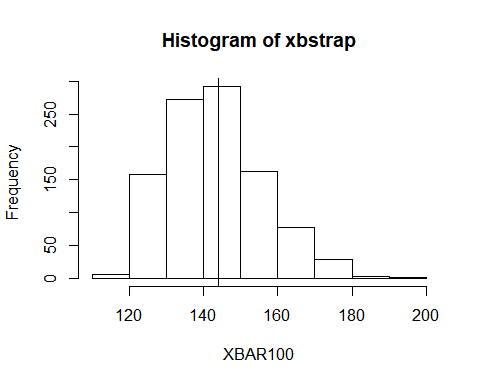
## [1] 2.40697

#~ 2.515  
  
#this is actually quite close to the SD in part 1 of 2.603.  
  
mean(staddev2)

## [1] 2.466816

# sd ~= 2.44  
  
#Again, very close. Our confidence intervals seem to samehwat accurately reflect the distribution, though as stated earlier most are centered away from teh true mean by a signifigant amount, which may be caused by the skewness of XBAR and weight. However, given under 90% of our 95% confidence intervals were in the right place, I would say we should reconsidder the confidence invertval or atleast note the way it was contructed does not fully match the reality of teh data.

bs2 = 0.0[-1]  
for(i in 1:12)  
 bs2 = c(s2$wt,bs2)  
  
xbstrap2 = 0.0[-1]  
sdbstrap2 = 0.0[-1]  
  
for (i in 1:1000){  
 sam = bs2[sample(na.omit(bs2), 100)]  
 usd = sqrt(unbvar(sam, length(bs2)))  
 sdbstrap2=c(sdbstrap2, usd)  
 xbstrap2 = c(xbstrap2, mean(sam))  
}  
  
hist(xbstrap, xlab = "XBAR100")  
abline(v=mean(bs))



sd(xbstrap2)

## [1] 2.607394

#~=2.57  
mean(sdbstrap2)

## [1] 2.582763

#~=2.58  
  
#again, these are very close to our values found in part a) and part c), more so for the answers in part a (this makes sense as our bootstrap was from the sample in part a). The bootstrap estimate for SD appears to be close to the true value as well (the one we estimated with the whole sample in part c)

quantile(xbstrap2,c(.025,.975))

## 2.5% 97.5%   
## 125.830 136.101

#~= (125.830,136.1012)  
# The ci found in part a was (125.87, 136.0699), which is slightly smaller, being .05 closer to the mean in either direction. .05 is about 2% of the estimated 2.6 SD for xbar, so the the confidence intervals are for most puposes almost identical.