STA 2210 Homework 4 (Due on Monday 6/22 by 11:59pm)

Software problems: The data set DOW, a .csv file, contains sixteen years of closing prices for the Dow Jones Industrial Average. Write your R codes, in addition to your answer, to the following problems. (Don’t forget to refer to the R reference card to find helpful commands.)

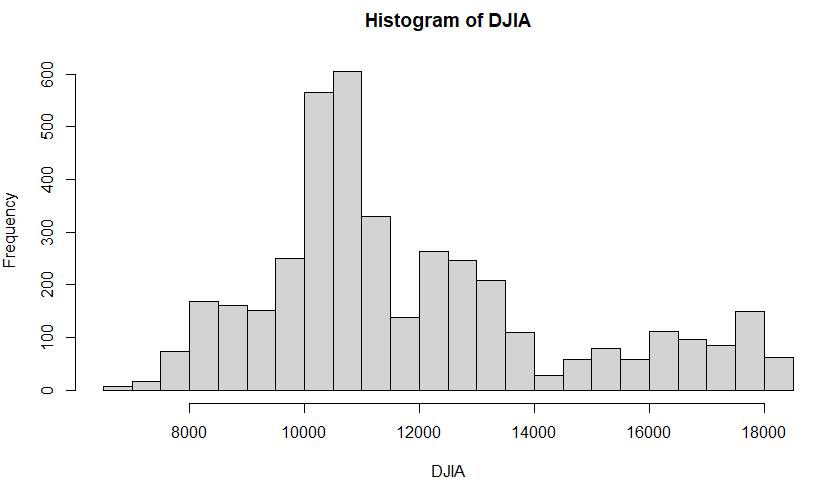
1. **Describe the population distribution, including its center, spread, modality, and shape.**

DJIA <- dow$DJIA

summary(DJIA)

Min. 1st Qu. Median Mean 3rd Qu. Max.

6547 10203 11011 11857 13096 18312

hist(DJIA, breaks = 25)

The distribution of DJIA is not a normal spread, and appears right skewed, as more data falls to the left side of the graph. The center appears around 11,000 on the histogram, the modality of the graph is unimodal, and the shape lacks in symmetry. The spread of the graph ranges from 6,547 to 18,312.

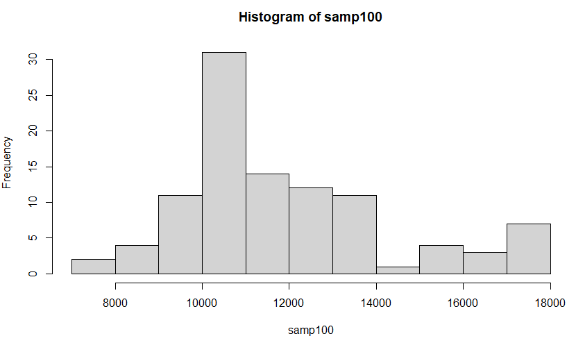
1. **Select one random sample of 100 closing prices (DJIA) and describe the distribution of the sample. How does it compare to the population distribution?**

samp100 <- sample(DJIA, 100)

summary(samp100)

Min. 1st Qu. Median Mean 3rd Qu. Max.

7553 10313 11092 11845 13019 17851

hist(samp100)

The shape of this histogram roughly resembles the population distribution, as the quartiles, means, and medians all are very close to each other. The only main difference is the 100-price sample has a smaller range because the random price samples didn’t include the mins and maxes. The histogram is not normal, lacks in symmetry, has a peak around 11,000 points, and is unimodal in shape, with a slightly smaller spread.

1. **Using your random sample of size 100, give point estimates for the population mean, median, and standard deviation. How close were your point estimates to the actual values?**

summary(DJIA)

Min. 1st Qu. Median Mean 3rd Qu. Max.

6547 10203 11011 11857 13096 18312

summary(samp100)

Min. 1st Qu. Median Mean 3rd Qu. Max.

7553 10313 11092 11845 13019 17851

sd(DJIA) = 2641.127

sd(samp100) = 2418.208

The sample has a very similar mean, median, and quartile values as the population. The only main difference is the mins and maxes are different because the random sample did not include these two values. The random sample has a smaller standard deviation by about 223 points. The point estimates are very close to the actual values, as shown by the summaries.

1. **Select 1000 samples of size 100 from the population and describe the sampling distribution. Be sure to specifically note its center.**

sample\_means\_100 <- rep(NA, 1000)

for(i in 1:1000){

samp <- sample(DJIA, 100)

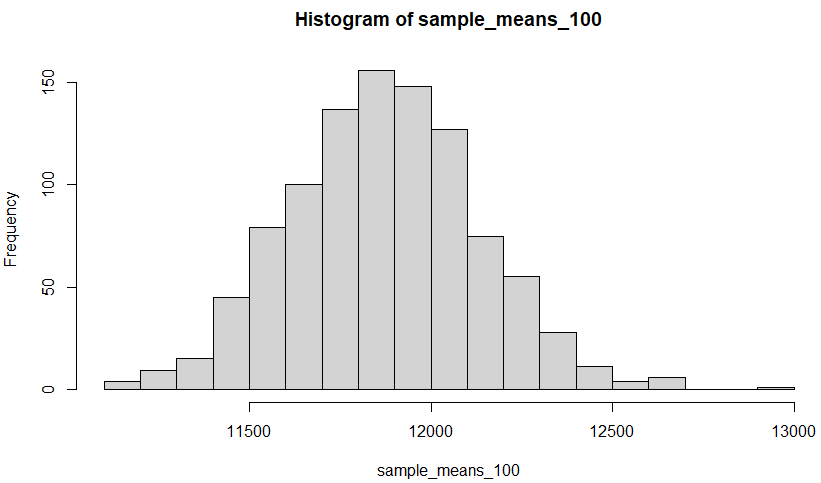
sample\_means\_100[i] <- mean(samp)

}

summary(sample\_means\_100)

Min. 1st Qu. Median Mean 3rd Qu. Max.

11133 11698 11872 11873 12041 12989

 hist(sample\_means\_100)

The 1000 sample histogram shows a normal distribution with a unimodal, symmetric shape, and a center appearing around 11,500-12,000. This lines up with the mean and median of the summary table, which show up around 11,872-11,873. The range goes from 11,133-12,989 which indicates smaller spread.

1. **Repeat number 4 using 3 different sample sizes: ones of size 50, ones of size 200, and ones of size 1000. What happens to the center and spread of each sampling distribution as the sample size increases?**

sample\_means\_50 <- rep(NA, 1000)

for(i in 1:1000){

samp <- sample(DJIA, 50)

sample\_means\_50[i] <- mean(samp)

}

sample\_means\_200 <- rep(NA, 1000)

for(i in 1:1000){

samp <- sample(DJIA, 200)

sample\_means\_200[i] <- mean(samp)

}

sample\_means\_1000 <- rep(NA, 1000)

for(i in 1:1000){

samp <- sample(DJIA, 1000)

sample\_means\_1000[i] <- mean(samp)

}

summary(sample\_means\_50)

Min. 1st Qu. Median Mean 3rd Qu. Max.

10753 11574 11833 11851 12119 13052

summary(sample\_means\_200)

Min. 1st Qu. Median Mean 3rd Qu. Max.

11223 11729 11859 11861 11991 12412

summary(sample\_means\_1000)

Min. 1st Qu. Median Mean 3rd Qu. Max.

11639 11803 11855 11855 11904 12078

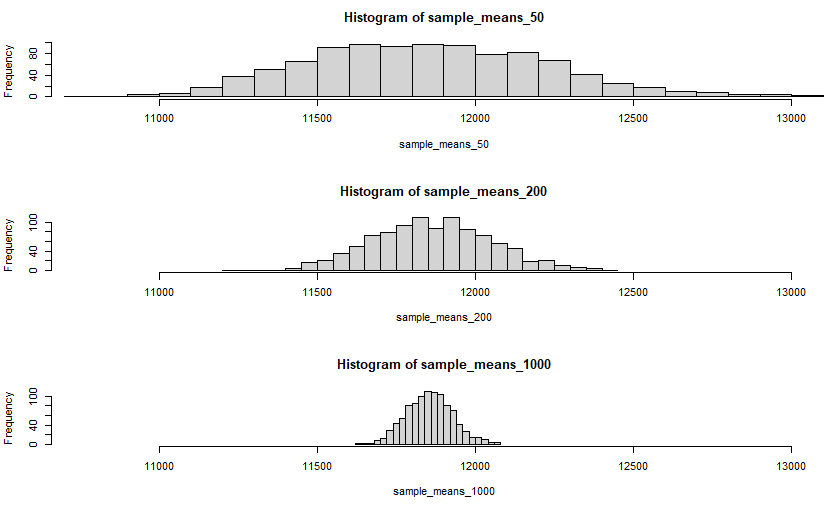
par(mfrow = c(3, 1))

xlimits <- range(sample\_means\_50)

hist(sample\_means\_50, breaks = 20, xlim = xlimits)

hist(sample\_means\_200, breaks = 20, xlim = xlimits)

hist(sample\_means\_1000, breaks = 20, xlim = xlimits)



As the sample sizes increase, the center of each histogram moves towards a point near 11,550-11,560, which is where the mean and median are near for all three histograms. The first histogram does not appear to have a concrete center, as there are many notable peaks between 11,500 and 12,000 points, but as the sample sizes increase, the center appears more apparent and begins revolving around a more singular point. The first histogram is the most variable because it has the smallest sample size, while the third histogram is the least variable because it has the largest sample size.

This is visualized by the spread, which decreases with an increase in the sample sizes. The graph of 1000 DJIA sample is much narrower than the 200 DJIA sample, which is also much narrower than the last 50 DJIA sample. This is because the standard error decreases as the sample size increases, which decreases spread.