Kartikeya Sharma

ASSIGNMENT 1 – Binomial Distribution

Exercise 1

Q) Generate 100 experiments of flipping 10 coins, each with 30% probability. What is the most common number? Why?

```
> rbinom(100,10,0.3) # no. of heads in 100 experiments
  [1] 4 2 6 2 2 5 3 4 0 2 4 3 4 2 2 2 1 2 3 3 2 2 3 5 2 4 5 1 2 4 3 2 0 4 5 2 4 0 2 3 4 4 4 3 4 0 2 2 3 4
 [51] 0 4 1 2 3 2 1 2 4 3 5 0 2 4 1 3 4 3 3 3 4 4 4 4 2 1 5 4 3 3 4 3 4 4 0 3 3 2 2 1 5 1 3 3 4 2 3 4 1 2

> rbinom(100,10,0.3) # no. of heads in 100 experiments
  [1] 3 4 3 2 3 4 4 2 4 3 4 2 2 3 3 5 6 3 5 2 4 3 4 1 4 5 2 1 0 6 1 4 4 2 3 5 1 5 2 3 4 2 1 3 6 5 3 4 1 1
 [51] 2 3 2 4 3 2 0 5 5 1 3 4 0 4 1 4 1 3 2 2 5 1 3 3 4 0 2 0 3 4 2 3 3 3 2 4 4 2 2 4 2 2 2 6 5 3 4 4 1 5

> rbinom(100,10,0.3) # no. of heads in 100 experiments
  [1] 5 3 4 3 3 4 2 1 4 3 2 2 1 5 3 2 4 2 3 3 3 2 5 3 2 1 2 3 5 4 3 3 3 2 1 3 4 4 5 2 2 1 3 6 3 3 1 4 0 3
 [51] 1 3 3 6 4 7 3 3 2 3 4 1 2 2 1 2 2 1 4 5 2 3 6 4 4 2 3 5 2 2 4 1 6 4 3 4 3 3 2 4 2 3 3 6 4 3 4 3 3 3
```

By looking at these three random generations, since the probability of getting a head is 0.3, almost all the numbers are less than 5. The majority are from range 2 to 4. However, as we generate more trails, the number of heads should get very close to its probable value of three (0.3 x 10). Thus, as we increase the experiments, the most common number of heads would be 3.

Exercise 2

Q) If you flip 10 coins each with a 30% probability of coming up heads, what is the probability exactly 2 of them are heads?

$\binom{10}{2}$ x $(0.3)^2$ x $(0.7)^8$ = 0.2334744 (← probability exactly 2 of them are heads)

```
> dbinom(2, 10, 0.3)
[1] 0.2334744
```

```
> mean(flips == 2)
[1] 0.2294
```

Q) Compare your simulation with the exact calculation

In comparison, the mean probability of seeing exactly 2 heads is 0.2294 which is only a 0.004 difference from the actual calculation.

Exercise 3

A) Use 10000 experiments and report the result

After using a 10000 experiments, each with 10 tosses, this was our result:

```
> mean(flips == 2)
[1] 0.2294
```

```
> dbinom(2, 10, 0.3)
[1] 0.2334744
```

B) Use 100000000 experiments and report the results

After using a 100000000 experiments, each with 10 tosses, this was our result:

```
> flips <- rbinom(100000000,10,0.3)
> mean(flips == 2)
[1] 0.2334864
```

CONCLUSION: The more experiments we do, the closer our probability gets to the probability of our exact value. This means that more trails would provide a less margin of error.

Exercise 4

Q) What is the expected value of a binomial distribution where 25 coins are flipped, each having a 30% chance of heads?

```
> mean(flips <- rbinom(100000,25, 0.3))
[1] 7.51261
> mean(flips <- rbinom(100000,25, 0.3))
[1] 7.50119
> mean(flips <- rbinom(100000,25, 0.3))
[1] 7.5032
> mean(flips <- rbinom(100000,25, 0.3))
[1] 7.50131
> mean(flips <- rbinom(100000,25, 0.3))
[1] 7.49214
```

The exact expected value is 7.5 (work shown after Exercise 5 in table)

Q) Compare your simulation with the exact calculation

All simulations are very close to the exact expected value since the mean of all the simulations above is 7.50209 which is only a 0.002 difference.

Exercise 5

Q) What is the variance of a binomial distribution where 25 coins are flipped, each having a 30% chance of heads?

```
> X <- rbinom(100000,25,0.3)
> var(X)
[1] 5.260221
> var(X)
[1] 5.260221
> var(X)
[1] 5.260221
> var(X)
[1] 5.260221
> var(X)
[1] 5.260221
```

The exact variance is 5.25 (work shown below)

Q) Compare your simulation with the exact calculation

The exact variance is very close to our simulated calculations however, not closer than our other calculations above. The difference between our above simulation for var(x) and the exact calculation is 0.010.

($n C k$)	(0.3)^a	(0.7)^b	a	b	a^2	($n C k$) x (0.3)^a x (0.7)^b	E[a]	E[a^2]
1	1	0.000134107	0	25	0	0.000134107	0	0
25	0.3	0.000191581	1	24	1	0.001436859	0.001436859	0.001436859
300	0.09	0.000273687	2	23	4	0.007389562	0.014779124	0.029558247
2300	0.027	0.000390982	3	22	9	0.024279989	0.072839966	0.218519898
12650	0.0081	0.000558546	4	21	16	0.057231402	0.228925608	0.915702431
53130	0.00243	0.000797923	5	20	25	0.103016524	0.515082618	2.575413088
177100	0.000729	0.00113989	6	19	36	0.147166462	0.882998773	5.297992639
480700	0.0002187	0.001628414	7	18	49	0.17119364	1.198355478	8.388488345
1081575	0.00006561	0.002326305	8	17	64	0.165079581	1.320636649	10.56509319
2042975	0.000019683	0.003323293	9	16	81	0.133635851	1.202722663	10.82450396
3268760	5.9049E-06	0.004747562	10	15	100	0.091636012	0.916360124	9.163601238
4457400	1.77147E-06	0.006782231	11	14	121	0.053553514	0.589088651	6.479975161
5200300	5.31441E-07	0.009688901	12	13	144	0.026776757	0.321321082	3.855852989
5200300	1.59432E-07	0.013841287	13	12	169	0.011475753	0.149184788	1.939402247
4457400	4.78297E-08	0.019773267	14	11	196	0.004215583	0.059018158	0.826254212
3268760	1.43489E-08	0.028247525	15	10	225	0.001324897	0.019873461	0.29810192
2042975	4.30467E-09	0.040353607	16	9	256	0.000354883	0.005678132	0.090850109
1081575	1.2914E-09	0.05764801	17	8	289	8.05197E-05	0.001368835	0.023270201
480700	3.8742E-10	0.0823543	18	7	324	1.53371E-05	0.000276068	0.004969217
177100	1.16226E-10	0.117649	19	6	361	2.42165E-06	4.60113E-05	0.000874214
53130	3.48678E-11	0.16807	20	5	400	3.11354E-07	6.22709E-06	0.000124542
12650	1.04604E-11	0.2401	21	4	441	3.17709E-08	6.67188E-07	1.4011E-05
2300	3.13811E-12	0.343	22	3	484	2.47565E-09	5.44643E-08	1.19822E-06
300	9.41432E-13	0.49	23	2	529	1.3839E-10	3.18298E-09	7.32086E-08
25	2.8243E-13	0.7	24	1	576	4.94252E-12	1.1862E-10	2.84689E-09
1	8.47289E-14	1	25	0	625	8.47289E-14	2.11822E-12	5.29555E-11
Var(x) = 61.5 - (7.5)^2 = 5.25							SUM = 7.5	SUM = 61.5

Kartikeya Sharma

ASSIGNMENT 2 – Distributions of Random Variables

Exercise 1

Q) Plot the histograms. How would you compare the various aspects of the two distributions?

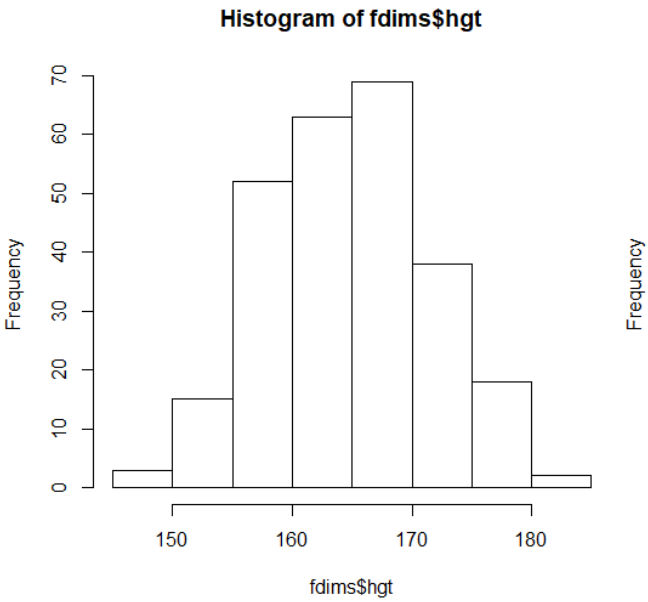


Image 1: Histogram of Women's Height

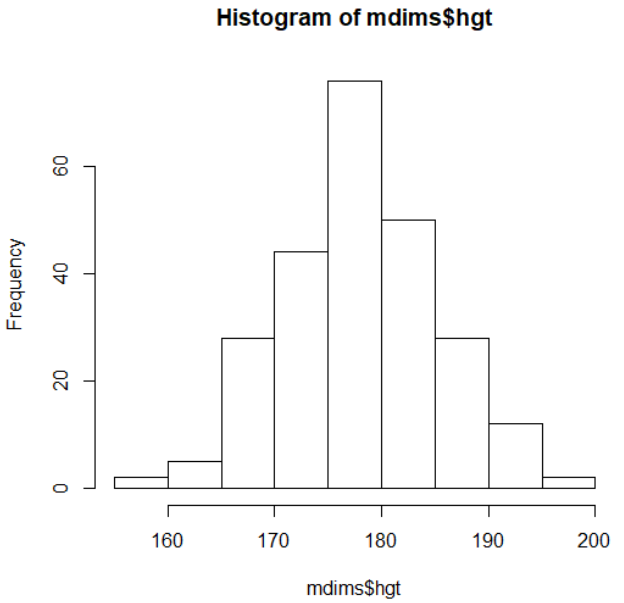


Image 2: Histogram of Men's Height

From *Image 1* and *Image 2*, we can see that both are visually similar to a normal distribution curve. However, women's height is skewed to the left (the frequency is higher on the left side of the histogram) while men's height is more even and closer resembles the bell-shaped curve. It is clear that the mean height for men is between 175-180 cm. Comparatively, women's height has an average value between 165-170 cm. This difference concludes that it was useful for us to create two additional data sets: one with only men and another with only women.

Exercise 2

Q) Based on this plot, does it appear that the data follows a nearly normal distribution?

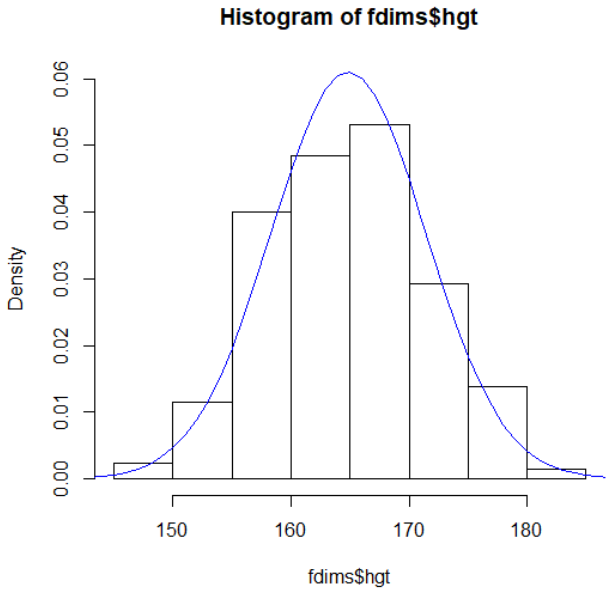


Image 3: Density Histogram of Women's Height with normal distribution curve

From *Image 3*, we can see that the data follows the normal distribution curve (blue line). This is because the data is, on the most part, symmetric and asymptotic. Even though we don't know how close the histogram is to the curve, each bar in the histogram seems to intersect with the normal distribution line, with a few outliers.

Exercise 3

Q) Make a normal probability plot of sim. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data?

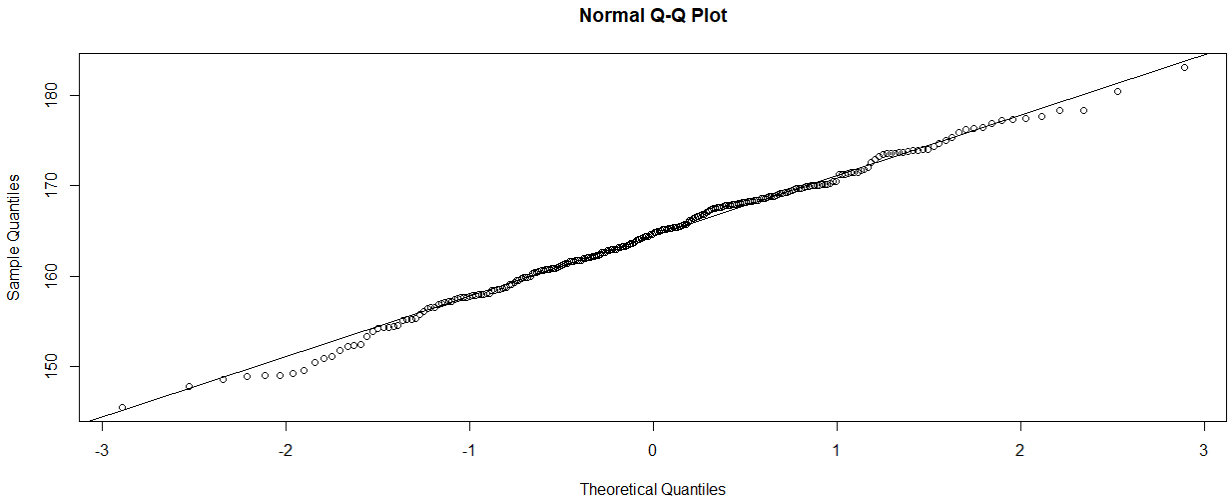


Image 4: Simulated Data Q-Q Plot

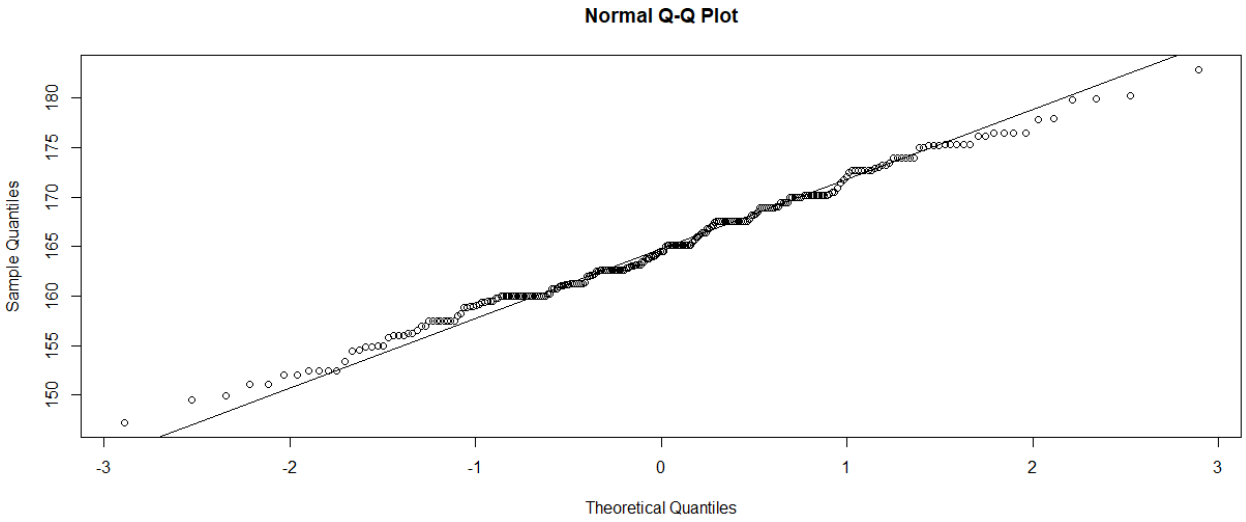


Image 5: Real Data Q-Q Plot

The majority of points on the simulation Q-Q plot are close to the Q-Q line. Upon inspection, we notice that that both ends, approximately from [-3,-2] and then from [2,3], of the plot are further away from the line. Likewise, this is the case for the real data. When comparing both, the real data and the simulation data, the real data progresses with a step-by-step trend whereas the simulation plot progresses in a linear fashion and is each point is closer to the Q-Q line.

Exercise 4

Q) Does the normal probability plot for fdims$hgt look similar to the plots created for the simulated data? That is, do plots provide evidence that the female heights are nearly normal?

When looking at the simulated plots, we see similarities with the female height plot. Most of the data in each plot is close to the Q-Q line in both plots. This suggests that the data from female height is normally distributed.

[look below for multiple Q-Q plots of female heights]

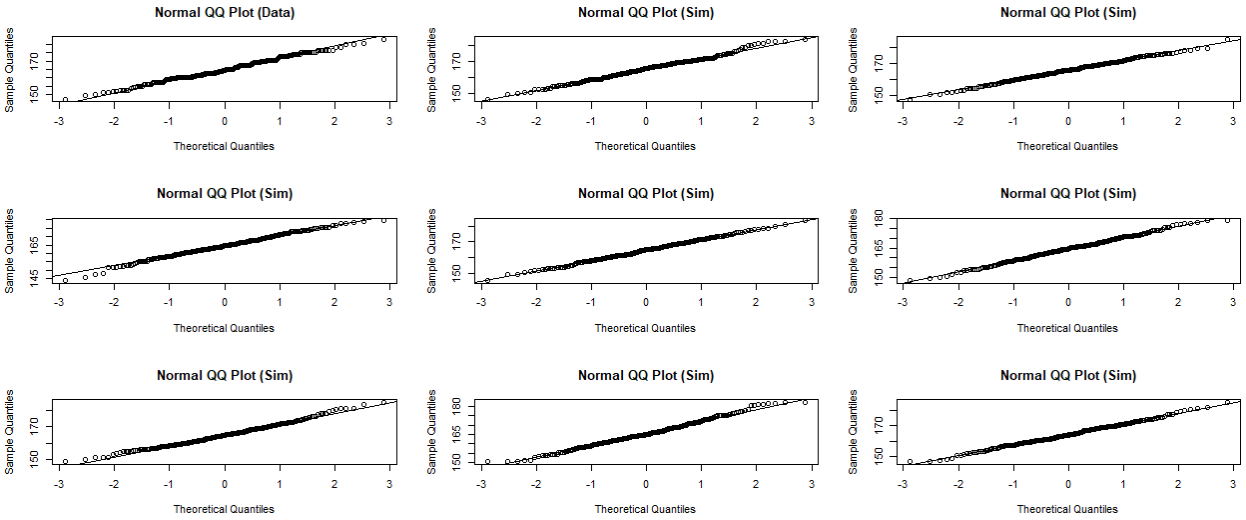


Image 6: Simulated data set vs. Sim vs. Normal probability plot

Exercise 5

Q) Using the same technique, determine whether or not female weights appear to come from a normal distribution. If not, how would you describe the shape of this distribution? Note: You may use a histogram to help you decide.

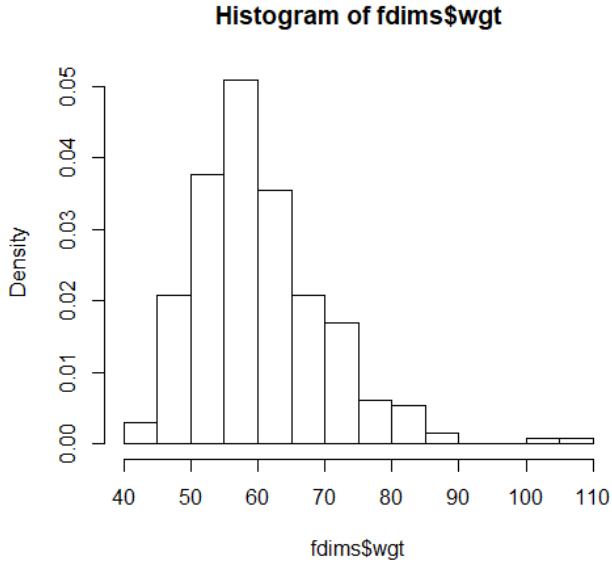


Image 7: Histogram for Female Weights

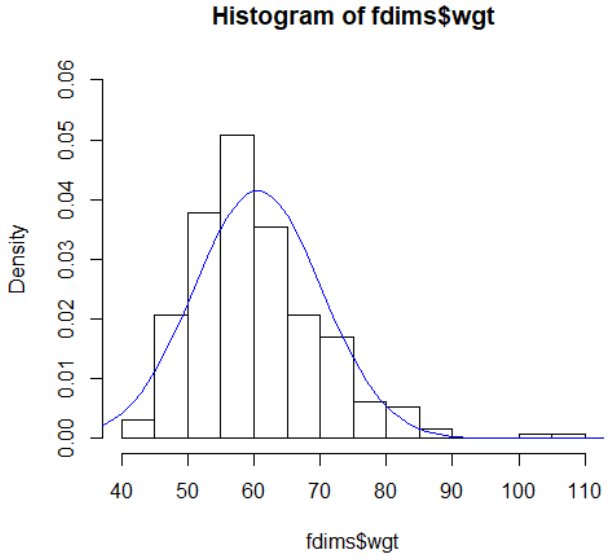


Image 8: Density Histogram of Female Weights with distribution curve

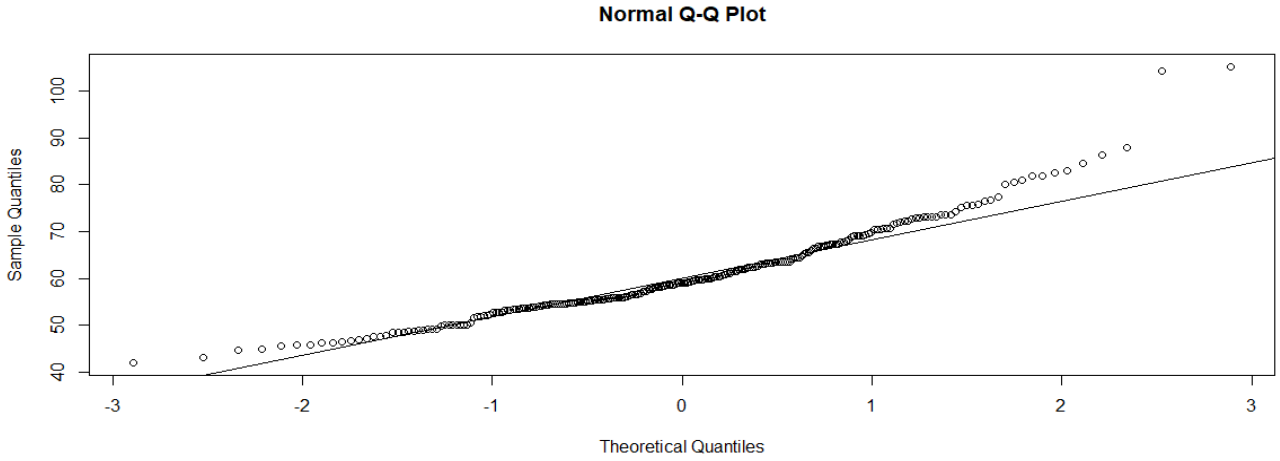


Image 9: Simulated Q-Q Data Plot

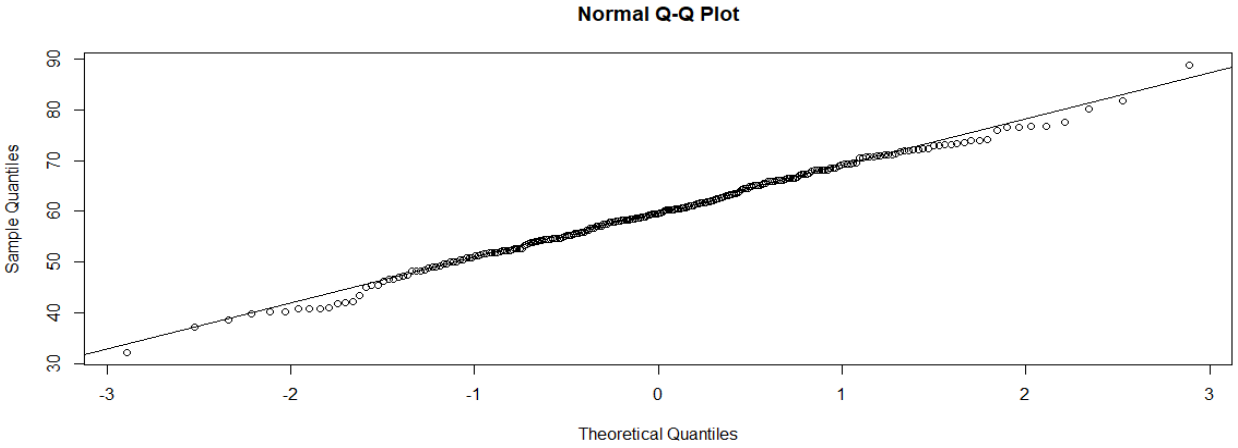


Image 10: Real Data Q-Q Plot

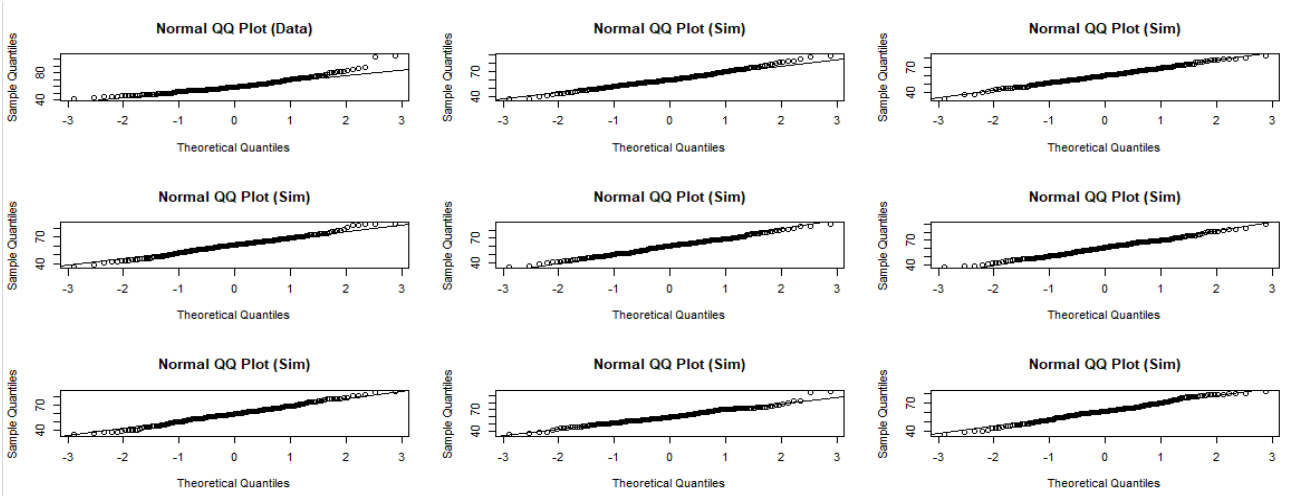


Image 11: Simulated data set vs. Sim vs. Normal probability plot

Image 7 shows us that since majority of the female in the data have a lower weight, the data is skewed to the right. Drawing the normal distribution curve in *Image 8* helps us confirm this hypothesis. By looking at *Image 11*, we notice that majority of the data points in each plot are close to the Q-Q line. This is similar to *Image 9*. Additionally, the simulated plot look similar to the real data plot (*Image 9* & *Image 10*). This leads to the conclusion that although there are some outliers near both ends, the data is organized linearly and ultimately follows a normal distribution.

Exercise 6

A. ***The histogram for female bi-iliac diameter (bii.di) belongs to normal probability plot letter*** B
B. ***The histogram for female elbow diameter (elb.di) belongs to normal probability plot letter*** C
C. ***The histogram for general age (age) belongs to normal probability plot letter*** D
D. ***The histogram for female chest depth (che.de) belongs to normal probability plot letter*** A

Exercise 7

Q) Note that normal probability plots C and D have a slight stepwise pattern. Why do you think this is the case?

The normal probability plots C and D have a slight stepwise pattern because the collected data represents age, which is in whole numbers. Displaying the age in whole numbers has allowed the values to repeat multiple times until it is rounded up again.

Exercise 8

Q) Make a normal probability plot for female knee diameter (kne.di). Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram to confirm your findings.

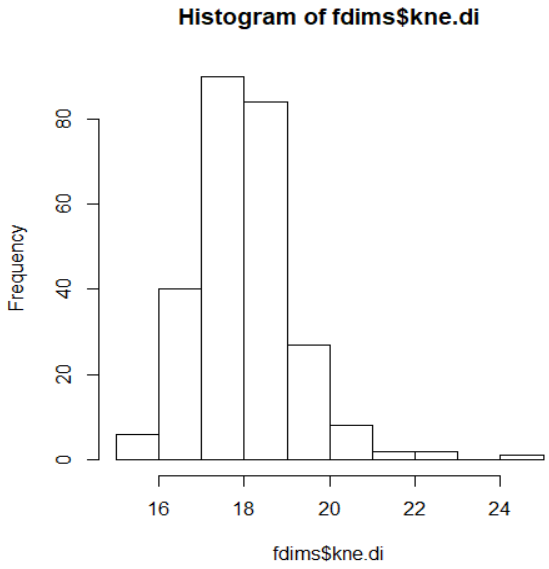


Image 12: Histogram of Female Knee Data

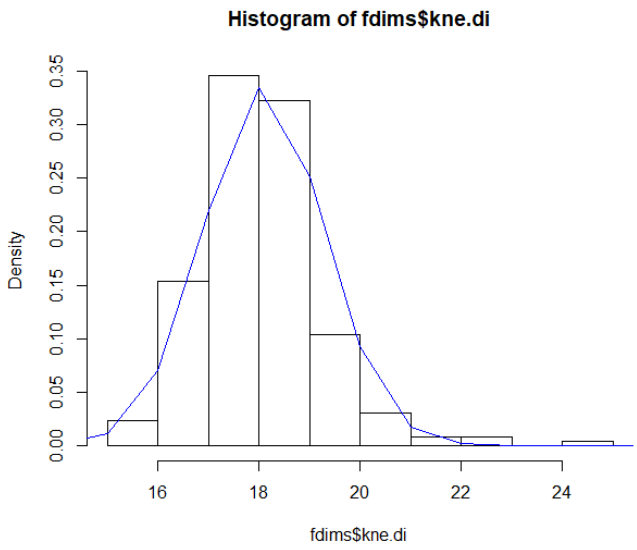


Image 13: Density Histogram of Female Knee Data with distribution curve

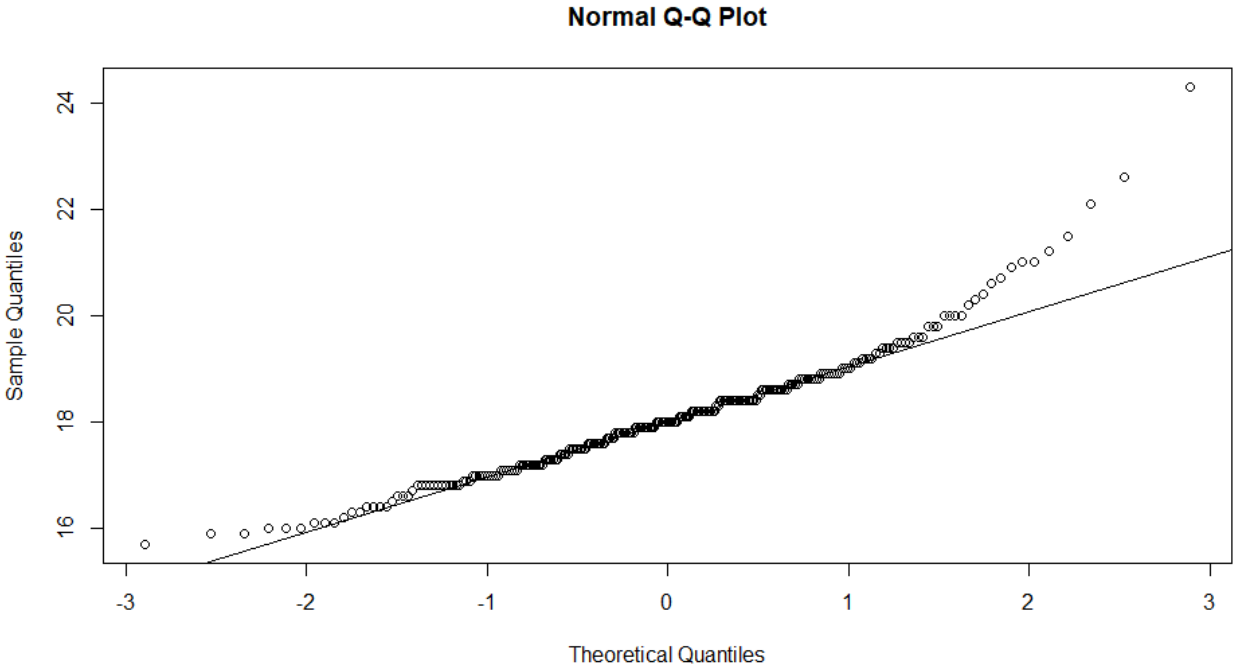


Image 13: Female Knee Normal Real Data Q-Q plot

```
> hist(fdims$kne.di)
> qqnorm(fdims$kne.di)
> qqline(fdims$kne.di)
> hist(fdims$kne.di, probability = TRUE)
> x = 10:30
> y = dnorm(x = x, mean = mean(fdims$kne.di), sd = sd(fdims$kne.di))
> lines(x = x, y = y, col = "blue")
```

Based on this normal probability plot (*Image 13*), we can see that the slope increase rapidly as the theoretical quantiles increase. Additionally, by looking at the density histogram, with the help of the blue distribution curve, we notice that there is no smooth bell-shaped distribution due to the rapid increase/decrease at each bar. This suggests that the data is skewed. Since the majority of the points are on the left, the normal probability plot suggests that the variable is right skewed.

Kartikeya Sharma

ASSIGNMENT 3 – Statistical Inference

Exercise 1

Q) Describe this population distribution. Be sure to include aa visualization in your answer.

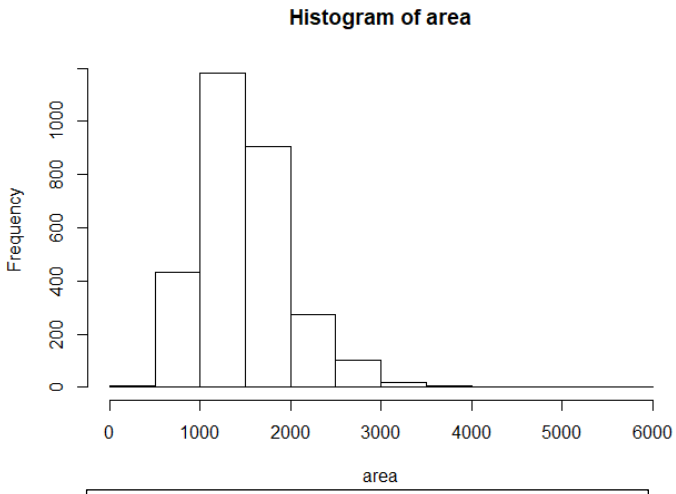


Image 1: Histogram for living Area of House

```
> summary(area)
   Min.  1st Qu.  Median
   334    1126    1442

   Mean   3rd  Qu.  Max.
    1500    1743    5642
>
> hist(area)
```

In Image 1, we can clearly see that the distribution is rightly skewed due to the majority of data falling between the interval [500, 2000].

Exercise 2

Q) In this exercise you will set a "random" seed. Use your unique id as the number you put in parentheses. For example, if your id was 123456789, you will enter 123456789.

```
> set.seed(823694285)
>
> samp1 = sample(area, 50)
>
> summary(samp1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    630    1128    1304    1426    1619    2872
>
> hist(samp1)
```

Exercise 3

Q) Describe the distribution of this sample? How does it compare to the distribution of the population? Be sure to include a visualization in your answer.

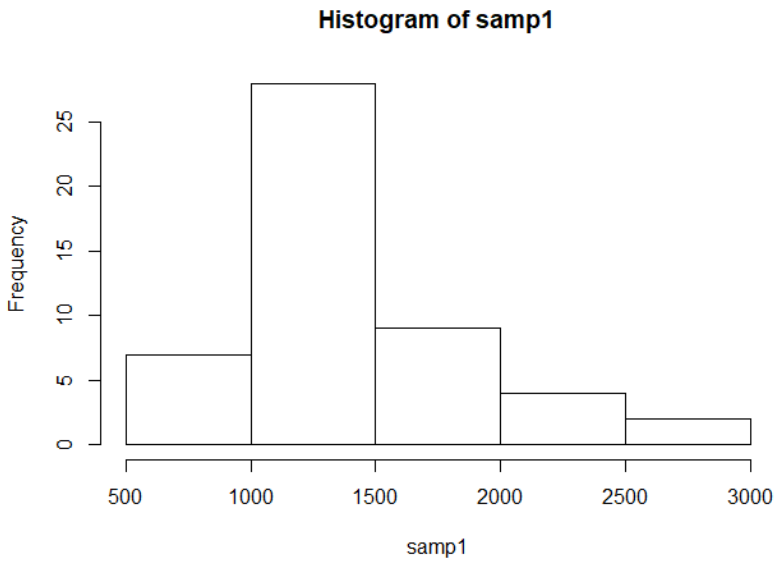


Image 2: Histogram for Simple Random Sample of size 50 from vector: Area (Exercise 2)

```
>
> summary(area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334    1126    1442    1500    1743    5642
>
> summary(samp1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    630    1128    1304    1426    1619    2872
>
```

As we can see from the summary of the area vs. the summary of samp1, the minimum and maximum values differ almost by a factor of 2. However, the median, mean, 1st quadrant, and 3rd quadrant values from both samples are almost similar. Looking visually at the histograms of both, the area sample vs. the random sample, our conclusion for the histogram of area being rightly skewed matches with this random sample (*Image 2*). However, *Image 2* seems more distributed compared to *Image 1*.

Exercise 4

Q) Take a second sample, also of size 50, and call it samp2. How does the mean of samp2 compare with the mean of samp1? Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean?

The mean of samp2 is closer to its median compared to samp1. By looking at the calculations below, we can also see that both samp1 and samp2 differ from the actual mean population considerably.

```
>
> summary(area)
   Min.  1st Qu.   Median  Mean   3rd Qu.  Max.
    334    1126    1442    1500    1743    5642
>
> summary(samp1)
   Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
    630    1128    1304    1426    1619    2872
>
> samp2 = sample(area, 50)
>
> summary(samp2)
   Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
    630    1180    1506    1529    1848    3447
>
> samp3 = sample(area, 100)
>
> summary(samp3)
   Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
    765    1178    1510    1539    1742    2634
>
> samp4 = sample(area, 1000)
>
> summary(samp4)
   Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
    407    1116    1452    1505    1771    4316
>
```

We don't see much difference in mean between samp2 and samp3, which only had an increased sample size of x2 between them. The minimum and maximum values differ considerably. The change is sample size from samp3 and samp4 is x10. Additionally, samp4 only differs by a value of 5 from the actual population mean. Thus, samp4, the largest sample, would provide a more accurate description of the estimate of the population mean.

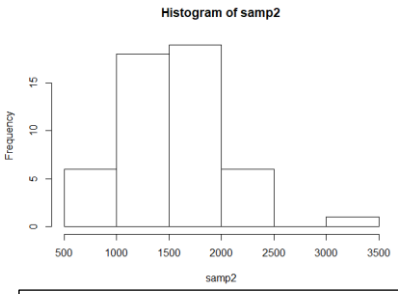


Image 3: Histogram for samp2

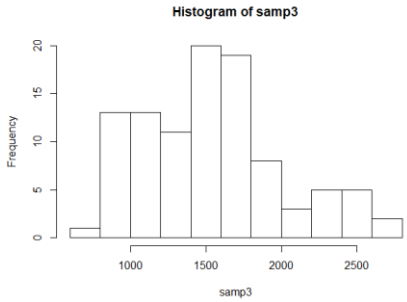


Image 4: Histogram for samp3

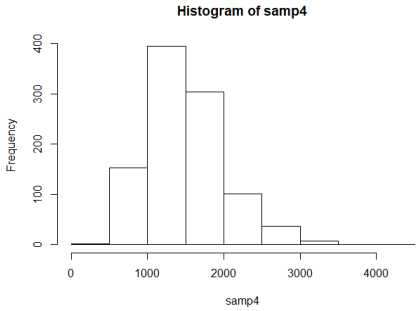


Image 5: Histogram for samp4

Exercise 5

Q) How many elements are there in sample means 50? Describe the sampling distribution and be sure to specifically note its center. Would you expect the distribution to change if we instead collected 50,000 sample means?

```
>
> summary(sample_means50)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1266    1451    1498    1499    1545    1753
>
```

There are 5000 elements in the sample means 50. This is because our for loop iterates from 1:5000 where 1 is included and 5,000 is excluded. By looking at *Image 6*, we can infer that the sampling distribution matches a normal distribution curve. Additionally, our inference from exercise 4, which stated that a larger sample size would provide a more accurate description of the population mean checks out with our current sample, which contains 5,000 sample means. The mean from the original 1,500 only differs by 1 from sample_means_50. Thus, if we were to take a sample mean of 50,000, we would get closer to the normal and the distribution would be a normal distribution.

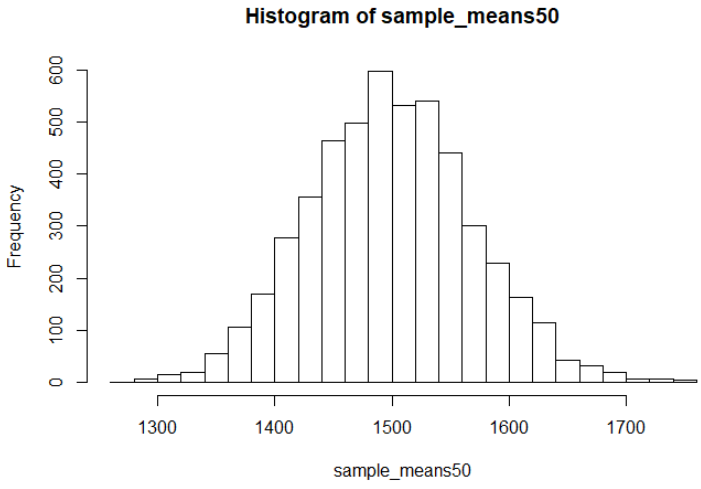


Image 6: Histogram for sample_means50

Exercise 6

Q) To make sure you understand what you've done in this loop, try running a smaller version. Initialize a vector of 100 zeros called sample_means_small. Run a loop that takes a sample of size 50 from area and stores the sample mean in sample_means_small, but only iterate from 1 to 100. Print the output to your screen (type sample means small into the console and press enter). How many elements are there in this object called sample_means_small? What does each element represent?

```
> sample_means_small = rep(0,100)
>
> for ( i in 1 : 100) {
+     samp = sample(area, 50)
+     sample_means_small[i] = mean(samp)
+ }
> sample_means_small
  [1] 1487.98 1505.56 1451.48 1595.92 1490.36 1506.60 1505.56
  [8] 1365.70 1520.34 1460.30 1495.18 1456.98 1414.02 1509.12
 [15] 1476.82 1493.66 1470.04 1381.22 1409.58 1524.18 1489.10
 [22] 1412.70 1575.04 1510.94 1395.38 1589.16 1559.32 1537.70
 [29] 1428.58 1440.58 1534.88 1645.44 1414.54 1508.68 1421.54
 [36] 1472.22 1463.58 1540.32 1506.34 1550.30 1483.80 1547.46
 [43] 1508.24 1651.36 1472.00 1398.62 1691.00 1508.60 1554.66
 [50] 1478.80 1571.24 1450.20 1523.60 1609.30 1554.48 1472.12
 [57] 1723.90 1445.30 1574.70 1460.62 1525.94 1544.94 1528.34
 [64] 1480.22 1395.74 1428.30 1483.84 1541.00 1518.78 1622.58
 [71] 1498.30 1428.78 1436.56 1506.70 1429.02 1489.68 1494.44
 [78] 1407.90 1482.60 1519.84 1430.88 1498.00 1512.08 1454.62
 [85] 1491.84 1475.36 1555.46 1557.74 1410.56 1485.72 1514.76
 [92] 1567.12 1426.98 1468.18 1453.52 1478.68 1520.24 1492.26
 [99] 1504.16 1524.54
>
```

By looking at the data inside sample_means_small, which contains 100 elements, each element represents the sample mean of the simple random sample of size 50 for the living area in a house. We can tell from "sample(area, 50)". We can also see that each value is close to 1500, which is the actual mean population. We can also see that the values are relatively off, due to the small sample size (only 100).

Exercise 7

Q) When the sample size is larger, what happens to the center? What about the spread?

```
> sample_means10 <- rep(NA, 5000)
>
> sample_means100 <- rep(NA, 5000)
>
> for(i in 1:5000){
+     samp <- sample(area, 10)
+     sample_means10[i] <- mean(samp)
+     samp <- sample(area, 100)
+     sample_means100[i] <- mean(samp)
+ }
>
> par(mfrow = c(3, 1))
>
> xlimits <- range(sample_means10)
>
> hist(sample_means10, breaks = 20, xlim = xlimits)
>
> hist(sample_means50, breaks = 20, xlim = xlimits)
>
> hist(sample_means100, breaks = 20, xlim = xlimits)
>
```

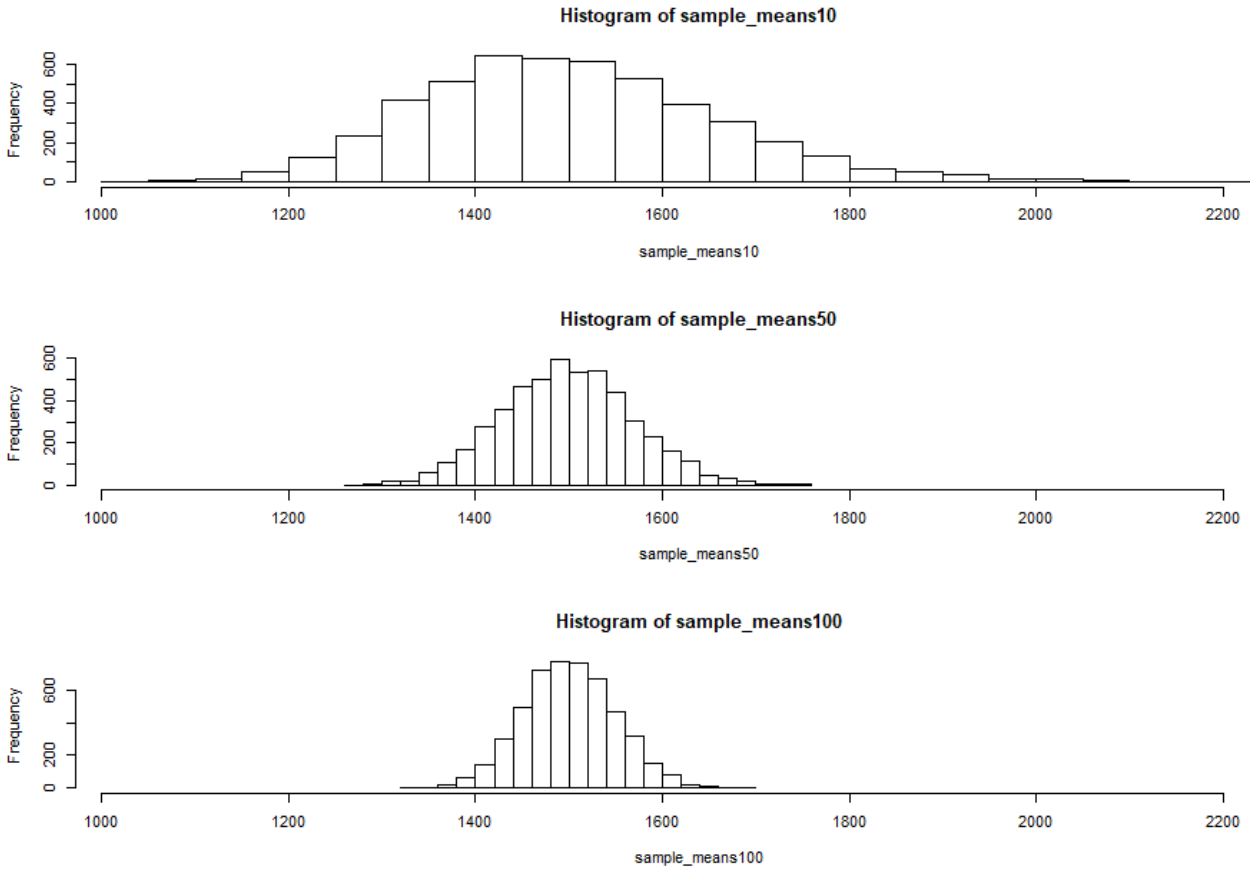


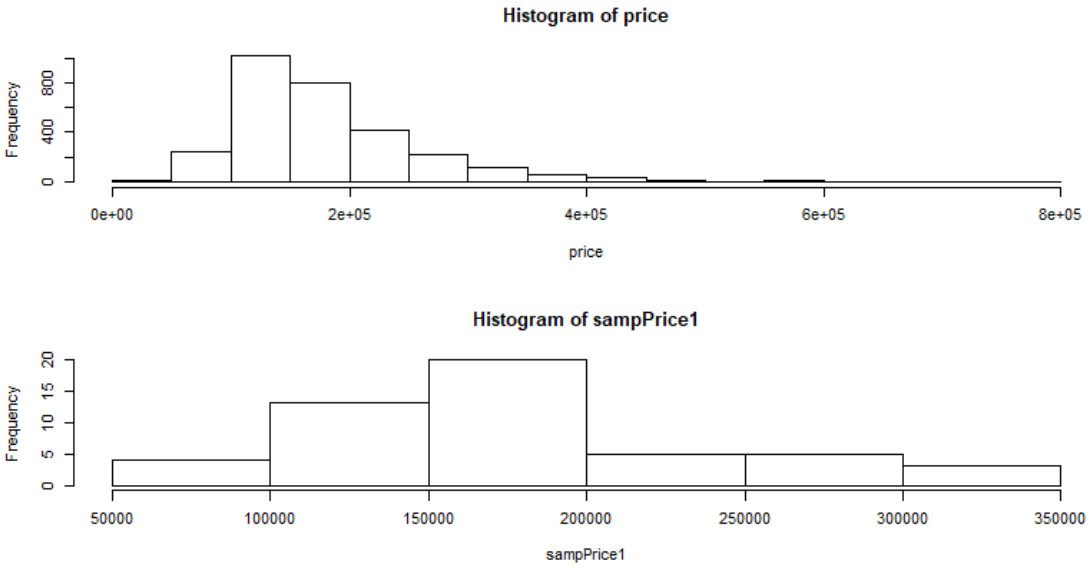
Image 7: Histogram for sample_means: 10 vs. 50 vs. 100

By looking at *Image 7*, the frequency of values around the center become larger. Each value starts converging to the center. In other words, the minimum and maximum values start coming closer to the median. We can also see that as the sample size becomes larger, the spread between each point becomes smaller. The reason behind this is because as we are increasing the sample size, the number of outliers start becoming insignificant.

Exercise 8

Q) Take a random sample of size 50 from price. Using this sample, what is your best point estimate of the population mean?

```
> summary(price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12789  129500  160000  180796  213500  755000
>
> hist(price)
>
> sampPrice1 = sample(price, 50)
>
> summary(sampPrice1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  95541  141675  173000  182945  205375  342643
>
> hist(sampPrice1)
>
```

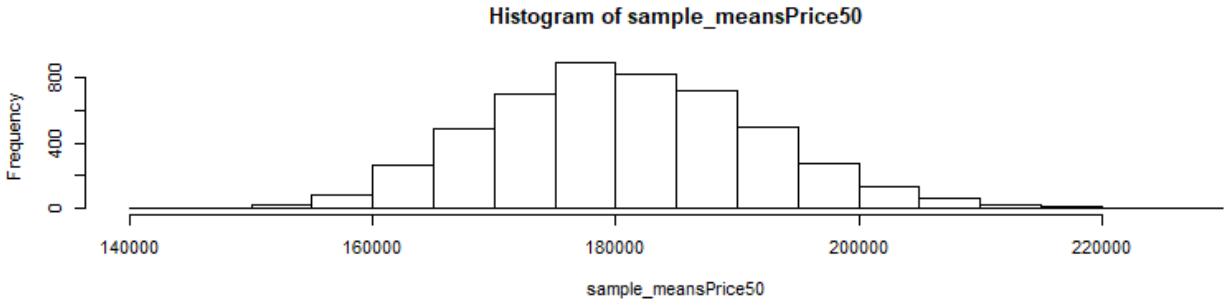


Based on the calculations above, I can estimate that the population mean is around 183,000. The mean of price, from where we took the random sample, has a population mean of around 180,000 which is close to our random sample of size 50.

Exercise 9

Q) Since you have access to the population, simulate the sampling distribution for price by taking 5000 samples from the population of size 50 and computing 5000 sample means. Store these means in a vector called sample_meansPrice50. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean home price of the population to be? Finally, calculate and report the population mean.

```
> sample_meansPrice50 = rep(0,5000)
>
> for(i in 1:5000){
+     samp = sample(price, 50)
+     sample_meansPrice50[i] = mean(samp)
+ }
> hist(sample_meansPrice50)
>
> hist(sample_meansPrice50, breaks = 25)
>
> summary(sample_meansPrice50)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 143606  172773  180314  180781  188101  227782
>
```

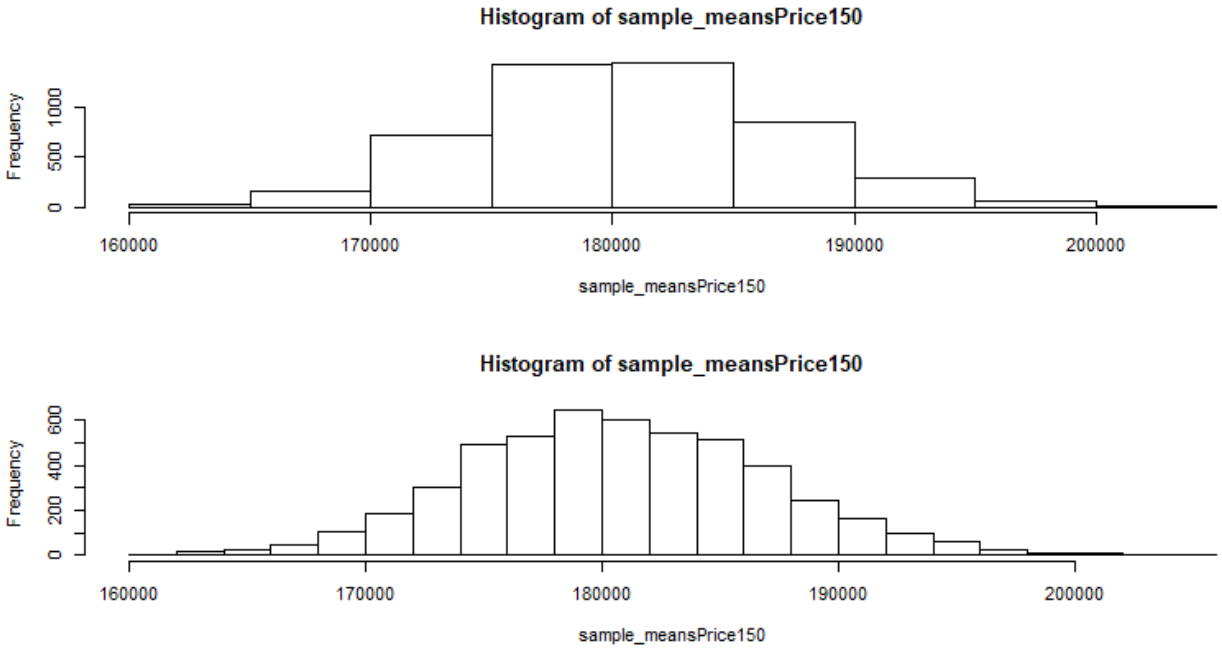


From looking at the histogram of sample_meanPrice50, we can see that the data is distributed normally. My guess is that the mean would be closer to the actual mean since we are taking the average of a sample which is x100 greater than the last random sample. From the calculations above, we can see that the mean is very close to 180,000. Comparing this to the mean of the random sample with only size 50, we can conclude that since we have taken the average of a larger sample size, the data (mean), is more accurate and closer to the original price mean. In fact, we are only about 14 off from the mean of the sample with 5000 elements and the real mean.

Exercise 10

Q) Change your sample size from 50 to 150, then compute the sampling distribution using the same method as above, and store these means in a new vector called sample_means150. Plot the data, then describe the shape of this sampling distribution, and compare it to the sampling distribution for a sample size of 50. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?

```
> sample_meansPrice150 = rep(0,5000)
>
> for(i in 1:5000){
+     samp = sample(price, 150)
+     sample_meansPrice150[i] = mean(samp)
+ }
>
> hist(sample_meansPrice150)
>
> hist(sample_meansPrice150, breaks = 25)
>
> summary(sample_meansPrice150)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 161486  176346  180566  180772  184920  204333
>
```



The distribution in the above histogram is relatively normal distribution. The shape is that of a bell-curve. If we compare it to the sampling distribution for a sample size of 50 from exercise 9, we can see that the values in the histogram of sample_meansPrice150 is more normally distributed compared to the histogram of sample_meansPrice50. The reason for this would most probably be the greater sample size. Additionally, based on the trend that a greater sample size would represent a stronger normal distribution, I would predict that the mean for sample_meansPrice150 will be closer to the actual mean than sample_meansPrice50. By looking at the calculations done in the code above, we can confirm these results (around 100 off).

Exercise 11

Q) Of the sampling distributions from 9 and 10, which has a smaller spread? If we're concerned with making estimates that are more often close to the true value, would we prefer a distribution with a large or small spread?

Comparing the sample distributions from exercise 9 and 10, we can conclude that exercise 10 has a smaller spread. Similarly, when making estimates that are more often close to the true value, we would prefer a smaller spread because as seen in

exercise 10, the mean of the data was closer to the mean of the actual population than the mean of exercise 9 was.