Abtahi Chowdhury

CSC 21700-D

R Assignments

R Assignment 1

1. **Generate 100 experiments of flipping 10 coins, each with 30% probability. What is the most common number? Why?**
   The most common number is 3 because the probability of heads is 30%, therefore statistically, there should be 3 heads for every 10 tosses.

```
> rbinom(100,10,0.3)
  [1] 6 2 5 1 6 6 4 5 2 1 4 1 3 6 4 1 3 2 2 1 2 0 3 3 1 5
 [27] 4 4 3 4 5 6 6 2 5 4 4 3 4 4 2 4 0 0 3 2 3 3 5 4 4 4
 [53] 2 4 1 3 3 2 1 4 0 4 4 4 3 3 4 3 4 4 1 4 1 2 5 7 2 6
 [79] 3 4 1 3 2 3 2 3 2 3 4 4 4 1 4 2 2 3 5 7 3 5
> rbinom(100,10,0.3)
  [1] 3 3 1 1 5 2 5 3 4 2 3 4 3 3 6 3 5 4 4 5 4 3 2 6 2 4
 [27] 1 6 2 2 4 5 6 1 4 1 4 4 1 3 3 3 3 0 5 2 0 2 3 1 6 1
 [53] 4 3 0 4 2 2 4 2 2 3 2 4 2 5 2 0 2 5 4 3 3 3 2 5 4 4
 [79] 4 1 3 2 2 4 3 4 7 4 0 4 3 5 0 3 1 3 4 6 3 4
> rbinom(100,10,0.3)
  [1] 2 4 4 2 6 3 2 3 5 3 2 1 3 1 4 1 0 2 3 1 1 5 2 2 5 1
 [27] 5 3 1 4 5 3 2 2 1 3 2 3 3 2 3 1 0 4 2 2 2 5 2 5 6 1
 [53] 2 3 3 3 5 1 2 4 2 5 2 4 4 1 4 4 4 5 5 3 4 3 5 3 4 3
 [79] 4 2 5 1 3 3 7 4 3 2 3 4 3 3 5 1 4 4 2 2 3 2
```

2. **If you flip 10 coins each with a 30% probability of coming up heads, what is the probability exactly 2 of them are heads? Compare your simulation with the exact calculation.**
   a. use 10000 experiments and report the result.
      Simulation: 0.2359
      Exact Calculation: 0.2334744
      Comparing the two results, it can be seen that the simulation is within 2 significant figures of accuracy with the exact calculation.

```
> flips <- rbinom(10000,10,0.3)
> mean(flips == 2)
[1] 0.2359
> dbinom(2,10,0.3)
[1] 0.2334744
```

   b. use 100000000 experiments and report the result.
      Simulation: 0.2334806
      Exact Calculation: 0.2334744
      Comparing the two results, it can be seen that the simulation is within 4 significant figures of accuracy with the exact calculation. This is much better than

the first part in which only 10000 experiments were done. This means that as the number of experiments goes up, so the accuracy of the result.

```
> flips <- rbinom(100000000,10,0.3)
> mean(flips == 2)
[1] 0.2334806
> dbinom(2,10,0.3)
[1] 0.2334744
```

3. **What is the expected value of a binomial distribution where 25 coins are flipped, each having a 30% chance of heads? Compare your simulation with the exact calculation.**
   Simulation: 7.507, 7.4745, 7.514
   Exact Calculation: 7.5
   It is seen that the experimental results have some error from the exact calculation of 7.5, however this is to be expected due to the random nature of the experiments. This can be fixed by simply increasing the number of experiments.

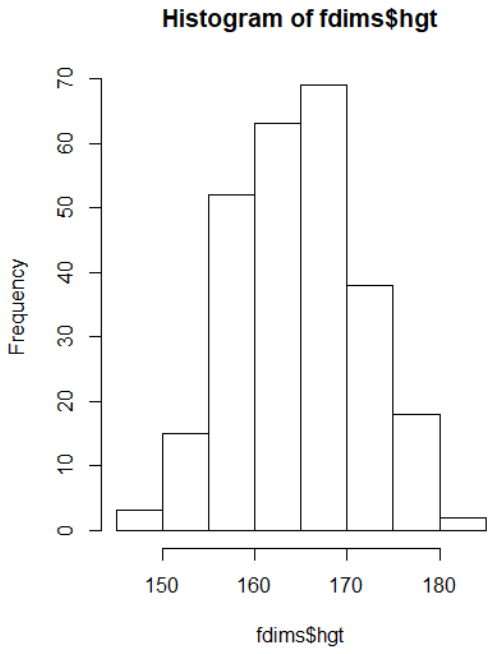
```
> mean(flips <- rbinom(10000,25,0.3))
[1] 7.507
> mean(flips <- rbinom(10000,25,0.3))
[1] 7.4745
> mean(flips <- rbinom(10000,25,0.3))
[1] 7.5154
```

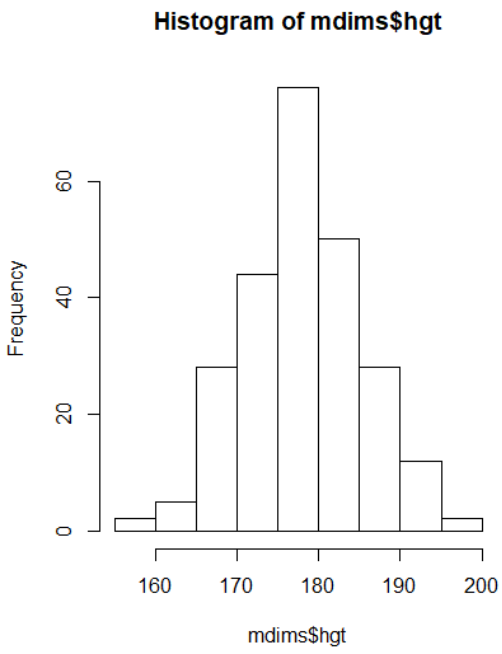
4. **What is the variance of a binomial distribution where 25 coins are flipped, each having a 30% chance of heads? Compare your simulation with the exact calculation.**
   Simulation: 5.220765, 5.221561, 5.289527
   Exact Calculation: 5.25
   It is seen that the experimental results have some error from the exact calculation of 5.25, however this is to be expected due to the random nature of the experiments. This can be fixed by simply increasing the number of experiments.

```
> var(rbinom(100000,25,0.3))
[1] 5.220765
> var(rbinom(100000,25,0.3))
[1] 5.221561
> var(rbinom(100000,25,0.3))
[1] 5.289527
```

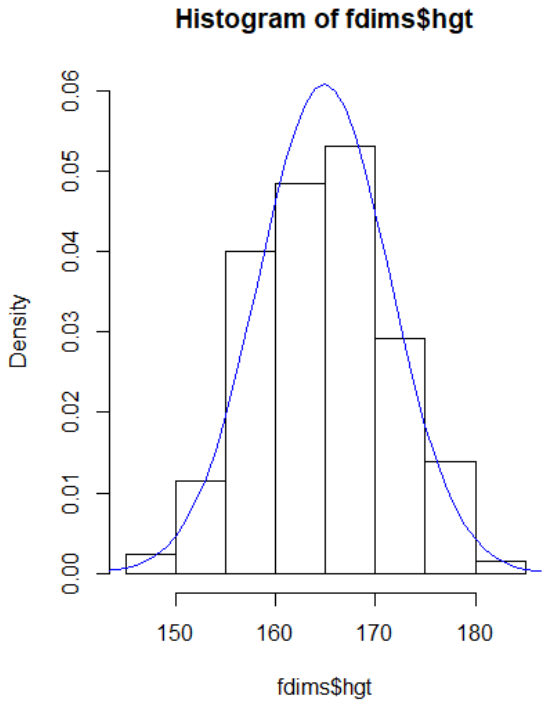
R Assignment 2

1. **Plot the histograms. How would you compare the various aspects of the two distributions?**
   Based off the two histograms, the male heights show a more bell curve distribution where as the female heights also show a bell curve distribution but is slightly left heavier on the left side.





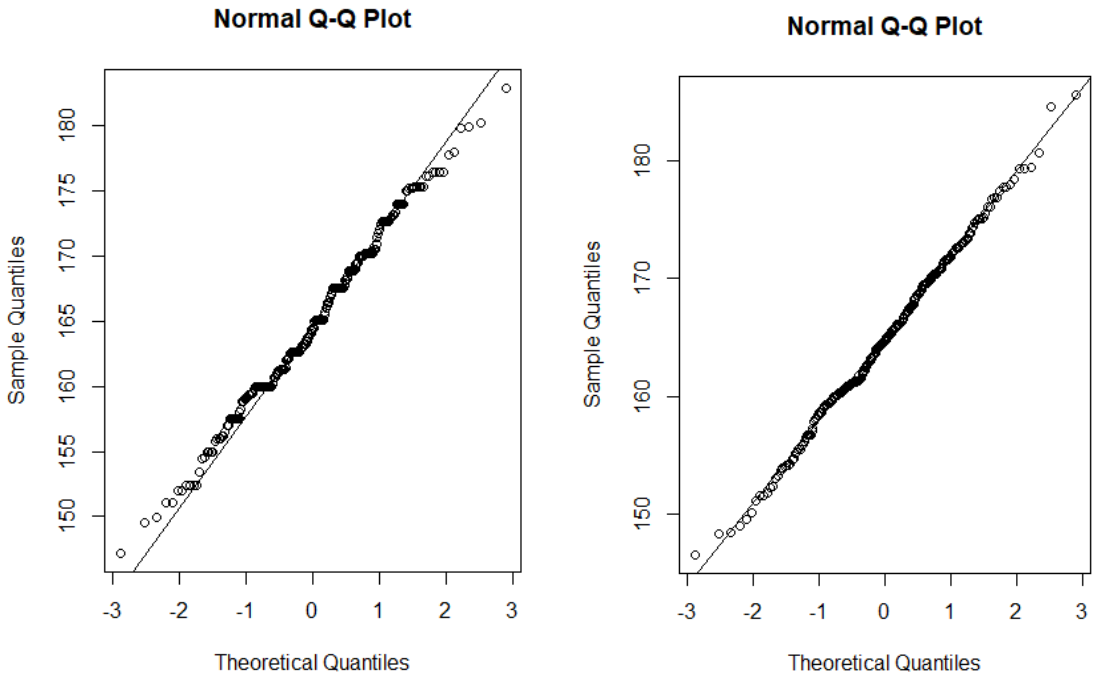
2. **Based on this plot, does it appear that the data follow a nearly normal distribution?**
   For this graph, it can be seen that the data does indeed form a nearly normal distribution.



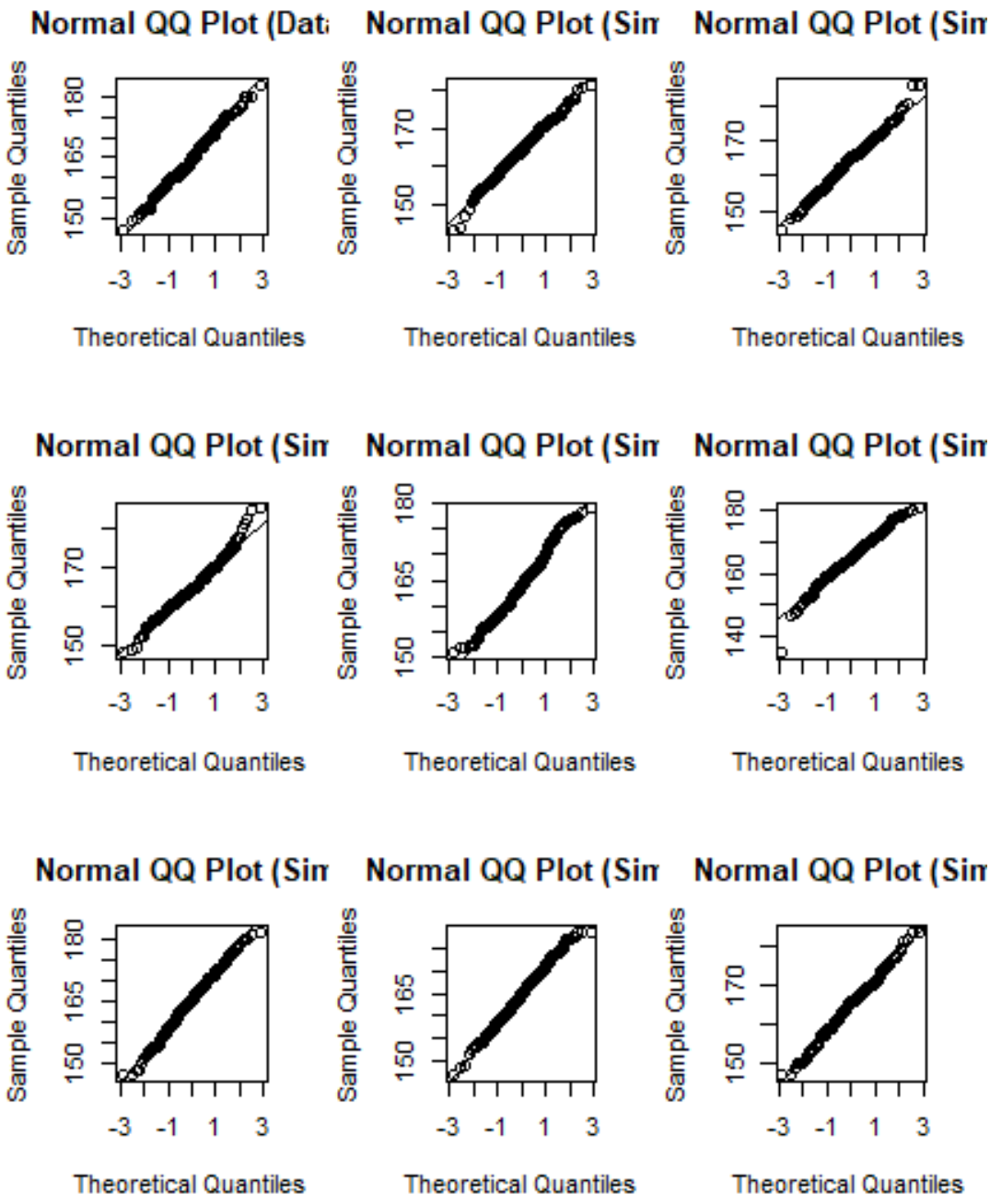
fhgtmean	164.872307692308
fhgtsd	6.54460213059717

3. **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?**

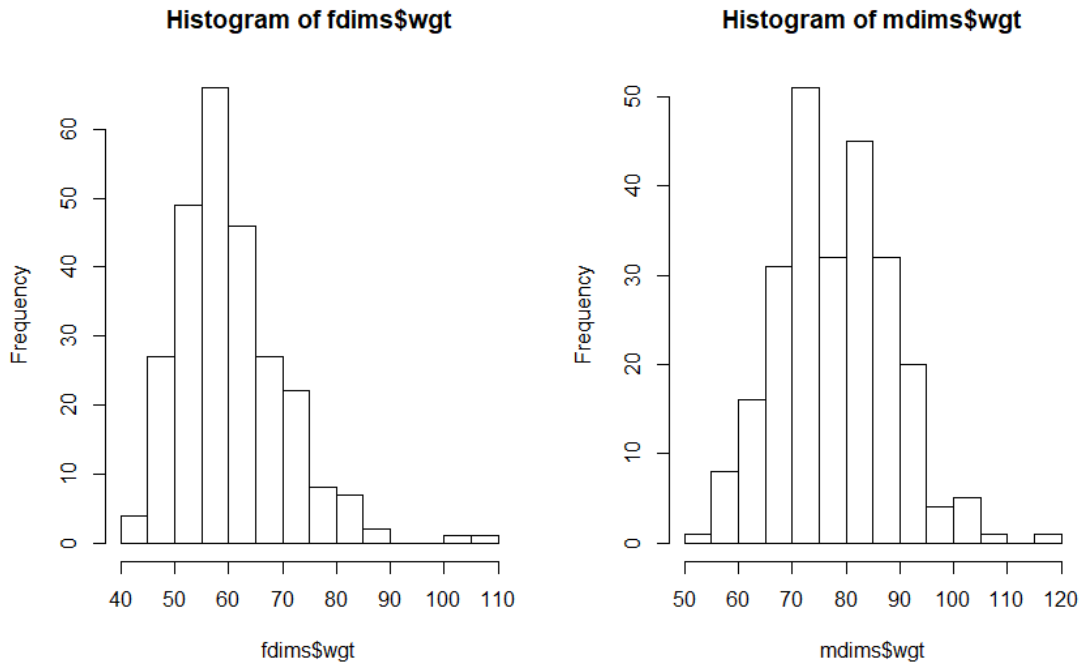
In the plot of sim, almost all the points fall within the line, however there are still a few points around the ends which aren't on the line. This is better than the plot of the real data as that one has more points that are outside the line.



4. **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?**
   The various plots created form the simulated data do indeed show that the female heights are nearly normal. This is because most of the simulated data form nearly normal plots which look similar to the original plot.

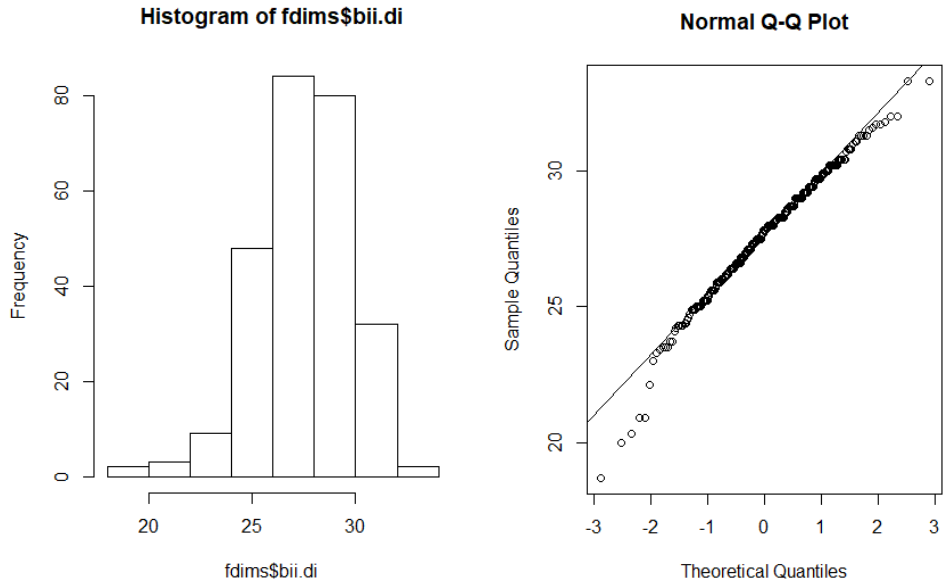


5. **Using the same technique, determine whether or not female weights appear to come from a normal distribution. If not, how would you describe the shape of this distribution? Note: You may use a histogram to help you decide**
   The female weights do form a slight normal distribution. To be more specific, it forms a normal distribution that is heavier to the left side.

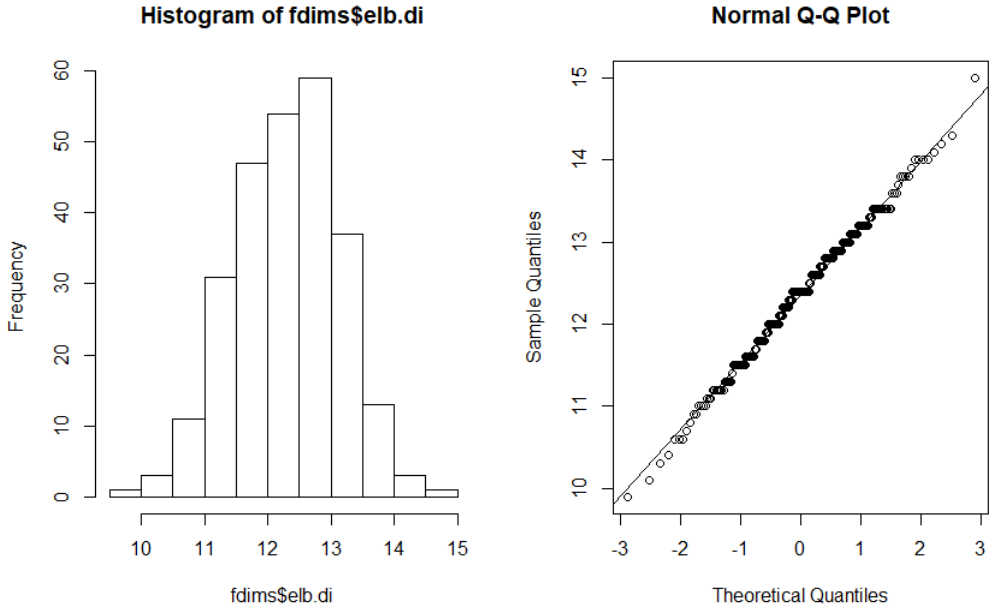



```
> 1 - pnorm(q = 182, mean = fhgtmean, sd = fhgtsd)
[1] 0.004434387
> sum(fdims$hgt > 182)/length(fdims$hgt)
[1] 0.003846154
```

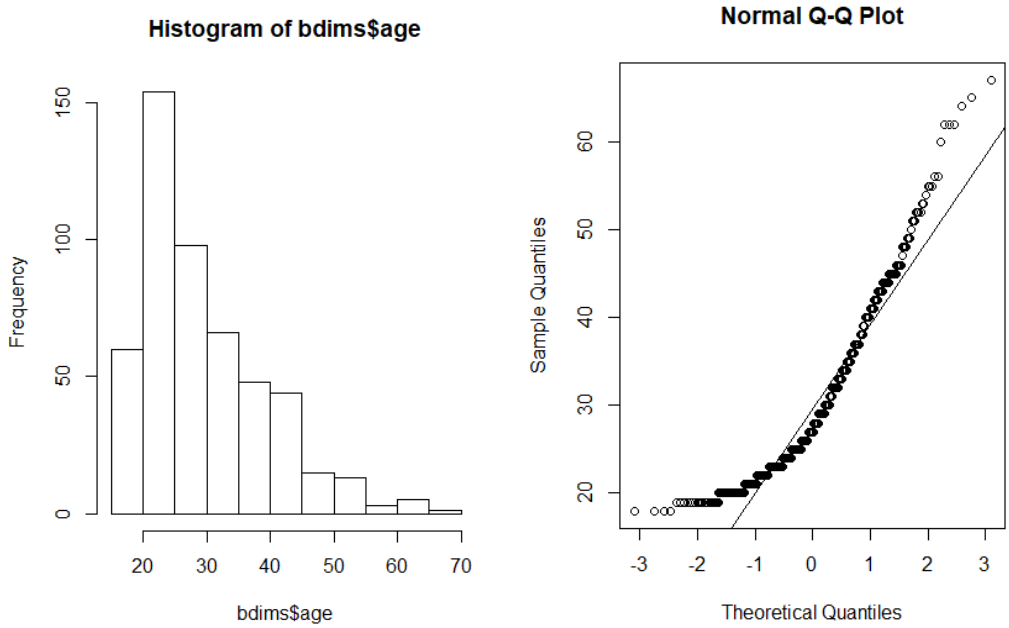
6. **Now let's consider some of the other variables in the body dimensions data set. Using the figures on the next page, match the histogram to its normal probability plot. All of the variables have been standardized (first subtract the mean, then divide by the standard deviation), so the units won't be of any help. If you are uncertain based on these figures, generate the plots in R to check.**
    a. **The histogram for female bi-iliac diameter (bii.di) belongs to normal probability plot letter:** B



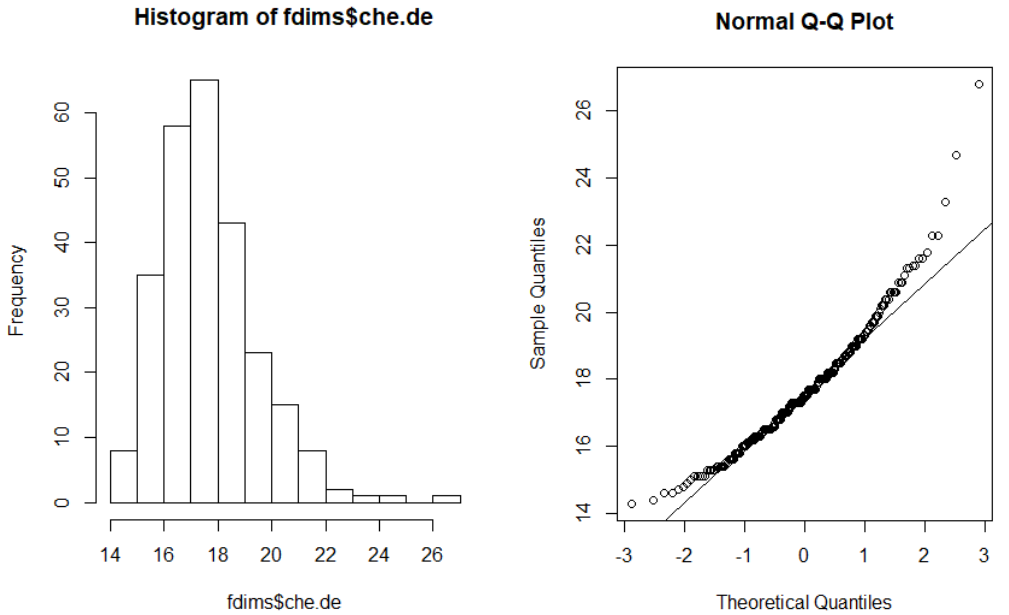
b. The histogram for female elbow diameter (elb.di) belongs to normal probability plot letter: C



c. The histogram for general age (age) belongs to normal probability plot letter: D

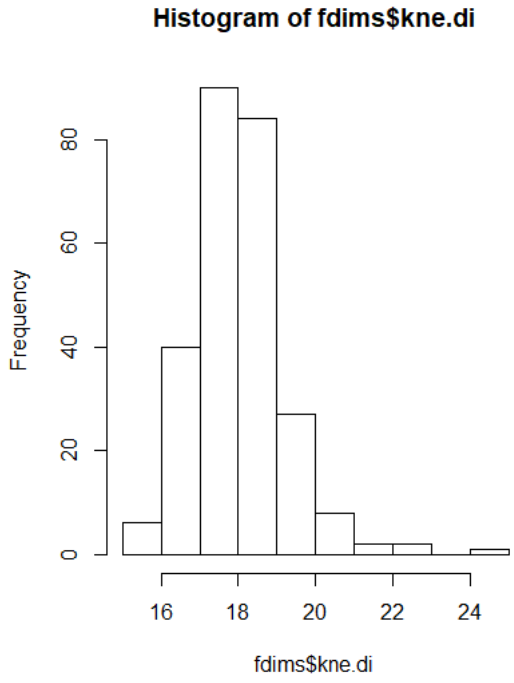


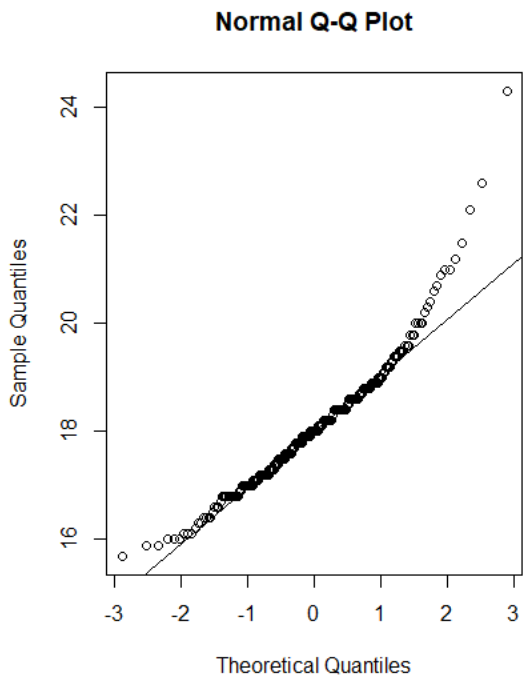
d. The histogram for female chest depth (che.de) belongs to normal probability plot letter: A



7. **Note that normal probability plots C and D have a slight stepwise pattern. Why do you think this is the case?**
   The stepwise pattern in the age plot might be due to the fact that if is recorded in whole numbers. The stepwise pattern in the female chest depth plot might be due to the age of the participant affecting it.
8. **As you can see, normal probably plots can be used both to assess normality and visualize skewness. Make a normal probability plot for female knee diameter**

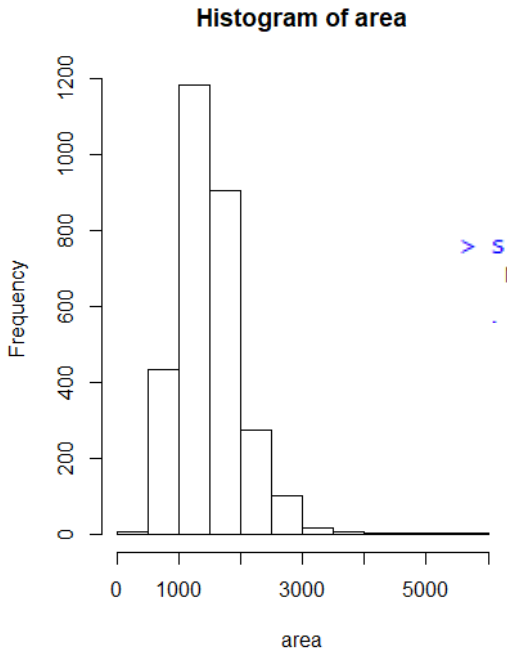
(kne.di). Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram to confirm your findings.
Based on the normal probability plot, the data for the female knee diameter is right skewed.





R Assignment 3

1. **Describe this population distribution. Be sure to include a visualization in your answer.**
   The population distributions based of the area of the lands show that a majority of the population use less than 2000 square meters of land. This is due to the fact that the data is left skewed.



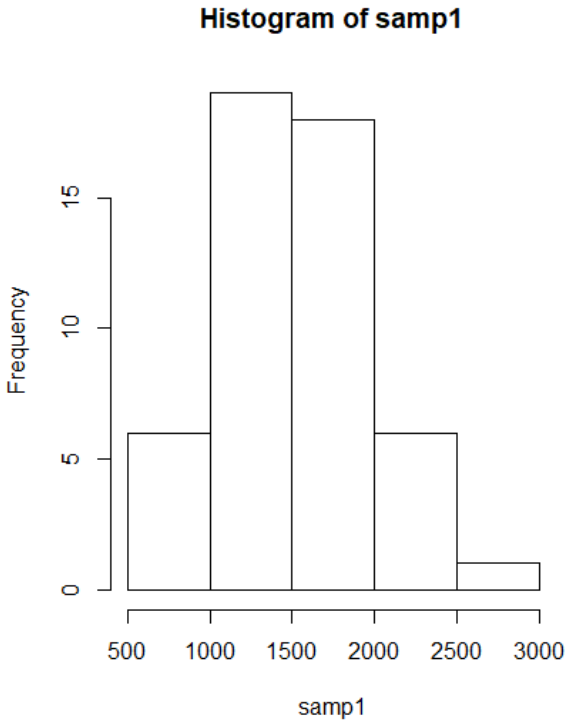

```
> summary(area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334    1126    1442    1500    1743    5642
```

2.

```
set.seed(489559603)
samp1 = sample(area, 50)
```

3. **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.**

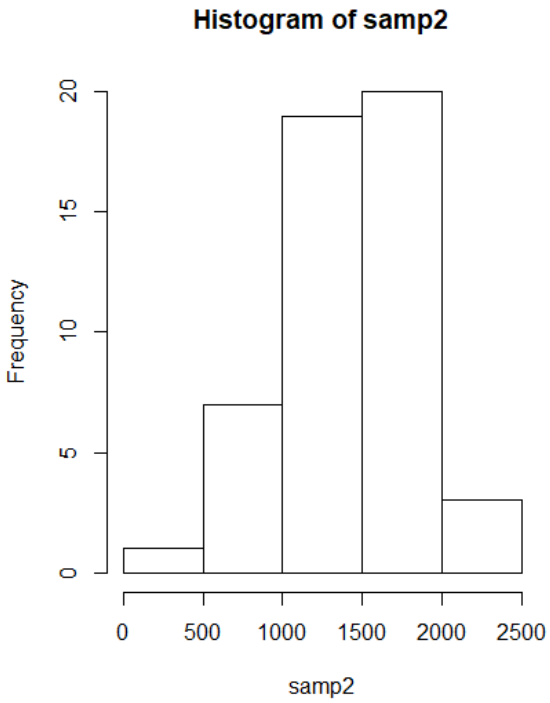
The distribution of the sample is roughly similar to the distribution of the entire population. Just like the other histogram, this one is also left skewed.




```
> mean(samp1)
[1] 1509.54
```

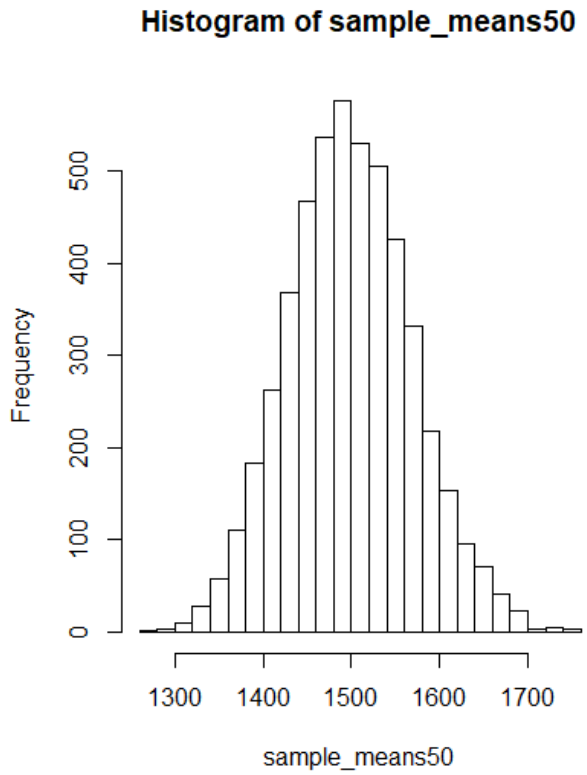
4. **Take a second sample, also of size 50, and call is samp2. How does the mean of samp2 compare with the mean of samp1? Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean?**

   This new sample ended up being different from the previous sample. Where the first sample was left skewed, this one is right skewed. The mean of the new sample is also lower than the mean of the first sample. If we took two more samples, the one of size 1000 would give us more accurate results.




```
> mean(samp2)
[1] 1424.94
```

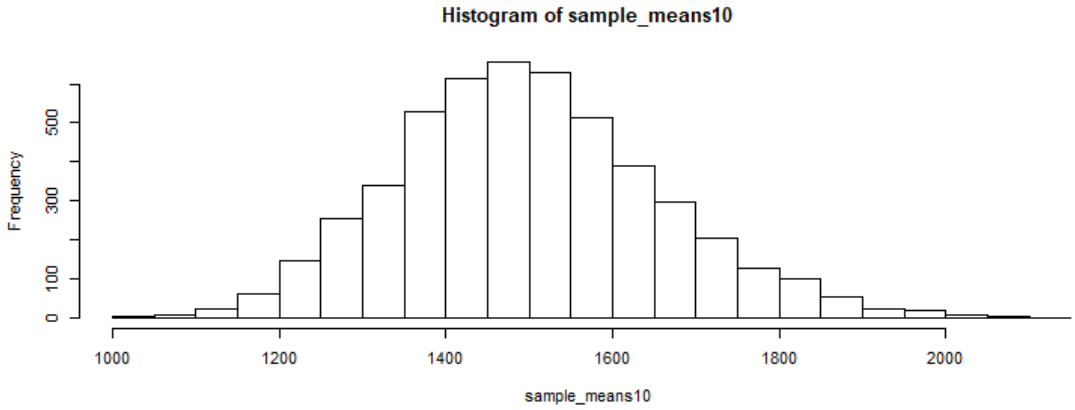
5. **How many elements are there in sample_means50? Describe the sampling distribution and be sure to specifically note its center. Would you expect the distribution to change if we instead collect 50000 sample means?**
   There are 5000 elements in sample_mean50. The distribution of the graph resembles a bell curve centerd around 1500. If we were to collect 50000 means as opposed to 5000 means, then the distrubution of the data will most likely not change much.

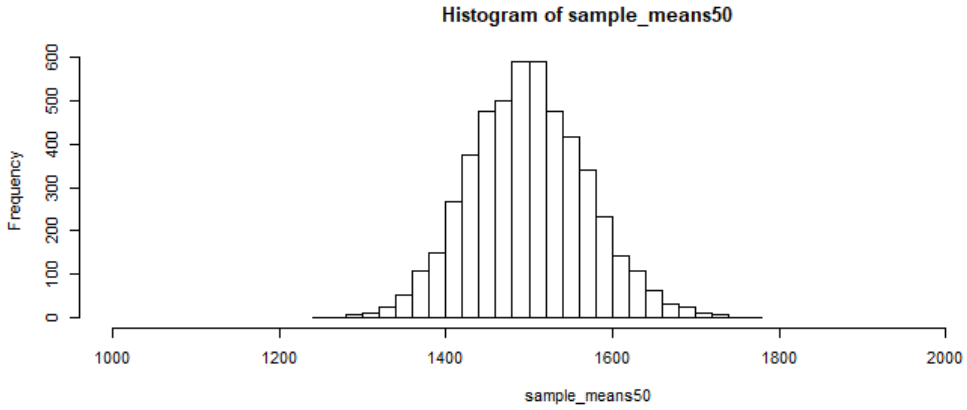


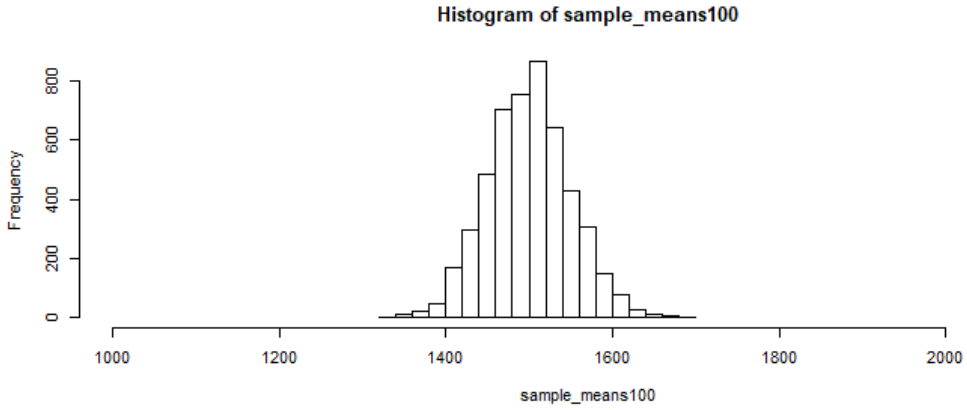
6. **How many elements are there in this object called sample_means_small? What does each element represent?**
   There are 100 elements in sample_means_small and each element of the vector represents the mean of a sample set of size 50.

```
> sample_means_small
  [1] 1543.98 1438.02 1477.10 1502.28 1564.14 1595.96
  [7] 1417.30 1558.72 1518.52 1398.08 1612.60 1575.94
 [13] 1547.68 1466.00 1541.62 1477.86 1516.40 1510.78
 [19] 1552.36 1467.94 1480.86 1601.64 1512.38 1466.90
 [25] 1483.12 1513.46 1516.10 1641.46 1446.62 1547.56
 [31] 1477.30 1495.56 1543.26 1491.24 1499.94 1671.54
 [37] 1595.24 1473.68 1466.56 1470.14 1413.78 1573.74
 [43] 1499.80 1546.92 1553.42 1440.52 1525.70 1409.54
 [49] 1558.24 1424.84 1511.78 1498.74 1464.38 1387.08
 [55] 1478.76 1572.96 1545.00 1372.60 1473.74 1428.34
 [61] 1624.66 1588.50 1649.90 1614.82 1471.66 1419.18
 [67] 1528.66 1421.70 1455.58 1508.74 1454.88 1535.42
 [73] 1408.94 1554.90 1390.34 1547.50 1487.40 1525.32
 [79] 1507.46 1512.22 1591.54 1464.26 1552.28 1468.28
 [85] 1454.62 1632.20 1527.12 1508.86 1471.86 1501.50
 [91] 1495.30 1405.74 1414.60 1469.16 1573.48 1599.46
 [97] 1446.06 1456.86 1537.00 1660.78
```

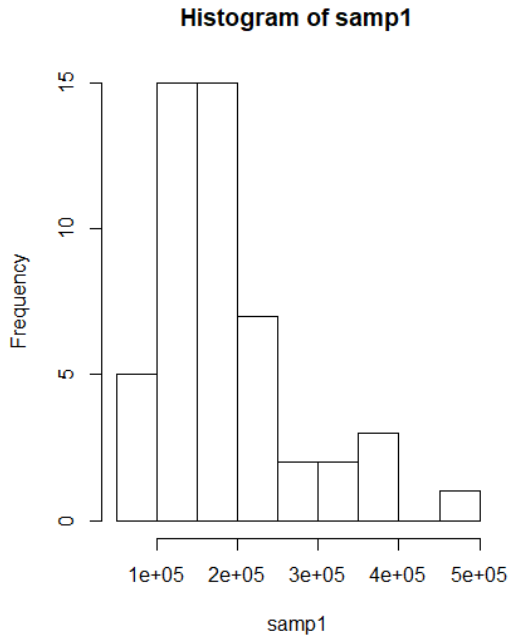
7. **When the sample size is larger, what happens to the center? What about the spread?**
   As the sample size gets larger the center stays the same but the spread decreases.





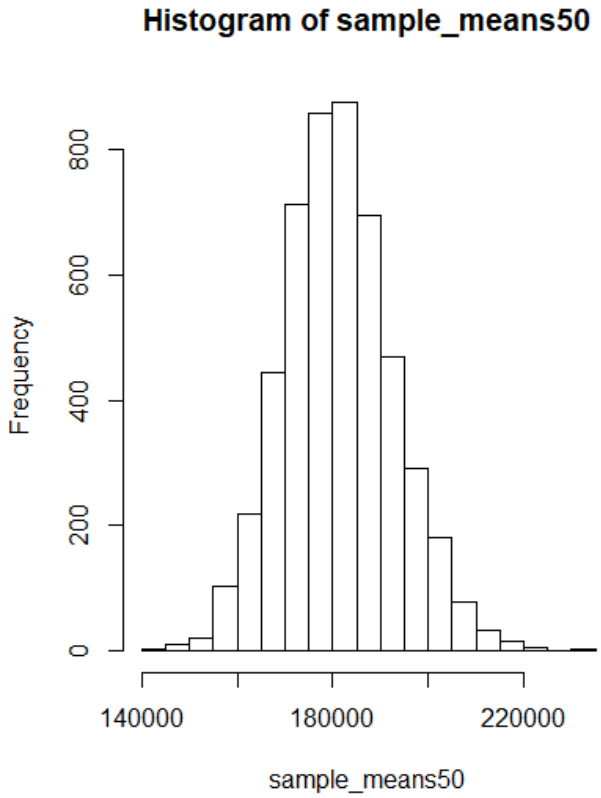


8. **Take a random sample of size 50 from price. Using this sample, what is your best point estimate of the population mean?**
   Based of this sample, the population mean would be about 1.75e5.



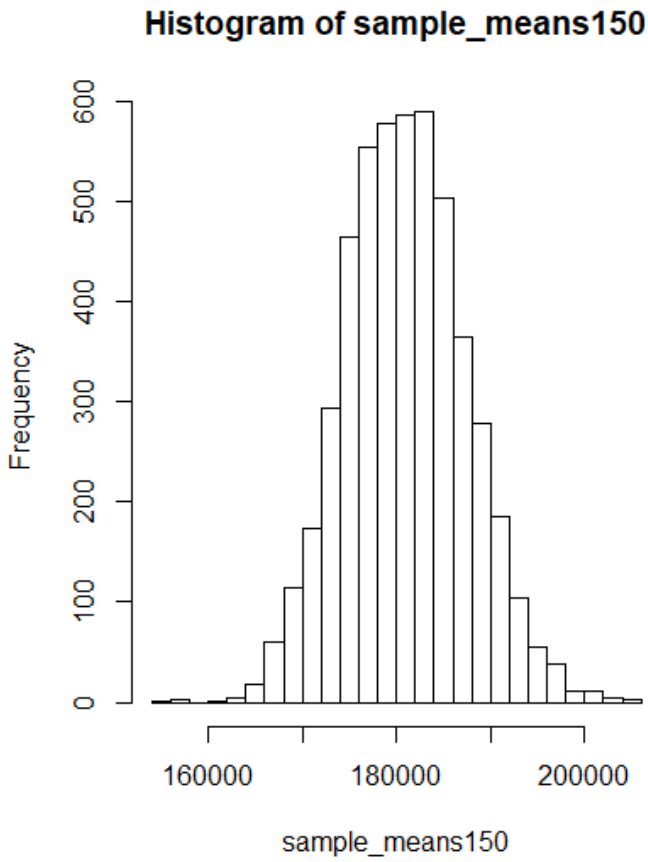
9. Describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?

The shape of this distribution is a bell curve. Based off the distribution, the mean sale prices of homes in Ames is around 180000.



10. Describe the shape of the sampling distribution and compare it to the sample distribution of sample size 50. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?

The shape of this distribution is a bell curve. Compared to the distribution of sample size 50, this one has more weight in the center. Based off this distribution, the mean sale prices of homes in Ames is about slightly higher 180000.



11. Of the sampling distributions from 9 to 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 larger or smaller spread?

Of the two sampling distributions, the one with 150 samples has the smaller spread. When we're trying to make estimates that are closer to the true value, we would want to use a sampling distribution with a smaller spread.