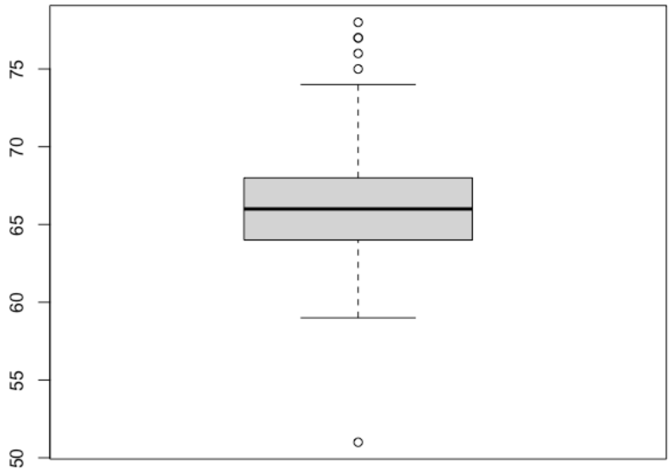
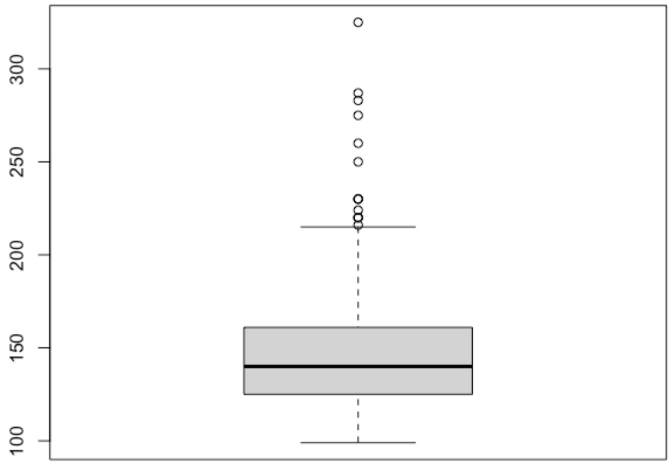
Dehshini Ali
Problem set 3

Box plot of heights
>boxplot(class621$height)



Box plot of weights
>boxplot(class621$weight)



*>class621bmi = mutate(class621, bmi=(weight/height^2)*704.5)*
>class621filtered = filter(class621bmi, bmi < 40, bmi > 15)
BMI values > 40 or <15 are implausible. These BMI values may have been from wrongly entered weight for that given height or wrongly entered height for the given weight. I decided to remove these individuals from the analysis.

Ii

>class621filtered.F = filter(class621filtered, gender==2)
>class621filtered.M = filter(class621filtered, gender==1)

>stem(class621filtered.M$bmi)

For gender = 1 (male)

The decimal point is at the |
16 | 8
18 | 5802266
20 | 24445678013577889
22 | 134125788
24 | 2334444458800111 2349
26 | 2346666713456
28 | 0569905669
30 | 02993
32 | 2
34 | 59
36 |
38 | 4

>summary(class621filtered.M$bmi)

summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17.79	21.75	24.62	24.97	27.36	38.43

> sd(class621filtered.M$bmi)

Sd = 4.084464

>stem(class621filtered.F$bmi)

For gender = 2 (female)

The decimal point is at the |

16 | 9
17 | 01334556889
18 | 0334555666999
19 | 00000111222233334445556678888889
20 | 0000001112222334444566667778888889999
21 | 000000001122223333455555555555777888888
22 | 0001222233445666677777889999999999
23 | 012222333333345555567777778889
24 | 001111222333344579
25 | 0223444456678899999
26 | 124677778
27 | 345555
28 | 004448
29 | 238
30 | 133578
31 | 5
32 | 1
33 | 89
34 |
35 | 06
36 | 7
37 | 48

> summary(class621filtered.F$bmi)

Summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16.91	20.16	22.05	22.76	24.22	37.84

> sd(class621filtered.F$bmi)

Sd = 3.704175

Assuming the distribution is a normal distribution

For males
Mean = 24.97, sd = 4.084464

Using the normal distribution applet,
Middle 50% of people
Mean +/- 0.67(sd)
24.97 +/- 0.67(4.084464)
22.215 – 27.725

Middle 95% of people
Mean +/- 1.96(sd)
24.97 +/- 1.96(4.084464)
16.964 – 32.975

For females
Mean = 22.76, sd = 3.704175

Using the normal distribution applet
Middle 50% of people
Mean +/- 0.67(sd)
22.76 +/- 0.67(3.704175)
20.262 – 25.258

Middle 95% of people
Mean +/- 1.96(sd)
22.76 +/- 1.96(3.704175)
15.500 – 30.020

Iii

>quantile(class621filtered.F$bmi, c(.005, .025, .25, .75, .975, .995))

Quantiles for females

0.5%	2.5%	25%	75%	97.5%	99.5%
17.01178	17.53750	20.15999	24.22376	32.37530	37.09679

<u>Middle 50%</u>
20.160 – 24.224
<u>Middle 95%</u>
17.538 – 32.375
<u>Middle 99%</u>
17.012 – 37.097

>quantile(class621filtered.M$bmi, c(.005, .025, .25, .75, .975, .995))

Quantiles for males

0.5%	2.5%	25%	75%	97.5%	99.5%
18.09055	18.85091	21.75464	27.36358	34.34723	36.94747

<u>Middle 50%</u>
21.755 – 27.364
<u>Middle 95%</u>
18.851 – 34.347
<u>Middle 99%</u>
18.091 – 36.947

Comparison of Empirical Intervals With Normal Distribution Intervals

	male		female	
	Normal dist	Actual dist	Normal dist	Actual dist
Middle 50%	22.215 – 27.725	21.755 – 27.364	20.262 – 25.258	20.160 – 24.224
Middle 95%	16.964 – 32.975	18.851 – 34.347	15.500 – 30.020	17.538 – 32.375

iv

For both males and females, the normal distribution slightly overestimates the middle 50% and slightly underestimates the middle 95%.

V

Assuming normal distribution

Using the normal distribution applet

For males
Pr(BMI < 25) = 0.50293
Pr(25 < BMI < 29.9) = 0.388
Pr(BMI > 30) = 0.10907

for females
Pr(BMI < 25) = 0.72732
Pr(25 < BMI < 29.9) = 0.24736
Pr(BMI > 30) = 0.02532

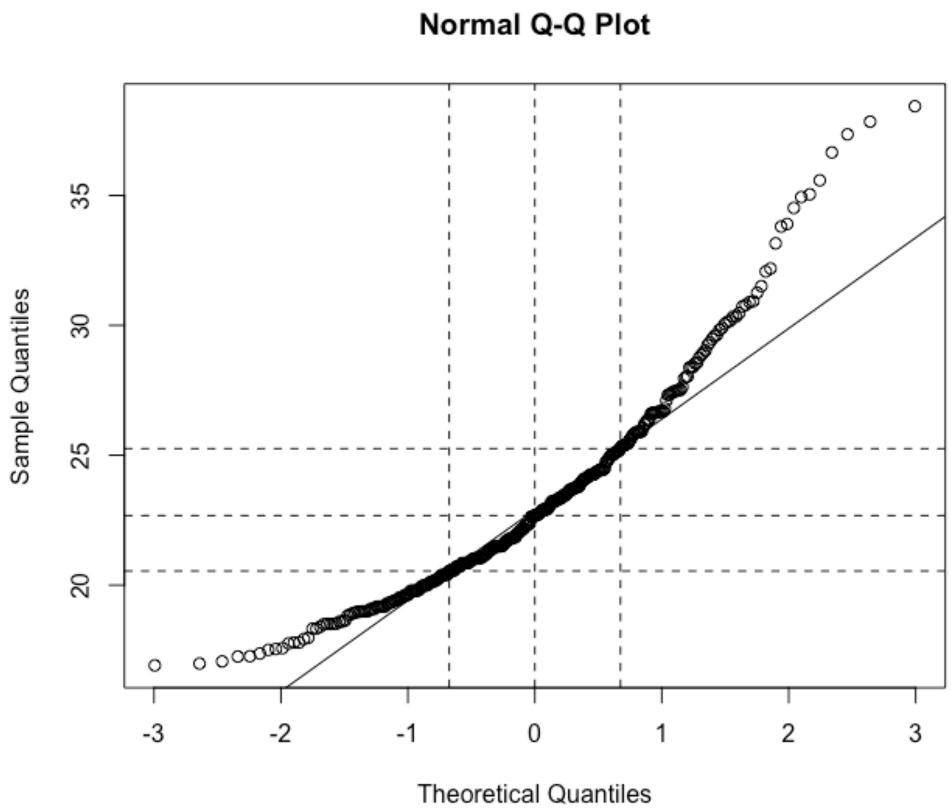
Using actual distribution
For males
Pr(BMI < 25) = 0.51807229
Pr(25 < BMI < 29.9) = 0.37349398
Pr(BMI >= 30) = 0.10843373

For females
Pr(BMI < 25) = 0.78571429
Pr(25 < BMI < 29.9) = 0.15714286
Pr(BMI >= 30) = 0.05714286

	males		**females**	
	Model based	Actual	Model based	Actual
Pr(BMI < 25)	0.503	0.518	0.727	0.786
Pr(25 < BMI < 29.9)	0.388	0.373	0.247	0.157
Pr(BMI >= 30)	0.109	0.108	0.025	0.057

vi.

>qqnorm(class621filtered$bmi)
>qqline(class621filtered$bmi)
>abline(h=quantile(class621filtered$bmi, c(.25,.5,.75), na.rm=TRUE), lty=2)
>abline(v=qnorm(c(.25,.5,.75)), lty=2)

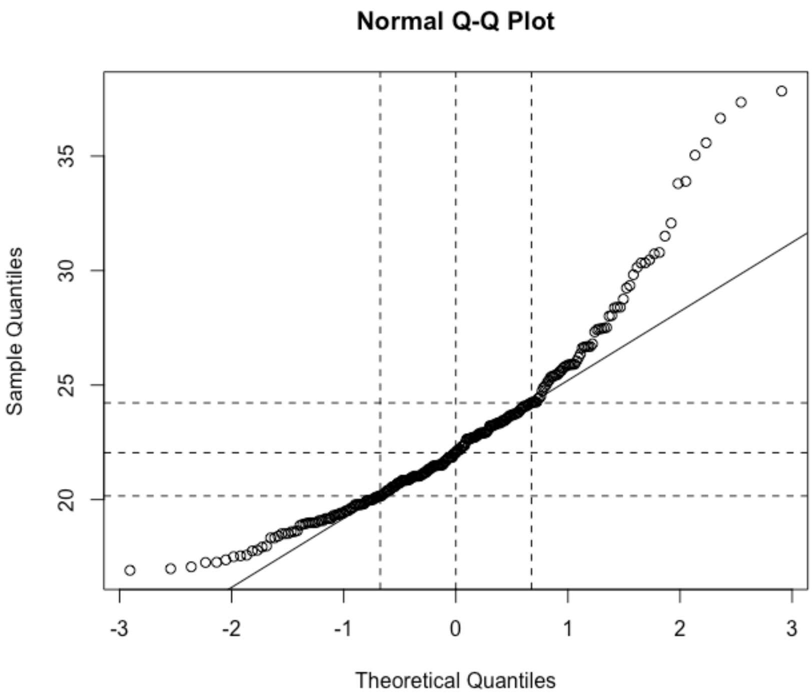


The normal distribution approximates the distribution of the BMI mainly in the middle 50% but underestimates the BMI values in the lower and upper extremes

Vii

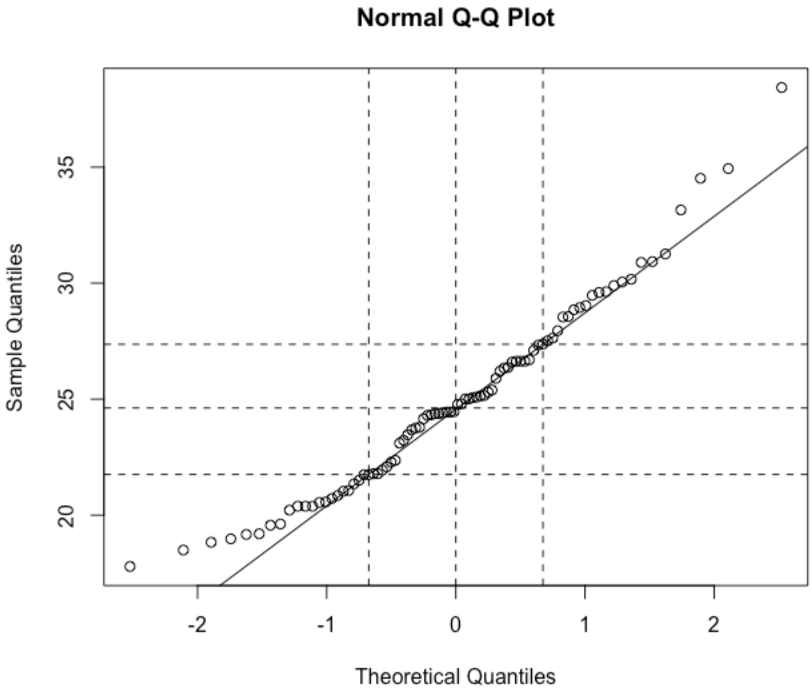
>qqnorm(class621filtered.F$bmi)
>qqline(class621filtered.F$bmi)
>abline(h=quantile(class621filtered.F$bmi, c(.25,.5,.75), na.rm=TRUE), lty=2)
>abline(v=qnorm(c(.25,.5,.75)), lty=2)

For females (gender=2)



```
>qqnorm(class621filtered.M$bmi)
>qqline(class621filtered.M$bmi)
>abline(h=quantile(class621filtered.M$bmi, c(.25,.5,.75), na.rm=TRUE), lty=2)
>abline(v=qnorm(c(.25,.5,.75)), lty=2)
```

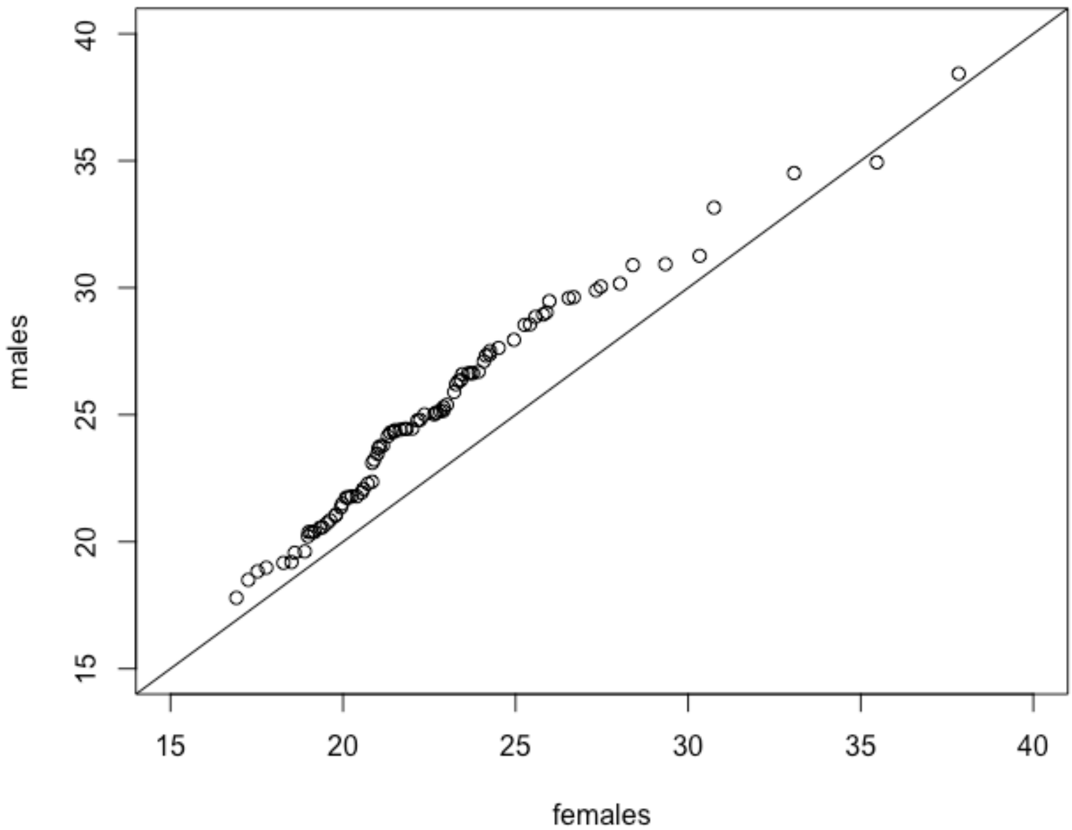
For males (gender=1)



The male BMI values are more well-approximated by the normal distribution for the middle 95% as compared to the female BMI values which are well-approximated by the normal distribution in the middle 50% of the distribution.

viii.

>qqplot(class621filtered.F$bmi, class621filtered.M$bmi, xlim=c(15,40), ylim=c(15,40), xlab = "females", ylab = "males")
>abline(a=0, b=1)

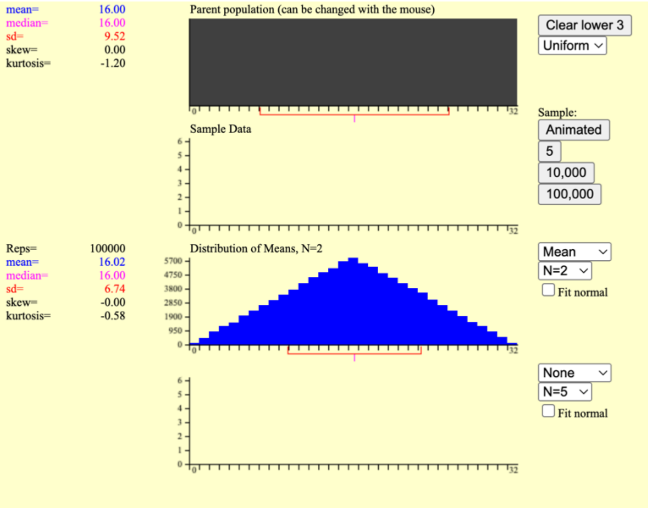


The values of BMI for males are consistently larger than the BMI values for females across the distribution of the sample.

Problem 2

ii.

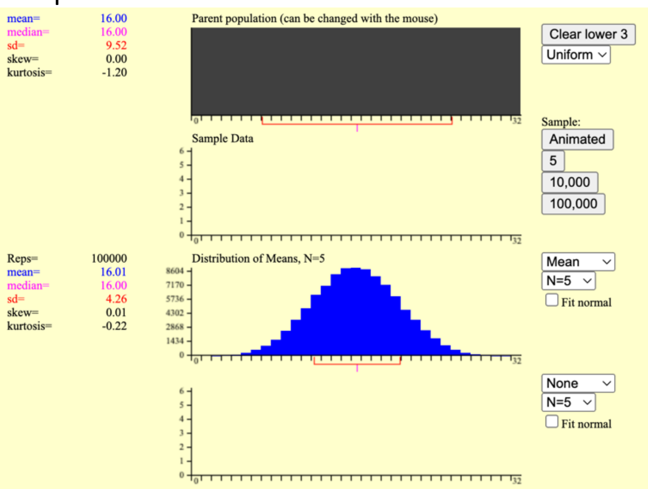
sample size n=2



The sampling distribution takes the shape of a normal distribution with a mean of 16.02 and a standard deviation of 6.74. It has a very broad base signifying a lot of variability.

iii.

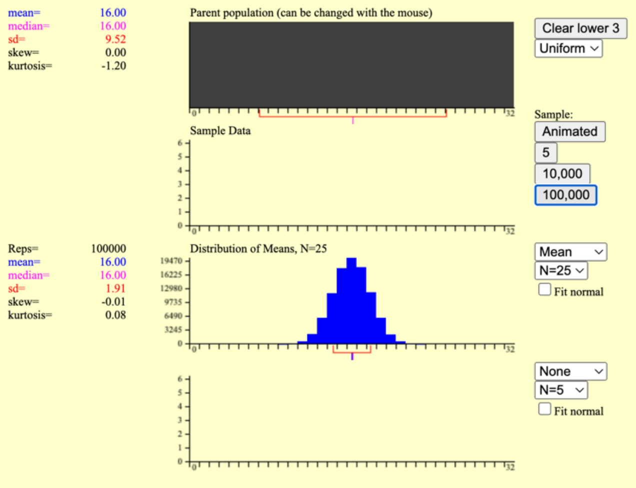
sample size n=5



The sampling distribution takes the shape of a normal distribution with a mean of 16.01 and a standard deviation of 4.26. It has a small base signifying less variability.

iv.

sample size n=25



The sampling distribution takes the shape of a normal distribution with a mean/median of 16 and a standard deviation of 1.91. It has a very narrow base signifying minimal variability.

v.

Estimated Means and Standard Deviations Of The Sampling Distribution(μ=16, var=90)

	Observed statistic for 100,000		Theoretical values for infinite replicates	
Size (n)	Mean	Standard deviation	Mean	Standard deviation
2	16.02	6.74	16	$\sqrt{\frac{90}{2}}$ = 6.708
5	16.01	4.26	16	$\sqrt{\frac{90}{5}}$ = 4.243
25	16.00	1.91	16	$\sqrt{\frac{90}{25}}$ = 1.897
100	NA	NA		

vi.

The sample means are relatively close to the population mean but approach the exact population mean as the sample size increases. The sample mean is directly proportional to the population variance and inversely proportional to the sample size(n)

The central limit theorem states that if you take a sample of size “n” from a population(mean=μ, s.d=σ, whether it is gaussian or not), the sampling distribution assumes a normal distribution if the sample (of size n) is large enough.