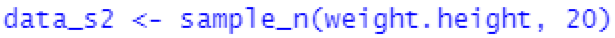
Khan Inan
Assignment 6

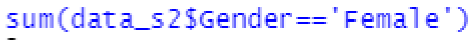
1. I used the “dplyr” data package



```
data_s2 <- sample_n(weight.height, 20)
```

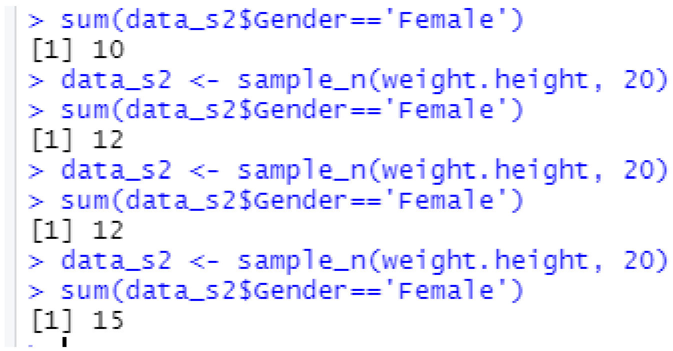
this was my code to create the tables of 20,40,60,80,100 people

And then this was my code to count the number of females, which also tells you the number of males



```
sum(data_s2$Gender=='Female')
```

For 20 sample size:



```
> sum(data_s2$Gender=='Female')
[1] 10
> data_s2 <- sample_n(weight.height, 20)
> sum(data_s2$Gender=='Female')
[1] 12
> data_s2 <- sample_n(weight.height, 20)
> sum(data_s2$Gender=='Female')
[1] 12
> data_s2 <- sample_n(weight.height, 20)
> sum(data_s2$Gender=='Female')
[1] 15
```

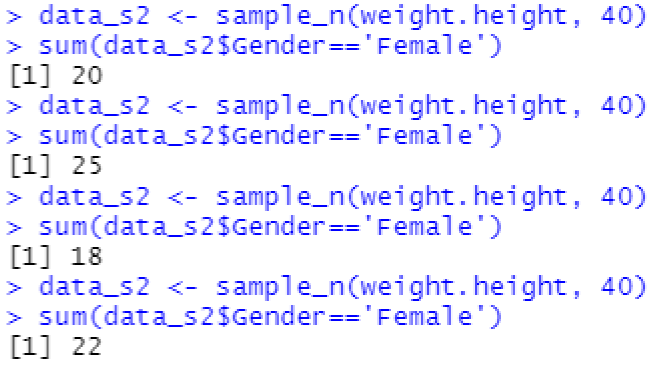
+ 10 men

+ 8 men

+ 8 men

+ 5 men

For 40 sample size



```
> data_s2 <- sample_n(weight.height, 40)
> sum(data_s2$Gender=='Female')
[1] 20
> data_s2 <- sample_n(weight.height, 40)
> sum(data_s2$Gender=='Female')
[1] 25
> data_s2 <- sample_n(weight.height, 40)
> sum(data_s2$Gender=='Female')
[1] 18
> data_s2 <- sample_n(weight.height, 40)
> sum(data_s2$Gender=='Female')
[1] 22
```

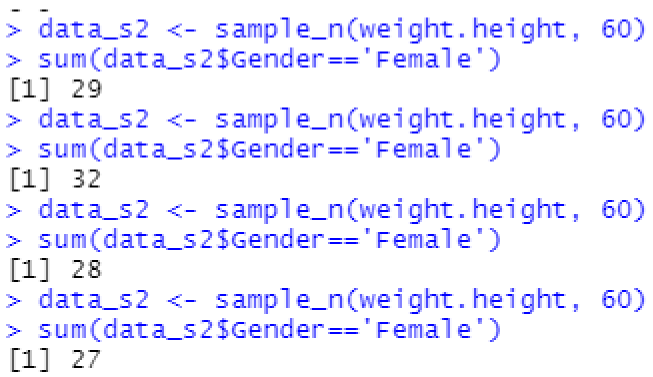
+ 20 men

+ 15 men

+ 12 men

+ 8 men

For 60 sample size



```
> data_s2 <- sample_n(weight.height, 60)
> sum(data_s2$Gender=='Female')
[1] 29
> data_s2 <- sample_n(weight.height, 60)
> sum(data_s2$Gender=='Female')
[1] 32
> data_s2 <- sample_n(weight.height, 60)
> sum(data_s2$Gender=='Female')
[1] 28
> data_s2 <- sample_n(weight.height, 60)
> sum(data_s2$Gender=='Female')
[1] 27
```

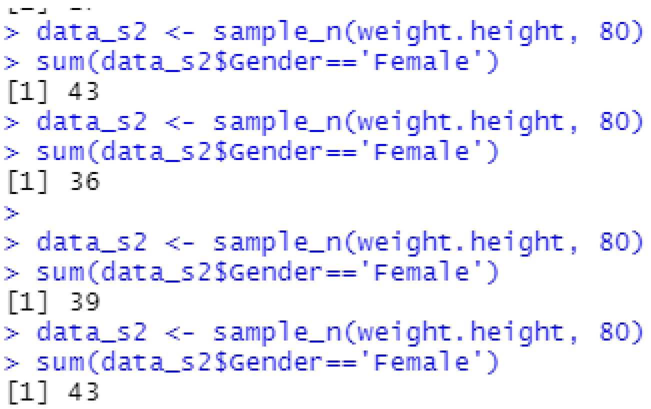
+ 11 men

+ 8 men

+ 12 men

+ 13 men

For 80 sample size



```
> data_s2 <- sample_n(weight.height, 80)
> sum(data_s2$Gender=='Female')
[1] 43
> data_s2 <- sample_n(weight.height, 80)
> sum(data_s2$Gender=='Female')
[1] 36
>
> data_s2 <- sample_n(weight.height, 80)
> sum(data_s2$Gender=='Female')
[1] 39
> data_s2 <- sample_n(weight.height, 80)
> sum(data_s2$Gender=='Female')
[1] 43
```

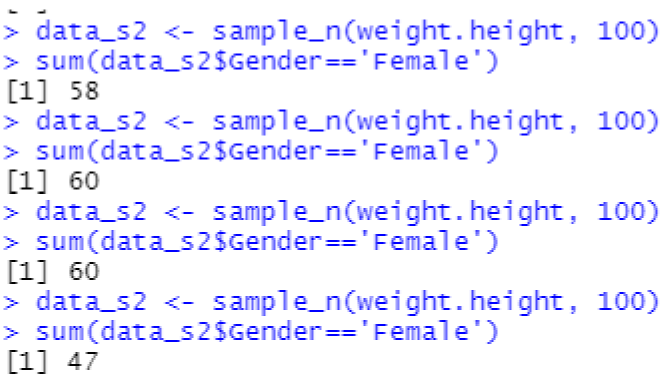
+ 37 men

+ 44 men

+ 41 men

+ 37 men

For 100 sample size



```
> data_s2 <- sample_n(weight.height, 100)
> sum(data_s2$Gender=='Female')
[1] 58
> data_s2 <- sample_n(weight.height, 100)
> sum(data_s2$Gender=='Female')
[1] 60
> data_s2 <- sample_n(weight.height, 100)
> sum(data_s2$Gender=='Female')
[1] 60
> data_s2 <- sample_n(weight.height, 100)
> sum(data_s2$Gender=='Female')
[1] 47
```

+ 42 men

+ 40 men

+ 40 men

+ 53 men

Based on these results and chi-squared test, you cannot accurately conclude that half the population in weight.height table is men and women

C. You would need atleast 1000 sample size to accurately determine the ratio of men to women in the original data set

D. since the population of NYC is 8 million, I would say you need about 10% sample size to accurately determine the ratio of male to females in the city, so about 800,000

2.A I got a p-value of about 2.448 for the genes, as a result you have around 8 different genes from the data set

B. if you relax the data set to around 20% you get a few more variety in the genes, like around 12. This is because more are being sampled

C. only 10 genes survive if you use the Bonferroni correction with a sig value of 0.05

D. For the most accurate results I would use the 20% value, and maybe even relax the FDR more to get an even more accurate result for the entire data set