CS 217
Fall 2018
Prof. Elahe Vahdani
Eftekher Husain
R Assignments

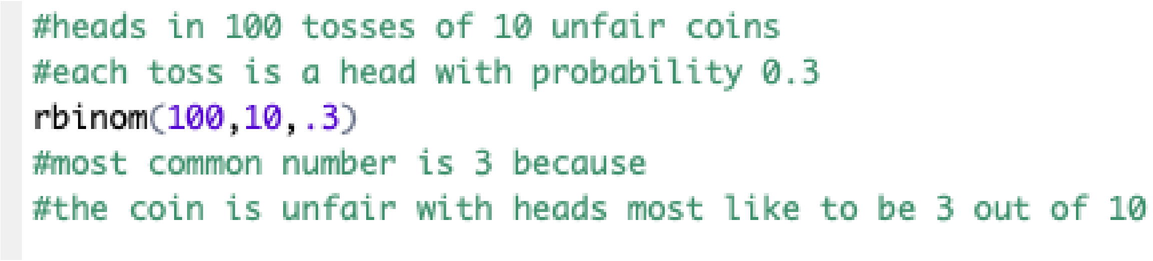
R Assignment 1

Exercise 1

Question: Generate 100 experiments of flipping 10 coins, each with 30% probability.

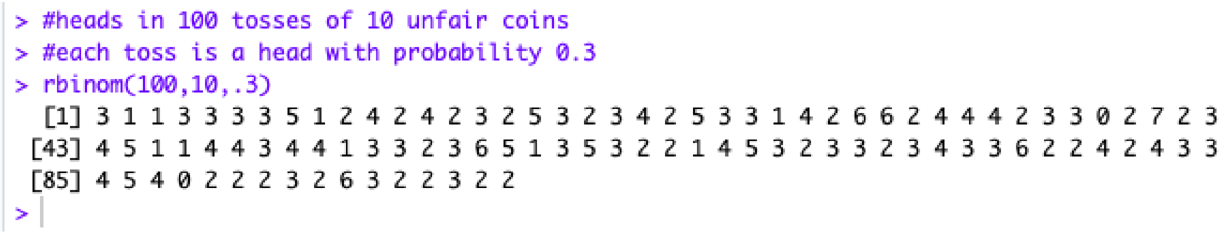
Answer: When we say that a coin has 30% probability, we are actually saying that the coins are unfair. In each toss, probability that a head might appear is .3. In R, to find the solution for the given problem, we type in rbinom () to get the outcomes. Inside rbinom (), we have the first parameter which is 100, describes the number of experiments. The second parameter inside rbinom (), describes the coin tosses. And Lastly the third parameter inside the rbinom () describes the probability. We want to do 100 experiments of flipping 10 coins, each with probability of .3. So, we write rbinom () as rbinom (100,10,.3). The outcomes are the numbers of heads in each experiment.

Code in R



```
#heads in 100 tosses of 10 unfair coins
#each toss is a head with probability 0.3
rbinom(100,10,.3)
#most common number is 3 because
#the coin is unfair with heads most like to be 3 out of 10
```

Outcome in Console



```
> #heads in 100 tosses of 10 unfair coins
> #each toss is a head with probability 0.3
> rbinom(100,10,.3)
  [1] 3 1 1 3 3 3 3 5 1 2 4 2 4 2 3 2 5 3 2 3 4 2 5 3 3 1 4 2 6 6 2 4 4 4 2 3 3 0 2 7 2 3
 [43] 4 5 1 1 4 4 3 4 4 1 3 3 2 3 6 5 1 3 5 3 2 2 1 4 5 3 2 3 3 2 3 4 3 3 6 2 2 4 2 4 3 3
 [85] 4 5 4 0 2 2 2 3 2 6 3 2 2 3 2 2
> 
```

Question: What is the most common number? Why?

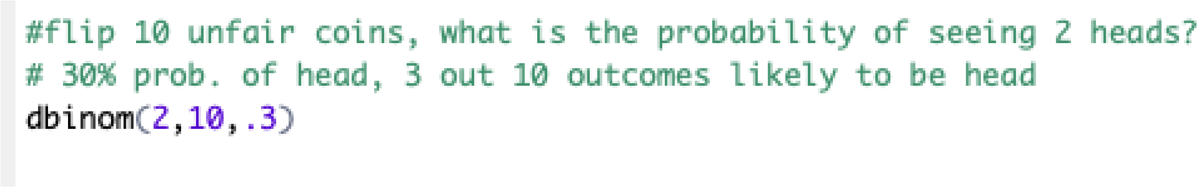
Answer: After analyzing the outcomes, we could see that the most common number is 3 because the coin tosses are unfair with probability .3 which means there are most likely to appear as the outcome.

Exercise 2

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

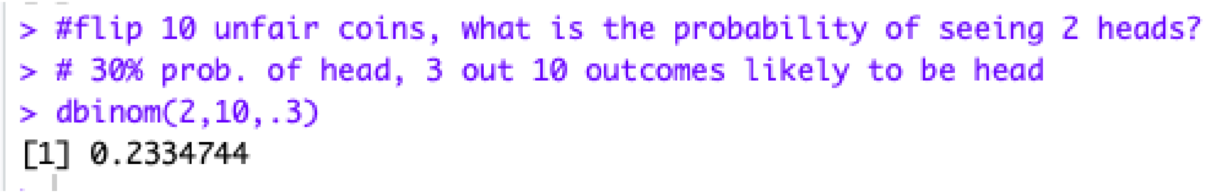
Answer: In this exercise, we use the dbinom(2,10,.3) command, where 2 is the head we want, 10 is the coin tosses and .3 is the probability of coming up heads. Using dbinom(2,10,.3), we get 0.2334744 as the probability.

Code in R



```
#flip 10 unfair coins, what is the probability of seeing 2 heads?
# 30% prob. of head, 3 out 10 outcomes likely to be head
dbinom(2,10,.3)
```

Outcome in Console



```
> #flip 10 unfair coins, what is the probability of seeing 2 heads?
> # 30% prob. of head, 3 out 10 outcomes likely to be head
> dbinom(2,10,.3)
[1] 0.2334744
```

Question: Compare your simulation with exact calculation.

Answer: The exact calculation is very close to our simulation.

Exercise 3

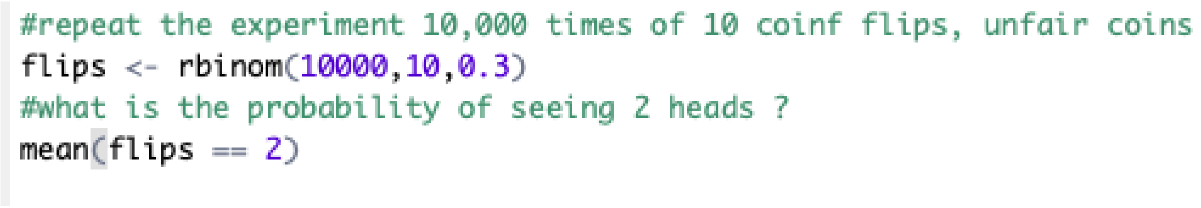
For Exercise 2,

Question:

Part a) Use 10,000 experiments and report the result.

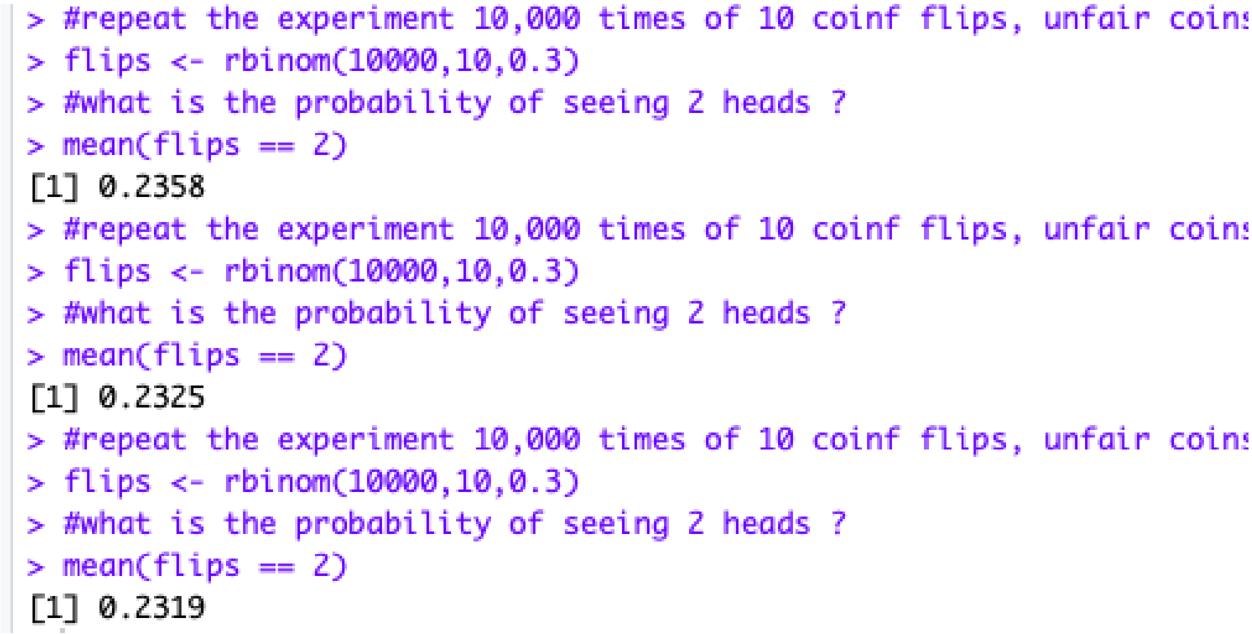
Answer: To experiment it, we use ribinom(10000,10,.3) and store it as flips, then we find the probability of seeing 2 heads by using mean(flips == 2) which gives us the outcome around 0.2358.

Code in R



```
#repeat the experiment 10,000 times of 10 coinf flips, unfair coins
flips <- rbinom(10000,10,0.3)
#what is the probability of seeing 2 heads ?
mean(flips == 2)
```

Outcome in Console

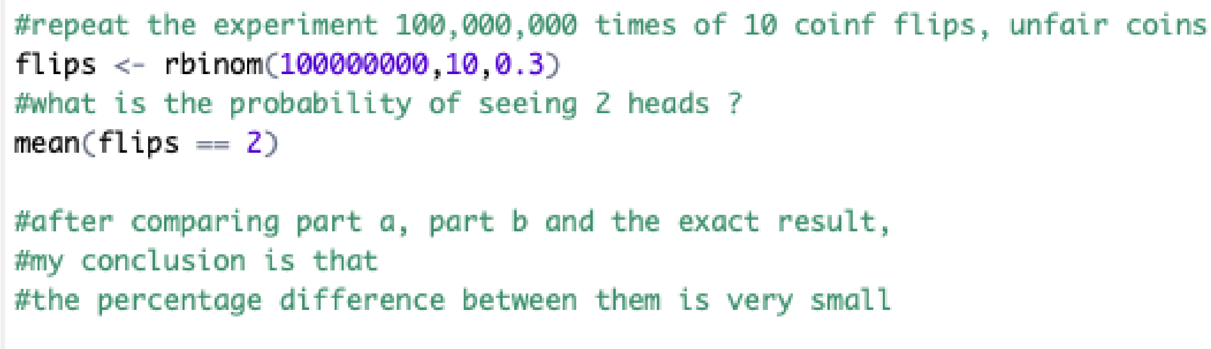


```
> #repeat the experiment 10,000 times of 10 coinf flips, unfair coin
> flips <- rbinom(10000,10,0.3)
> #what is the probability of seeing 2 heads ?
> mean(flips == 2)
[1] 0.2358
> #repeat the experiment 10,000 times of 10 coinf flips, unfair coin
> flips <- rbinom(10000,10,0.3)
> #what is the probability of seeing 2 heads ?
> mean(flips == 2)
[1] 0.2325
> #repeat the experiment 10,000 times of 10 coinf flips, unfair coin
> flips <- rbinom(10000,10,0.3)
> #what is the probability of seeing 2 heads ?
> mean(flips == 2)
[1] 0.2319
```

Part b) Use 100,000,000 experiments and report the result.

Answer: To experiment it, we use ribinom(100000000,10,.3) and store it as flips, then we find the probability of seeing 2 heads by using mean(flips == 2) which gives us the outcome 0.2335164

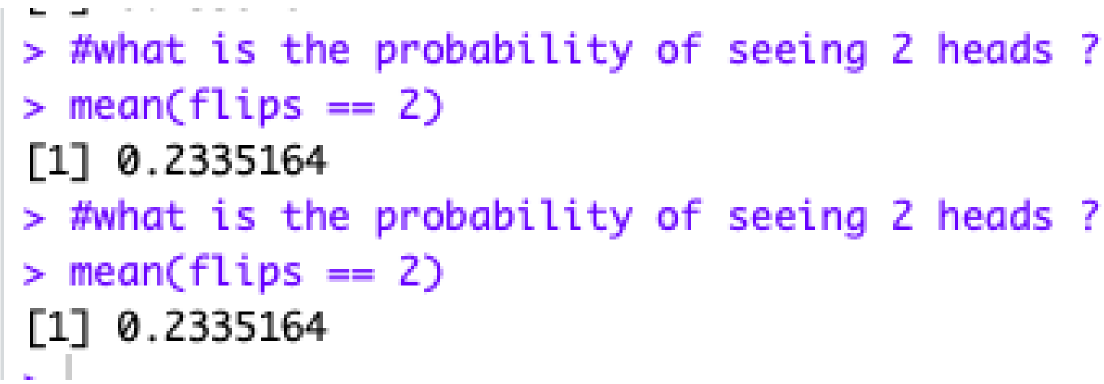
Code in R



```
#repeat the experiment 100,000,000 times of 10 coinf flips, unfair coins
flips <- rbinom(100000000,10,0.3)
#what is the probability of seeing 2 heads ?
mean(flips == 2)

#after comparing part a, part b and the exact result,
#my conclusion is that
#the percentage difference between them is very small
```

Outcome in Console



```
> #what is the probability of seeing 2 heads ?
> mean(flips == 2)
[1] 0.2335164
> #what is the probability of seeing 2 heads ?
> mean(flips == 2)
[1] 0.2335164
```

Question: Compare the result of part a and part b, with the exact calculation. What is your conclusion?

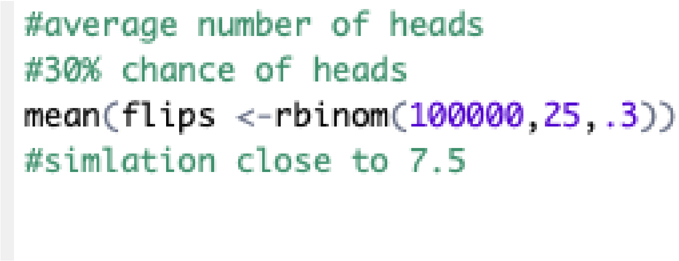
Answer: The outcome of one is very close to the other. However, with 10,000 experiments the probability was in four decimal places but with 100,000,000 experiments, the probability was in seven decimal places showing a clear and concise approximations.

Exercise 4

Question: What is the expected value of a binomial distribution where 25 coins are flipped, each having a 30% chance of heads.

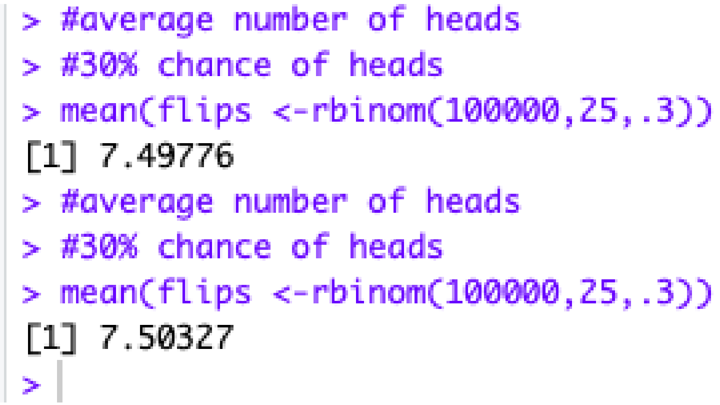
Answer: Using calculation, E[X] = 7.5 or we could say that the result simulation is closed to 7.5.

Code in R



```
#average number of heads
#30% chance of heads
mean(flips <-rbinom(100000,25,.3))
#simlation close to 7.5
```

Outcome in Console



```
> #average number of heads
> #30% chance of heads
> mean(flips <-rbinom(100000,25,.3))
[1] 7.49776
> #average number of heads
> #30% chance of heads
> mean(flips <-rbinom(100000,25,.3))
[1] 7.50327
> 
```

Question: Compare your simulation with the exact calculation.

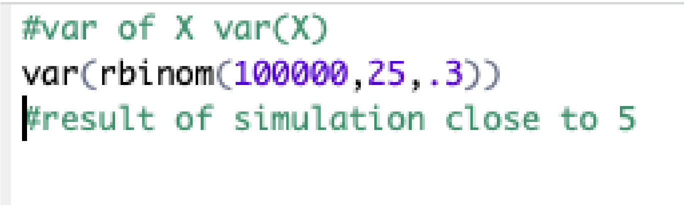
Answer: They have very little difference in terms of value.

Exercise 5

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

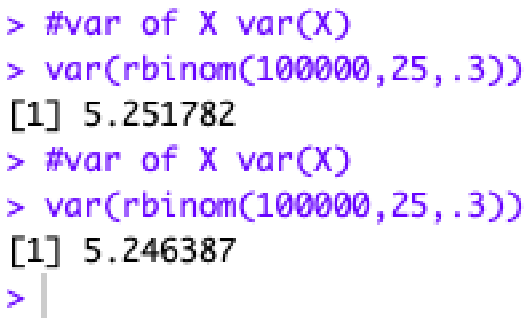
Answer: The variance, var(X) = 5.

Code in R



```
#var of X var(X)
var(rbinom(100000,25,.3))
#result of simulation close to 5
```

Outcome in Console



```
> #var of X var(X)
> var(rbinom(100000,25,.3))
[1] 5.251782
> #var of X var(X)
> var(rbinom(100000,25,.3))
[1] 5.246387
> 
```

Question: Compare your simulation with the exact calculations.

Answer: The result of simulation is close to 5.

CS 217
Fall 2018
Prof. Elahe Vahdani
Eftekher Husain
R Assignments

R Assignment 2

Exercise 1

Question: Plot the histograms. How would you compare the various aspects of two distributions?

Histograms

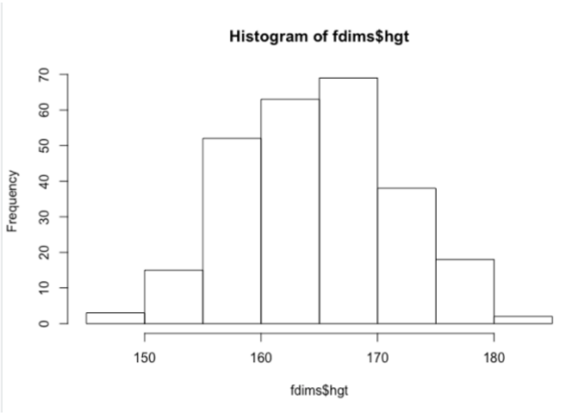


Figure 1: Histogram of women's height

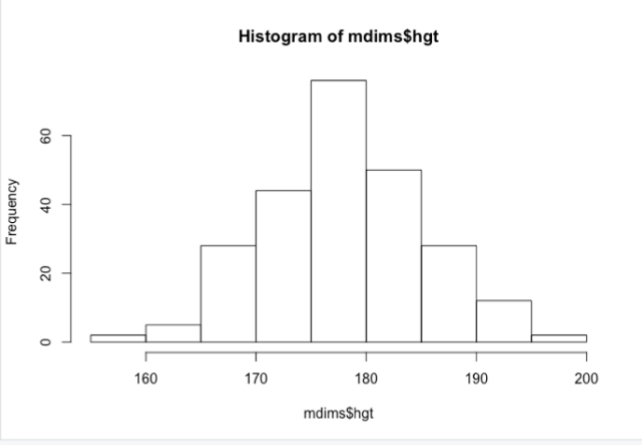


Figure 2: Histogram of men's height

Answer: When comparing various aspects of the two distributions, the shapes of the two distributions are quite similar. The spread of the two distribution is also similar with most of the observations falling within an interval spanning 25 cm. However, they differ most notably in the centers with a mean/median/mode of 178 cm for men and 165 cm for women.

Exercise 2

Question: Based on this plot, does it appear that the data follow a nearly normal distribution?

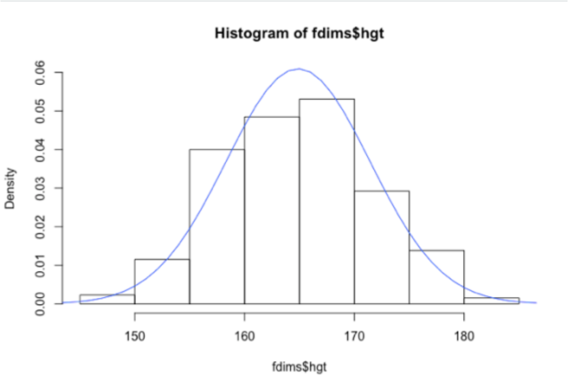


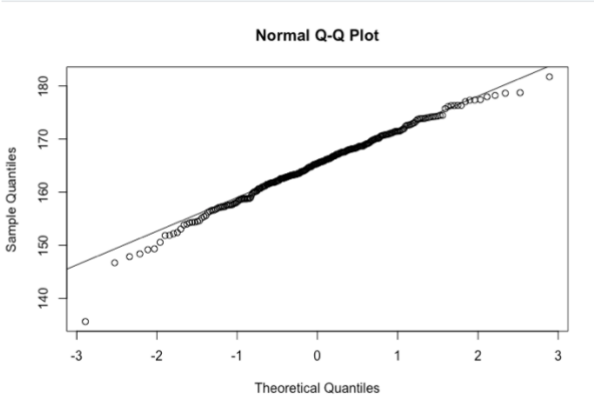
Figure 3: Plotting a Normal Distribution Curve on the Histogram

Answer: Based on this plot, it is also difficult to tell if that data follows a nearly normal distribution. The histogram does seem bell – shaped but has a higher concentration in the middle of the distribution. The normal curve does seem like a reasonable approximation.

Exercise 3

For Exercise 2,

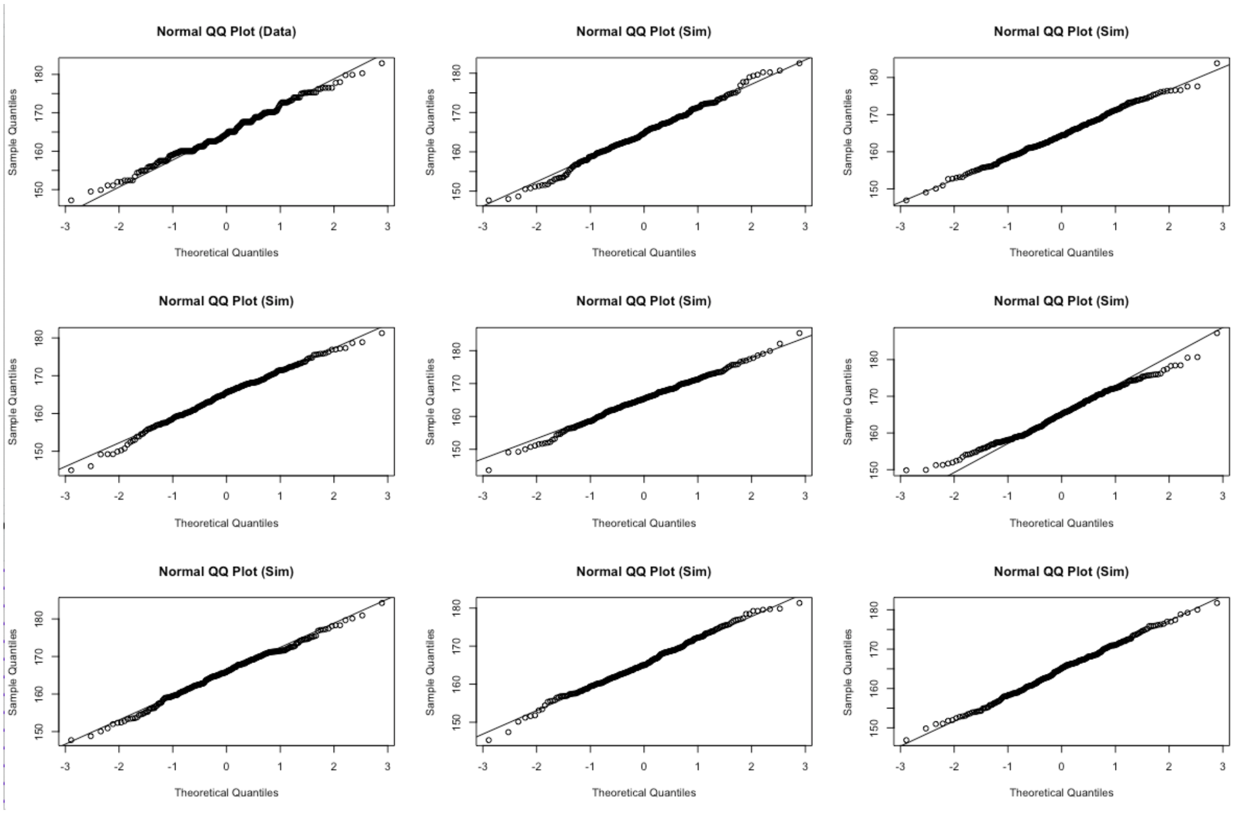
Question: Make a normal probability plot of sim. Do all the points fall on the line? How does this plot compare to the probability plot for the real data?



Answer: Not all the points fall on the line, but they are quite close. The largest deviation is in the tail of the distribution. The simulated plot is smoother in linearity than the data due to a more accurate approximation.

Exercise 4

Question: 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?



Answer: The QQ Plot for female heights is very similar to that from simulated normal data set. Several plots show a great deviation from linearity in the tails as the original data. Main difference seems to be the stair case shape of the data.

Exercise 5

Question: 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.

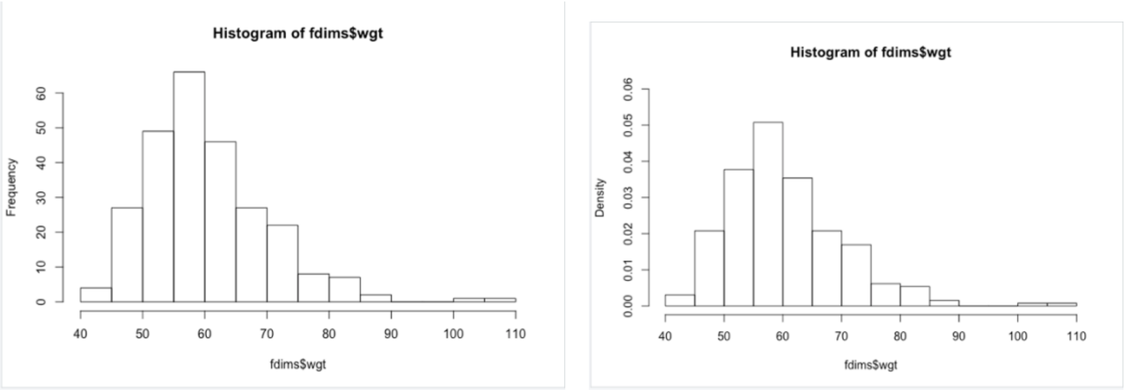
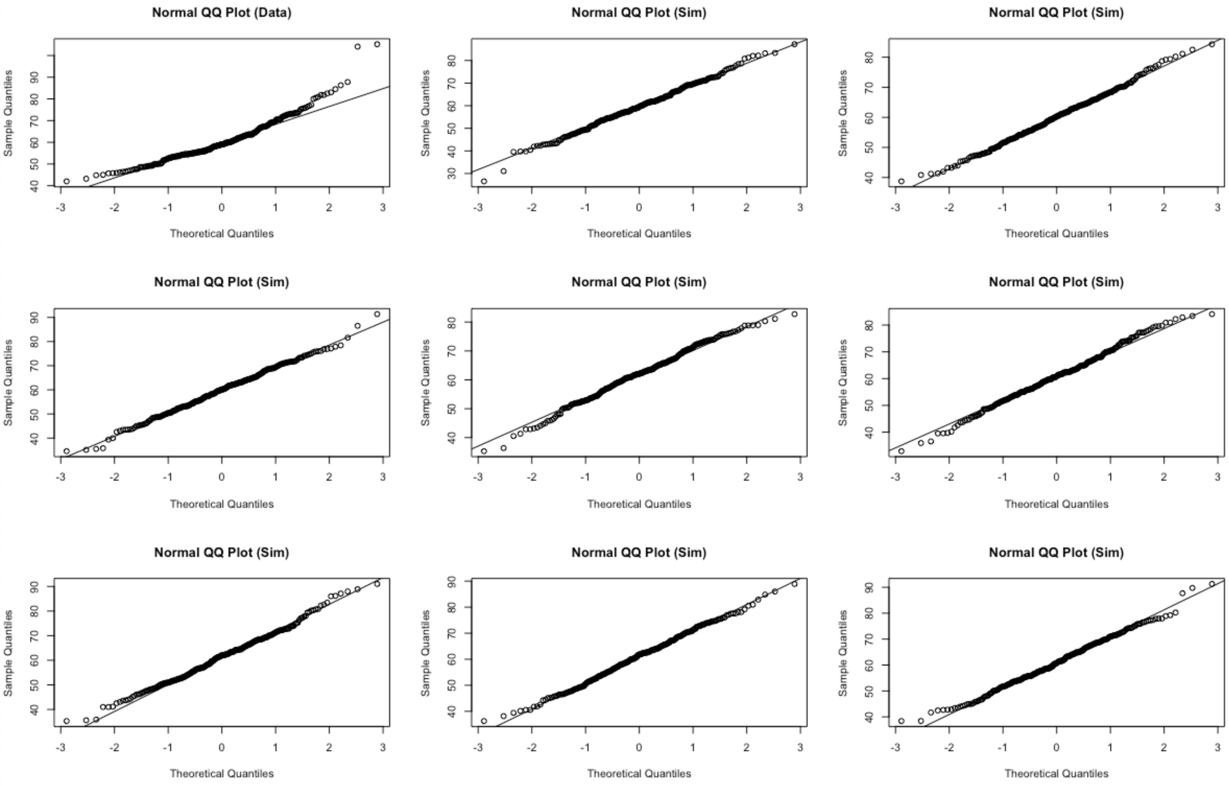


Figure 4: Histogram of Women's weight



Answer: After analyzing the histograms and plotting normal distribution curve, the normal approximation appears to be less appropriate for **wgt** than for **hgt**. The data shows some curvature in the shape of the QQ Plot that suggest a longer right tail that a normal data would show and it shows two notable outliners.

Exercise 6

Question:

Now let’s consider some of the other variables in the body dimensions. 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

___

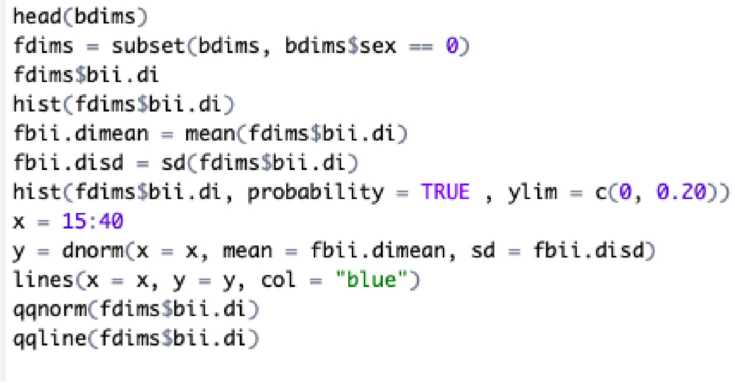
(b) The histogram for female elbow diameter (elb.di) belongs to normal probability plot letter

___

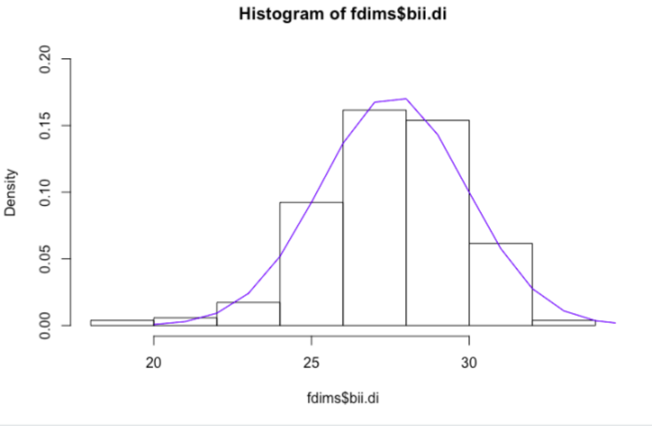
(c) The histogram for general age (age) belongs to normal probability plot letter ___

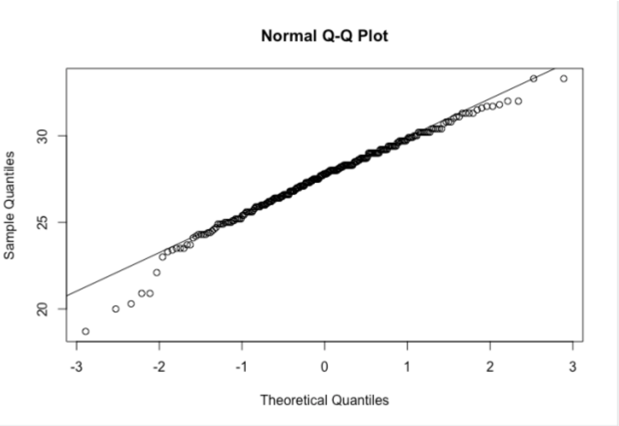
(d) The histogram for female chest depth (che.de) belongs to normal probability plot letter ___

Answer: (a) The histogram for female bi-iliac diameter (bii.di) belongs to normal probability plot letter **B(Normal Q – Q Plot B).**

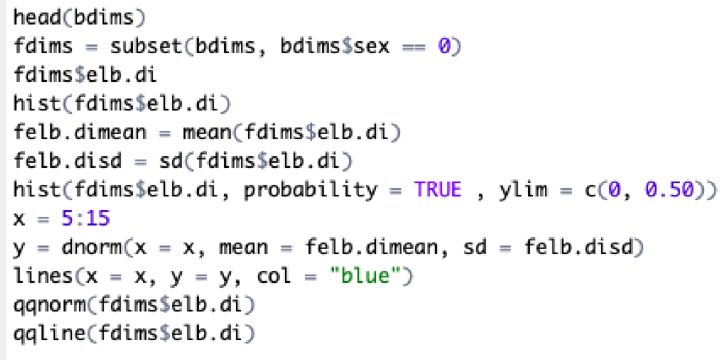


```
head(bdims)
fdims = subset(bdims, bdims$sex == 0)
fdims$bii.di
hist(fdims$bii.di)
fbii.dimean = mean(fdims$bii.di)
fbii.disd = sd(fdims$bii.di)
hist(fdims$bii.di, probability = TRUE , ylim = c(0, 0.20))
x = 15:40
y = dnorm(x = x, mean = fbii.dimean, sd = fbii.disd)
lines(x = x, y = y, col = "blue")
qqnorm(fdims$bii.di)
qqline(fdims$bii.di)
```

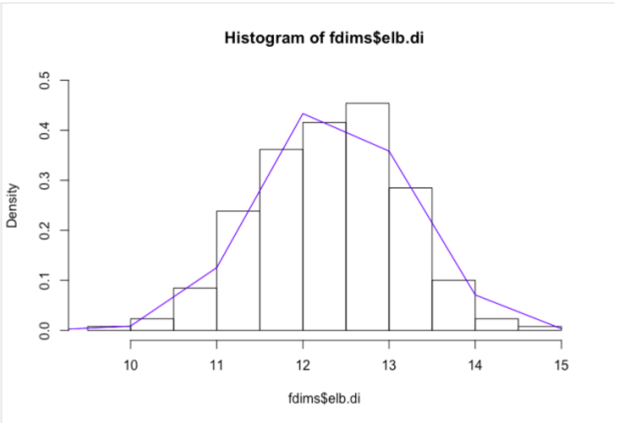


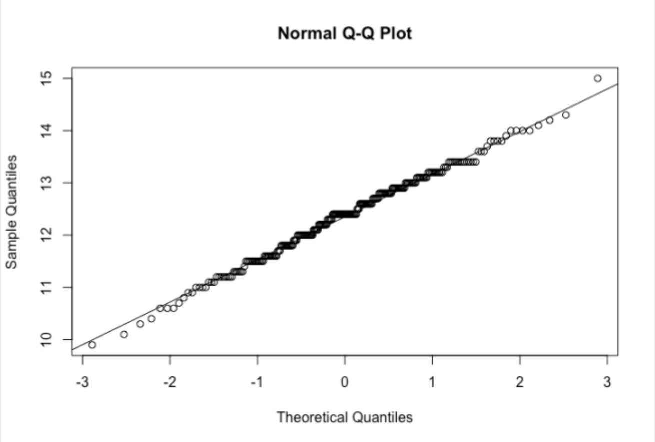


Answer: (b) The histogram for female elbow diameter (elb.di) belongs to normal probability plot letter ___**C(Normal Q – Q Plot C)**

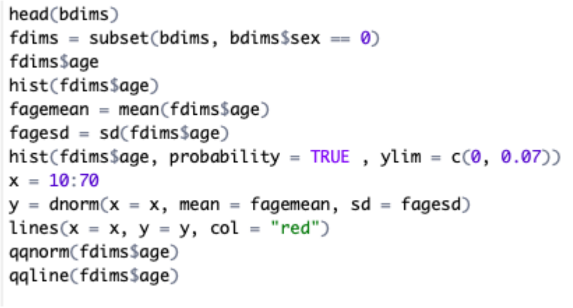


```
head(bdims)
fdims = subset(bdims, bdims$sex == 0)
fdims$elb.di
hist(fdims$elb.di)
felb.dimean = mean(fdims$elb.di)
felb.disd = sd(fdims$elb.di)
hist(fdims$elb.di, probability = TRUE , ylim = c(0, 0.50))
x = 5:15
y = dnorm(x = x, mean = felb.dimean, sd = felb.disd)
lines(x = x, y = y, col = "blue")
qqnorm(fdims$elb.di)
qqline(fdims$elb.di)
```

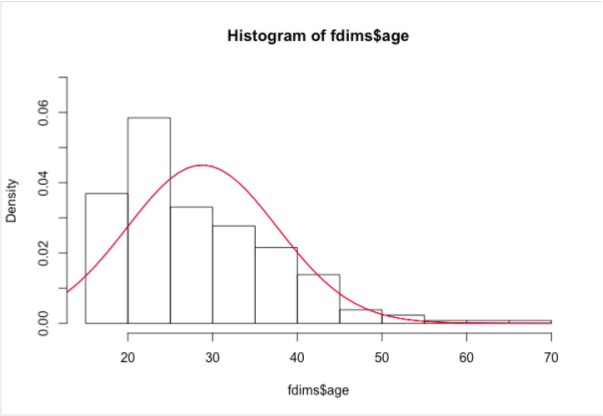


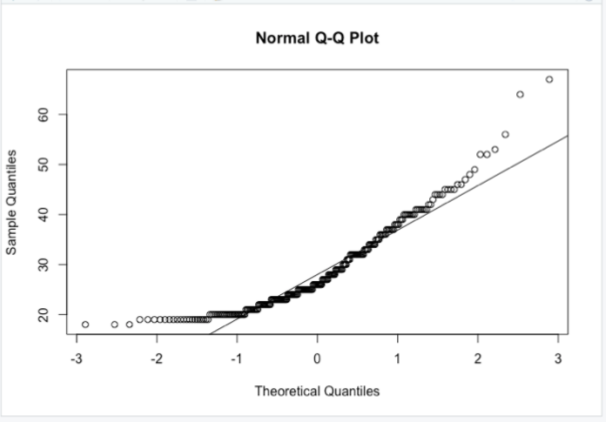


Answer: (c) The histogram for general age (age) belongs to normal probability plot letter ___**D (Normal Q-Q Plot D)**

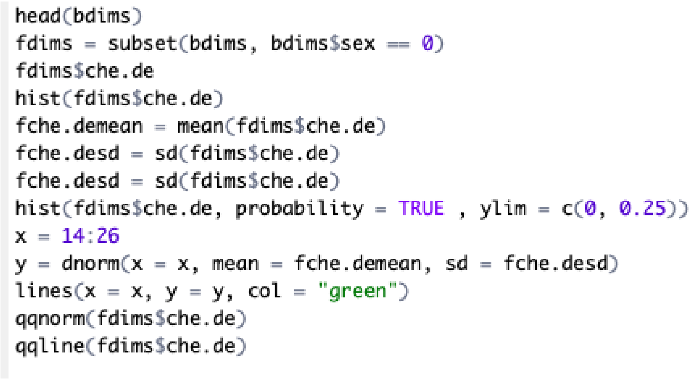


```
head(bdims)
fdims = subset(bdims, bdims$sex == 0)
fdims$age
hist(fdims$age)
fagemean = mean(fdims$age)
fagesd = sd(fdims$age)
hist(fdims$age, probability = TRUE , ylim = c(0, 0.07))
x = 10:70
y = dnorm(x = x, mean = fagemean, sd = fagesd)
lines(x = x, y = y, col = "red")
qqnorm(fdims$age)
qqline(fdims$age)
```

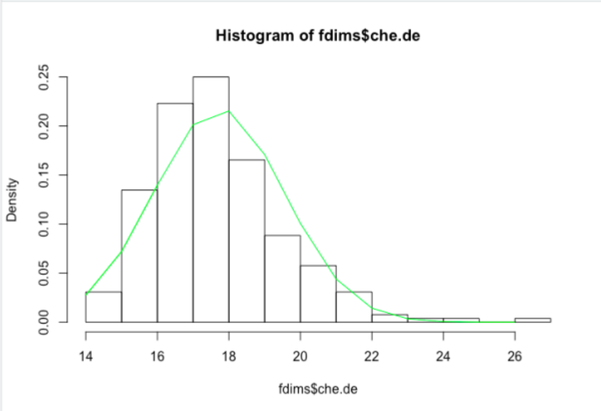


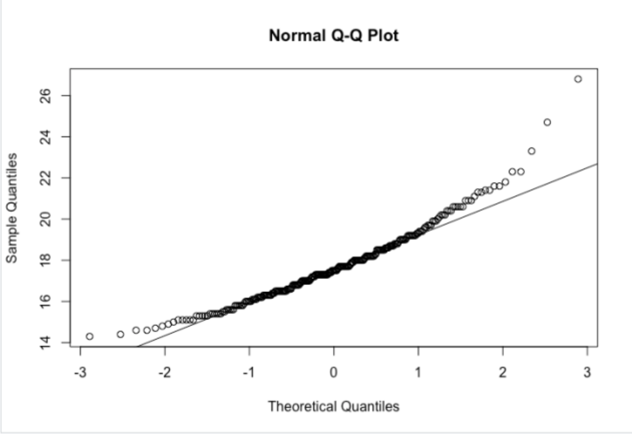


Answer: (d) The histogram for female chest depth (che.de) belongs to normal probability plot letter ___**A (Normal Q-Q Plot A)**



```
head(bdims)
fdims = subset(bdims, bdims$sex == 0)
fdims$che.de
hist(fdims$che.de)
fche.demean = mean(fdims$che.de)
fche.desd = sd(fdims$che.de)
fche.desd = sd(fdims$che.de)
hist(fdims$che.de, probability = TRUE , ylim = c(0, 0.25))
x = 14:26
y = dnorm(x = x, mean = fche.demean, sd = fche.desd)
lines(x = x, y = y, col = "green")
qqnorm(fdims$che.de)
qqline(fdims$che.de)
```





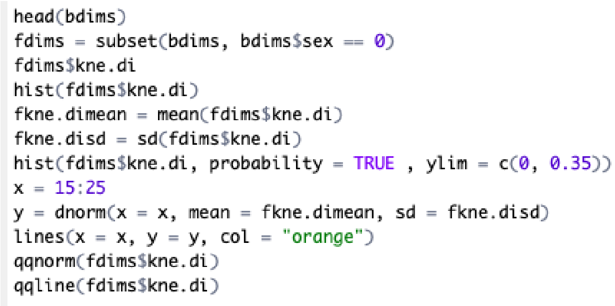
Exercise 7

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

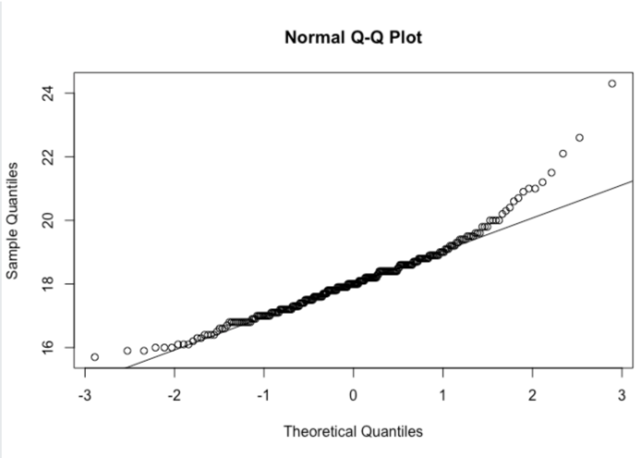
Answer: The reason for this maybe that the data was measured in discrete scale. For example, when people report their age, they usually only provide an integer value and not some decimal number like 25.8384 years. This creates a stepwise pattern in the variable on the y-axis. The x-axis is showing the percentiles of the normal distribution which is why the plots are continuous in their x-axis.

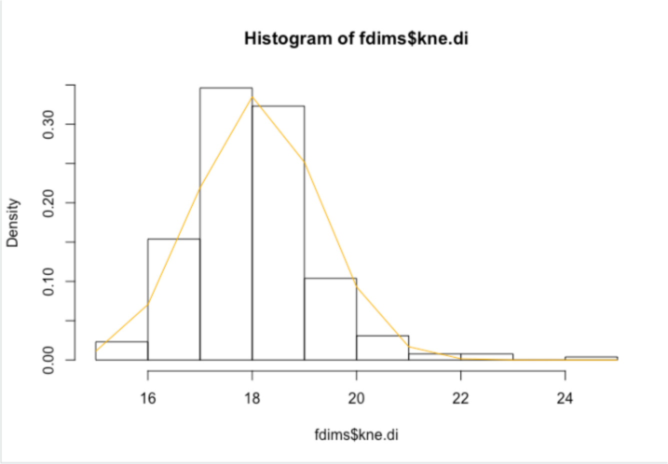
Exercise 8

Question: 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.



```
head(bdims)
fdims = subset(bdims, bdims$sex == 0)
fdims$kne.di
hist(fdims$kne.di)
fkne.dimean = mean(fdims$kne.di)
fkne.disd = sd(fdims$kne.di)
hist(fdims$kne.di, probability = TRUE , ylim = c(0, 0.35))
x = 15:25
y = dnorm(x = x, mean = fkne.dimean, sd = fkne.disd)
lines(x = x, y = y, col = "orange")
qqnorm(fdims$kne.di)
qqline(fdims$kne.di)
```





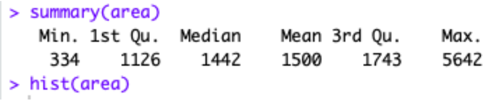
Answer: After making the normal probability plot, it is clear that the normal curve is a poor approximation to female knee diameter. The qqplot shows strong deviation from linearity in the right tail, suggesting that it's longer that normal distribution which makes us conclude that it is right-skewed.

CS 217
Fall 2018
Prof. Elahe Vahdani
Eftekher Husain
R Assignments

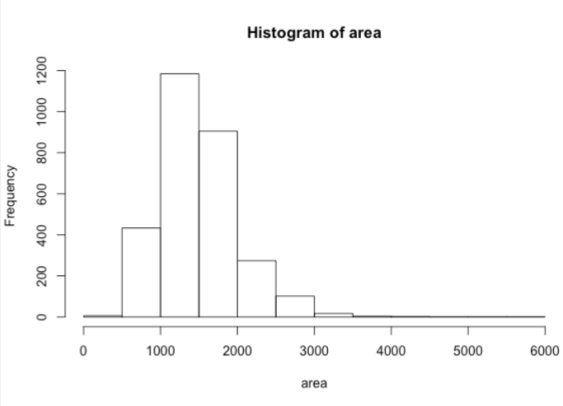
R Assignment 3

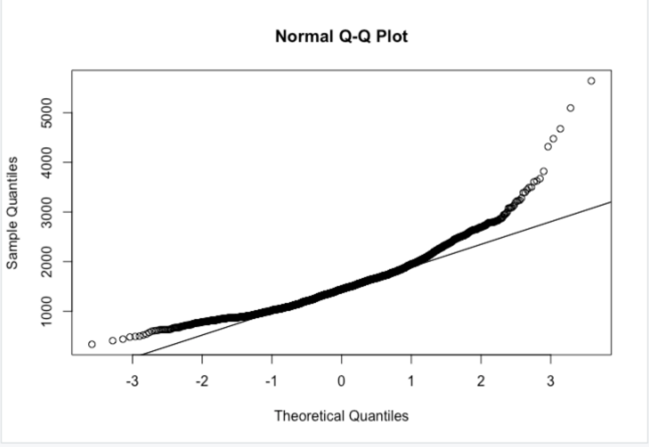
Exercise 1

Question: Describe this population distribution.



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



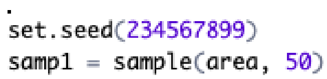


Answer: The Distribution is right skewed.

Exercise 2

Question: Set a unique “random” seed

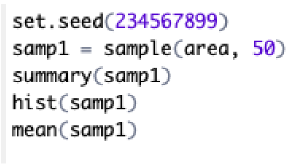
Answer:



```
set.seed(234567899)
samp1 = sample(area, 50)
```

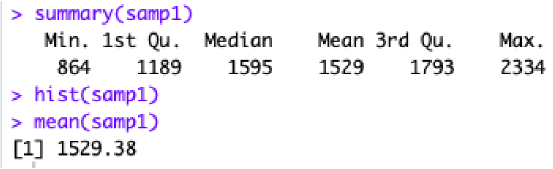
Exercise 3

Question: Describe the distribution of this sample? How does it compare to the distribution of the population?

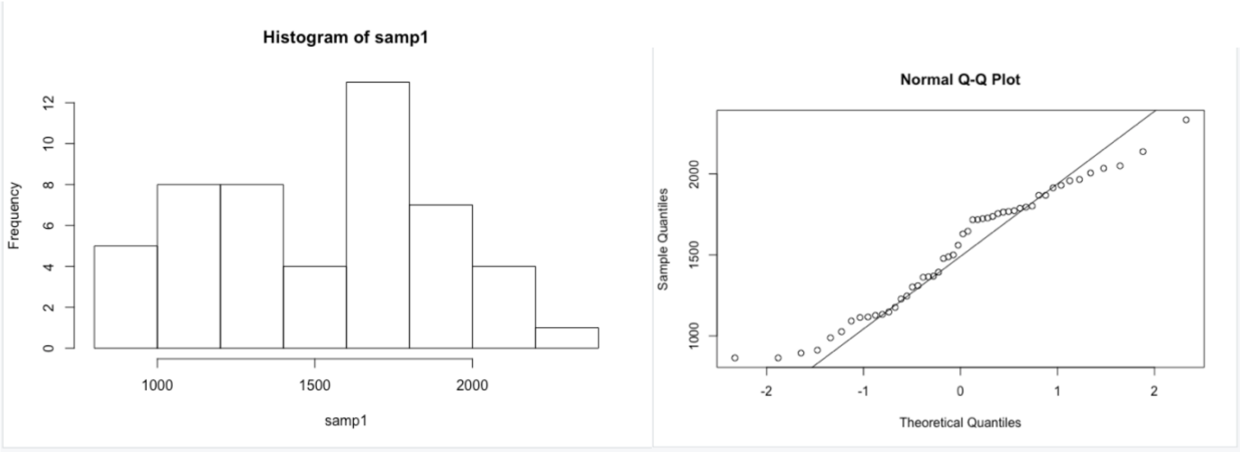


```
set.seed(234567899)
samp1 = sample(area, 50)
summary(samp1)
hist(samp1)
mean(samp1)
```

→ →



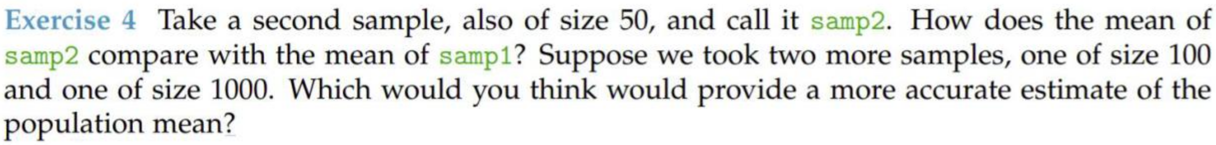
```
> summary(samp1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    864    1189    1595    1529    1793    2334 
> hist(samp1)
> mean(samp1)
[1] 1529.38
```



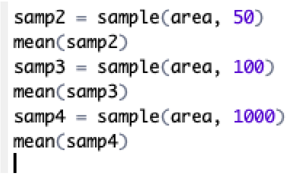
Answer: The sample distribution is also right skewed. It has a big irregular tail compared to distribution of population.

Exercise 4

Question:



Exercise 4 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?

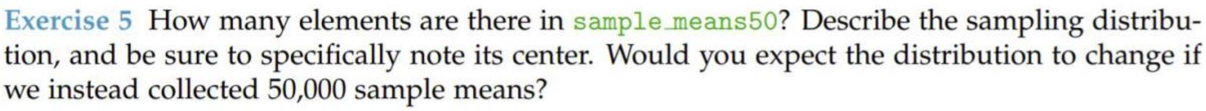


```
samp2 = sample(area, 50)
mean(samp2)
samp3 = sample(area, 100)
mean(samp3)
samp4 = sample(area, 1000)
mean(samp4)
```

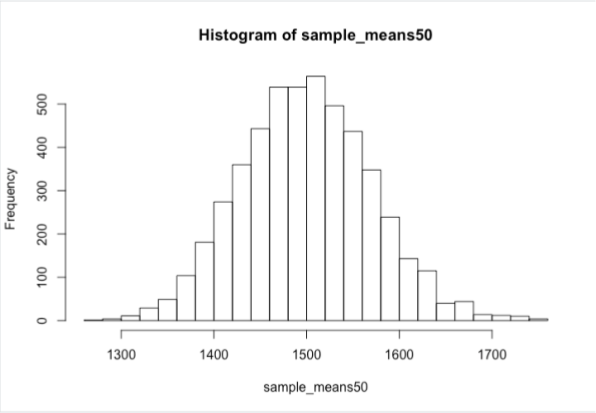
Answer: Mean of sample 2 is 1475.42 which is much closer to population mean compared to mean of samp1(which is 1431.62). The mean of sample size 1000 provides more accurate estimate of the population mean.

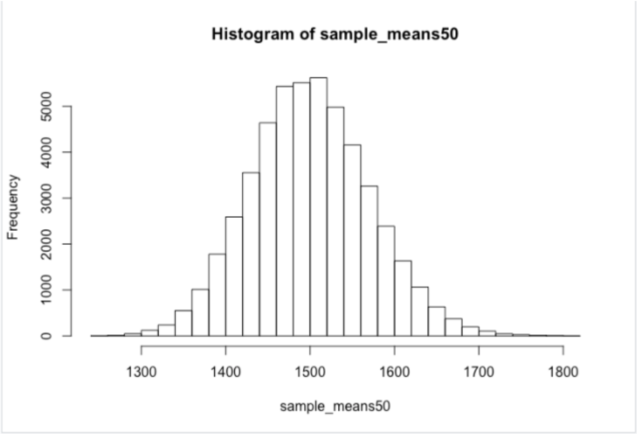
Exercise 5

Question:



Exercise 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 collected 50,000 sample means?

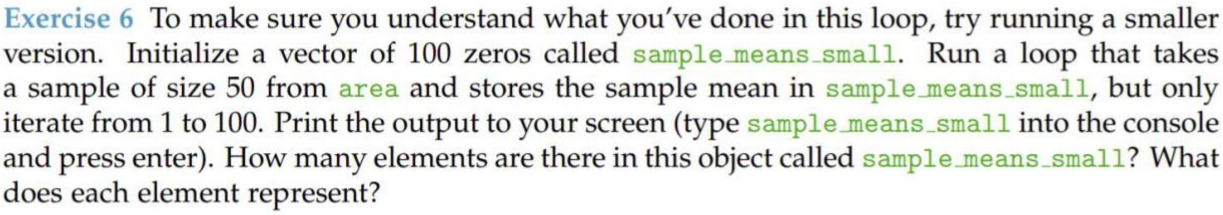




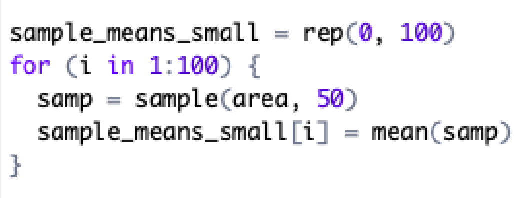
Answer: There are 5000 elements in the sample_means50. The sampling distribution of 50000 samples looks almost the same as distribution of mean of 5000. However, the 50000 mean distribution is closer to the theoretical normal distribution.

Exercise 6

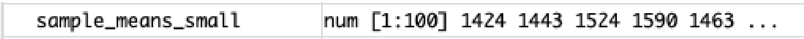
Question:



Exercise 6 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)
}
```

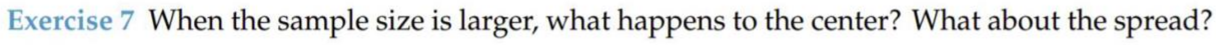


```
sample_means_small      num [1:100] 1424 1443 1524 1590 1463 ...
```

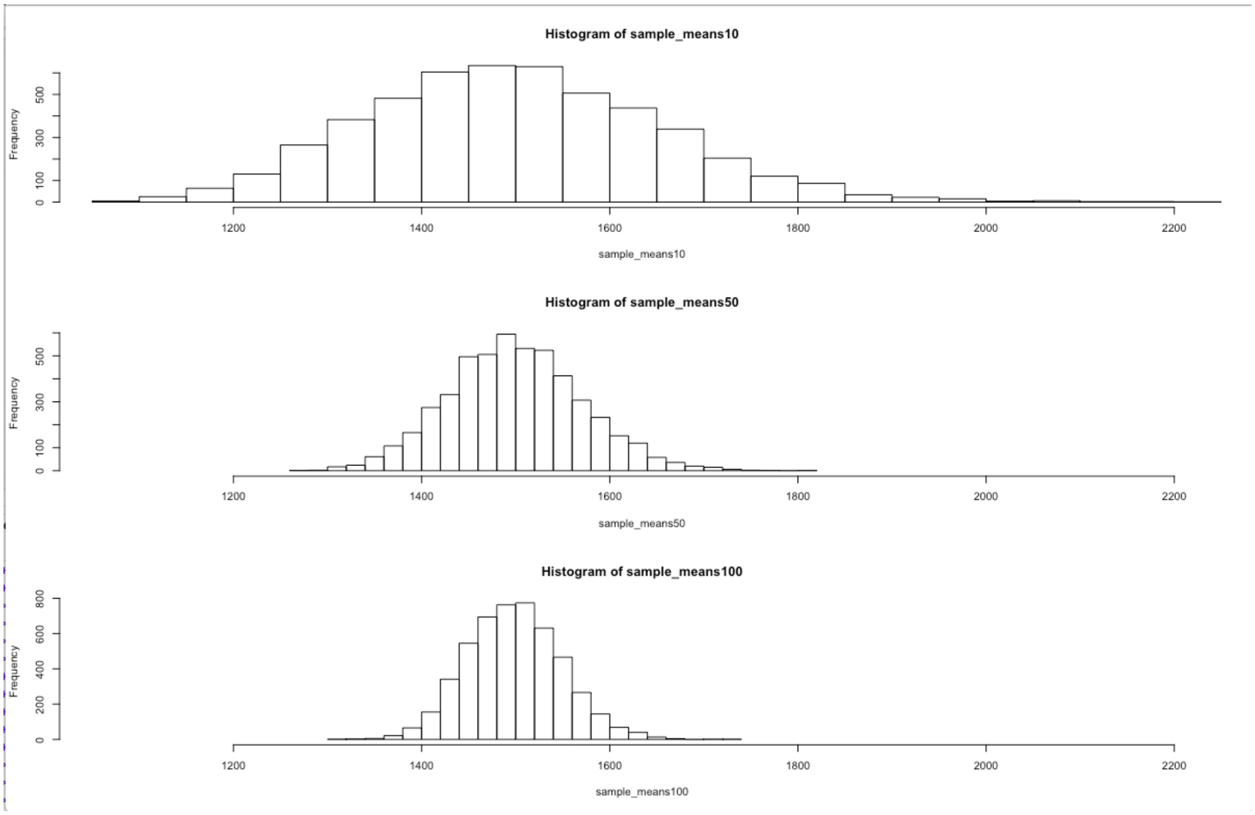
Answer: There are 100 elements in the object and each of them represents a sample mean.

Exercise 7

Question:



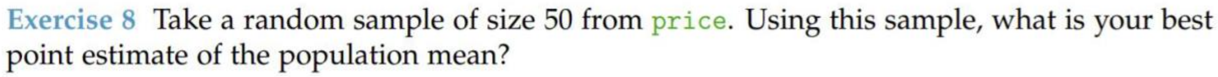
Exercise 7 When the sample size is larger, what happens to the center? What about the spread?



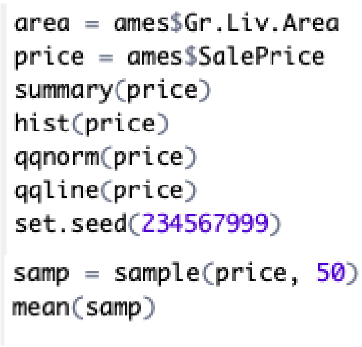
Answer: The center is closer to population mean and there are more observations closer to population mean as the sample size gets larger. The spread of the distribution gets smaller as the sample size increases.

Exercise 8

Question:

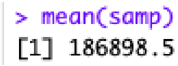


Exercise 8 Take a random sample of size 50 from `price`. Using this sample, what is your best point estimate of the population mean?



```
area = ames$Gr.Liv.Area
price = ames$SalePrice
summary(price)
hist(price)
qqnorm(price)
qqline(price)
set.seed(234567999)

samp = sample(price, 50)
mean(samp)
```

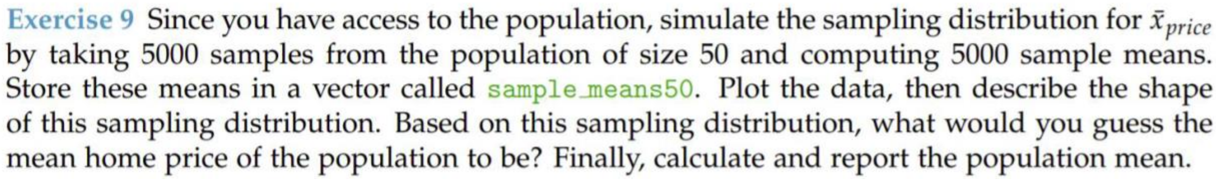


```
> mean(samp)
[1] 186898.5
```

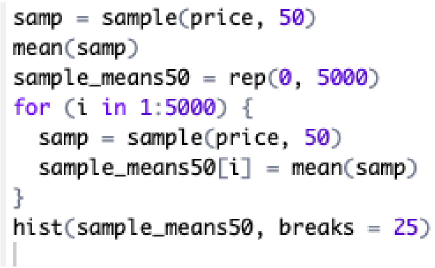
Answer: The best point estimate of the population mean from sample mean is 186898.5

Exercise 9

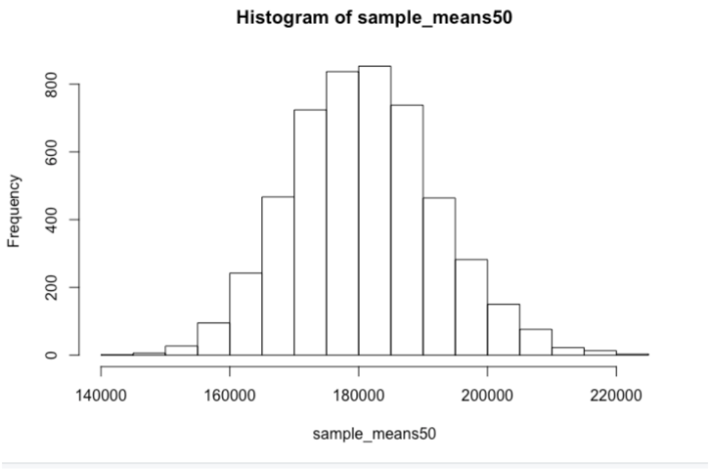
Question:

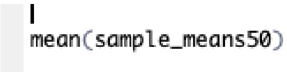


Exercise 9 Since you have access to the population, simulate the sampling distribution for $\bar{x}_{price}$ by taking 5000 samples from the population of size 50 and computing 5000 sample means. Store these means in a vector called `sample_means50`. Plot the data, then describe the shape of this sampling 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.

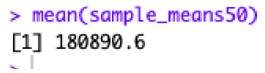


```
samp = sample(price, 50)
mean(samp)
sample_means50 = rep(0, 5000)
for (i in 1:5000) {
  samp = sample(price, 50)
  sample_means50[i] = mean(samp)
}
hist(sample_means50, breaks = 25)
```





```
mean(sample_means50)
```

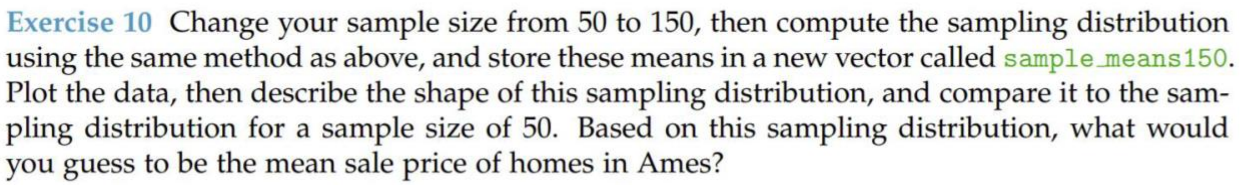


```
> mean(sample_means50)
[1] 180890.6
```

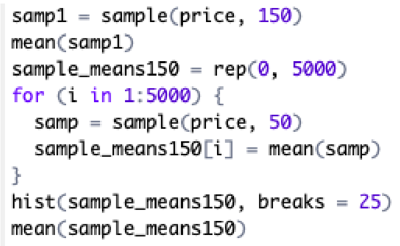
Answer: The distribution of sample mean is inverted bell shaped and is similar to normal distribution. The guess of the mean home price from sample is 180890.9. Actual population is 180890.6

Exercise 10

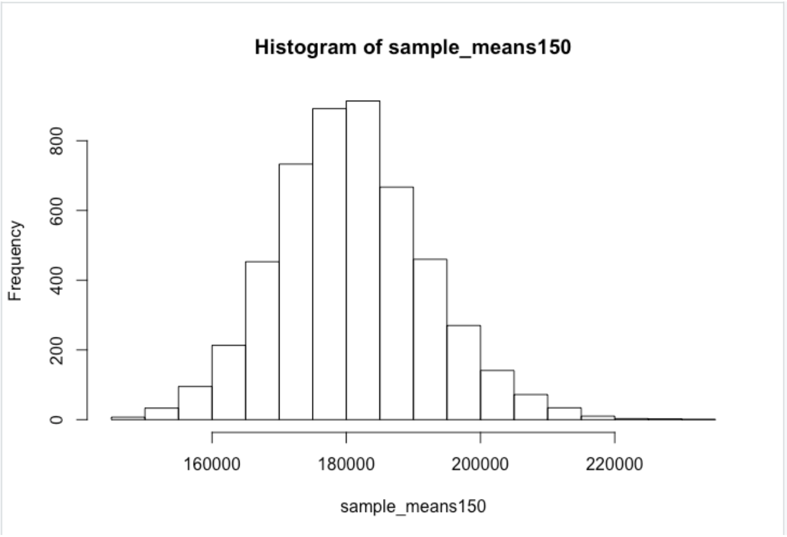
Question:

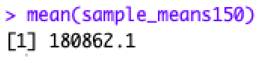


Exercise 10 Change your sample size from 50 to 150, then compute the sampling distribution using the same method as above, and store these means in a new vector called sample_means150. Plot the data, then describe the shape of this sampling distribution, and compare it to the sampling distribution for a sample size of 50. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?



```
samp1 = sample(price, 150)
mean(samp1)
sample_means150 = rep(0, 5000)
for (i in 1:5000) {
  samp = sample(price, 50)
  sample_means150[i] = mean(samp)
}
hist(sample_means150, breaks = 25)
mean(sample_means150)
```



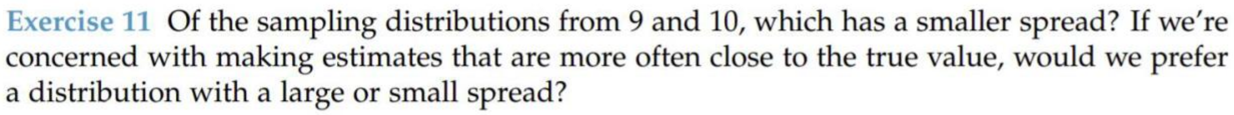


```
> mean(sample_means150)
[1] 180862.1
```

Answer: The shape of the distribution of sample mean of sample size is inverted bell shaped and resembles normal distribution normal distribution. There are more observations near center compared to the distribution of means of the sample size 50. From the sample we could guess the mean sale price to be 180885.6

Exercise 11

Question:



Exercise 11 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?

Answer: The distribution of means of sample size 150 has lesser spread compared to sample size 50. We would prefer a distribution with small spread, if we are interested in making estimate that are closer to the true value