Preyasi Gaur

705704939

Stats 10

Lab 4

**Exercise 1**

1. Question:  
   Place the flags 1 standard deviation on either side of the mean. What is the area between these two values? What does the empirical rule say this area is?

Answer:

Placing the flags 1 standard deviation away from the mean, we see that the area between the two values is approximately 0.6822. From the empirical rule, we know that the area should be 68% of the data.

1. Question:  
   Repeat for 2 and 3 standard deviations on either side of the mean. Again compare the empirical rule with the area given in the applet.   
     
   Answer:   
   Placing the flags 2 standard deviations on either side of the mean, we see that the area between the two values is approx 0.9544. From the empirical rule, we know that the area should be 95% of the data. Placing the flags 3 standard deviations on either side of the mean, we see that the area between the two values is approx 0.9973. From the empirical rule, we know that the area should be 99.7% of the data.

**Exercise 2**

Question:  
Using the applet, how many standard deviations above and below the mean do the quartiles of any normal distribution lie? Use the closest available values (the applet can't hit every value exactly).

Answer:  
Using the applet we see that,

Q1 = -0.671

Q3 = 0.671

Thus, the quartiles of any Normal distribution are about 0.671 standard deviations above and below the mean.

**Exercise 3**

Question:

Adult male height (X) follows (approximately) a normal distribution with a mean of 69 inches and a standard deviation of 2.8 inches.

1. Question:

What proportion of males are less than 65 inches tall? In other words, what is P(X < 65)?   
  
Answer:   
The proportion of males that are less than 65 inches tall are:  
P(X < 65) = pnorm(65, mean = 69, sd = 2.8)

1. Question:

What proportion of males are more than 75 inches tall? In other words, what is P(X > 75)?   
  
Answer:   
The proportion of males that are more than 75 inches tall are:  
P(X > 75) = 1 - pnorm(75, mean = 69, sd = 2.8)

1. Question:

What proportion of males are between 66 and 72 inches tall? In other words, what is P(66 < X < 72)?   
  
Answer:   
The proportion of males that are between 66 and 72 inches tall are:  
P ( 66 < X < 72) = P(X<72) - P(X<66) = pnorm(72, mean = 69, sd = 2.8) - pnorm(65, mean = 69, sd = 2.8)

**Exercise 4**

Question:

Suppose adult male height follows a normal distribution with a mean of 69 inches and a standard deviation of 2.8 inches.

1. Question:   
   How tall must a male be in order to be among the shortest 0.5% of males?

Answer:  
To find how tall a male must be in order to be among the shortest 0.5% of males we use the formula, qnorm(0.005, mean = 69, sd = 2.8). Thus, male must be shorter than 61.79 inches

1. Question:  
   How tall must a male be in order to be among the tallest 0.25% of males?

Answer:   
To find how tall a male must be in order to be among the tallest 0.25% of males we use the formula, qnorm(1 - 0.0025, mean = 69, sd = 2.8). Thus, male must be at least 76.85inches

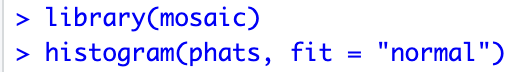
**Exercise 5**

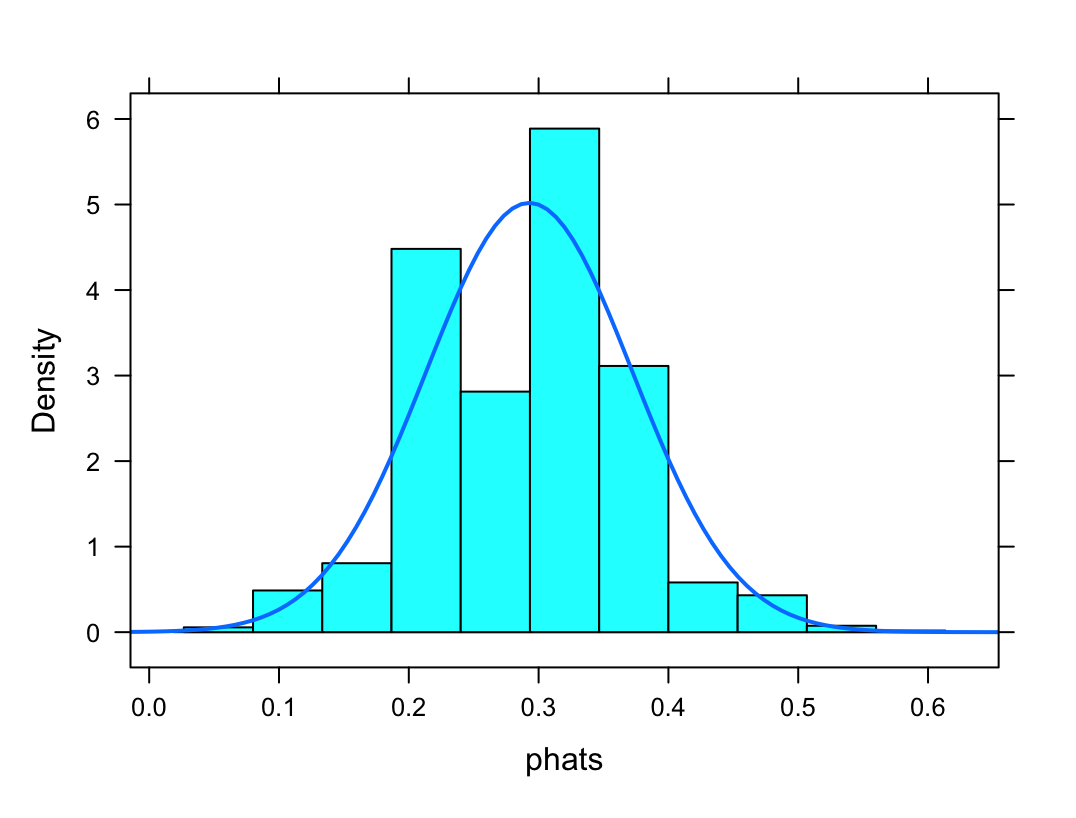
pawnee <- read.csv("/Users/preyasigaur/Desktop/pawnee.csv")  
# We first create objects for common quantities we will use for this exercise.  
 n <- 30 # The sample size  
N <- 541 # The population size  
M <- 1000 # Number of samples/repetitions

# Create vectors to store the simulated proportions from each repetition. phats <- numeric(M) # for sample proportions  
# Set the seed for reproduceability  
set.seed(123)

# Always set the seed OUTSIDE the for loop.  
# Now we start the loop. Let i cycle over the numbers 1 to 1000 (i.e., iterate 1000 times). for(i in seq\_len(M)){

# The i-th iteration of the for loop represents a single repetition.  
 # Take a simple random sample of size n from the population of size N.  
 index <- sample(N, size = n)  
 # Save the random sample in the sample\_i vector.  
 sample\_i <- pawnee[index, ]  
 # Compute the proportion of the i-th sample of households with a new health issue. phats[i] <- mean(sample\_i$New\_hlth\_issue == "Y")  
}

1. Question:  
   Run the entire chunk of code in the previous page to run a for loop that creates a vector of sample proportions. Using the results, create a relative frequency histogram of the sampling distribution of sample proportions.   
   Superimpose a normal curve to your histogram with following instructions:   
      
    • If you use the histogram() function from the mosaic package, add the argument: fit = "normal".   
    • If you use the hist() function from base R, add the argument: prob = TRUE, then run the command: curve(dnorm(x, mean(phats), sd(phats)), add = TRUE).   
     
   Prompt Code:  
   

Output:   


1. Question:   
   What is the mean and standard deviation of the simulated sample proportions?

Prompt code:

> mean (phats)  
> sd(phats)  
  
Output:   
Mean:    
Standard Deviation: 

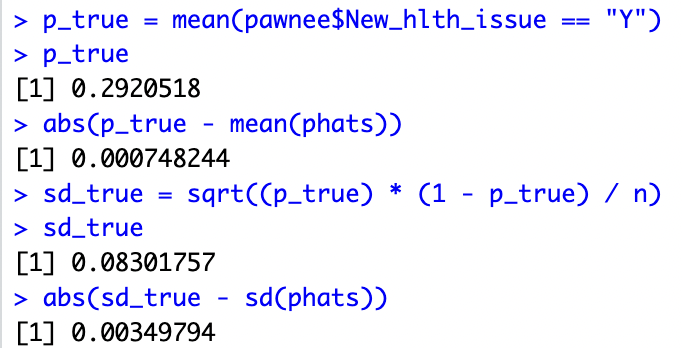
1. Question:   
   Do you think the simulated distribution of sample proportions is approximately normal? Explain why or why not.   
     
   Answer:

Yes, we see that the histogram is approximately normal. We also notice that the normal curve is roughly bell-shaped, and the histogram is unimodal and symmetric around the center.

1. Question:   
   Using the theory-based method (i.e., normal approximation by invoking the Central Limit Theorem), what would you predict the mean and standard deviation of the sampling distribution of sample proportions to be? How close are these predictions to your answers from Part B?   
     
   Prompt Code:  
   > p\_true = mean(pawnee$New\_hlth\_issue == "Y")  
   > p\_true  
   > abs(p\_true - mean(phats))  
     
   > sd\_true = sqrt((p\_true) \* (1 - p\_true) / n)

> sd\_true

> abs(sd\_true - sd(phats))

Output:   


**Exercise 6**

1. Question:

Create a new for loop to create a vector of sample means of the arsenic levels. Use n = 30, N = 541, and M =1000 just like before, and set the seed to 123.   
  
Prompt Code:  
n <- 30 # The sample size

N <- 541 # The population size

M <- 1000 # Number of samples/repetitions

xbars <- numeric(M)

set.seed(123)

for(i in seq\_len(M)){

index <- sample(N, size = n)

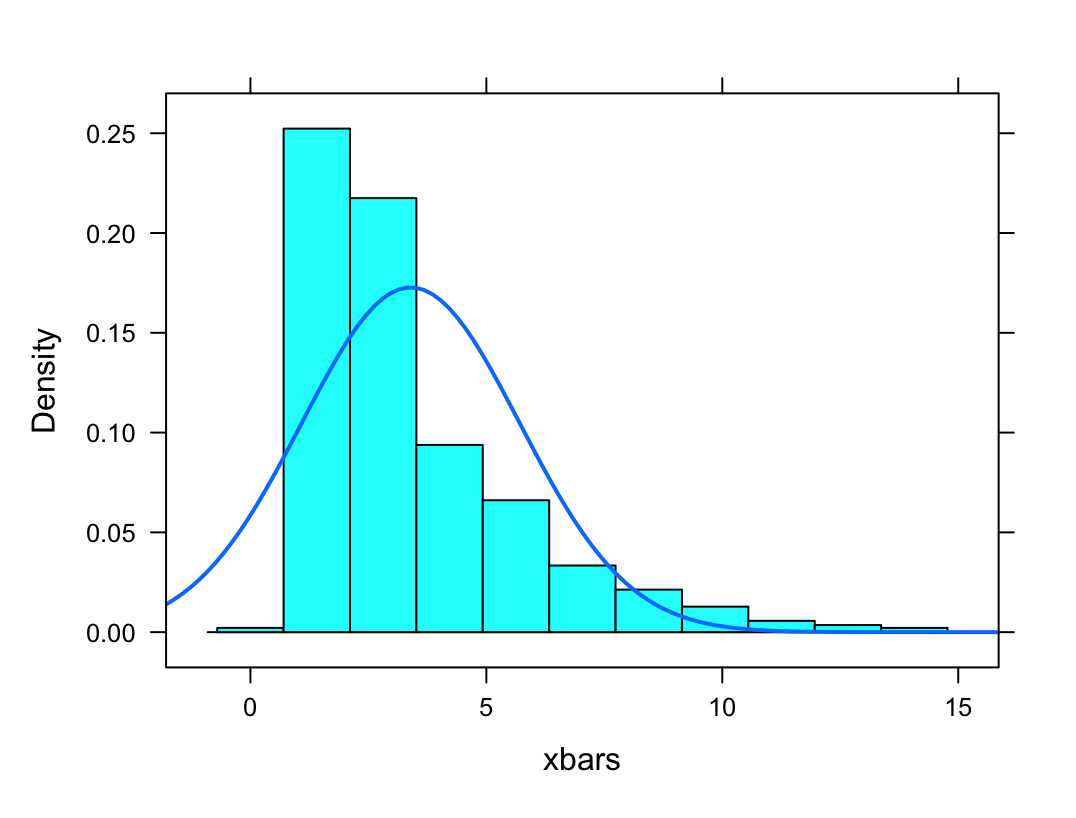
sample\_i <- pawnee[index, ]

xbars[i] <- mean(sample\_i$Arsenic)

}

1. Question:   
   Create a relative frequency histogram of the sampling distribution of sample means for arsenic. Superimpose a normal curve by following the instructions given in 2(a).   
     
   Prompt Code:   
   > library(mosaic)

> histogram(xbars, fit = "normal")

Output:   


1. Question:  
   Do you think the simulated distribution of sample arsenic means is approximately normal? Explain why or why not. If your answer was different from your answer to Exercise 2(c), why do you think this is the case?   
     
   Answer:  
   The histogram that we get is not approximately normal. Clearly, we see that it is neither unimodal nor symmetric. We see that the left tail is missing as well.  
   There is discrepancy between this histogram and the one obtained in 2(c). The reason for that is because the sample size was just 30, which is too small. Also, over half the pawnee data has Arsenic level = 0. Moreover, there are several extreme outliers, especially the five data points around/above 100.