Lab 3

2020-11-14

# Question 1

set.seed(137)  
  
sample\_index <- sample(2919, 30)  
sample\_0 <- college[sample\_index, ]  
head(sample\_0, 3)

OnCampus CumGpa  
1595 N 3.81  
2466 Y 3.75  
2568 Y 3.53

# Question 2

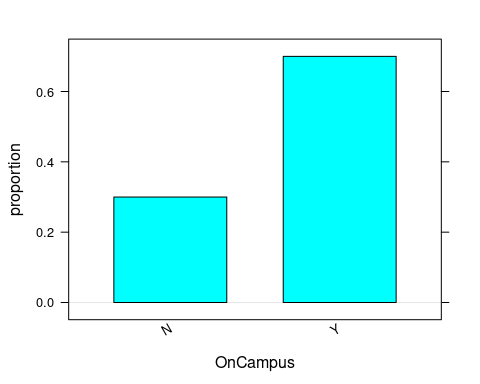
population\_ratio <- mean(college$OnCampus == "Y")  
sample\_ratio <- mean(sample\_0$OnCampus == "Y")  
population\_ratio;sample\_ratio

[1] 0.7759507

[1] 0.7

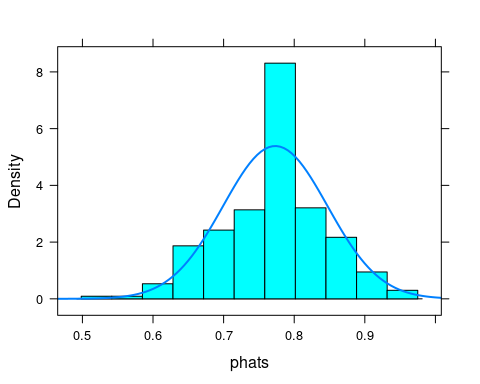
# Question 3

bargraph(~ OnCampus,   
 data = sample\_0,   
 type = "proportion")



# Question 4

n <- 30  
N <- 2919  
M <- 1000  
  
phats <- numeric(M)  
set.seed(9001)  
  
for(i in 1:M){  
 index <- sample(N, n)  
 sample\_i <- college[index, ]  
 phats[i] <- mean(sample\_i$OnCampus == "Y")  
}  
  
histogram(phats, fit = "normal")



# Question 5

sample\_prop\_mean <- mean(phats)  
sample\_prop\_se <- sd(phats)/sqrt(M-1)  
  
sample\_prop\_mean;sample\_prop\_se

[1] 0.7732

[1] 0.002344135

# Question 6

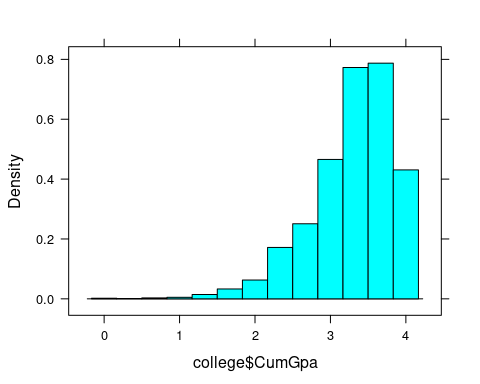
up\_lim <- sample\_prop\_mean + sample\_prop\_se \* 1.96  
down\_lim <- sample\_prop\_mean - sample\_prop\_se \* 1.96  
  
c(down\_lim, up\_lim)

[1] 0.7686055 0.7777945

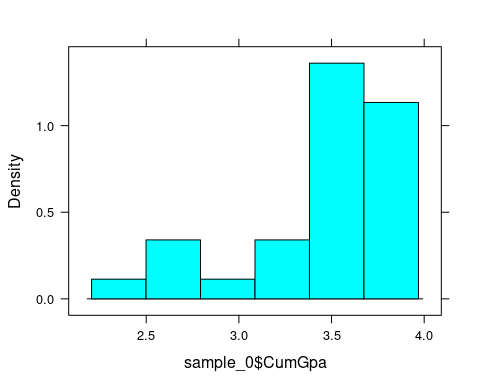
Answer.  
Yes, the population parameter is among the 95% confidence interval.

# Question 7

histogram(college$CumGpa)



histogram(sample\_0$CumGpa)



Answer.  
As the CumGpa is an quantitative variable, it’s distribution can be inspected by using histogram rather than bargraph.

The population distribution of the cumulative GPA is an positive skewed distribution. The distribution of sample cumulative GAP largely captured the distribution shape of population, as the smaller sample size, variable exist, such as the distribution range is smaller.

# Question 8

population\_gpa\_mean <- mean(college$CumGpa)  
sample\_gpa\_mean <- mean(sample\_0$CumGpa)  
  
population\_gpa\_mean; sample\_gpa\_mean

[1] 3.287955

[1] 3.428333

# Question 9

population\_gpa\_sd <- sd(college$CumGpa)  
sample\_gpa\_sd <- sd(sample\_0$CumGpa)  
  
population\_gpa\_sd; sample\_gpa\_sd

[1] 0.550747

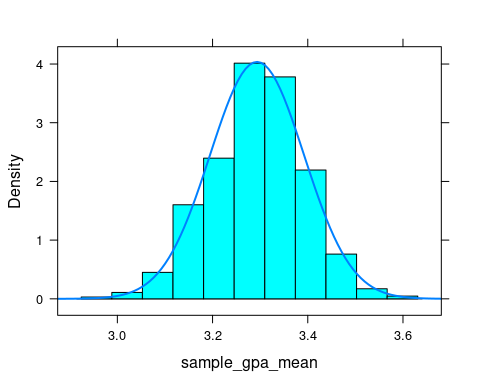
[1] 0.3893769

# Question 10

set.seed(1234)  
  
n <- 30 # the size of sample  
N <- 2919 # the size of population  
M <- 1000 # round of samples  
sample\_gpa\_mean <- numeric(M)  
  
for(i in 1:M){  
 sample\_index <- sample(N, n)  
 sample\_i <- mean(college[sample\_index, "CumGpa", drop = T])  
 sample\_gpa\_mean[i] <- sample\_i  
}

# Question 11

histogram(sample\_gpa\_mean,   
 fit = "normal",   
 density = T)



# Question 12

gpa\_mean\_sample <- mean(sample\_gpa\_mean)  
gpa\_se\_sample <- sd(sample\_gpa\_mean)/sqrt(M - 1)

# Question 13

up\_lim <- gpa\_mean\_sample + gpa\_se\_sample \* 1.96  
down\_lim <- gpa\_mean\_sample - gpa\_se\_sample \* 1.96  
  
c(down\_lim, up\_lim)

[1] 3.287256 3.299522

Answer.  
Yes, the 95% confidence interval contains the population parameter.

# Question 14

Answer.  
According to the Central Limit Theorem, the mean of the sample distribution of sample proportions should be very close to the population proportion. And the standard deviation of the standard deviation of sample proportion should be smaller than the population standard deviation of proportion.

The mean of the sample distribution of sample proportion is very close to the population mean in Questiuon 5, and the standard deviation of sample proportion is about 3 times smaller than the population standard deviation in Question 5.

# Question 15

Answer.  
The distribution of sample mean and sample standard deviation should be narrow than the population mean and standard deviation.

# Question 16

Answer.  
Yes, the sample distribution of sample proportions is approximately normal.

# Question 17

Answer.  
Yes, the sample distribution of sample means is approximately normal.

# Question 18

cat("lower boundary:")

lower boundary:

sample\_prop\_mean - sample\_prop\_se \* 1.96

[1] 0.7686055

cat("upper boundary:")

upper boundary:

sample\_prop\_mean + sample\_prop\_se \* 1.96

[1] 0.7777945

Answer.  
1. If we select multiple groups of students from the college population, each group with 30 students, the proportion of live on campus percentage between 76.86% and 77.78% with overall about 95% correct ratio.

1. Yes, the interval contains the population parameter.

# Question 19

cat("lower boundary:")

lower boundary:

gpa\_mean\_sample - gpa\_se\_sample \* 1.96

[1] 3.287256

cat("upper boundary:")

upper boundary:

gpa\_mean\_sample + gpa\_se\_sample \* 1.96

[1] 3.299522

Answer.  
Randomly select 30 students from the college population, we can assert that their mean CumGpa is between 3.282803 and 3.295301 with 95% percent correct ratio.