Lesson 2 HW Submission

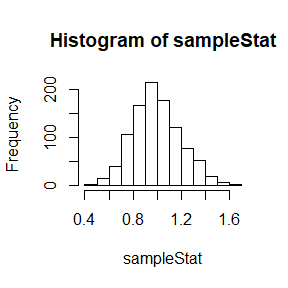
Josh Jarvey

05/25/2020

## Exercise 1

The following block of code will produce a simulated sampling distribution of the sample mean for 1,000 samples of size 23 drawn from an exponential distribution and then make a histogram of the results. Some R programmers insist that for loops should be avoided because they are slow, but we’re aiming for transparency at the moment, not efficiency.

# r defaults to a 7 by 7 figure (units?), use fig.width and fig.height to adjust  
reps <- 1000  
n <- 23  
sampleStat <- numeric(reps) # initialize the vector  
for (i in 1:reps){  
 sampleStat[i] <- mean( rexp(n) )  
}  
hist(sampleStat)



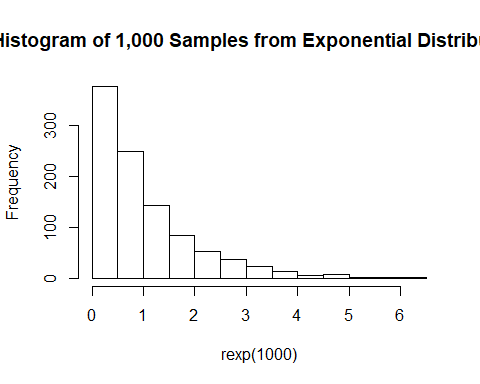
You are going to demonstrate the Central Limit Theorem, and gain some practice with loops in R, by showing that distribution of the sample means becomes increasingly normal as the sample size increases.

### Part 1a

First, draw a random sample of 1000 observations from the exponential distribution and make a histogram to illustrate just how skewed the exponential distribution is. You shouldn’t need a for loop or mean() to do this bit. (You’re not taking means of anything and you don’t need a loop. Recall that rnorm(500) would generate 500 observations from a standard normal distribution.)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

#generate 1000 samples from exponential distribution. Wrapped in hist() to display it as a histogram.  
hist(rexp(1000), main = "Histogram of 1,000 Samples from Exponential Distribution")

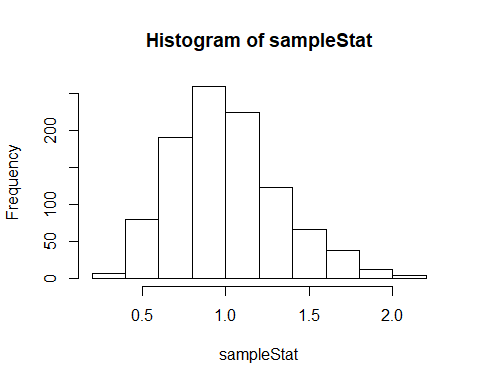


### Part 1b

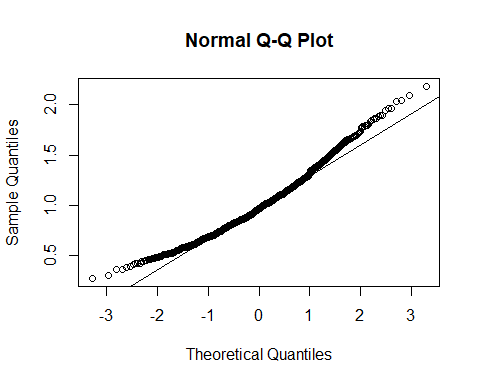
Copy the block of code at the top and use it to simulate the sampling distribution of sample means for 1000 samples of size 10. After the histogram, use qqnorm() to make a normal probability plot of sampleStat. Add a fit line to the plot with qqline(). Do the sample means appear to be normally distributed? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

#setting 1000 reps of sample size 10.  
reps <- 1000  
n <- 10  
  
 #using the for-loop to calculate the mean of the current rep and store it into a vector  
sampleStat <- numeric(reps) # initialize the vector  
for (i in 1:reps){  
 sampleStat[i] <- mean( rexp(n) )  
}  
hist(sampleStat)



#plotting a QQ plot to access normality.   
qqnorm(sampleStat)  
qqline(sampleStat)



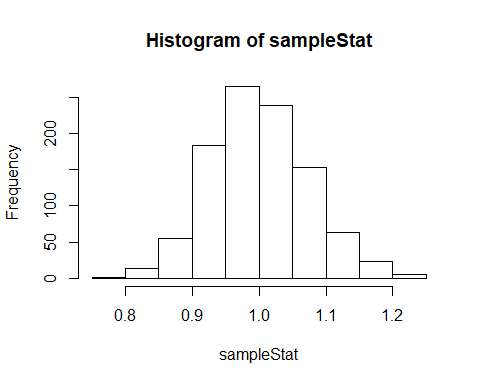
Based on the QQPlot, the data does appear to be skewed to the right (lower & upper tail above the line).

### Part 1c

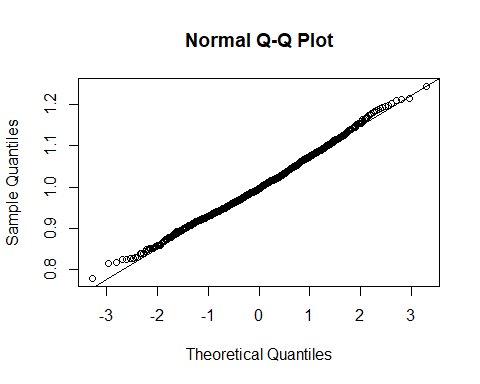
Repeat 1b for samples of size 200. Do the sample means appear to closer to normally distributed than they were for n = 10? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

#setting 1000 reps of sample size 200.  
reps <- 1000  
n <- 200  
  
 #using the for-loop to calculate the mean of the current rep and store it into a vector  
sampleStat <- numeric(reps) # initialize the vector  
for (i in 1:reps){  
 sampleStat[i] <- mean( rexp(n) )  
}  
hist(sampleStat)



#plotting a QQ plot to access normality.   
qqnorm(sampleStat)  
qqline(sampleStat)



Based on the QQPlot, when the sample size is greater (200 in this case), the data is a much better fit to the theoritical normal as its more closely aligned with the theoritical normal line.

## Exercise 2

This problem is modeled loosely after problem 5.63 on page 297 of Ott.

### Part 2a

Using the data we can determine that the 95% -CI for is about (4.8,5.2) with margin of error 0.2. For large samples and . Use the formula on page 241 to estimate the sample size required to get margin of error . Always round up for sample size. Read Section 5.3 in Ott if you need to review this material.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

#first, find the t-value using the qnorm at 97.5%  
 #second, using the formula n = t^2\*s^2/E^2, we find the sample size required is 784.  
ceiling(qnorm(.975)^2)\*(0.7^2)/(0.05^2)

## [1] 784

### Part 2b

Suppose you now have a sample with size as determined in 2a that yields and   
Use R to build a fake data set with exactly the same statistics (as shown in the swirl lesson T Confidence Intervals in the UW\_Stat\_Methods course or consider the command scale() in R). The idea is to create a sample with exactly the right statistics so that we can use R functions to perform the analysis. Now apply t.test to your constructed sample to find the 95% CI for the population mean. (Note: rnorm(50, mean = 5.03, sd = 0.68) is not right as it produces a sample that has close to, but not exactly the right statistics … try it.)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

#generate 50 observations from the normal distribution using mean 0, sd 1.  
fakeData = rnorm(50,0,1)  
 #"standardize" the observations using the z-score formula: (obs - mean(dataset))/sd(dataset).  
 #the scale() function does this automatically, just provide it the vector of observations.   
fakeData = scale(fakeData)  
 #now, rescale and shift the data to match the dataset described in the paragraph above.   
fakeData = fakeData\*0.68 + 5.03  
  
 #finally, perform the t.test and pull out the 95% confidence interval. 4.8 to 5.2, which matches the above in part a.  
t.test(fakeData, conf.level = .95)$conf.int

## [1] 4.836746 5.223254  
## attr(,"conf.level")  
## [1] 0.95

## Exercise 3

For this exercise and the rest of the exercises in this homework set you are going to use the data from problem 5.27 on page 289 of Ott. This data is saved in the file ‘ex5-29.TXT’ that you downloaded along with this submission file. You can load the data like this (assumes data file is the same directory as this Rmd file)

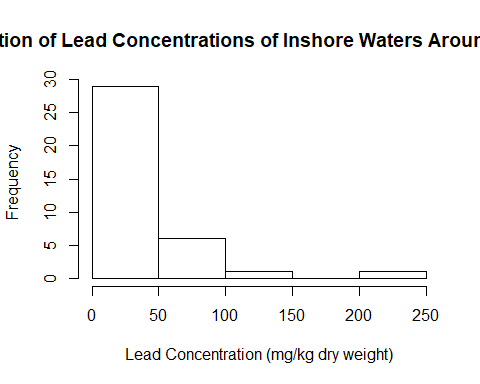
# load csv into data frame  
df <- read.csv('ex5-29.TXT')  
# put data for lead concentrations in vector called "lead"  
lead <- df$Lead   
# delete the data frame since we no longer need it  
rm(df)

### Part 3a

Before using a t distribution based procedure we need to determine if it is plausible that the data is sampled from a normally distributed random variable. Make a histogram of the data. Based on the histogram is it reasonable to say that lead concentrations are normally distributed?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3a -|-|-|-|-|-|-|-|-|-|-|-

hist(lead, main = "Distribution of Lead Concentrations of Inshore Waters Around Mombasa", xlab = "Lead Concentration (mg/kg dry weight)")



Based on the histogram, the data does not appear to be normally distributed.

### Part 3b

In spite of the fact that the lead concentration sample is clearly not from a mound-shaped distribution, we’ll apply based procedures to start so we can compare them to a bootstrap approach a little later. Use t.test in R to construct a 80% CI for the population mean lead concentration. Write a sentence interpreting the result. (80% is a low confidence level, but for this problem were mostly interested in the lower bound so we’re 90% confident that the pop mean is greater than the lower bound.)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3b -|-|-|-|-|-|-|-|-|-|-|-

t.test(lead, conf.level = 0.80)$conf.int

## [1] 29.27551 45.21098  
## attr(,"conf.level")  
## [1] 0.8

At an 80% significance level, we expect the true population mean of lead concentration levels of inshore waters around Mombasa to fall within 29.28 mg/kg - 45.21 mg/kg (dry weight).

### Part 3c

Note that your 80% CI for the population mean also gives you a 90% lower confidence bound for the population mean concentration. Does this lower confidence bound suggest that the population mean lead concentration is larger than 30 mg/kg at the significance level? Explain your answer (think of the equivalence of confidence intervals and hypothesis tests).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3c -|-|-|-|-|-|-|-|-|-|-|-

It does not suggest that the true population mean is larger than 30 mg/kg, since the lower bounds of the confidence interval is below 30, it could take on any value in the range of 29.28 to 45.21 kg/mg.

### Part 3d

Use R to conduct a one-sample -test to determine if the population mean lead concentration is larger than 30 mg/kg at the significance level. Fill in all the steps in hypothesis test. Use conf.level = 0.9 in the t.test if you want it to also produce the one-sided confidence bound you saw above.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3d -|-|-|-|-|-|-|-|-|-|-|-

(Step 1) The hypotheses and (change!)

(Step 2) Already did this in 3a.

(Step 3) Compute:

#perform a t.test() to see if population mean is greater than 30, at an alpha of 0.10.  
t.test(lead, mu = 30, alternative = "greater", conf.level = 0.9)

##   
## One Sample t-test  
##   
## data: lead  
## t = 1.1868, df = 36, p-value = 0.1215  
## alternative hypothesis: true mean is greater than 30  
## 90 percent confidence interval:  
## 29.27551 Inf  
## sample estimates:  
## mean of x   
## 37.24324

(Step 4) Conclusion: At a 90% significance level, there is not enough evidence to reject that the true population mean of lead concentration of inshore water around Mombasa is greater than 30 mg/kg (p-value=0.1215).

### Part 3e

Since the lead concentrations are very skewed and the sample size is relatively small one should suspect the validity of the based procedures above. Follow the code that uses replicate() in the presentations to make an 80% percentile boostrap CI for the population mean lead concentration. Use 5000 resamples. Don’t forget to adjust the confidence level.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3e -|-|-|-|-|-|-|-|-|-|-|-

#set number of bootstraps to generate  
B = 5000  
 #sampling with replacement from the lead data set. Replicate this B number of times (5000). This is my resampling dataset.  
rs = replicate(B, sample(lead,replace = TRUE))  
 #apply the mean() function to the columns in the resampled matrix.   
bd = apply(rs,2,mean)  
 #pull the quantile values from the 10th and 90th percentile, thus covering 80% of the range. This is my confidence interval.  
quantile(bd, c(0.10, 0.90))

## 10% 90%   
## 29.86486 45.16216

### Part 3f

Use the boot package to make an 80% percentile BCa CI for the population mean lead concentration. Use 5000 resamples. Make sure to install the boot package in the R console one time using install.packages('boot') and then library(boot) or require(boot) to the top of your code below. Don’t forget to adjust the confidence level.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3f -|-|-|-|-|-|-|-|-|-|-|-

#load the necessary boot library  
library(boot)

## Warning: package 'boot' was built under R version 3.6.3

#create the auxillary function that returns the mean of the passed in replicated sample.  
bootMean <- function(x, i){  
 return(mean(x[i]))  
}  
  
 #complete 5000 sampling distributions of the mean statistic from the lead data set.   
bootObject = boot(lead,bootMean,R=5000)  
 #create an 80% confidence interval, of type BCa (bias corrected & accelerated) from the bootstrapped sampling distribution.  
boot.ci(bootObject,conf = 0.80,type = 'bca')

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 5000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = bootObject, conf = 0.8, type = "bca")  
##   
## Intervals :   
## Level BCa   
## 80% (31.14, 47.81 )   
## Calculations and Intervals on Original Scale

### Part 3g

Based on your BCa interval in Part 3f, what is the the lower 90% confidence bound for the population mean lead concentration?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3g -|-|-|-|-|-|-|-|-|-|-|-

The lower 90% bound (one-sided) for the population mean lead concentration is 31.19 mg/kg (dry weight). It is the same thing as the two-sided 80%’s lower bound (since 10% discard is in the lower tail).

### Part 3h

We’d like to do a hypothesis test at the 10% significance level () to determine if the population mean lead concentration is greater that 30 mg/kg. Using your 90% lower confidence bound from Part 3g, what conclusion would you reach for the hypothesis test? Write a complete conclusion as you would for any hypothesis test.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3h -|-|-|-|-|-|-|-|-|-|-|-

At a 90% significance level, there is enough evidence to support that the true population mean of lead concentration of inshore water around Mombasa is greater than 30 mg/kg (dry weight).

### Part 3i

Which do you trust more and why? The results of the bootstrap procedures or the result of the theory based distribution procedures? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3i -|-|-|-|-|-|-|-|-|-|-|-

## I trust the bootstrapped version a bit more because the original data set was not normally distributed, therefore in using a t.test() we violate the normality assumption. Bootstrapping the data provides a more normal sampling distribution and statistic (mean), and therefore a stronger measure of the true population parameter.

### Part 3j

Maybe we shouldn’t trust t.test() here, but for practice we’ll do some power calculations in 3j and 3k.

Suppose it would be worthwhile to be able to detect an alternative mean lead concentration of mg/kg when testing . If using a sample of size 37 with an estimated standard deviation of 37.1 and , use R to approximate the power of the -test in this situation.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3j -|-|-|-|-|-|-|-|-|-|-|-

#calculate the power of the t.test(), using a significance at 0.10.   
 #so about 63% chance that population mean is 40 mg/kg.  
power.t.test(n = 37, sd = 37.1, delta = 10, sig.level = 0.10,  
 type = "one.sample",  
 alternative = "one.sided")$power

## [1] 0.632762

### Part 3k

Suppose we need the power to be 0.9 to be able to detect an alternative mean lead concentration of mg/kg when testing . Again, with estimated standard deviation of 37.1 and , what sample size is required for power = 0.9?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3k -|-|-|-|-|-|-|-|-|-|-|-

#this time placing in a power, but leaving sample size blank.  
 #if we want a 90% chance that population mean is 40 mg/kg, then we need a sample size of 92 to confirm this delta  
 #using ceiling since you always want the nearest integer rounded up for sample size.   
ceiling(power.t.test(power = 0.90, sd = 37.1, delta = 10, sig.level = 0.10,  
 type = "one.sample",  
 alternative = "one.sided")$n)

## [1] 92