Lesson 2 HW Submission

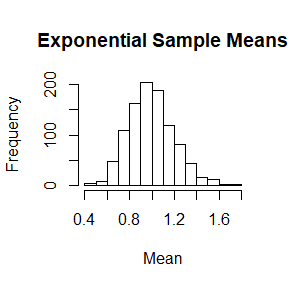
Matt Allen

06/11/2018

## 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  
sampleExponentialMean <- function(reps, n) {   
 sampleStat <- numeric(reps) # initialize the vector  
 for (i in 1:reps){  
 sampleStat[i] = mean( rexp(n) )  
 }  
 sampleStat  
}  
sampleStat <- sampleExponentialMean(1000, 23)  
hist(sampleStat, xlab = "Mean", main = "Exponential Sample Means")



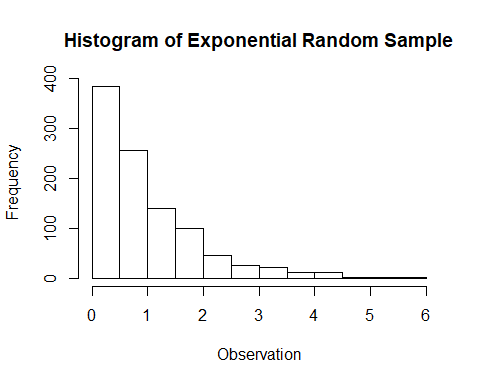
You are going to demonstrate the Central Limit Theorem, and gain some practice with loops in R, by showing that distribution of the sample means becomes increasingly normal as the sample size increases.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Draw a random sample of 1000 observations from an exponential distribution and assign it to exp\_sample\_1000.  
exp\_sample\_1000 <- rexp(1000)  
hist(exp\_sample\_1000, xlab = "Observation", main = "Histogram of Exponential Random Sample")

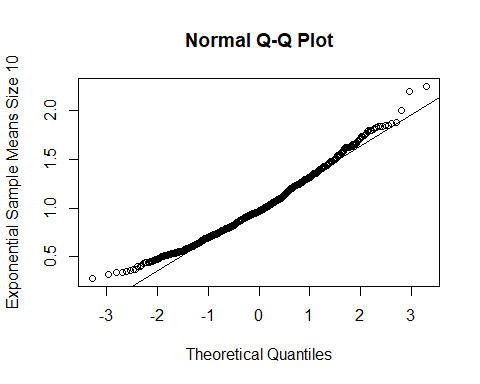


### Part 1b

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 -|-|-|-|-|-|-|-|-|-|-|-

# Create 1000 samples of size 10 from an exponential distribution.  
sampleStat <- sampleExponentialMean(1000, 10)  
  
# Make normal probability plot of sampleStat and add fit line.  
{  
 qqnorm(sampleStat, ylab="Exponential Sample Means Size 10")   
 qqline(sampleStat)  
}



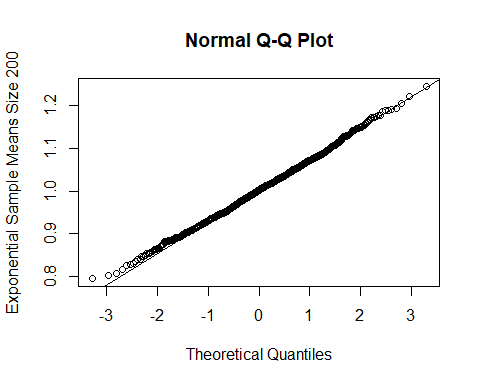
The sample means do not appear to be normally distributed. From the normal probability plot, the distribution of sampleMean appears to have a short tail on the left and is right skewed.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Create 1000 samples of size 200 from an exponential distribution.  
sampleStat <- sampleExponentialMean(1000, 200)  
  
# Make normal probability plot of sampleStat and add fit line.  
{  
 qqnorm(sampleStat, ylab="Exponential Sample Means Size 200")   
 qqline(sampleStat)  
}



The sample means from samples of size 200 do appear to be closer to normally distributed than the samples of size 10. The distribution for sample means from samples of size 200 is less right skewed and the left tail looks more inline with a normal distribution. The larger sample being closer to normally distributed is a consequence of the Central Limit Theorem.

## Exercise 2

This problem is modeled loosely after problem 5.70 on page 287 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 231 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 -|-|-|-|-|-|-|-|-|-|-|-

# Use formula n = ((z\_alpha/2)^2 \* sigma^2)/E^2  
sample\_size <- (1.96^2 \* 0.7^2)/0.05^2  
sample\_size

## [1] 752.9536

# Sample Size rounded up is 753.

### 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 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 -|-|-|-|-|-|-|-|-|-|-|-

# Compute sample and shift to mean 5.03 and stddev of 0.68, then use t test to compute 95% confidence interval.  
z <- rnorm(sample\_size); fakeData <- 0.68\*(z-mean(z))/sd(z) + 5.03  
t.test(fakeData,conf.level=.95)$conf.int

## [1] 4.98132 5.07868  
## 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 5.29 on page 279 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)

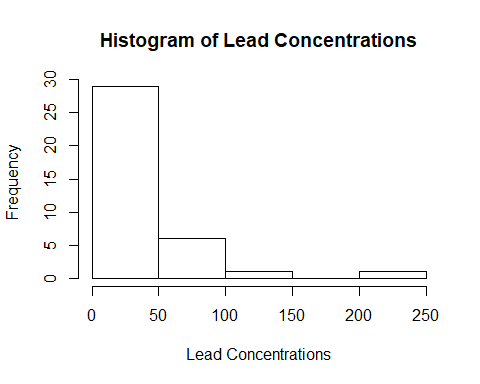
# load csv into data frame  
df <- read.csv('ex5-29.TXT')  
# put data for lead concentrations in vector called "lead"  
lead <- df$Lead   
# delete the data frame since we no longer need it  
rm(df)

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Make a histogram of the data to see if it looks normally distributed.  
hist(lead, xlab = "Lead Concentrations", main = "Histogram of Lead Concentrations")



Based on the histogram, lead concentrations definitely do not look to be normally distributed. The histogram shows that lead concentrations are right skewed.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# replace with your code  
t.test(lead,conf.level=.80)$conf.int

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

We are 80% confident that the population mean lead concentraion is between 30 and 46.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

Yes. The lower bound of the confidence interval suggests that the population mean lead concentration is larger that 30 mg/kg at a significance level. The 90% lower confidence bound is the same as an alternative hypothesis of greater than 30 mg/kg at a 10% level of significance.

### 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

(Step 2) Already did this in 3a.

(Step 3) Compute:

#Ho: mu = 30 and Ha: mu > 30   
t.test(x = lead,  
 alternative = "greater",  
 mu = 30,  
 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 10% level of significance, we fail to reject the null hypothesis that lead concentration is less than or equal to 30 mg/kg.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

B <- 5000  
rs <- replicate(B, sample(lead, replace = T))  
bd <- apply(rs, 2, mean)  
quantile(bd, c(.1,.9))

## 10% 90%   
## 29.89189 45.13514

### 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 -|-|-|-|-|-|-|-|-|-|-|-

library(boot)  
bootMean <- function(x, i) {  
 # x is a numeric vector  
 # i is a vector of indices for the resampled elements of x  
 mean(x[i])  
}  
set.seed(NULL)  
boot.object <- boot(lead, bootMean, R=5000)  
boot.ci(boot.object, conf = .8)

## Warning in boot.ci(boot.object, conf = 0.8): bootstrap variances needed for  
## studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 5000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = boot.object, conf = 0.8)  
##   
## Intervals :   
## Level Normal Basic   
## 80% (29.47, 45.09 ) (28.97, 44.73 )   
##   
## Level Percentile BCa   
## 80% (29.76, 45.51 ) (31.03, 47.89 )   
## 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% confidence bound is 32 mg/kg.

### 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 10% sigificance level, we reject the null hypothesis that the mean lead concentration is less than or equal to 30 mg/kg in favor of the alternative hypothesis that mean lead concentration is greater than 30 mg/kg.

### 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 bootstrap procedure more. The t test is intended for distributions that are roughly normal. Lead Concentrations are right skewed, and so a t test is not the best choice. The bootstrap procedure transforms the lead concentrations into a distribution that is approximately normal.

### 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 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 -|-|-|-|-|-|-|-|-|-|-|-

# Approximate the power of the t test.  
power.t.test(n = 37, delta = 10, sd = 37.1, sig.level = 0.1, 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 when testing . Again, with estimated standard deviation of 37.1 and $= 0.10, what sample size is required for power = 0.9?

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

# Find the sample size required for power = 0.9  
power.t.test(power = .9, delta = 10, sd = 37.1, sig.level = 0.1, type = "one.sample", alternative = "one.sided")$n

## [1] 91.25578