Week 2 HW Submission

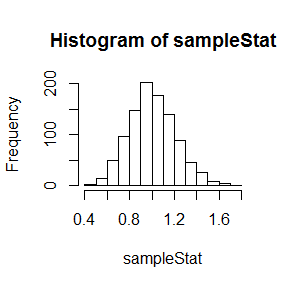
Spencer Swartz

January 31, 2017

## 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)



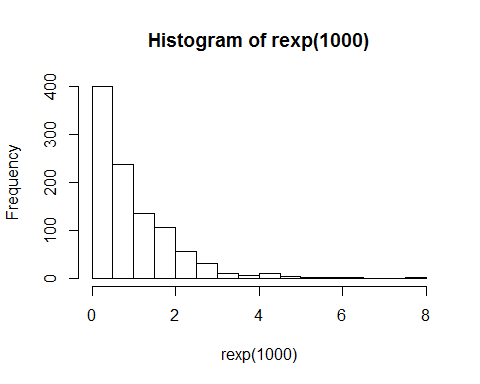
You are going to demonstrate the Central Limit Theorem, and gain some practice with loops in R, by showing that distribution of the sample means becomes increasingly normal as the sample size increases.

### 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 ... )

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

# create a histogram of 1000 random samples of the exponential distribution.  
hist(rexp(1000))

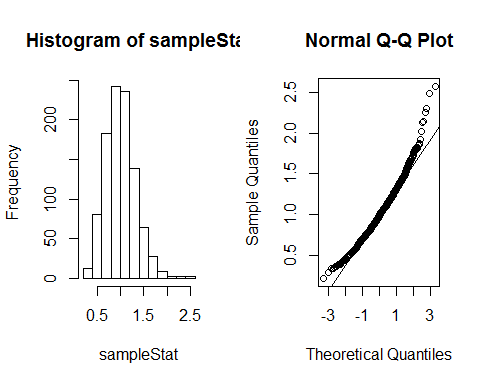


### Part 1b

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

# replace with your code  
par(mfrow=c(1,2))  
reps <- 1000  
n <- 10  
sampleStat <- numeric(reps) # initialize the vector  
for (i in 1:reps){  
 sampleStat[i] <- mean( rexp(n) )  
}  
hist(sampleStat)  
qqnorm(sampleStat)  
qqline(sampleStat)



#reset to 1 plot  
par(mfrow=c(1,1))

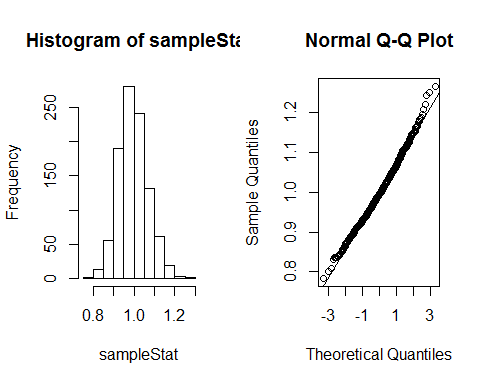
It looks as if there may be some fat tails in this distrobution so it does not seem to be normaly distributed

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

# replace with your code.  
par(mfrow=c(1,2))  
reps <- 1000  
n <- 200  
sampleStat <- numeric(reps) # initialize the vector  
for (i in 1:reps){  
 sampleStat[i] <- mean( rexp(n) )  
}  
hist(sampleStat)  
qqnorm(sampleStat)  
qqline(sampleStat)



#reset to 1 plot  
par(mfrow=c(1,1))

This looks to be more closely to a normal distrobution, the fat tails are not as pronounced

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

# replace with your code  
sig <- (5.2-4.8)/4  
ceiling((((1.96)^2)\*((sig)^2))/(.05^2))

## [1] 16

### 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.

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

# replace with your code  
fakeData <- rnorm(50)  
fakeData <- (fakeData - mean(fakeData))/sd(fakeData)  
fakeData <- .68\*fakeData + 5.03  
print(mean(fakeData))

## [1] 5.03

print(sd(fakeData))

## [1] 0.68

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 5.29 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 file)

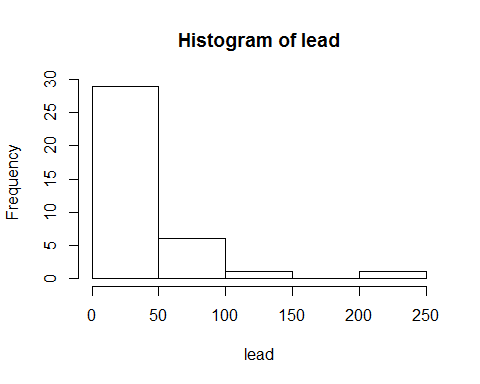
# load csv into data frame  
df <- read.csv('ex5-29.TXT')  
# put data for lead concentrations in vector lead  
lead <- df$Lead   
# delete the data frame since we no longer need it  
rm(df)

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

# replace with your code  
hist(lead)



No, it looks as if the data is not a normal distribution and it is skewed to the left.

### 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 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

There is an 80% chance that a given lead concentration level will be between 29.27 and 45.21.

### 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 discussed in the class notes and NEWS anncouncement).

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

No, this simpily means that there is a 90% chance that a given lead concentration level will be higher than 29.27.

### 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.

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

(Step 1) The hypotheses and

(Step 2) Already did this in 3a.

(Step 3) Compute:

t.test(lead, mu = 30, alternative = "greater")

##   
## One Sample t-test  
##   
## data: lead  
## t = 1.1868, df = 36, p-value = 0.1215  
## alternative hypothesis: true mean is greater than 30  
## 95 percent confidence interval:  
## 26.93932 Inf  
## sample estimates:  
## mean of x   
## 37.24324

(Step 4) Conclusion:

We must accept the null Hypothesis (H0) as the p-value is greater than .1

### 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. The code below was used to bootstrap a CI for men's cholesterol levels in the notes. Modify it produce a 80% CI for the population mean lead concentration using the lead data above. You'll have to modify the data at the beginning, the confidence level, and change the degrees of freedom when plotting a true t distribution for comparison. Comment on differences between the theory and bootstrap intervals. Is the bootstrap sampling distribution in good agreement with the t distribution from theory?

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

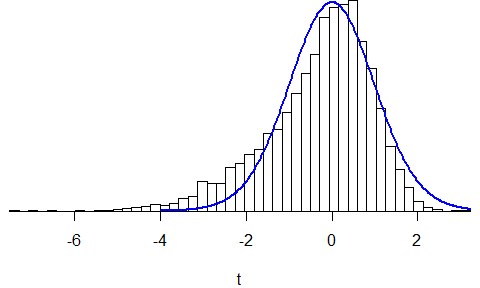
chol <- lead  
n <- length(chol)  
xbar <- mean(chol)  
s <- sd(chol)  
  
reps <- 10000  
t\_sim <- numeric(reps)  
  
# generate bootstrapped sampling distribution  
for (i in 1:reps){  
 x <- sample(chol,n,replace=TRUE)  
 t\_sim[i] <- (mean(x)-xbar)/(sd(x)/sqrt(n))  
}  
  
# build boostrapped "Studentized" t-inverval for mu  
alpha <- .20  
prbs <- c( 1-alpha/2, alpha/2 )  
  
tcrit\_est <- unname(quantile(t\_sim, prob = prbs ))  
t\_ci\_bootstrap <- xbar - tcrit\_est\*s/sqrt(n)  
t\_ci\_bootstrap

## [1] 30.86431 49.65803

# for comparison here is the t-interval from theory  
t\_ci\_theory <- as.vector(t.test(chol)$conf.int)  
t\_ci\_theory

## [1] 24.86550 49.62099

# Compare the bootstrapped sampling distribution to the  
# t-distribution from theory. If the sample were from a normal distribution  
# these would be in good agreement, instead we see some skewness in the smapling distribution ...  
par(mar=(c(4,.5,0,.5)),yaxs="i",bty="l",xaxs="i",yaxt="n")  
brks <- seq(min(t\_sim),max(t\_sim),length=50)  
hist(t\_sim,prob=TRUE,breaks=brks,main='',xlab='t',ylab='')  
x0 <- seq(-4,4,length=400)  
y0 <- dt(x0,df=n-1)  
points(x0,y0,type='l',lwd=2,col='blue')



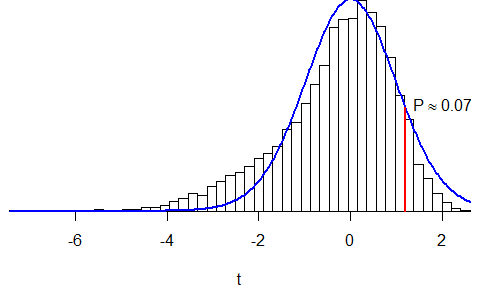
The low end of the boot strap does not quite fit with the low end in theory. We are not quite taking care of the fat tails.

### Part 3f

Using the bootstrapped sampling distribution approximate the -value of the hypothesis test. Modify the code below that was used to approximate the -value in the class notes. How does it compare to the -value from theory? If testing at the 10% significance level, does your conclusion change? If so, give a new conclusion.

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

## [1] 0.0741



We can now reject the null Hypothesis (H0) as the p-value is less than .1

### Part 3g

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

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

Since the original data was very skewed the t-test was inaccurate. But since bootstraping accounts for the skewness we were able to reject H0, meaning the averge lead level is greater than 30.

### Part 3h

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 3h -|-|-|-|-|-|-|-|-|-|-|-

# your code goes here  
power.t.test(n = n,delta = 10,sd = s, sig.level = .1, type = "one.sample", alternative = "one.sided")

##   
## One-sample t test power calculation   
##   
## n = 37  
## delta = 10  
## sd = 37.12397  
## sig.level = 0.1  
## power = 0.6323678  
## alternative = one.sided

### Part 3i

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

power.t.test(power = .9,delta = 10,sd = 37.1, sig.level = .1, type = "one.sample", alternative = "one.sided")$n

## [1] 91.25578