stat interference

I D

```
# Exponential Distribution compare to Central limit Theorem
# The project consists of two parts:
# 1. A simulation exercise.
# 2. Basic inferential data analysis.
# 1: A simulation exercise
#  Overview
# In this project the exponential distribution is investigated in R and compare it with Central Limit T
# Simulations
# A series of 1000 simulations is run to create a data set for comparison purpose. Each simulation cont
# Given data: n = 40; simNum = 1000; lambda = 0.2
# For reproducibility, set seed = 10000
# Exponential sampling parameters
# set seed for reproducability
set.seed(31)
# set lambda to 0.2
lambda <- 0.2
# 40 samples
n <- 40
# 1000 simulations
simulations <- 1000
```

```
# simulate
simulated_exponentials <- replicate(simulations, rexp(n, lambda))
simExp = function(n, lambda){
    mean(rexp(n,lambda))
    }

simul = data.frame(ncol=2,nrow=simulations)
names(simul) = c("Sample","Mean")
for (i in 1:simulations)
{
    simul[i,1] = i
    simul[i,2] = simExp(n,lambda)
}
```

```
# calculate mean of exponentials
means_exponentials <- apply(simulated_exponentials, 2, mean)
```

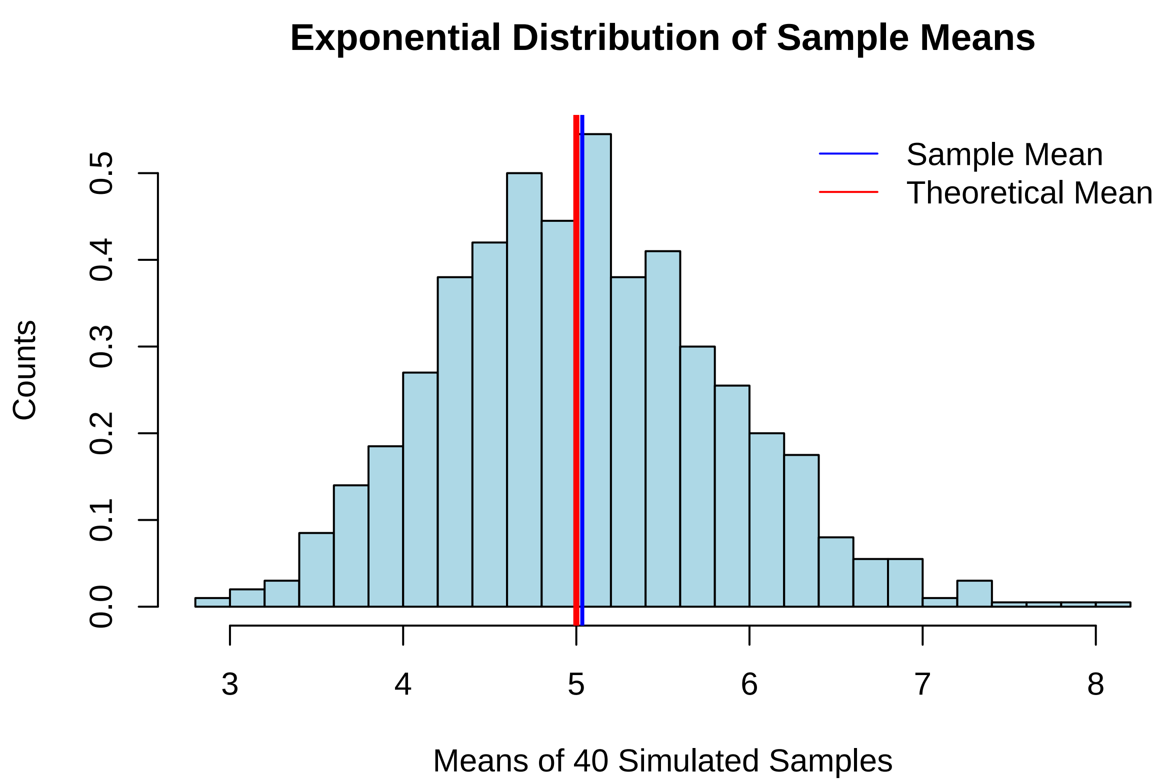
```
#Question 1

#Show where the distribution is centered at and compare it to the theoretical center of the distributio
analytical_mean <- mean(means_exponentials)
analytical_mean
```

```
## [1] 4.993867
```

```r
# analytical mean
theory_mean <- 1/lambda
theory_mean
```

```
## [1] 5
```

```r
# visualization
meanSample = mean(simul$Mean)
meanTheory = 1/lambda
hist(simul$Mean, breaks = 30, prob = TRUE,col = "lightblue",
     main="Exponential Distribution of Sample Means",
     xlab="Means of 40 Simulated Samples", ylab = "Counts")
abline(v = meanTheory, col= "red", lwd = 3)
abline(v = meanSample, col = "blue",lwd = 2)
legend('topright', c("Sample Mean", "Theoretical Mean"),
       bty = "n",
       lty = c(1,1),
       col = c(col = "blue", col = "red"))
```



The analytics mean is 5.006 the theoretical mean 5. The center of distribution of averages of 40 exponentials is very close to the theoretical center of the distribution.

Question 2

Show how variable it is and compare it to the theoretical variance of the distribution..

```
# standard deviation of distribution
standard_deviation_dist <- sd(means_exponentials)
standard_deviation_dist
```

```
## [1] 0.7931608
```

```
# standard deviation from analytical expression
standard_deviation_theory <- (1/lambda)/sqrt(n)
standard_deviation_theory
```

```
## [1] 0.7905694
```

```
# variance of distribution
variance_dist <- standard_deviation_dist^2
variance_dist
```

```
## [1] 0.6291041
```

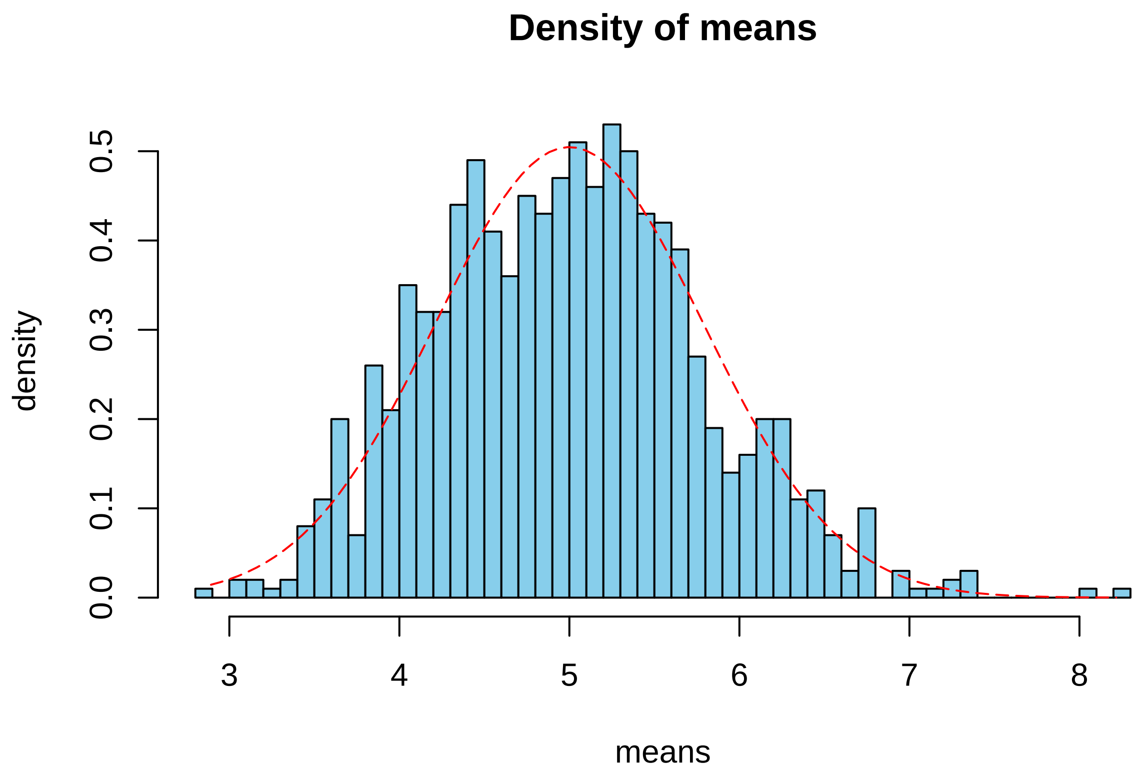
```
# variance from analytical expression
variance_theory <- ((1/lambda)*(1/sqrt(n)))^2
variance_theory
```

```
## [1] 0.625
```

Standard Deviation of the distribution is 0.7931608 with the theoretical SD calculated as 0.7905694. The Theoretical variance is calculated as ((1 / ??) * (1/???n))2 = 0.625. The actual variance of the distribution is 0.6291041

Question 3

```
# Show that the distribution is approximately normal.
xfit <- seq(min(means_exponentials), max(means_exponentials), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(n)))
hist(means_exponentials,breaks=n,prob=T,col="skyblue",xlab = "means",main="Density of means",ylab="dens
lines(xfit, yfit, pch=22, col="red", lty=5)
```



compare the distribution of averages of 40 exponentials to a normal distribution qqnorm(means_exponentials) qqline(means_exponentials, col = 2) # Due to Due to the central limit theorem (CLT), the distribution of averages of 40 exponentials is very close to a normal distribution. # 2. Basic inferential data analysis. # Load the ToothGrowth data and perform some basic exploratory data analyses

```
# load the data ToothGrowth
data(ToothGrowth)
# preview the structure of the data
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
# preview first 5 rows of the data
head(ToothGrowth, 5)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
```

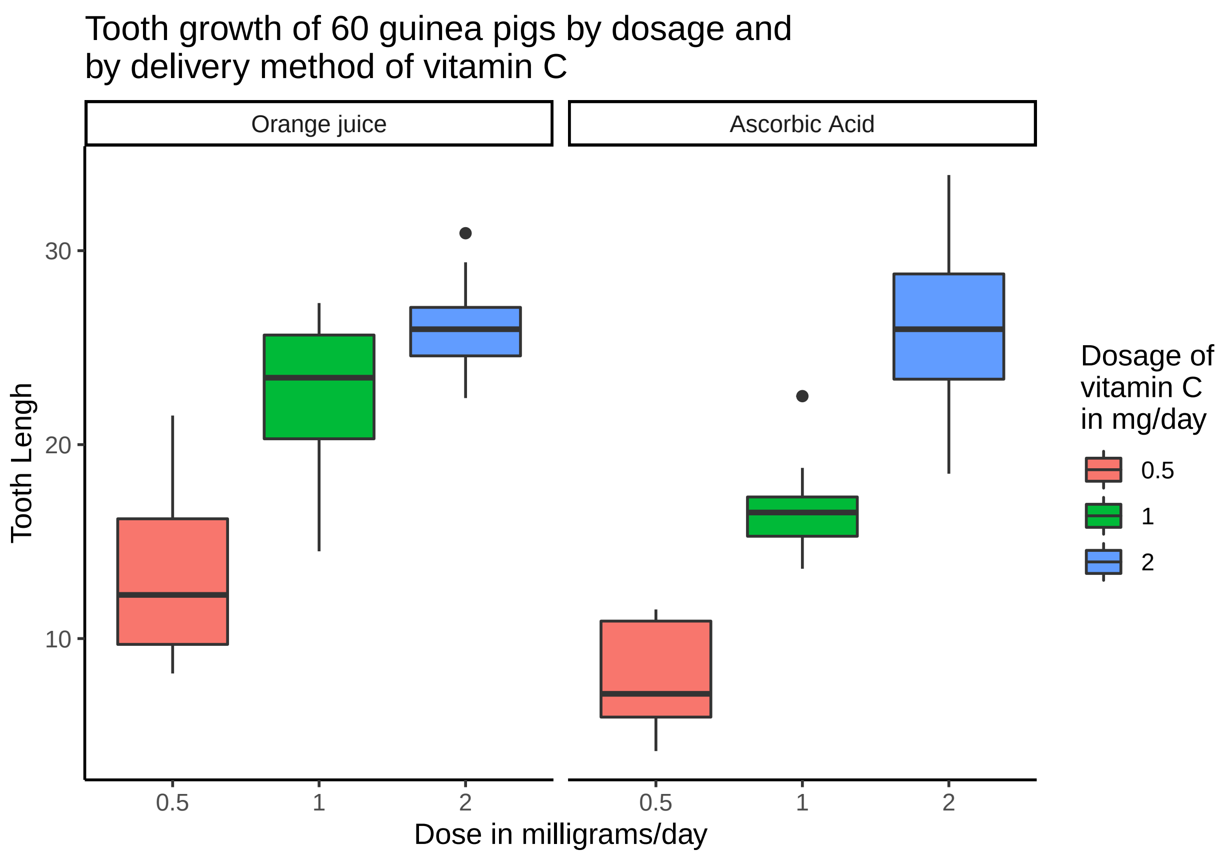
```
# Provide a basic summary of the data.
# data summary
summary(ToothGrowth)
```

```
##       len         supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000
```

```
# compare means of the different delivery methods
tapply(ToothGrowth$len,ToothGrowth$supp, mean)
```

```
##       OJ       VC 
## 20.66333 16.96333
```

```
# plot data graphically
library(ggplot2)
ggplot(ToothGrowth, aes(factor(dose), len, fill = factor(dose))) +
  geom_boxplot() +
  # facet_grid(.~supp)+
  facet_grid(.~supp, labeller = as_labeller(
    c("OJ" = "Orange juice",
      "VC" = "Ascorbic Acid"))) +
  labs(title = "Tooth growth of 60 guinea pigs by dosage and\nby delivery method of vitamin C",
       x = "Dose in milligrams/day",
       y = "Tooth Lengh") +
  scale_fill_discrete(name = "Dosage of\nvitamin C\nin mg/day") +
  theme_classic()
```




```
# Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.
# comparison by delivery method for the same dosage
t05 <- t.test(len ~ supp,
             data = rbind(ToothGrowth[(ToothGrowth$dose == 0.5) &
                                          (ToothGrowth$supp == "OJ"),],
                          ToothGrowth[(ToothGrowth$dose == 0.5) &
                                          (ToothGrowth$supp == "VC"),]),
             var.equal = FALSE)

t1 <- t.test(len ~ supp,
            data = rbind(ToothGrowth[(ToothGrowth$dose == 1) &
                                         (ToothGrowth$supp == "OJ"),],
                         ToothGrowth[(ToothGrowth$dose == 1) &
                                         (ToothGrowth$supp == "VC"),]),
            var.equal = FALSE)

t2 <- t.test(len ~ supp,
            data = rbind(ToothGrowth[(ToothGrowth$dose == 2) &
                                         (ToothGrowth$supp == "OJ"),],
                         ToothGrowth[(ToothGrowth$dose == 2) &
                                         (ToothGrowth$supp == "VC"),]),
            var.equal = FALSE)
```

summary of the conducted t.tests, which compare the delivery methods by dosage,

take p-values and CI

```
summaryBYsupp <- data.frame(
  "p-value" = c(t05$p.value, t1$p.value, t2$p.value),
  "Conf.Low" = c(t05$conf.int[1],t1$conf.int[1], t2$conf.int[1]),
  "Conf.High" = c(t05$conf.int[2],t1$conf.int[2], t2$conf.int[2]),
  row.names = c("Dosage .05","Dosage 1","Dosage 2"))
```

```
# show data table
summaryBYsupp
```

```
##                 p.value  Conf.Low Conf.High
## Dosage .05 0.006358607  1.719057  8.780943
## Dosage 1   0.001038376  2.802148  9.057852
## Dosage 2   0.963851589 -3.798070  3.638070
```

Conclusion

For dosage of .5 milligrams/day and 1 milligrams/day does matter the delivery method. the delivery method for 2 milligrams/day. For dosage of 2 milligrams/day the delivery method doesn't matter.