Statistical Inference Course Project

Maxim Bulanov

Monday, March 10, 2015

Comparing normal and uniform distribution

Intro

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should 1. Show the sample mean and compare it to the theoretical mean of the distribution. 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. 3. Show that the distribution is approximately normal.

Analysis

First lets create the data :

```r
set.seed(123)
lamda=0.2
n=40
sim=1000
data <- replicate(sim, rexp(n, lamda))
```

Now let's answer the questions :

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution

```r
means <- apply(data, 2, mean)
mean(means)

## [1] 5.011911
```

Which is very close to :

```r
norm_mean<-1/lamda
norm_mean
```

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

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

```
sd(means)

## [1] 0.7749147
```

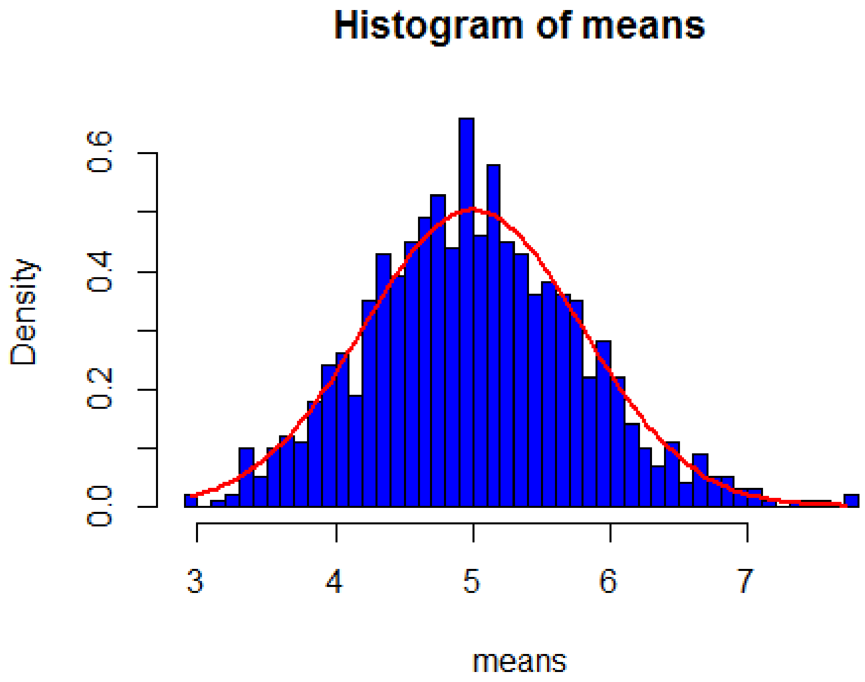
Which is very close to :

```
norm_sd<-(1/lamda)/sqrt(n)
norm_sd

## [1] 0.7905694
```

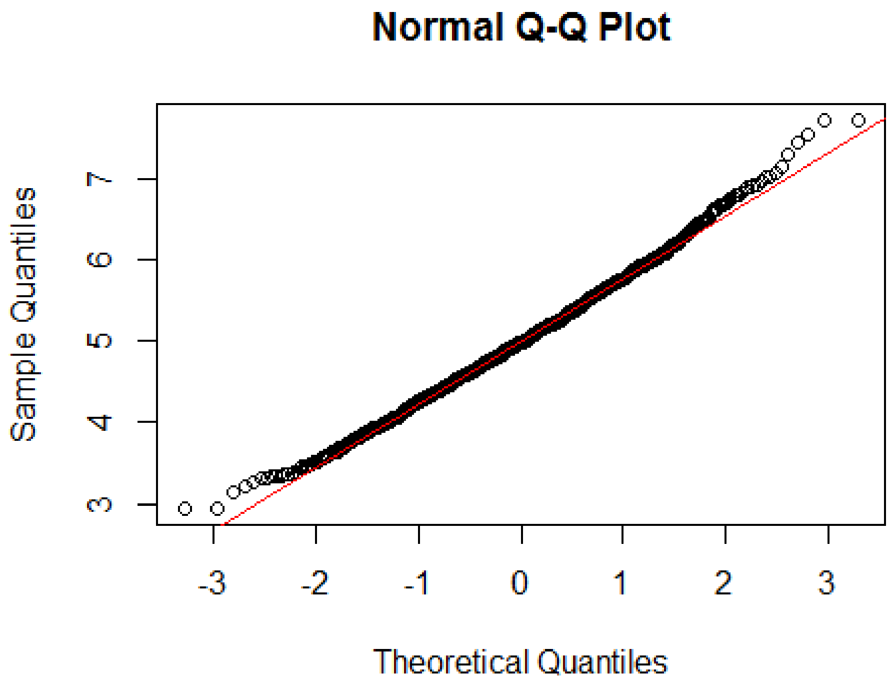
3. Show that the distribution is approximately normal.

```
hist(means,breaks=n,prob=TRUE,col="blue")
x <- seq(min(means), max(means), length=100)
y <- dnorm(x, mean=norm_mean, sd=norm_sd)
lines(x, y,lwd=2, col="red")
```



And using qqnorm :

```
qqnorm(means)
qqline(means, col = 2)
```



So we can see that the distribution is very like normal distribution .

Investigating ToothGrowth data

Intro

Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.

1.Load the ToothGrowth data and perform some basic exploratory data analyses 2.Provide a basic summary of the data. 3.Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering) 4. State your conclusions and the assumptions needed for your conclusions.

Some criteria that you will be evaluated on

- Did you perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data?
- Did the student perform some relevant confidence intervals and/or tests?
- Were the results of the tests and/or intervals interpreted in the context of the problem correctly?
- Did the student describe the assumptions needed for their conclusions?

Analysis

1.Load the ToothGrowth data and perform some basic exploratory data analyses

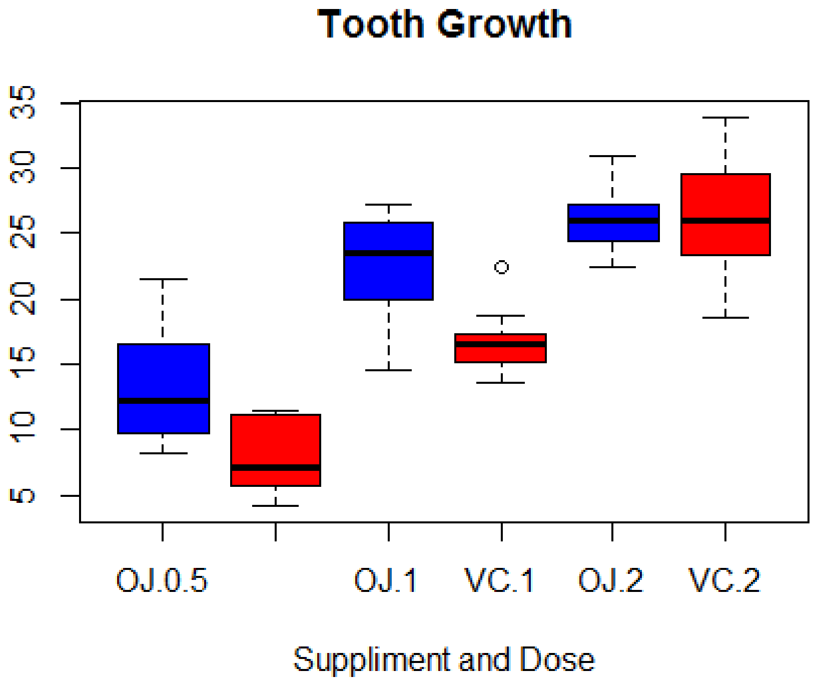
```
if (!(require("ggplot2", character.only=T, quietly=T))) {
    install.packages("ggplot2")
    library("ggplot2", character.only=T)
}

if (!(require("gridExtra", character.only=T, quietly=T))) {
    install.packages("gridExtra")
    library("gridExtra", character.only=T)
}

data(ToothGrowth)
str(ToothGrowth)
```

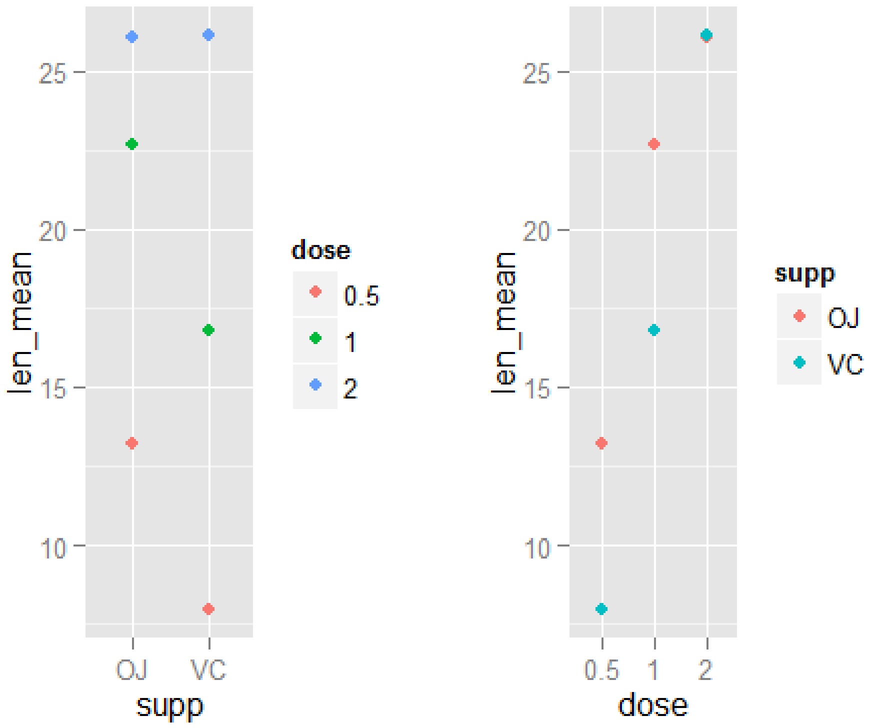
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
boxplot(len~supp*dose, data=ToothGrowth,col=(c("blue","red")),main="Tooth
Growth", xlab="Suppliment and Dose")
```

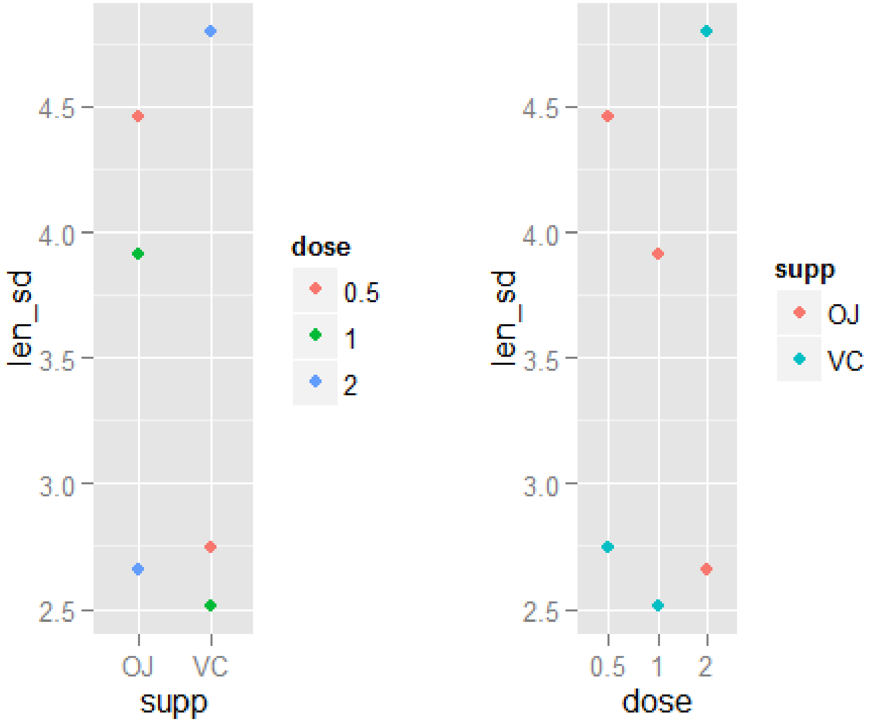


2.Provide a basic summary of the data.

```
n_by_type<-aggregate(ToothGrowth$len,
list(ToothGrowth$supp,ToothGrowth$dose), length)
mean_by_type<-aggregate(ToothGrowth$len,
list(ToothGrowth$supp,ToothGrowth$dose), mean)
names(mean_by_type)<-c("supp","dose","len_mean")
sd_by_type<-aggregate(ToothGrowth$len,
list(ToothGrowth$supp,ToothGrowth$dose), sd)
names(sd_by_type)<-c("supp","dose","len_sd")
mean_by_supp <- qplot(supp,len_mean, data = mean_by_type,color = dose)
mean_by_dose<- qplot(dose,len_mean, data = mean_by_type,color = supp)
grid.arrange(mean_by_supp, mean_by_dose, ncol = 2)
```



```
sd_by_supp <- qplot(supp,len_sd, data = sd_by_type,color = dose)
sd_by_dose<- qplot(dose,len_sd, data = sd_by_type,color = supp)
grid.arrange(sd_by_supp, sd_by_dose, ncol = 2)
```



3.Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)

Compare tooth growth by supplement

```
t.test(len ~ supp, paired = F, var.equal = F, data = ToothGrowth)

## 
##  Welch Two Sample t-test
## 
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC 
##         20.66333         16.96333
```

Since 0 is in the conf. interval we can't say for sure that there is a difference in tooth len among the two supplements. Another support for this is |t| value less than 2.

Compare tooth growth by dose

```
t.test(len ~ dose, paired = F, var.equal = F, data =
subset(ToothGrowth,ToothGrowth$dose!="0.5"))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

Since 0 is not in the conf. interval it's safe to assume that dosage of 1 vs. 2 does have influence on tooth len. Another support for this is |t| value which is high enough.

```
t.test(len ~ dose, paired = F, var.equal = F, data =
subset(ToothGrowth,ToothGrowth$dose!="1"))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

Since 0 is not in the conf. interval it's safe to assume that dosage of 0.5 vs. 2 does have influence on tooth len. Another support for this is |t| value which is high enough.

```
t.test(len ~ dose, paired = F, var.equal = F, data =
subset(ToothGrowth,ToothGrowth$dose!="2"))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

Since 0 is not in the conf. interval it's safe to assume that dosage of 0.5 vs. 1 does have influence on tooth len. Another support for this is |t| value which is high enough.

4. State your conclusions and the assumptions needed for your conclusions.

Generally speaking we've seen that supplement type probably does not have a big influence on tooth growth while dosage does have an impact(though it seems that orange juice does generates longer len teeth in average but it's not enough to conclude it as a better supplement) . The bigger the dosage is , the longer the teeth are.

The following assumptions were made : 1. The variance of all tested groups are different . 2. The tested groups are not paired (meaning it's not same subjects with getting both supplements or all the dosages). 3. Since the number of tested subjects is relatively small , we assume it's random enough to be used as testing subjects