Simulation Exercise and Basic Data Analysis

Jean-Paul Courneya

8/6/2017

This class project consists of two parts:

1. A simulation exercise.

2. Basic inferential data analysis.

First you will set a working directory and load dependencies

```
setwd()
library(knitr)
library(ggplot2)
```

Part 1: Simulation Exercise

In this part of the project we are investigating the exponential distribution in R and comparing 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 = s for all of the simulations.

Investigating the distribution of averages of 40 exponentials.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

Use set.seed so that the simulation can be repeated. Setting the seed provides the same starting value to the random number generating function.

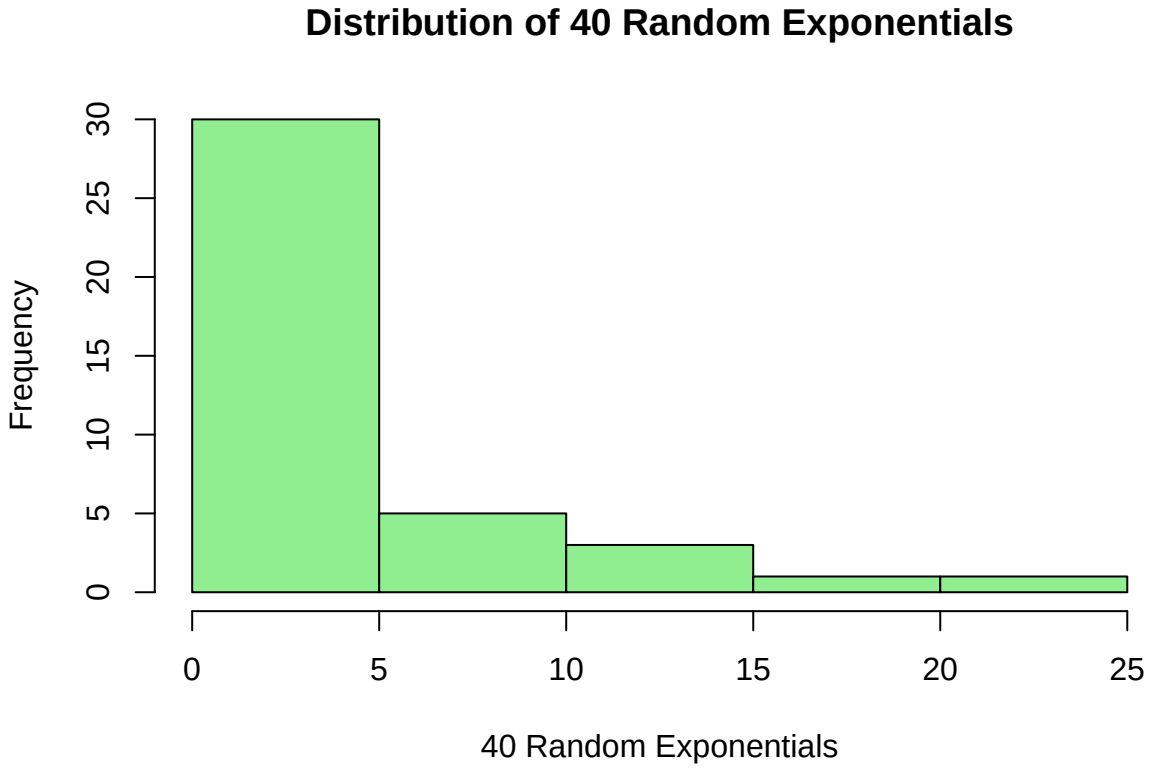
```
set.seed(15)
```

set some constants

```
lambda <- 0.2
sampleSize <- 40
simulations <- 1000
```

Start by taking a look at the distribution of a sample of 40 random exponentials

```
SimDist <- rexp(sampleSize,0.2)
hist(SimDist, xlab = "40 Random Exponentials", ylab = "Frequency",main = "Distribution of 40 Random Exp
```



Clearly we see that 40 random exponentials chosen look exponentially distributed.

Calculate the Sample vs Theoretical Mean

Next we will evaluate the mean of the sample data set from above.

```
Sample.Mean<- mean(SimDist)
cat("Sample Mean: ", Sample.Mean)
```

```
## Sample Mean:  4.506131
```

The theoretical mean for an eponential distribution as stated in the instructions is 1/lambda

```
Theoretical.Mean <- 1/lambda
cat("Theoretical Mean: ", Theoretical.Mean)
```

```
## Theoretical Mean:  5
```

The sample mean approximates the theoretical mean.

Calculate the Sample vs Theoretical Variance

```
Sample.Variance <- var(SimDist)
cat("Sample Variance: ", Sample.Variance)
```

```
## Sample Variance:  21.66123
```

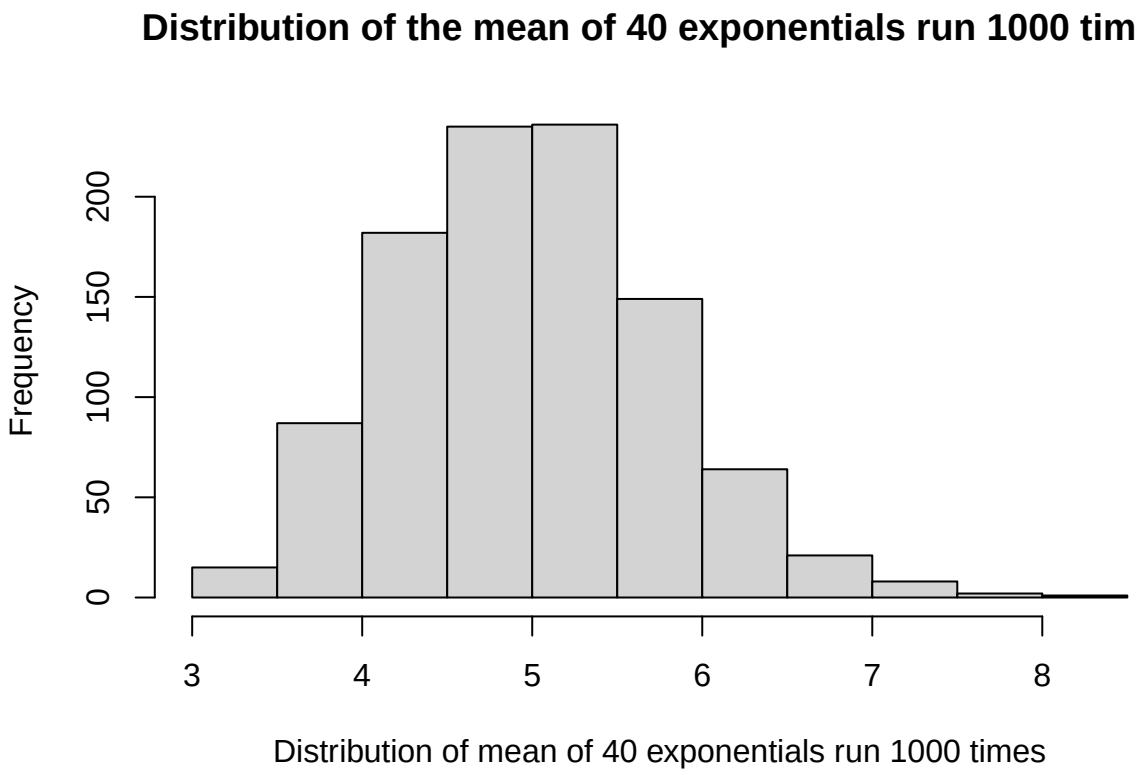
```
Theoretical.Variance <- (1/lambda)^2/sampleSize
cat("Theoretical Variance: ", Theoretical.Variance)
```

```
## Theoretical Variance:  0.625
```

There is a big difference between the sample variance and theoretical variance.

Now that we have seen the mean and variance of 40 random exponentials and compared it to the theoretical mean and variance. Next we will look at the distribution of the mean of 40 exponentials run 1000 times.

```
Simulation <- replicate(1000, mean(rexp(40,0.2)))
hist(Simulation, xlab = "Distribution of mean of 40 exponentials run 1000 times", ylab = "Frequency", ma
```



The data show a distribution around a central point of 5. Next we will evaluate the mean of the replicated sample data set.

```
mean(replicate(1000, rexp(40,0.2)))
```

```
## [1] 4.976328
```

This distribution looks far more Gaussian than the original exponential distribution!

Part 2: Basic inferential data analysis.

```
library(ggplot2)
```

Load the ToothGrowth data and perform some basic exploratory data analyses

```
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 ...
```

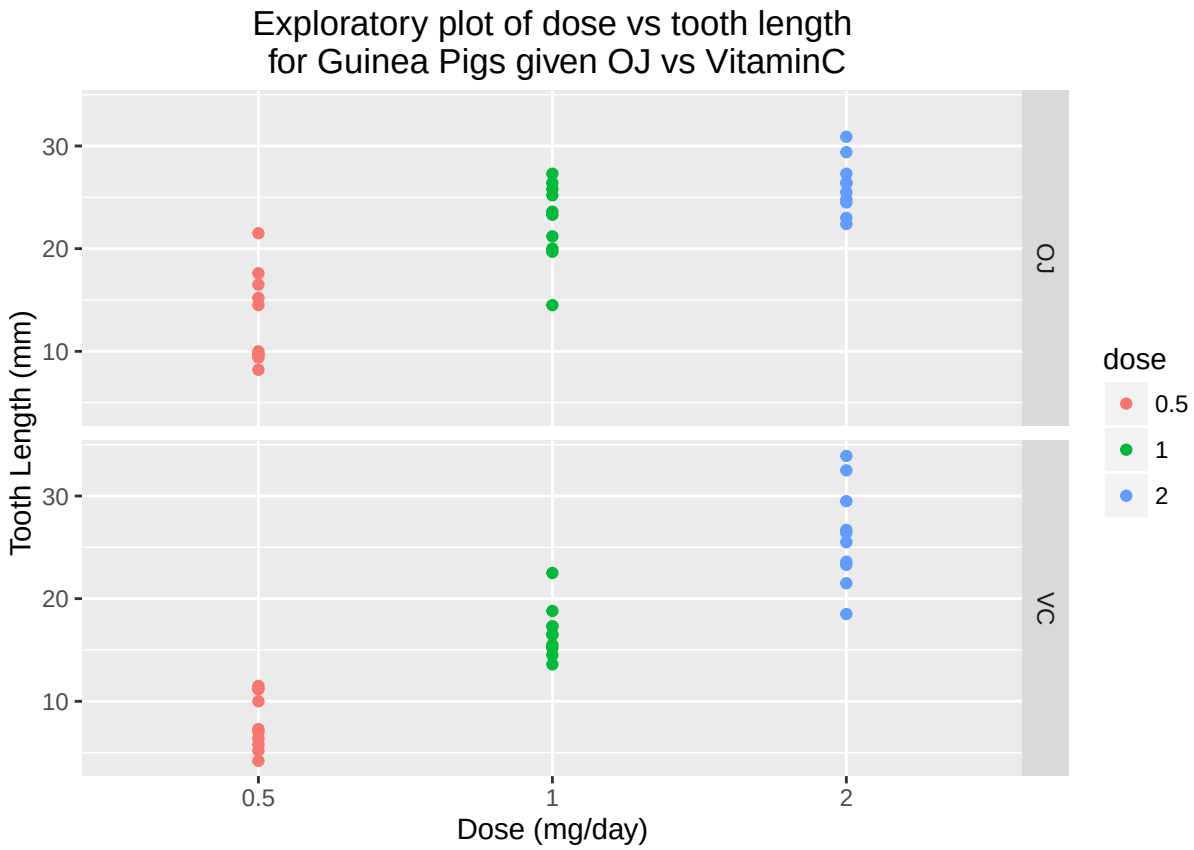
The data for the dose is currently numeric. Since the doses are factors they will be changed to factors.

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
str(ToothGrowth)
```

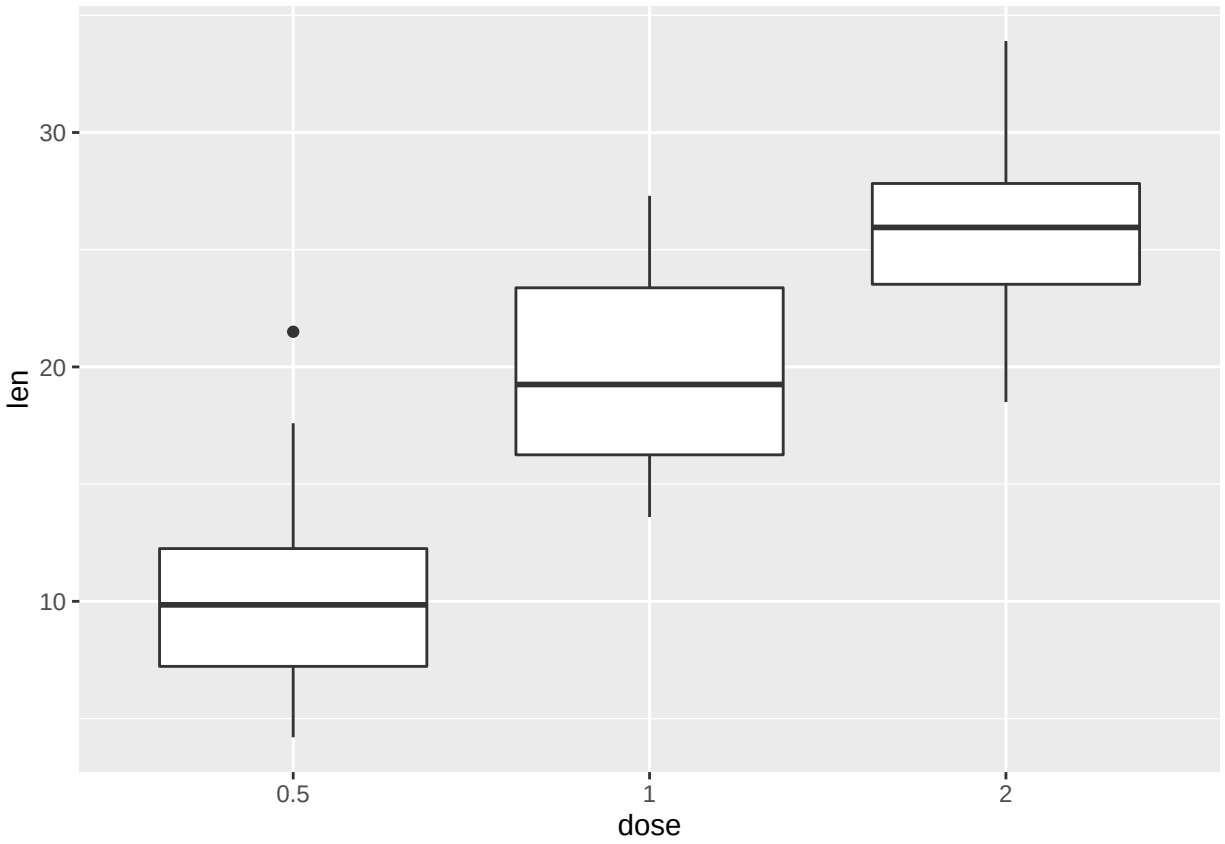
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

What does the data look like?

```
qplot(dose, len, data = ToothGrowth, facets = supp ~ ., color = dose, xlab = "Dose (mg/day)", ylab = "T
```




```
f <- ggplot(ToothGrowth, aes(dose,len))
f + geom_boxplot()
```



A basic summary of the data

```
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

compare tooth growth by supp and dose using confidence intervals and/or hypothesis tests