Statistical Inference (statinference-011) Course Project

Esmit Pérez

February 16, 2015

Part 1

Overview

This is the report for the project for the statistical inference class. In it, I will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

The format and formulas here included are based off the outline of the project.

Simulations

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.

```
# load libraries, setup variables
library(ggplot2)
lambda <- 0.2
mu <- 1/lambda
stdDev <- 1/lambda
numExponentials <- 40
numSimulations <- 1:1000

# we're dealing with random data, so always set seed to make it reproducible.
set.seed(909)

# obtains the mean of running rexp with 40 exponentials and given lambda
cfunc <- function(v) {mean(rexp(numExponentials, lambda))}

#  for each entry in array of size 1000, run the function
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(cfunc()))

dat <- data.frame(x = mns)
```

Sample Mean versus Theoretical Mean

Theoretical mean is 1/lambda:

```
mu
```

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

The sample mean is:

```
mean(dat$x)
```

```
## [1] 4.960913
```

Sample Variance versus Theoretical Variance

Theoretical variance is $\mu/\sqrt(n)$:

```
mu/sqrt(numExponentials)
```

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

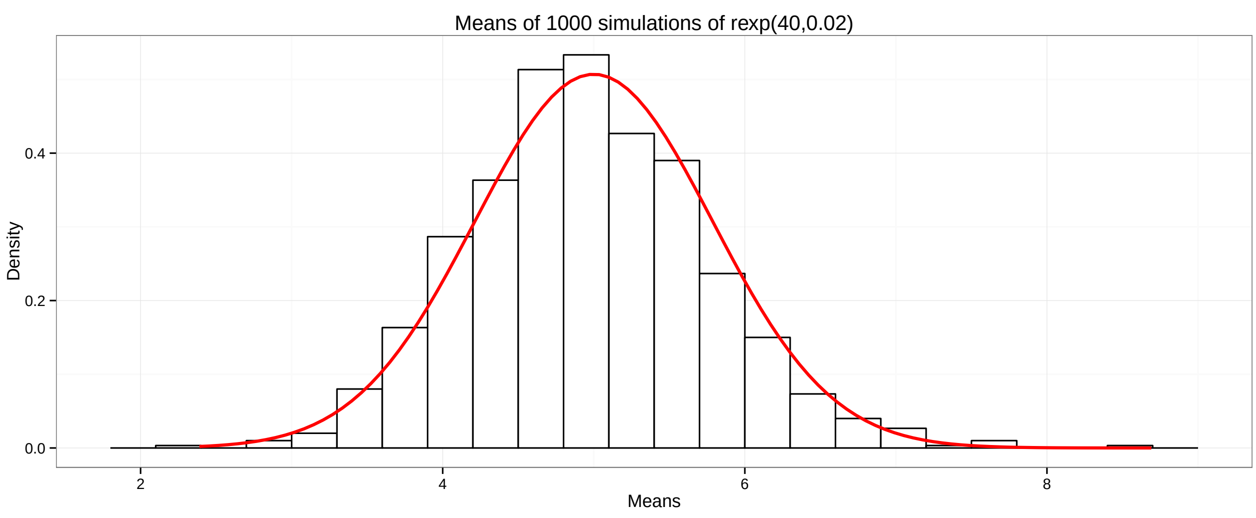
The sample variance S is:

```
var(dat$x)
```

```
## [1] 0.619083
```

Distribution

The following graph shows how the mean values of 1000 simulations approximate the normal distribution (curve in red). For complete code and output see appendix, fig 1.1



Part 2

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

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

```r
#load data set from R's provided datasets
data(ToothGrowth)
```

Provide a basic summary of the data.

Note: Please refer to appendix for a complete summary.

The following plot shows the growth length, per supplement at varying dosages (see R code and graph in ap-

pendix, fig 2.1):

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)

H_0 : Supplement *OJ (Orange Juice)* leads to more growth.

State your conclusions and the assumptions needed for your conclusions.

Based on (Week 3 video) (https://class.coursera.org/statinference-011/lecture/243) we can perform a `t.test` to determine whether this is true, assuming variances are equal (see complete output in appendix, fig 2.2):

```
##             [,1]     [,2]
## [1,] -0.1670064 7.567006
## [2,] -0.1670064 7.567006
```

Factors in `supp` show `OJ` comes first (see appendix, fig 2.3) and in both tests the result, is the same, column `1`, correspoding to `OJ` (*Orange Juice*) is below zero, meaning it does not lead to more length.

Appendix

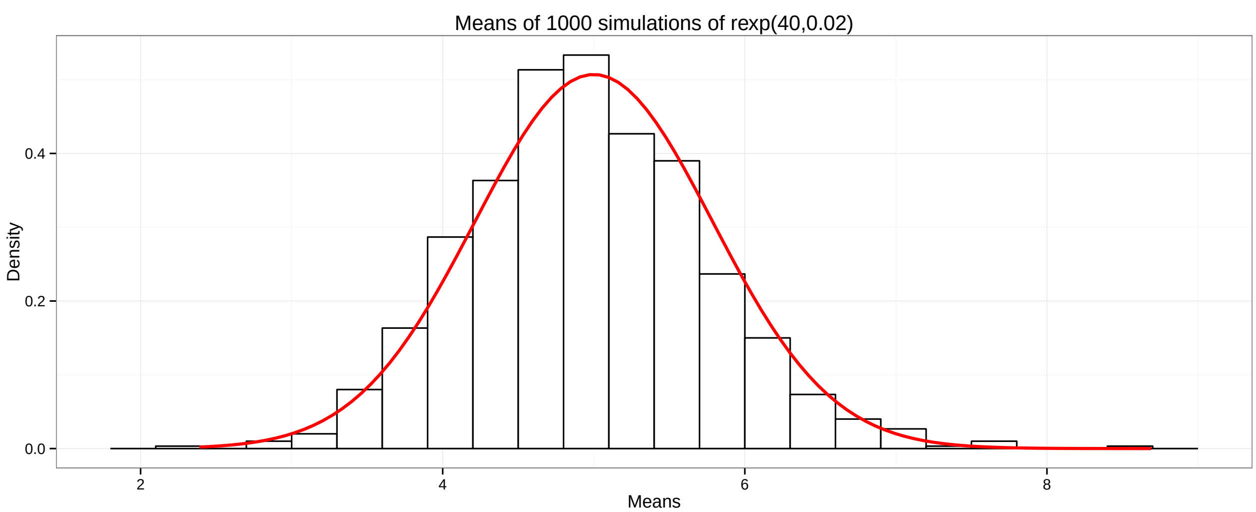
Part 1 - Supporting figures.

Figure 1.1

```r
# plot and customize graph
g <- ggplot(dat, aes(x = x))
g <- g + geom_histogram(binwidth=.3, colour = "black", fill="white",
        aes(y = ..density..))
g <- g + ggtitle("Means of 1000 simulations of rexp(40,0.02)")
g <- g + xlab("Means") + ylab("Density")

g <- g + stat_function(fun = dnorm, arg=list( mean= mu, sd=sd(dat$x) ),
        color="red", size=1)
g <- g + theme_bw()

g
```



Part 2 - Supporting figures and exploratory analysis, ToothGrowth data.

Description

"The Effect of Vitamin C on Tooth Growth in Guinea Pigs"

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

Summary information for `ToothGrowth`:

```r
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 ...
```

```r
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
```

```r
head(ToothGrowth)
```

```
##    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
## 6 10.0   VC  0.5
```

Example coplot as per help(ToothGrowth):

```r
require(graphics)
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```

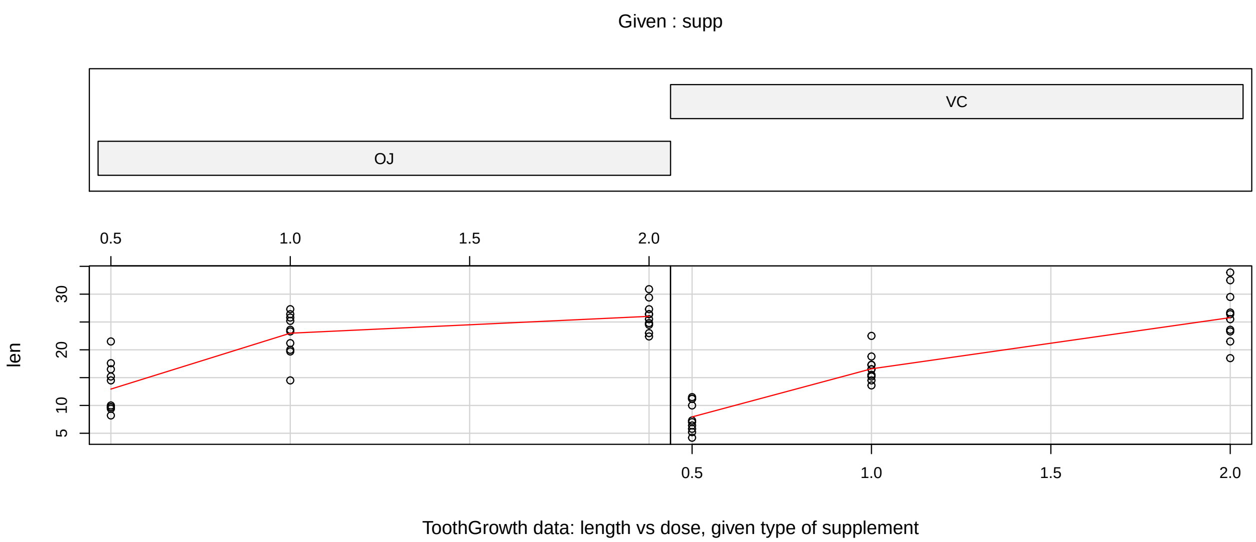


Figure 2.1

```
#plot and customize graph
g <- ggplot(ToothGrowth,aes(x = dose, y = len, color=supp))
g <- g + ggtitle("The Effect of Vitamin C on Tooth Growth in Guinea Pigs")
g <- g + xlab("Dose level (in mg)") + ylab("Length")
g <- g + scale_color_discrete(name="Delivery Method (Supplement)",
                        breaks=c("OJ", "VC"),
                        labels=c("Orange Juice", "Ascorbic Acide"))
g <- g + geom_boxplot()
g <- g + facet_grid(. ~ dose , labeller = function(variable, value){ paste0(value," mg") })
g
```

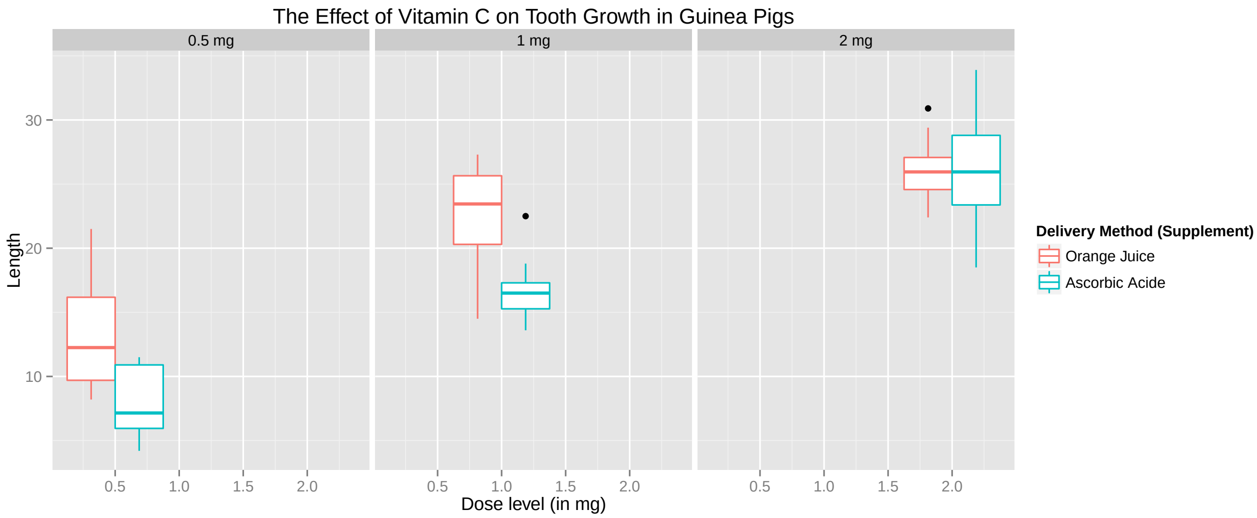


Figure 2.2

```
# Assume variances ARE equal
rbind(t.test(len ~ supp, paired=FALSE, var.equal=T, data=ToothGrowth)$conf,
      t.test(len ~ supp, paired=FALSE, var.equal=T, data=ToothGrowth)$conf
      )
```

```
##             [,1]     [,2]
## [1,] -0.1670064 7.567006
## [2,] -0.1670064 7.567006
```

```
factor(ToothGrowth$supp)
```

Figure 2.3

```
##  [1] VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC
## [24] VC VC VC VC VC VC VC OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ
## [47] OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ
## Levels: OJ VC
```