Inferential Data Analysis

Hasan Shojaei

20 June 2018

Overview

In Part 1, we investigate the exponential distribution and compare it with Central Limit Theorem. We show that the distribution of sample means is centered at theoretical mean of the popilation, and has a variance that is equal to the variance of the population divided by sample size. In Part 2, we analyze the ToothGrowth data and perform a hypothesis test, which indicates the use of two different supplement types (OJ or VC) does not cause a true difference in the average tooth growth.

Part 1: Simulation Exercise

We use rexp(n, lambda) to simulate exponential distribution in R. We investigate the distribution of averages of 40 exponentials by performing a thousand simulations.

```
n1 <- 1000       # number of simulations
n2 <- 40         # sample size
lambda <- 0.2    # exponential rate, also 1/mean and 1/sd
set.seed(1)      # set seed to ensure reproducibility
# drwa 40 random exponentials and take their average. repeat 1000 times
mns = NULL
for (i in 1 : n1) mns = c(mns, mean(rexp(n2,lambda)))
# calculate mean and variance of averages
avg <- mean(mns); var <- var(mns)
print(avg)
```

```
## [1] 4.990025
```

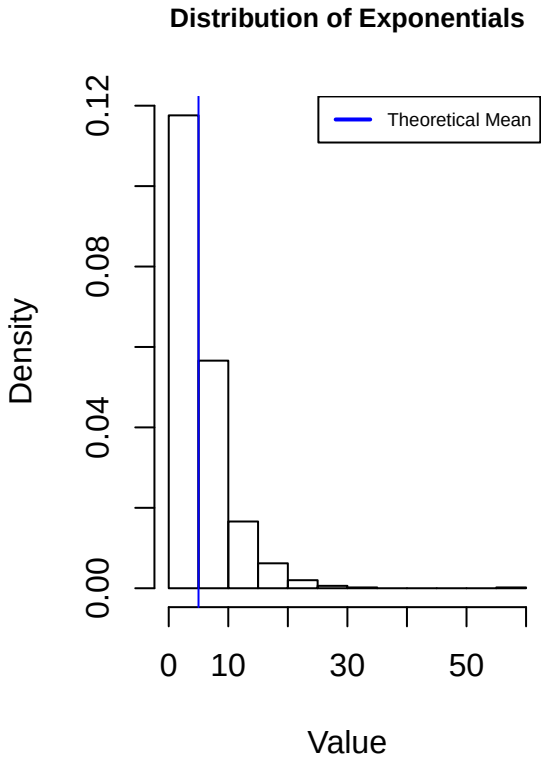
```
print(var)
```

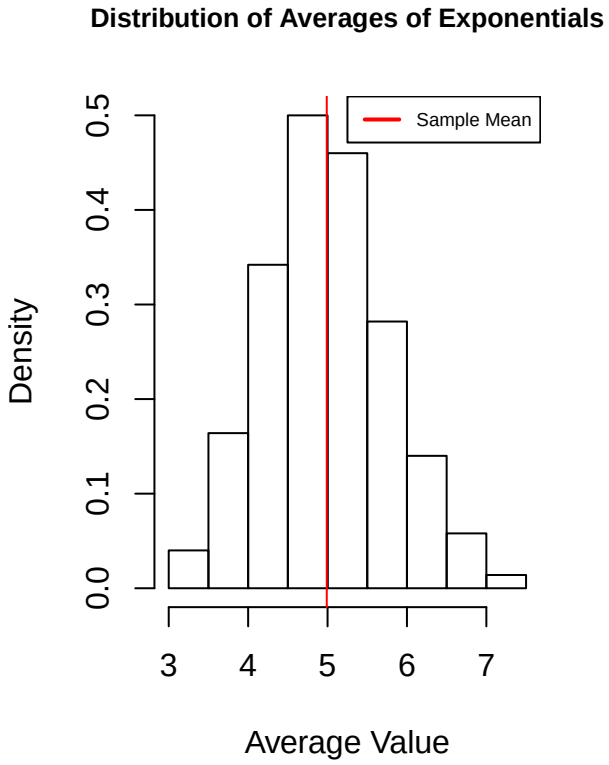
```
## [1] 0.6111165
```

Sample Mean vs. Theoretical Mean

We compare sample mean and theoretical mean in this section. Let's plot the distribution of a large collection of random exponentials (left figure below) and the distribution of a large collection of averages of 40 exponentials (right figure below) first. The vertical lines denote mean values.

```
par(mfrow=c(1,2))
# draw 1000 random exponentials and plot their distribution
hist(rexp(n1,lambda), prob=T,
     main="Distribution of Exponentials", xlab="Value", cex.main=0.85)
abline(v=1/lambda, col="blue")
legend("topright", legend="Theoretical Mean", col="blue", lwd=2, cex=0.6)
# plot the distribution of sample means
hist(mns, prob=T,
     main="Distribution of Averages of Exponentials",
     xlab="Average Value", cex.main=0.85)
abline(v=avg, col="red")
legend("topright", legend="Sample Mean", col="red", lwd=2, cex=0.6)
```





Sample mean is **4.99**, which is very similar to the theoretical mean of the distribution: $\mu = 1/\lambda = \mathbf{5}$.

Sample Variance vs. Theoretical Variance

Sample variance is **0.61**, which is much smaller than the theoretical variance of the distribution: $\sigma^2 = 1/\lambda^2 = \mathbf{25}$. This is evident from the figure above too. According to Central Limit Theorem, if we multiply sample variance (**0.61**) by sample size (**40**), we obtain a good estimate of the distribution variance: **24.44**.

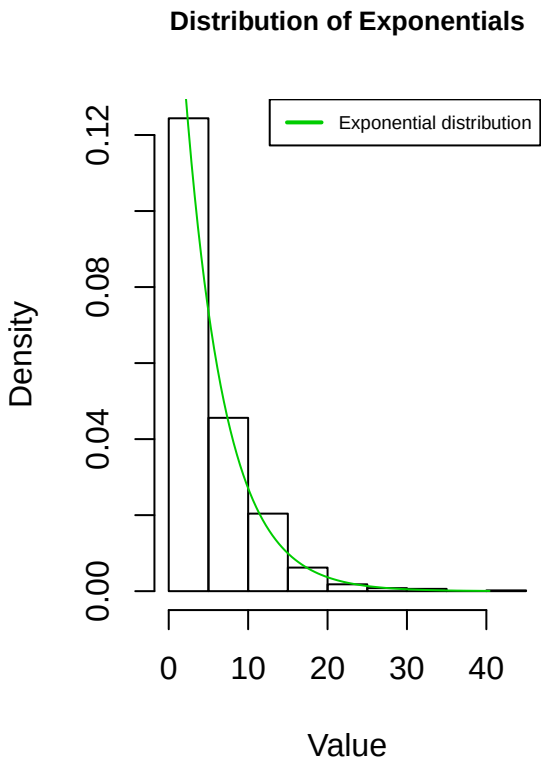
Sampling Distribution of the Mean vs. Population Distribution

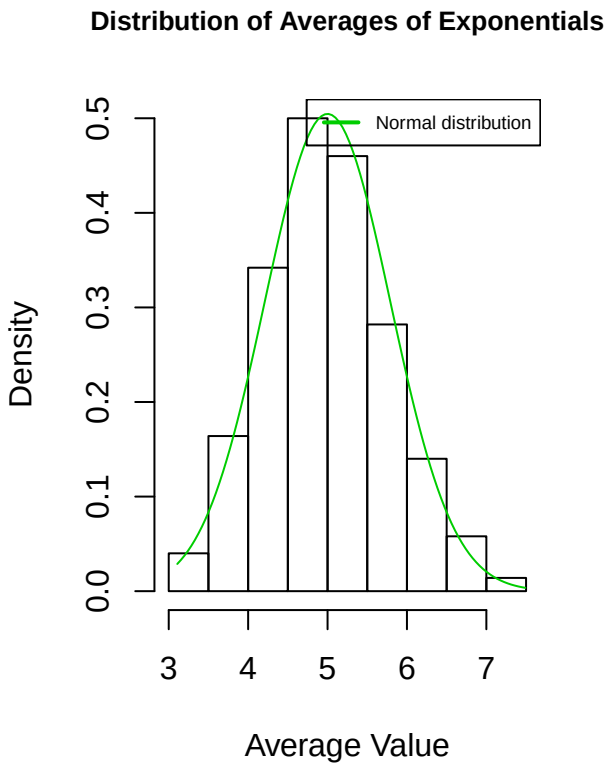
In this section we compare sampling distribution of the means (right figure below) with population distribution (left figure below). The sampling distribution of the mean looks far more Gaussian than the original exponential distribution! For reference, we have added Exponential (with rate=lambda) and Gaussian (with mean=1/lambda and sd=1/lambda/sqrt(n2)) probability density plots to the left and right figures respectively.

```
par(mfrow=c(1,2))
# draw 1000 random exponentials and plot their distribution
exp_distr <- rexp(n1,lambda)
hist(exp_distr, prob=T,
     main="Distribution of Exponentials", xlab="Value", cex.main=0.85)
x <- seq(min(exp_distr), max(exp_distr), length = 100)
lines(x, dexp(x, rate=lambda), pch = 25, col = "green3")
legend("topright", legend="Exponential distribution", col="green3", lwd=2, cex=0.6)

# plot the distribution of sample means
```

```r
hist(mns, prob=T,
     main="Distribution of Averages of Exponentials",
     xlab="Average Value", cex.main=0.85)
x <- seq(min(mns), max(mns), length = 100)
lines(x, dnorm(x, mean = 1/lambda, sd = (1/lambda/sqrt(n2))), pch = 25, col = "green3")
legend("topright", legend="Normal distribution", col="green3", lwd=2, cex=0.6)
```





Part 2: Basic Inferential Data Analysis

In the second portion of the project, we analyze the ToothGrowth data in the R datasets package.In this dataset, the response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

Loading the Data and Providing a Summary of the Data

We will first need to load the data into R environment. The structure of the data frame and a summary of the data is provided here.

```r
library(datasets)
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 ...
```

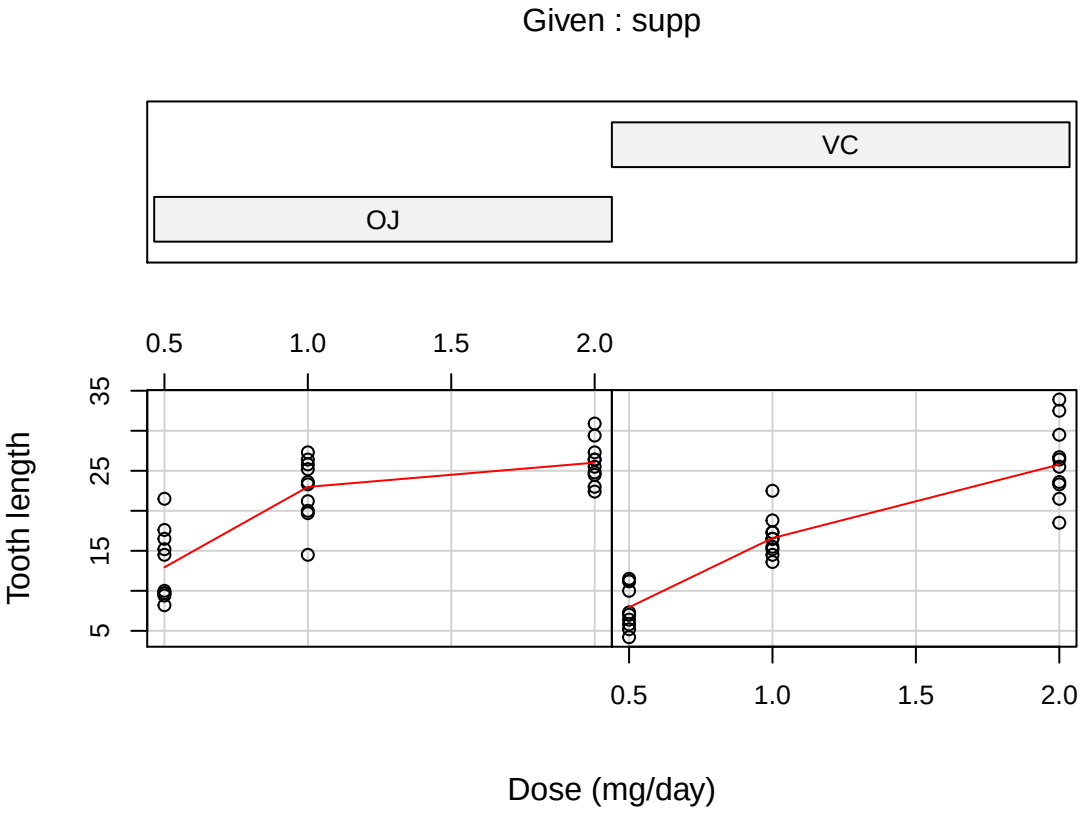
```
summary(ToothGrowth)
```

```
##       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
```

Performing Exploratory Data Analysis

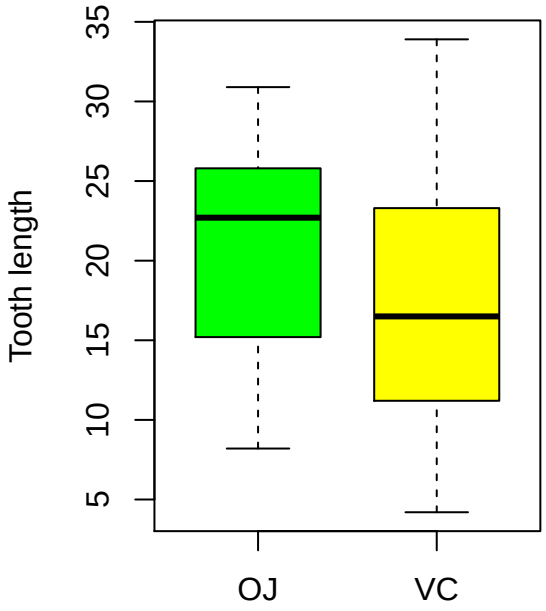
We first use the graphics package to create a coplot of length vs dose, given supplement type.

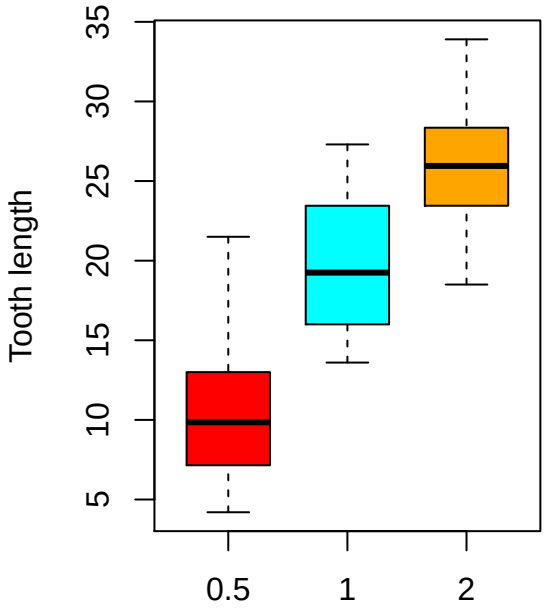
```
library(graphics)
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "Dose (mg/day)", ylab="Tooth length")
```



We can also use boxplots with respect to supplement type and dose to further explore the data.

```
par(mfrow=c(1,2))
boxplot(len~supp,data=ToothGrowth,boxwex=0.7,
        col=c("green", "yellow"), xlab="Supplement type", ylab="Tooth length")
boxplot(len~as.factor(dose),data=ToothGrowth,boxwex=0.7,
        col=c("red", "cyan", "orange"), xlab="Dose (mg/day)", ylab="Tooth length")
```





The main observations from this section is that the tooth length increases with higher doses of both supplement types. It's also observed that the min, Q1, median and Q3 are smaller for VC supplement compared to OJ supplement, while VC supplement results in a larger max value of tooth length.

Performing Hypothesis Test

In this section we perform a t-test to compare tooth growth by supplement type. The null hypothesis is that the true difference in means is equal to 0. The alternative hypothesis is that the true difference in means is NOT equal to 0. Since there are two independent group of pigs in this study, we set paired=FALSE. It's also assumed that the vairances are different.

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

```
##
##  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
```

Conclusion

The p-value is larger than 0.05 (equivalently 0 is within 95% confidence interval). This means we have weak evidence against the null hypothesis, so we fail to reject the null hypothesis.