Statistical Inference Project

Greta Garcia

11/12/2022

A simulation exercise

Show the sample mean and compare it to the theoretical mean of the distribution

```
lambda <- 0.2
smn <- mean(rexp(40, lambda))
mn <- 1/lambda

compMeans <- data.frame(Types=c("Sample Mean", "Theorical Mean"),
                        Values=c(smn, mn))
p <- ggplot(data=compMeans, aes(x=Types, y=Values)) +
        geom_bar(stat="identity", fill="steelblue", width=0.5) +
        ggtitle("Compare Means")
```

Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```
svariance <- var(rexp(40, lambda))
variance <- mn^2

compVariance <- data.frame(Types=c("Sample Variance", "Theorical Variance"),
                           Values=c(svariance, variance))
p <- ggplot(data=compVariance, aes(x=Types, y=Values)) +
        geom_bar(stat="identity", fill="steelblue", width=0.5) +
        ggtitle("Compare Variances")
```

Show that the distribution is approximately normal

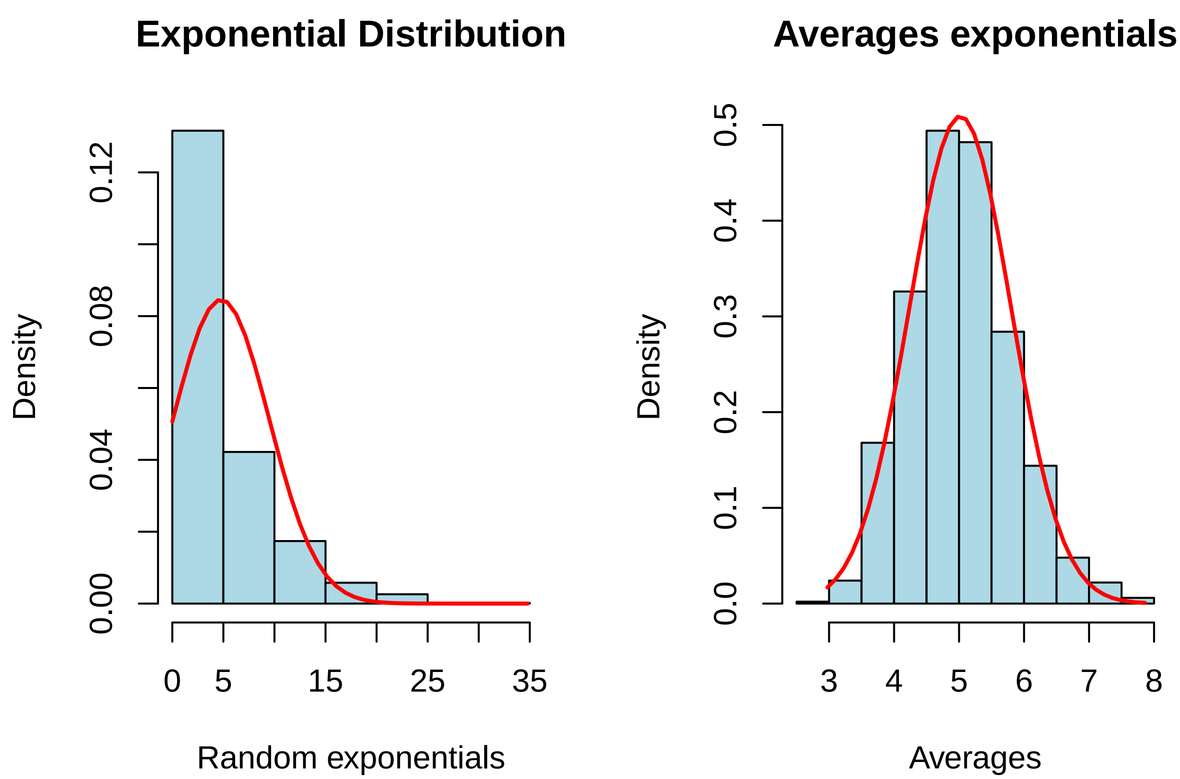
```
random <- rexp(runif(1000), lambda)
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(runif(40), lambda)))
```

Draw graphics on the same pane to compare them and add a Normal of the Distribution to see if it's similar or not. To draw the normal, we calculate the mean and the standard Deviation of the sample.

```
par(mfrow=c(1,2))
hist(random, prob=TRUE, main="Exponential Distribution",
     xlab="Random exponentials", col = "lightblue")
x <- seq(min(random), max(random), length = 40)
f <- dnorm(x, mean = mean(random), sd = sd(random))
```

```
lines(x, f, col = "red", lwd = 2)

hist(mns, prob=TRUE, main="Averages exponentials",
     xlab="Averages", col = "lightblue")
x <- seq(min(mns), max(mns), length = 40)
f <- dnorm(x, mean = mean(mns), sd = sd(mns))
lines(x, f, col = "red", lwd = 2)
```



The Means distribution of a random sample of enough variables create a Normal Distribution for the Central Limit Theorem and in this exercise we have a demostration of this Theorem.

Basic inferential data analysis

Load the ToothGrowth data and perform some basic exploratory data analyses

```
tg <- ToothGrowth
str(tg)
```

```
## '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 ...
```

Provide a basic summary of the data.

```r
summary(tg)
```

```
##       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
```

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

For Central limit Theorem we know that the means of this sample have a Normal distribution.

```r
tg_oj <- tg[tg$supp=="OJ",]
tg_vc <- tg[tg$supp=="VC",]
```

In all cases, variables are Paired and supose that variances are equal * H0: alpha1 = alpha2 * Ha: alpha1 != alpha2

```r
t.test(tg_oj$len, tg_vc$len, paired=TRUE, var.equal=TRUE)
```

```
## 
##  Paired t-test
## 
## data:  tg_oj$len and tg_vc$len
## t = 3.3026, df = 29, p-value = 0.00255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.408659 5.991341
## sample estimates:
## mean of the differences 
##                     3.7
```

```r
v_dose <- unique(tg$dose)
tg_supp1 <- tg[tg$dose==v_dose[1],]
tg_supp2 <- tg[tg$dose==v_dose[2],]
tg_supp3 <- tg[tg$dose==v_dose[3],]
```

```r
t.test(tg_supp1$len, tg_supp2$len, paired=TRUE, var.equal=TRUE)
t.test(tg_supp1$len, tg_supp3$len, paired=TRUE, var.equal=TRUE)
t.test(tg_supp2$len, tg_supp3$len, paired=TRUE, var.equal=TRUE)
```

State your conclusions and the assumptions needed for your conclusions We deny H0 because: * Statistic t is equal to 3.3026 so is greater than quartile t at 95% (1.75305) * p-value is 0.00255 so is lower than 5% In conclusion mean from supp OJ is different than mean from supp VC with a Type I error lower than 5%.