Statistical Inference Course Project

Oleksandr Fialko

11/26/2016

Contents

Part 1: Central Limit Theorem 1
Part 2: Basic Inferential Data Analysis 2
Results: 4

Part 1: Central Limit Theorem

The Central Limit Theorem (CLT) states that given a sufficiently large sample size from a population with finite variance, the mean of all samples from the same population will be approximately equal to the mean of the population. The samples means will follow an approximate normal distribution pattern centered at the population mean and the variance being approximately equal to the variance of the population divided by each sample's size.

In this part I investigate the exponential distribution in R and compare it with the CLT. 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`.

Here, I investigate the distribution of averages of 40 exponentials and do a thousand simulations for `lambda=0.2`:

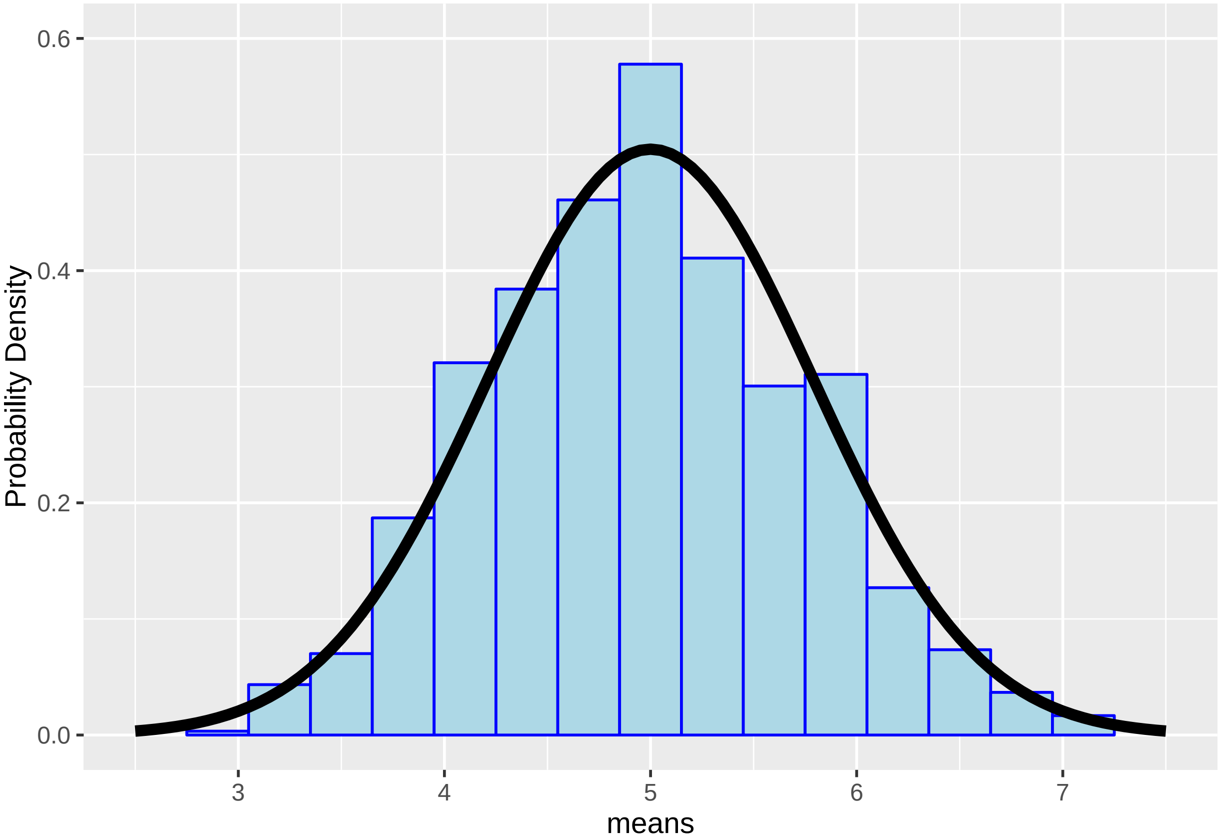
```
set.seed(12345)
lambda <- 0.2  # rate parameter
n <- 40        # a sample size
n_sim <- 1000  # number of simulations
# each row is a sample
sims <- matrix(rexp(n*n_sim,lambda),ncol = n)
```

According to the CLT, the samples means are distributed approximately normally. I create a function `norm_approx` to compare the distribution of the means with:

```
norm_approx <- function(x,mean,sd,n){
    dnorm(x,mean = mean,sd=sd/sqrt(n))
}
```

Here, I plot the normalized histogram of the means and compare it with `norm_approx`:

```
means <- apply(sims, 1, mean)
library(ggplot2)
g<-ggplot(data = data.frame(means=means),aes(x=means))
g<-g+geom_histogram(binwidth=0.3,col='blue',
                    fill='lightblue',center=1/lambda,
                    aes(y=..density..))
g+stat_function(fun=norm_approx,
                args = list(mean=1/lambda,sd=1/lambda,n=n),geom='line',size=2)+
    xlim(2.5,7.5) + ylim(0,0.6) + ylab('Probability Density')
```



Samples means and their standard deviation are

```
c(mean(means),sd(means))
```

```
## [1] 4.9719720 0.7847246
```

These values are in good agreement with the mean and the standard deviation of the `norm_approx`, namely

```
c(1/lambda,1/lambda/sqrt(40))
```

```
## [1] 5.0000000 0.7905694
```

Part 2: Basic Inferential Data Analysis

`ToothGrowth` dataset contains 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 (coded as OJ) or ascorbic acid (a form of vitamin C and coded as VC).

```
data("ToothGrowth")
str(ToothGrowth)
```

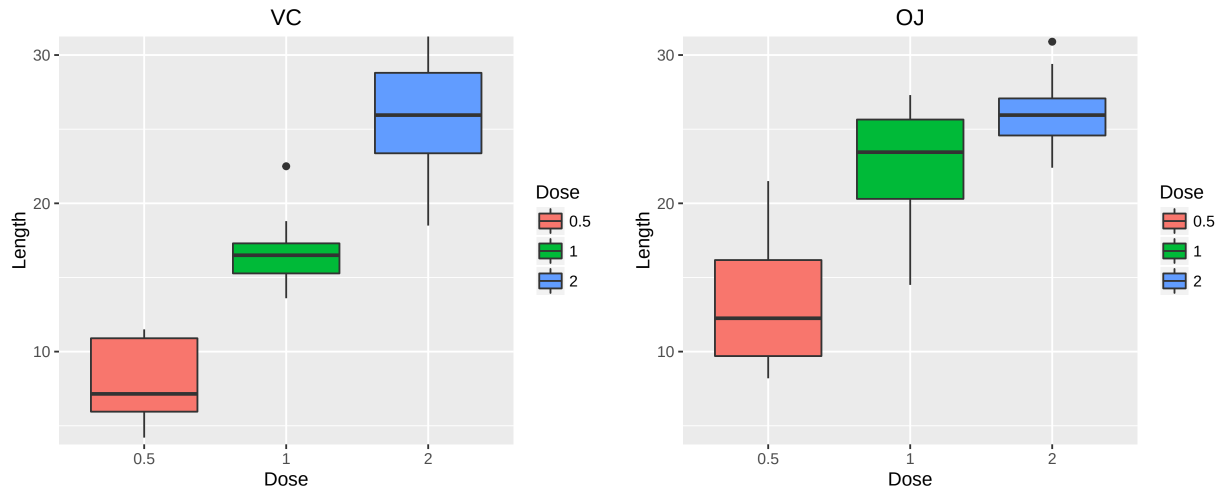
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Subsetting on `OJ` and `CV`:

```
data_OJ <- subset(ToothGrowth,supp=='OJ')
data_VC <- subset(ToothGrowth,supp=='VC')
```

Visualizing the resulting datasets:

```
library(ggplot2)
library(gridExtra)
g1<-ggplot(data_VC,aes(x=as.factor(dose),y=len,fill=as.factor(dose)))
plot1<-g1+geom_boxplot()+labs(fill='Dose',x='Dose',y='Length',title='VC')+
    coord_cartesian(ylim = c(5,30))
g2<-ggplot(data_OJ,aes(x=as.factor(dose),y=len,fill=as.factor(dose)))
plot2<-g2+geom_boxplot()+labs(fill='Dose',x='Dose',y='Length',title='OJ')+
    coord_cartesian(ylim = c(5,30))
grid.arrange(plot1,plot2,ncol=2)
```



Below I run several `t.test` to check different hypothesis derived from the naive visual analysis. I will report p-values. If a p-value is below 0.05, then a hypothesis is correct.

Hypothesis 1: Higher Dose of VC increases the length of odontoblasts

```
d05_VC <- data_VC[data_VC$dose==0.5,]$len
d10_VC <- data_VC[data_VC$dose==1.0,]$len
test<-t.test(d05_VC,d10_VC,var.equal = TRUE,paired = FALSE)
test$p.value
```

```
## [1] 6.492265e-07
```

```
d20_VC <- data_VC[data_VC$dose==2.0,]$len
test<-t.test(d10_VC,d20_VC,var.equal = TRUE,paired = FALSE)
test$p.value
```

```
## [1] 3.397578e-05
```

Hypothesis 2: Higher Dose of OJ increases the length of odontoblasts

```
d05_OJ <- data_OJ[data_OJ$dose==0.5,]$len
d10_OJ <- data_OJ[data_OJ$dose==1.0,]$len
test<-t.test(d05_OJ,d10_OJ,var.equal = TRUE,paired = FALSE)
test$p.value
```

```
## [1] 8.357559e-05
```

```
d20_OJ <- data_OJ[data_OJ$dose==2.0,]$len
test<-t.test(d10_OJ,d20_OJ,var.equal = TRUE,paired = FALSE)
test$p.value
```

```
## [1] 0.0373628
```

Hypothesis 3: OJ is more efficient than VC in growing odontoblasts

```
test<-t.test(d05_OJ,d05_VC,var.equal = TRUE,paired = FALSE)
test$p.value
```

```
## [1] 0.005303661
```

```
test<-t.test(d10_OJ,d10_VC,var.equal = TRUE,paired = FALSE)
test$p.value
```

```
## [1] 0.0007807262
```

```
test<-t.test(d20_OJ,d20_VC,var.equal = TRUE,paired = FALSE)
test$p.value
```

```
## [1] 0.9637098
```

Results:

All three hypothesis seem to be correct except one statement: when the dose is 2 mg/day, the length of odontoblasts seem do not depend on the method of delivery.