Statistical Inference Course Project

KKher

7/18/2020

Overview

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

```
library(ggplot2)
library(dplyr)
```

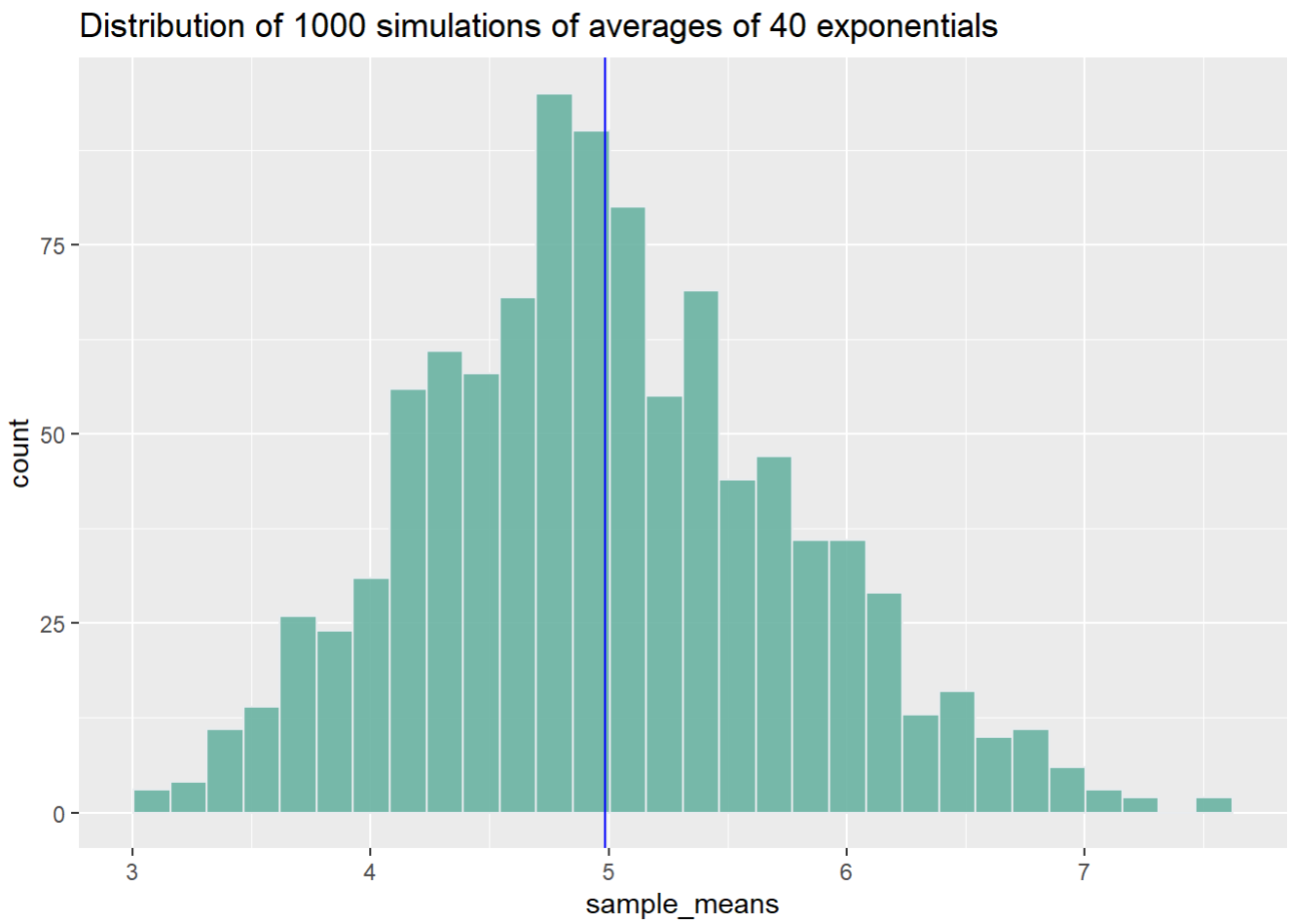
Part 1: Simulation Exercise Instructions

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. - Has mean = 1/lambda & standard deviation = 1/lambda. - Set lambda = 0.2 - Investigate the distribution of averages of 40 exponentials. - Note that you will need to do a 1000 simulations.

```
lambda = 0.2

# theoretical Mean & standard deviation
actual_mean = 1/0.2
actual_sd = 1/0.2

# simulate 1000
sample_means = NULL
for (i in 1 : 1000) sample_means = c(sample_means, mean(rexp(40, lambda)))
ggplot() +
  aes(sample_means) +
  geom_histogram(fill="#69b3a2", color="#e9ecef", alpha=0.9) +
  ggtitle("Distribution of 1000 simulations of averages of 40 exponentials") +
  geom_vline(xintercept = mean(sample_means), color="blue")
```



Actual Mean = `5` and the simulated mean = `4.9871958`

Actual sd = `5` and the simulated mean = `0.7817004`

Part 2: Basic Inferential Data Analysis Instructions

Load the ToothGrowth data and perform some basic exploratory data analyses

```
data(ToothGrowth)
```

Provide a basic summary of the data.

```
head(ToothGrowth)
```

	len	supp	dose
	<dbl>	<fctr>	<dbl>
1	4.2	VC	0.5
2	11.5	VC	0.5
3	7.3	VC	0.5

	len <dbl>	supp <fctr>	dose <dbl>
4	5.8	VC	0.5
5	6.4	VC	0.5
6	10.0	VC	0.5

6 rows

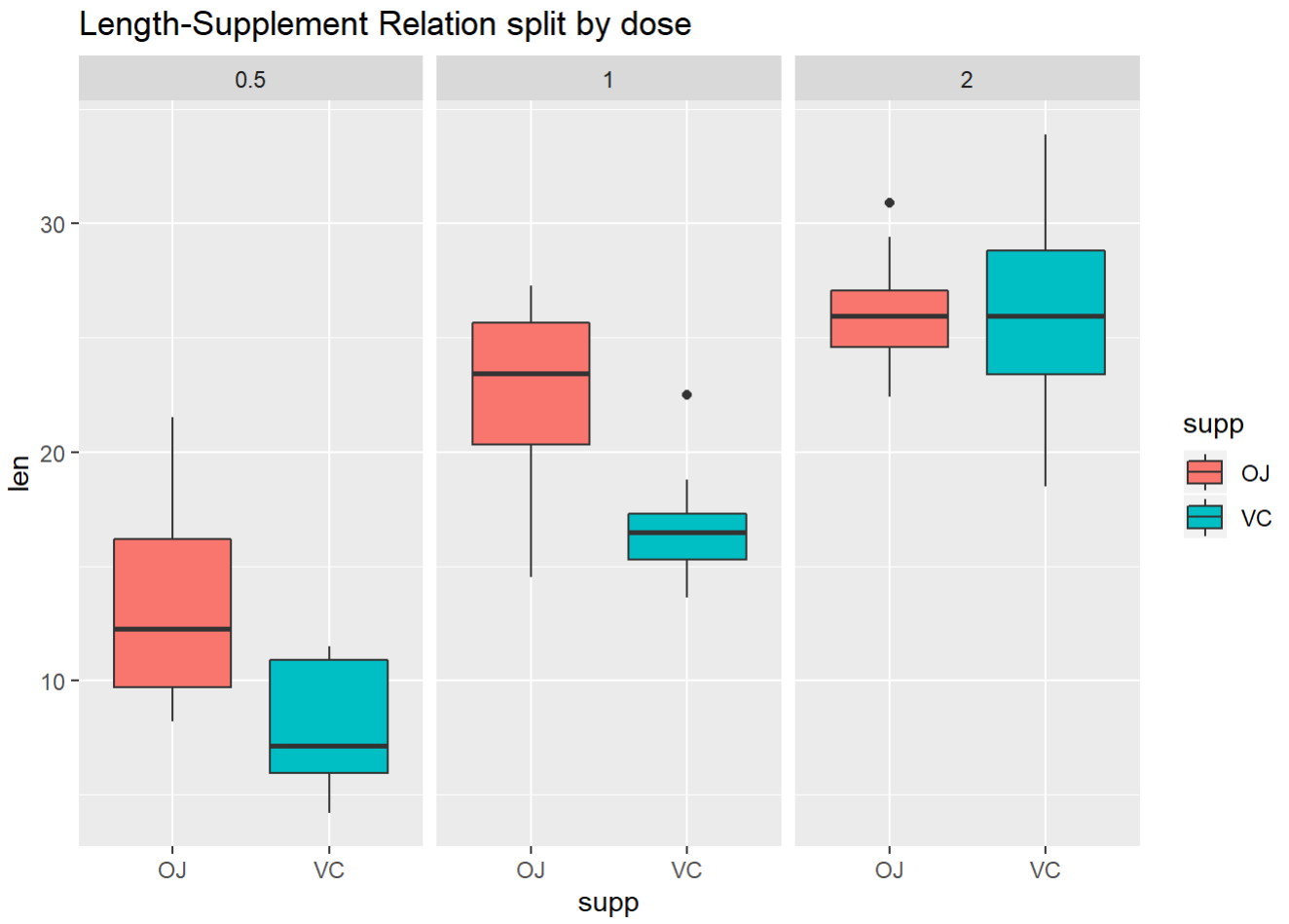
```
summary(ToothGrowth)
```

```
##       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
```

```
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 ...
```

```
# plot len & supp
ggplot(data=ToothGrowth) +
  aes(x=supp, y=len)+
  geom_boxplot(aes(fill=supp)) +
  facet_grid(cols = vars(dose)) +
  ggtitle("Length-Supplement Relation split by dose")
```



From Figure above, mean seems to be equal for both supp only for dose=2.

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)

```
# create t test

# perform t test between supp types where dose = 2
t.test(ToothGrowth$len[ToothGrowth$supp=="OJ" & ToothGrowth$dose==2], ToothGrowth$len[ToothGrowth$supp=="VC" & ToothGrowth$dose==2], paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[To
othGrowth$supp == "VC" & ToothGrowth$dose ==     2] and     2]
## t = -0.042592, df = 9, p-value = 0.967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of the differences
##                   -0.08
```

```
# perform t test between supp types where dose != 2
t.test(ToothGrowth$len[ToothGrowth$supp=="OJ" & ToothGrowth$dose!=2], ToothGrowth$len[ToothGrowt
h$supp=="VC" & ToothGrowth$dose!=2], paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose !=  and ToothGrowth$len[To
othGrowth$supp == "VC" & ToothGrowth$dose !=     2] and     2]
## t = 4.6042, df = 19, p-value = 0.0001936
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.048852 8.131148
## sample estimates:
## mean of the differences
##                    5.59
```

State your conclusions and the assumptions needed for your conclusions.

If we consider Null Hypothesis (H0) to be; mean is almost equal per supp per dose

1- We Fail to reject H0 where dose = 2, as t-value is very small and is equal to `-0.042592`

2- We reject H0 where dose does not equal to 2, as t-value is large enough and is equal to `4.6042202`