Simulation and inferential data analysis

Kirill Setdekov

13 may 2019

Overview

In the first part of this assignment, I will be exploring results of the simulation of exponential distribution and properties of the sample mean. In the second part of the assignment, I will explore the tooth growth dataset and apply statistical inference tools learned from the course.

Part 1.A simulation exercise.

Simulations:

I simulate 1000 means of 40 exponential distribution and save them in a vector **mns**.

```
#load libraries silently
require(knitr)
require(ggplot2)
require(ggpubr)
```

```
set.seed(42)
n <- 40
lambda <- 0.2
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(n, lambda)))
```

Sample Mean versus Theoretical Mean:

Comparison of Sample mean an theoretical mean in the table below:

```
theory_mean <- 1/lambda
sample_mean <- mean(mns)
kable(cbind(theory_mean,sample_mean), digits = 3)
```

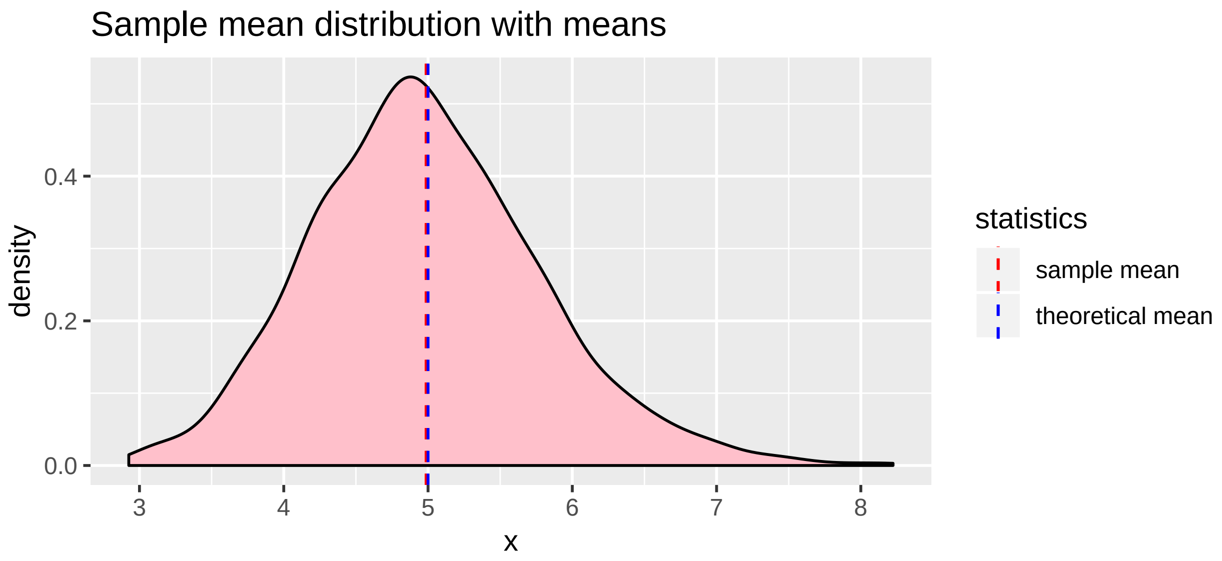
theory_mean	sample_mean
5	4.987

They are very close.

For more detail, I plot the distribution of simulated means together with lines for sample and theoretical mesns below:

```
ggplot(data.frame(x = mns), aes(x = x)) +
geom_density(fill = "pink") +
geom_vline(aes(xintercept = mean(mns),color = "sample mean"),
           linetype="dashed") +
geom_vline(aes(xintercept = 1 / lambda,color = "theoretical mean"),
           linetype="dashed")+
scale_color_manual(name = "statistics",
                   values = c('sample mean' = "red", 'theoretical mean' = "blue")) +
```

```r
labs(title = "Sample mean distribution with means")
```



Sample Variance versus Theoretical Variance:

Below is a calculation and comparison of sample variance and theoretical variance. They are close as the number of simulations (1000) is significant.

```r
theory_variance <- (1/lambda)/sqrt(40)
sample_variance <- sd(mns)
kable(cbind(theory_variance,sample_variance), digits = 3)
```

theory_variance	sample_variance
0.791	0.797
## Distribution	

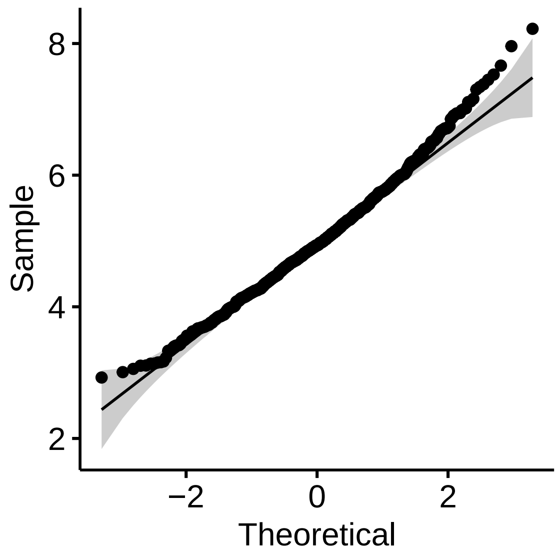
Based on the **Shapiro-Wilk's test** the distribution of the means of 40 exponential random variables are different from normal as p-value<0.05.

```r
shapiro.test(mns)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mns
## W = 0.99084, p-value = 6.987e-06
```

Also, looking at the **qq-plot** below, we can see, that the right tail of the distribution is significantly different from the normal one. We have a large sample and we can not assume normality.

```r
ggqqplot(mns)
```



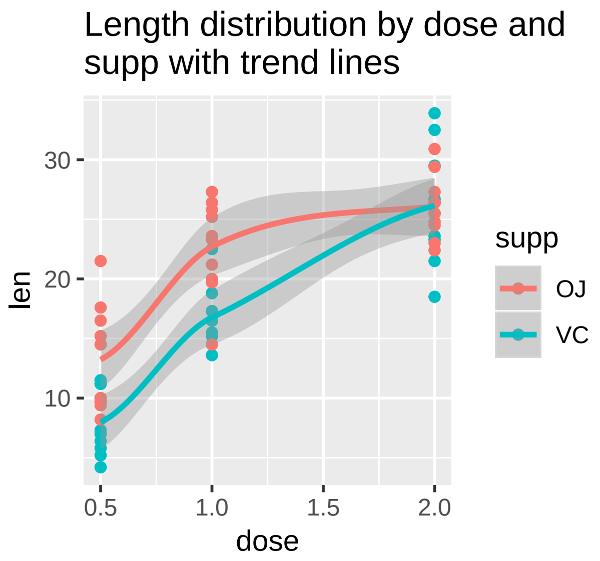
Basic inferential data analysis.

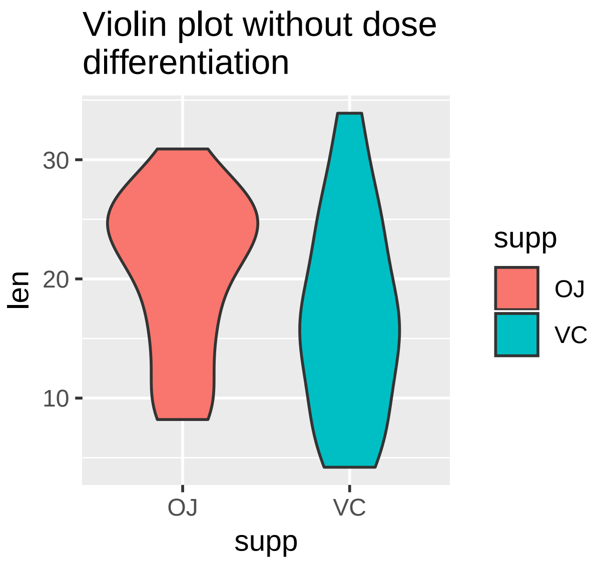
```
tooth <- datasets::ToothGrowth

summary(tooth)
```

```
##       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  
```

```
library(ggplot2)
library(gridExtra)
p1 <-
    ggplot(tooth, aes(x = dose, y = len, color = supp)) +
    geom_point() + geom_smooth() +
    labs(title = "Length distribution by dose and \nsupp with trend lines")
p2 <-
    ggplot(tooth, aes(y = len, x = supp, fill = supp)) + geom_violin() +
    labs(title  = "Violin plot without dose \ndifferentiation")
grid.arrange(p1, p2, nrow = 1)
```


Length distribution by dose and supp with trend lines


Violin plot without dose differentiation

testing

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)

conclusions

State your conclusions and the assumptions needed for your conclusions.

review criteria

- Did you show where the distribution is centered at and compare it to the theoretical center of the distribution?
- Did you show how variable it is and compare it to the theoretical variance of the distribution?
- Did you perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data?
- Did the student perform some relevant confidence intervals and/or tests?
- Were the results of the tests and/or intervals interpreted in the context of the problem correctly?
- Did the student describe the assumptions needed for their conclusions?