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)
require(gridExtra)
require(dplyr)
```

```
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)
```

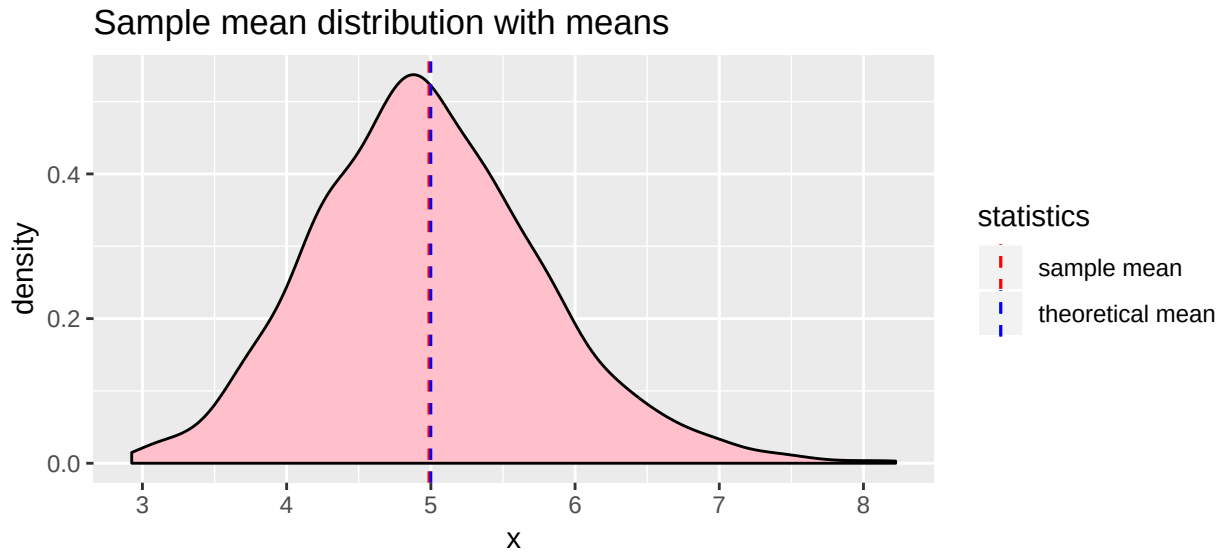
theory_mean	sample_mean
5	4.987

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")) +
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.

```
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

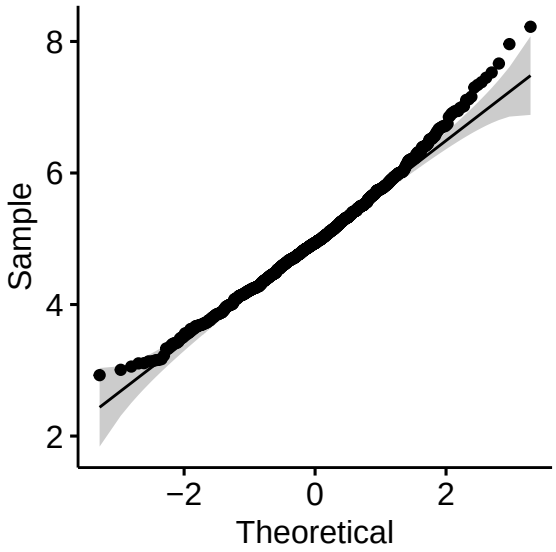
Based on the **Shapiro-Wilk's test** the distribution of the means of 40 exponential random variables are different from normal as p-value<0.05.

```
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.

```
ggqqplot(mns)
```



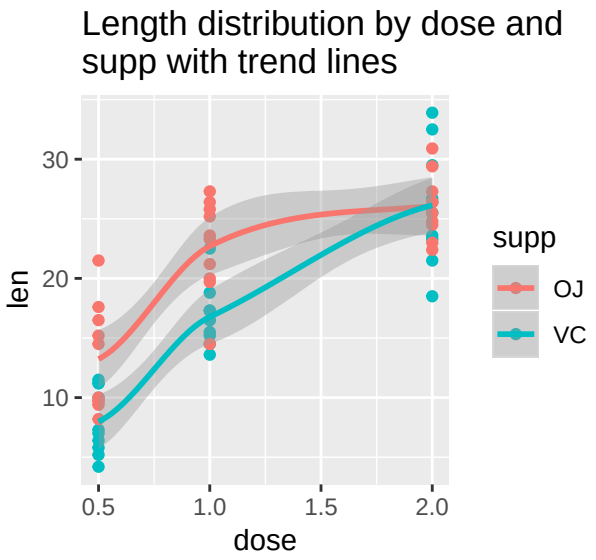
Part 2. Basic inferential data analysis.

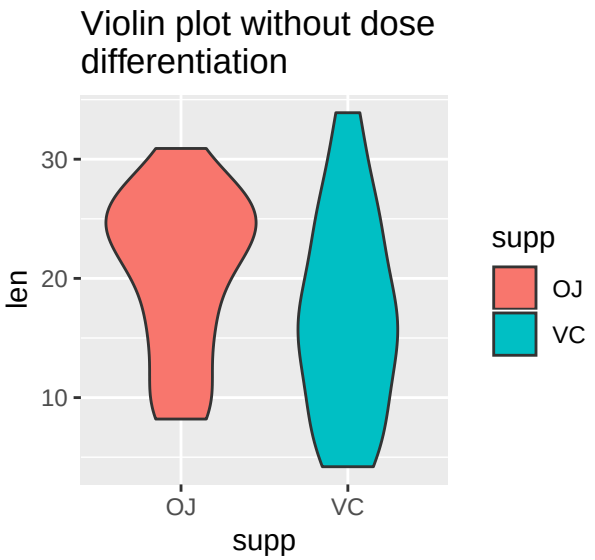
"Tooth" dataset consists of 60 observatiosn of measurements of tooth growth related cells in 60 guinea pigs. There were 3 dose levels of vitamin C (encoded as "dose") and 2 types of delivery methods, encoded as "supp", where VC is vitamin C and OJ is orange juice.

```
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
```

```
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)
```





On the plots we can see that Vitamin C leads to higher variance in tooth growth in total. If we look at mean length by supp and dose, we can see that it increases with dose and is higher for OJ than for VC for all doses, excep 2.0 mg.

```
kable(tooth %>% group_by(supp, dose) %>% summarise(mean(len)))
```

supp	dose	mean(len)
OJ	0.5	13.23
OJ	1.0	22.70
OJ	2.0	26.06
VC	0.5	7.98
VC	1.0	16.77
VC	2.0	26.14

testing

Tooth growth vs supp testing

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

```
##
##  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
```

We conduct t-test for difference in means between group OJ and VC. We can see that the confidence interval includes zero and p-value > 0.05. We cannot reject the H_0 of no difference between these groups in terms of mean tooth growth.

However, if we look at permutation testing and compare mean tooth growth for OJ and VC, we can reject H_0 at $\alpha = 0.05$.

```
toothshort <- tooth
y <- toothshort$len
group <- as.character(toothshort$supp)
testStat <- function(w, g)
    mean(w[g == "OJ"]) - mean(w[g == "VC"])
observedStat <- testStat(y, group)
permutations <-
    sapply(1:10000, function(i)
        testStat(y, sample(group)))
# difference in average count between b and c
observedStat
```

```
## [1] 3.7
```

```
# share of more extreme
mean(permutations > observedStat)
```

```
## [1] 0.0318
```

```
# pvalue is very low - reject H0
```

Tooth growth vs dose testing

```
tooth2doses <- subset(tooth, dose %in% c(0.5,2))
t.test(len~dose, paired = FALSE, var.equal = FALSE, data = tooth2doses)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

We conduct t-test for difference in means between group doses of 0.5 and 2 mg. We can see that p-value $<$ 0.01. We can reject the H_0 of no difference between mean tooth growh for low and high dosages of Vitamin C.

conclusions

Based on t-test and visual analysis of the data, we conclude that the amount of vitamin C consumed is linked to tooth growth, however the type of delivery in genereal does not lead to a statistically significant difference in tooth growth.

For these two tests I used the following assumptions:

- These are unpaired observations;
- Variance in unequal between tested groups.

Using perputation testing, we can say that at $\alpha = 0.05$, there is higher mean tooth growth for OJ, but this is not significant at a higher confidence level.