Statistical Inference Course Project

E. Slavyaninov

01.07.2019

Part 1. Simulation

In this part I investigate the distribution of averages from exponential distribution. The mean and stadart deveation of the distribution correspond to the theoretical values and the distribution itself is normal.

At first, dowmload required libraries and genetate sample: 1000 avearges of 40 exponenential from exponrntial distribution with lambda = 0.2. To do this, use the function sapply and save the result to the variable Sample

```
library(tidyverse)
lambda = 0.2
Sample<-sapply(1:1000, function(i) mean(rexp(40,lambda)))
```

1.1 Sample mean exploration

Copmare sample mean (from generated sample) with theoretical mean (1/lambda)

```
mean(Sample)
```

```
## [1] 4.981403
```

```
1/lambda
```

```
## [1] 5
```

As we can see values are very close

1.2 Sample standard deviation exploration

Copmare sample standard deviation with theoretical standard deviation (1/lambda)/sqrt(number of observations)

```
sd(Sample)
```

```
## [1] 0.7887039
```

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

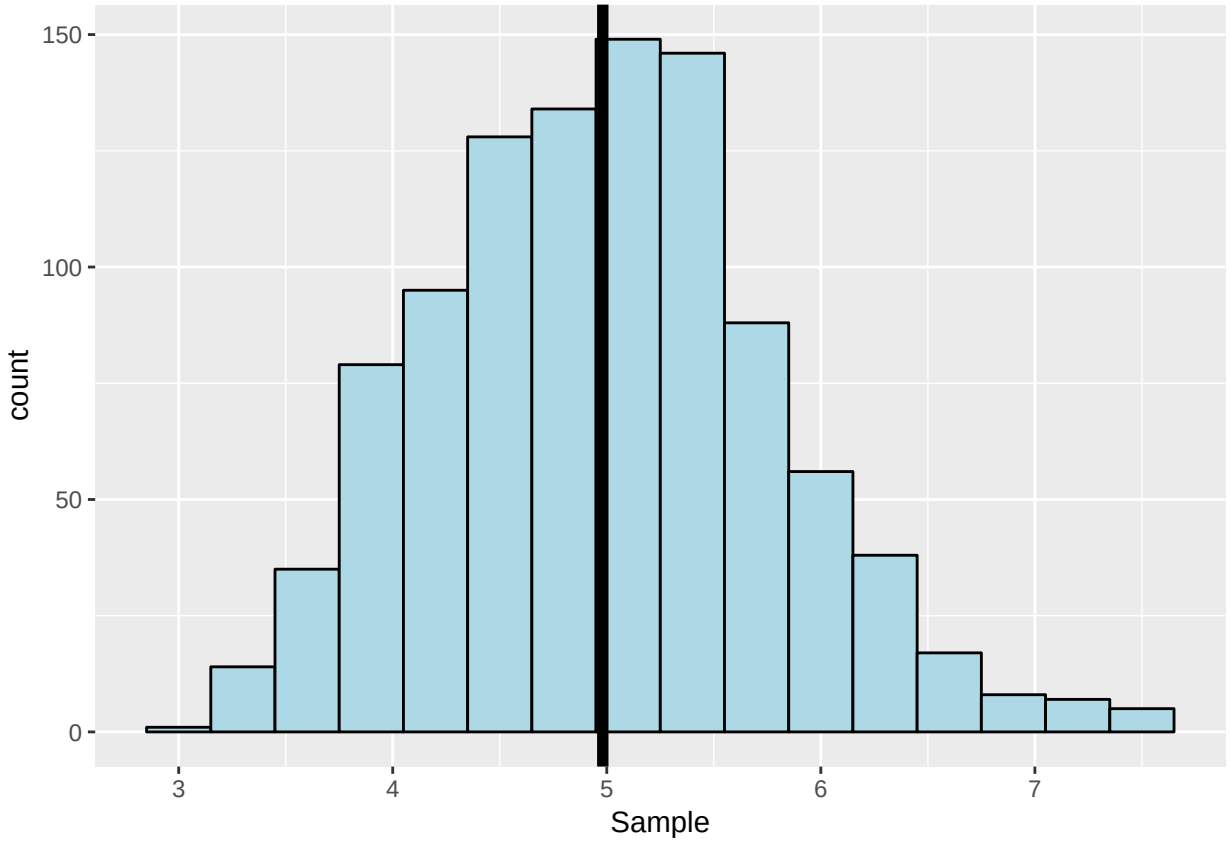
```
## [1] 0.7905694
```

As we can see values are very close too

1.3 Sample normality exploration

Check out the visualisation of the data to normality by histogram

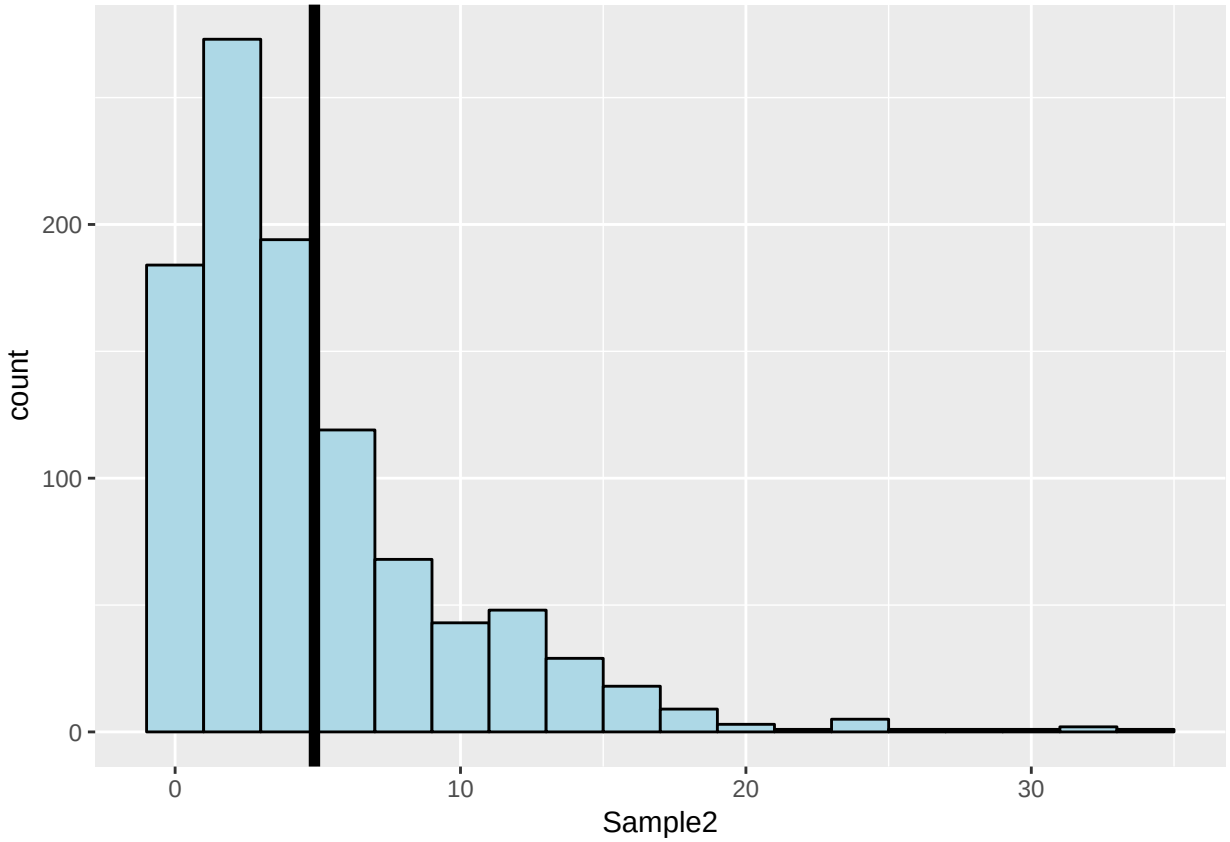
```
ggplot(as.data.frame(Sample), mapping = aes(Sample))+
    geom_histogram(col="black", fill="lightblue",binwidth = 0.3)+
    geom_vline(xintercept = mean(Sample),lwd=2)
```



The distribution of data looks absolutely normal and symmetrical

Compare sample distribution with distribution of a large collection of random exponentials (without averegies): Sample2. Generate it and check out the visualisation of the data to normality by histogram

```
Sample2<-rexp(1000,0.2)
ggplot(as.data.frame(Sample2), mapping = aes(Sample2))+
    geom_histogram(col="black", fill="lightblue",binwidth = 2)+
    geom_vline(xintercept = mean(Sample2),lwd=2)
```



The distribution does not look normal and symmetrical. **This shows that averaging the data leads the distribution to the normality.**

Part 1. Inferential Data Analysis

2.1 Basic exploratory data analyses

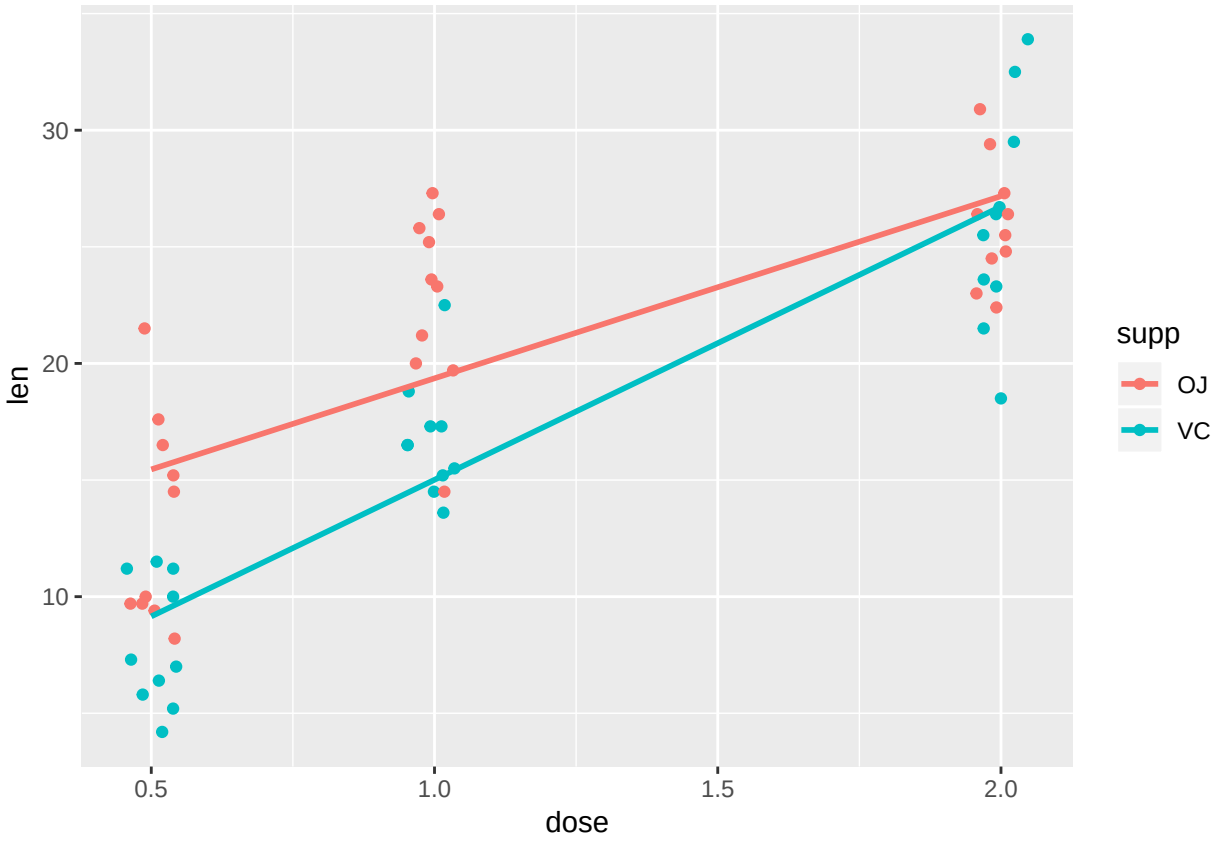
Load the ToothGrowth data and discover it structure

```
Data<-ToothGrowth
str(Data)
```

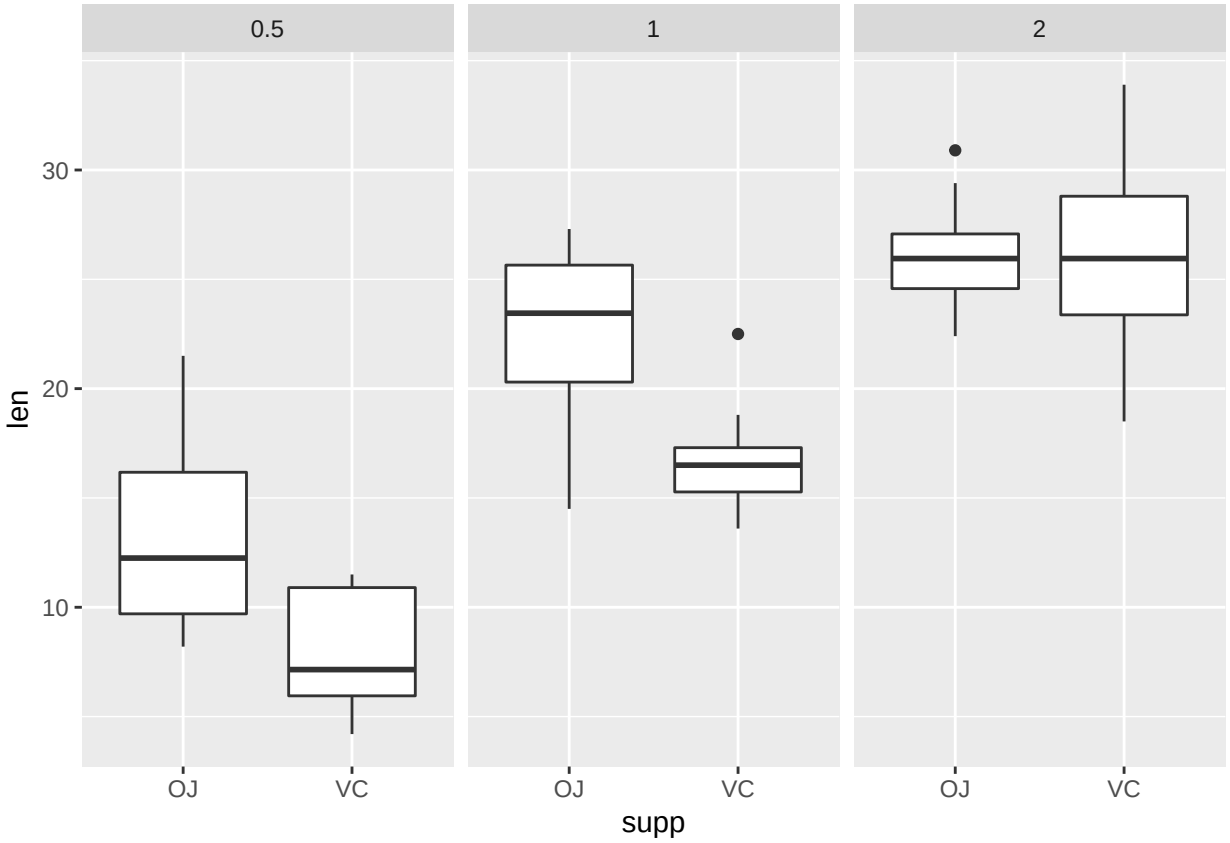
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Make same visualisations

```
ggplot(Data,aes(dose,len))+geom_jitter(aes(color=supp),width = 0.05,height = 0)+
    geom_smooth(aes(color=supp),method = "lm",se = FALSE)
```




```
ggplot(Data,aes(supp,len))+geom_boxplot(aes(group=supp))+facet_grid(.~dose)
```



2.2 Basic summary of the data

Show the number of observations, minimum, maximum, mean, and standard deviation of teeth length by Supplement type and Dose

```
group_by(Data,supp,dose)%>%summarise(NObs=n(),Min=min(len), max = max(len),Mean=mean(len),Sd=sd(len))
```

```
## # A tibble: 6 x 7
## # Groups:   supp [?]
##   supp   dose  NObs   Min   max  Mean    Sd
##   <fct> <dbl> <int> <dbl> <dbl> <dbl> <dbl>
## 1 OJ      0.5    10   8.2  21.5 13.2   4.46
## 2 OJ      1      10  14.5  27.3 22.7   3.91
## 3 OJ      2      10  22.4  30.9 26.1   2.66
## 4 VC      0.5    10   4.2  11.5  7.98  2.75
## 5 VC      1      10  13.6  22.5 16.8   2.52
## 6 VC      2      10  18.5  33.9 26.1   4.80
```

2.3 Hypothesis testing

At first, compare the growth of teeth by Supplement type. Our null hypothesis is that there is no difference between them. Use t.test function

```
t.test(Data$len[Data$supp=="OJ"],Data$len[Data$supp=="VC"])
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  Data$len[Data$supp == "OJ"] and Data$len[Data$supp == "VC"]
## 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 of x mean of y
##  20.66333  16.96333
```

We fail to reject null hypothesis: p-value exceeds 5 percents and 95 percent confidence interval includes 0.

Then compare the growth of teeth by Supplement type, when dose of vitamin C equal 0.5 milligrams/day. Calculate the required values without using t.test function. Denote two sets of data for comparison d1 and d2.

```
d1<-Data$len[Data$supp=="OJ"& Data$dose==0.5]
d2<-Data$len[Data$supp=="VC"& Data$dose==0.5]
```

Find mean, standart deviation and length of the sets, group standart deviation and t-statistic

```
m1<-mean(d1); m2<-mean(d2); sd1<-sd(d1); sd2<-sd(d2); n1<-length(d1); n2<-length(d2)
sd<-sqrt((sd1^2*(n1-1)+sd2^2*(n2-1))/(n1+n2-2))
t<-(m1-m2)/(sd*sqrt(1/n1+1/n2))
```

Finally, calculate p-value and confidence interval

```
pt(t,18,lower.tail = FALSE)*2
```

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

```
(m1-m2)+c(-1,1)*qt(0.975,18)*sd*sqrt(1/n1+1/n2)
```

```
## [1] 1.770262 8.729738
```

As we can see, our p-value is small and confidence interval doesn't include 0, so we rejecte null hypothesis

2.4 Conclusions and the assumptions

When comparing an entire data set by supplement type we fail to reject null hypothesis (there is no differencies) with alpha = 0.05. But when we compared inside dose = 0.5 milligrams/day null hypothesis was rejected.

Our calculation included the assupmtionr, that data includes Independent and identically distributed variables.