Statistical Inference Course Project

Yoni Wainsztok

04/01/2018

This report is a key requirement/assignment of the Statistical Inference Course by Johns-Hopkins University. It consists of two parts: A simulation exercise and a basic inferential data analysis. In order to fulfill this report several packages must be loaded:

```
#Loading necessary packages
library(datasets)
library(ggplot2)
library(knitr)
library(dplyr)
```

Part 1: Simulation Exercise

Comparison between Exponential Distribution (EXPD) and the Central Limit Theorem (CLT).

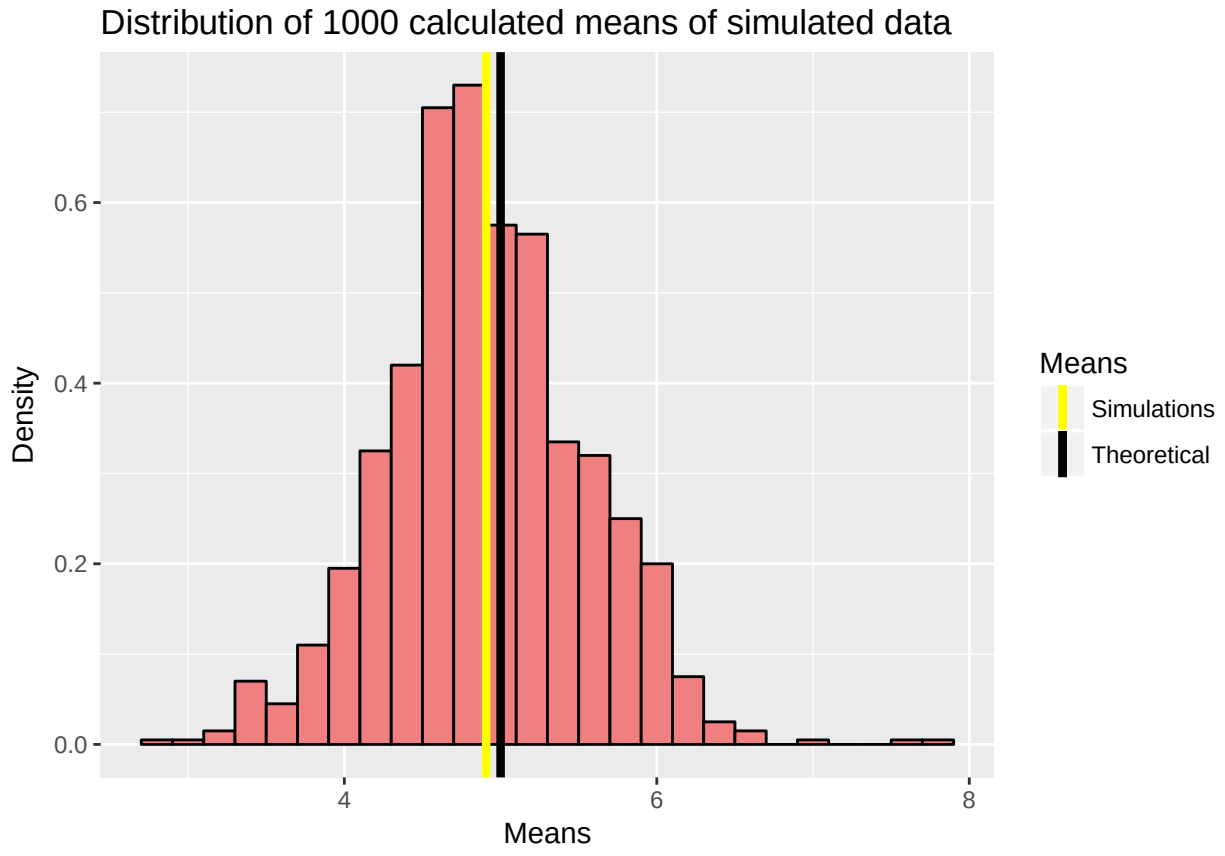
The simulations described in the code below are based on a random exponential distribution of 40 values generated by rexp(n, lambda) command:
1. Generating known variables (lambda, n, Theoretical Mean, Theoretical Standard Deviation, Theoretical Variance).
2. Creating the original data (random exponential distribution of 40 values).
3. Define simulation properties (1000 simulations) and execution (implementation into a matrix - each row is considered as one simulation).

```
#Defining variables & defaults of mean, sd and variance
lambda<-0.2
n<-40
teoreticMean<-1/0.2
teoreticSd<-teoreticMean/sqrt(n)
teoreticVar<-teoreticSd^2
#Creating the random exponential distribution of 40 values
originalData<-rexp(n, lambda)
#Simulation properties & execution (based on the originalData)
simNum<-1000
#each row consists one simulation values
resamples<-matrix(sample(originalData,n*simNum,replace = T),simNum,n)
#Extracting mean from each simulation
means<-apply(resamples,1,mean)
```

Mean

The next steps are to show in a histogram what is the distribution of 1000 mean values calculated from each simulation and compare this distribution to the theoretical mean (1/lambda). The graph below also shows vertical lines which represent the theoretical mean (5) and the calculated mean of the generated distribution (4.91).

```r
#Creating the histogram using ggplot
meansDf<-as.data.frame(means)
histm<-ggplot(meansDf, aes(x=means, color=means))
histm<-histm+geom_histogram(binwidth = lambda,fill="lightcoral",color="black",
                            aes(y=..density..))
histm<-histm+labs(title="Distribution of 1000 calculated means of simulated data",
                  x="Means", y="Density")
histm<-histm+geom_vline(aes(xintercept = mean(meansDf$means),
                            color="Simulations"), size=1.5) +
  geom_vline(aes(xintercept = teoreticMean,color="Theoretical"),
             size=1.5)+
  scale_color_manual(name = "Means", values = c(Simulations = "yellow",
                                                Theoretical = "black"))
histm
```



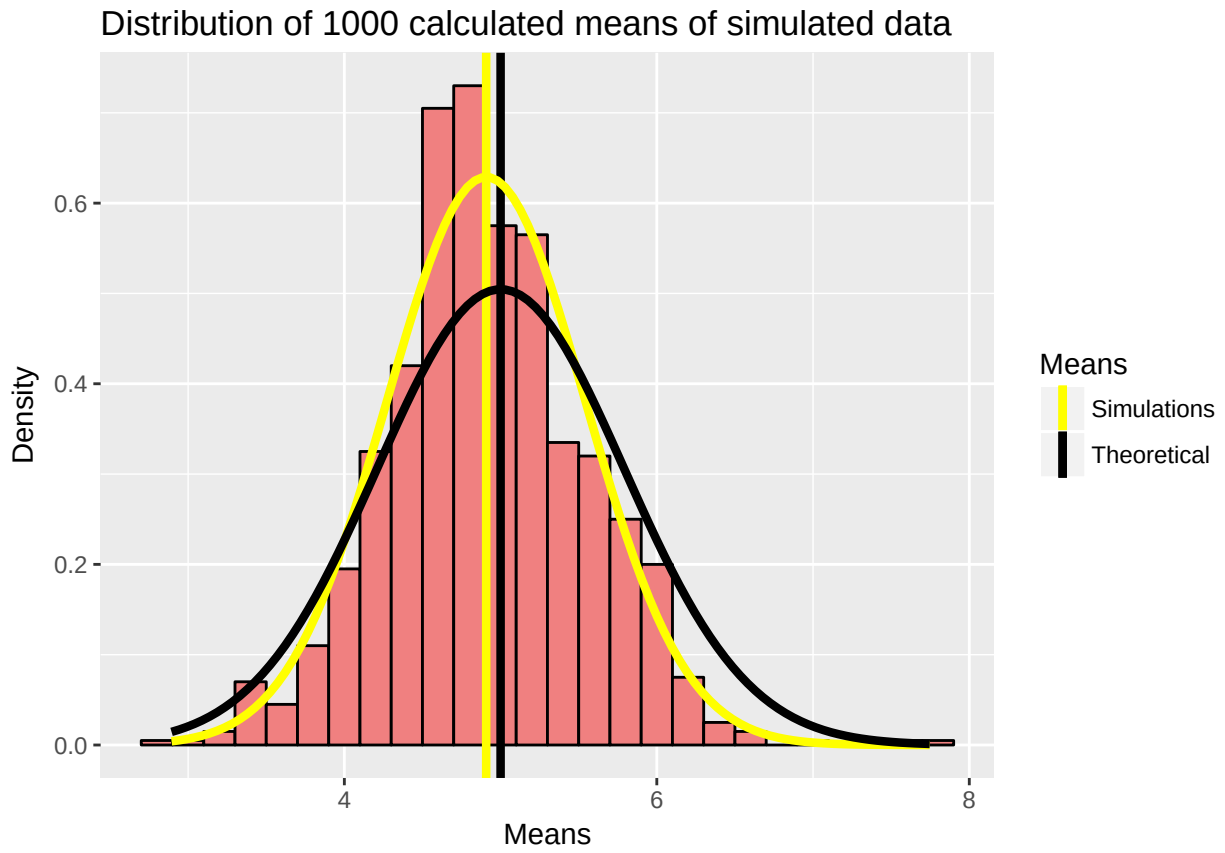
Variance

Addition of the distribution curves of the theoretical values (mean and standard deviation) and the simulated calculated values can show the difference of the variance and standard deviation between the two distributions (see modified graph below). Broader distribution (and lower peak) means higher variance and vice versa. The standard deviation values are 0.79 (Theoretical) and

0.63

(Simulated). It is clear that the two distributions are approximately normal and are following the CLT.

```r
#Adding distributions to the previous graph.
histm<-histm+stat_function(fun=dnorm,
                           args=list(mean=mean(meansDf$means),
                                     sd=sd(meansDf$means)),color = "yellow",
                           size = 1.5)
histm<-histm+stat_function(fun=dnorm,
                           args=list(mean=teoreticMean, sd=teoreticSd),
                           color = "black", size = 1.5)
histm
```



Summary between the two distributions is given in the table below:

Variable	Theoretical	Simulated
Mean	5	4.91
Standard Deviation	0.79	0.63
Variance	0.62	0.4

Part 2: Basic Inferential Data Analysis

This section focuses on basic data and statistical analysis of the ToothGrowth dataset (The Effect of Vitamin C on Tooth Growth in Guinea Pigs):

1. Basic analysis and summary of the data.
2. Use of confidence intervals and hypothesis tests to compare tooth growth by supp and dose.
3. Based conclusions.

Basic analysis and summary of the data

The next tables show statistic summary of the ToothGrowth dataset. This dataset consists 60 observations of 3 variables:

- **len** - Length of odontoblasts (teeth).
- **supp** - Supplement type of vitamin C (2 levels only): orange juice (OJ) or ascorbic acid (VC).
- **dose** - Dose in milligrams/day (values of 0.5, 1, and 2 only).

```
#Loading ToothGrowth dataset and its summary
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 ...
```

```
kable(summary(ToothGrowth), align = 'c')
```

len	supp	dose
Min. : 4.20	OJ:30	Min. :0.500
1st Qu.:13.07	VC:30	1st Qu.:0.500
Median :19.25	NA	Median :1.000
Mean :18.81	NA	Mean :1.167
3rd Qu.:25.27	NA	3rd Qu.:2.000
Max. :33.90	NA	Max. :2.000

```
suppGrouped <- group_by(ToothGrowth, supp)
summary <- summarise(suppGrouped, count= n(),
                     "Mean"=mean(len),
                     "Median"=median(len),
                     "Standard deviation" = sd(len))
summData <- as.data.frame(summary)
kable (summData,digits = 3,align = 'c')
```

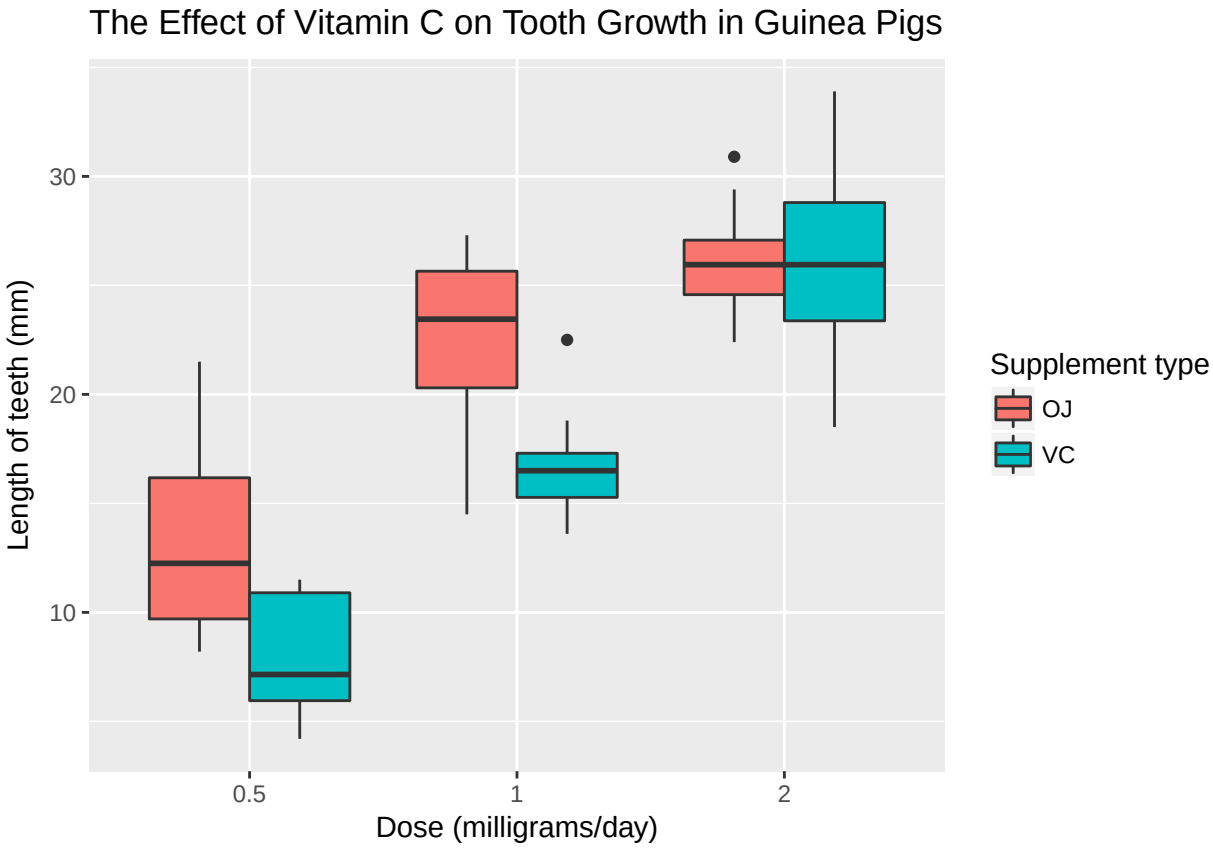
supp	count	Mean	Median	Standard deviation
OJ	30	20.663	22.7	6.606
VC	30	16.963	16.5	8.266

The tables shows several key points:

- The average lenth of teeth is 18.81mm.
- Observations by supp are divided equally.
- dose variable requires more investigation due to its data-type.
- The average teeth length of OJ (Orange juice) is gretear than VC (ascorbic acid). Also, the standard deviation is narrower in the case of OJ.

The graph below was created in order to show these findings visually and to add the dimension of the dose quantity. It is clear that increase in dose quantity results in longer teeth. Also obvious is the difference between the supplement types, especially in the smaller doses.

```
g<-ggplot(ToothGrowth,aes(x=as.factor(dose),y=len, fill=supp))
g<-g+geom_boxplot()
g<-g+labs(title="The Effect of Vitamin C on Tooth Growth in Guinea Pigs",
          x="Dose (milligrams/day)", y="Length of teeth (mm)")
g$labels$fill <- "Supplement type"
g
```



Comparison of tooth growth by supplement type and dose quantity (statistical testing)

Since the information on this study is limited and the number of observations is relatively low, a number of assumptions (which are not necessarily correct) should be taken into consideration:

- The data was collected from a representative, randomly selected portion of the total population.
- The data is distributed normally or close to it.
- Homogeneity of variance.

Comparison by Dosage

Comparison of the teeth length between the different dose quantites was done by three t-tests (Alpha=0.05) below (p-values and confidence intervals are summarised afterwards). The null hypothesis is that there is no difference of the teeth length between the different dose quantities.

```r
# subsets of each dose type
dose0.5v1<-filter(ToothGrowth,dose<2)
dose0.5v2<-filter(ToothGrowth,dose!=1)
dose1v2<-filter(ToothGrowth,dose>0.5)
# t-tests between all dose types
t.test(len~dose, dose0.5v1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```r
t.test(len~dose, dose0.5v2)
```

```
##
##  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
```

```r
t.test(len~dose, dose1v2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

Comparison of dosage	Confidence Interval	P-value
0.5 vs 1.0	$-11.9837813, -6.2762187$	1.2683007×10^{-7}
0.5 vs 2.0	$-18.1561665, -12.8338335$	4.397525×10^{-14}
1.0 vs 2.0	$-8.9964805, -3.7335195$	1.9064295×10^{-5}

The calculated confidence intervals and p-values of the t-tests clearly show that there is difference between the dose quantities regarding the teeth length. Hence we can reject the null hypothesis. This also can be seen

in the graph above.

Comparison by Supplement type

Comparison of the teeth length between the different supplement types (OJ and VC) was done by a single t-test (Alpha=0.05) which can be seen below. The null hypothesis is that there is no difference of the teeth length between the different supplement types.

```
t.test(len~supp, ToothGrowth)
```

```
##
##  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
```

Results of the above t-test show that we cannot reject the null hypothesis. Hence, there is no significant difference between the supplement types regarding teeth length. This significance can be achieved by decreasing the confidence interval (which is usually not recommended).

Conclusion

According to the t-tests applied in the previous section and according to the summarised data, it is clear that dose quantity of vitamin C affects the measured length of teeth in Guinea pigs - length increases with dose quantity. On the other hand, no difference was found between the supplement types. Therefore, teeth length of Guinea pigs is not affected by this factor.
It is important to mention that the influence of both dose quantity and supplement types on the teeth length was not tested. In such scenario, supplement type of low dosage may have effect on the length of teeth.
It is also important to note that the tested dataset has only few observations (60) which can contradict some of the assumptions made for this analysis (Normality, Representation of the population, Homogeneity of variance).