Peer-graded Assignment: Statistical Inference Course Project

Marco Pasin

19 aprile 2017

Instructions

The project is part of the course “Statistical Inference” offered by Johns Hopkins University through Coursera. It consists of two parts:

- A simulation exercise.
- Some basic inferential data analysis.

Part 1: Simulation Exercise

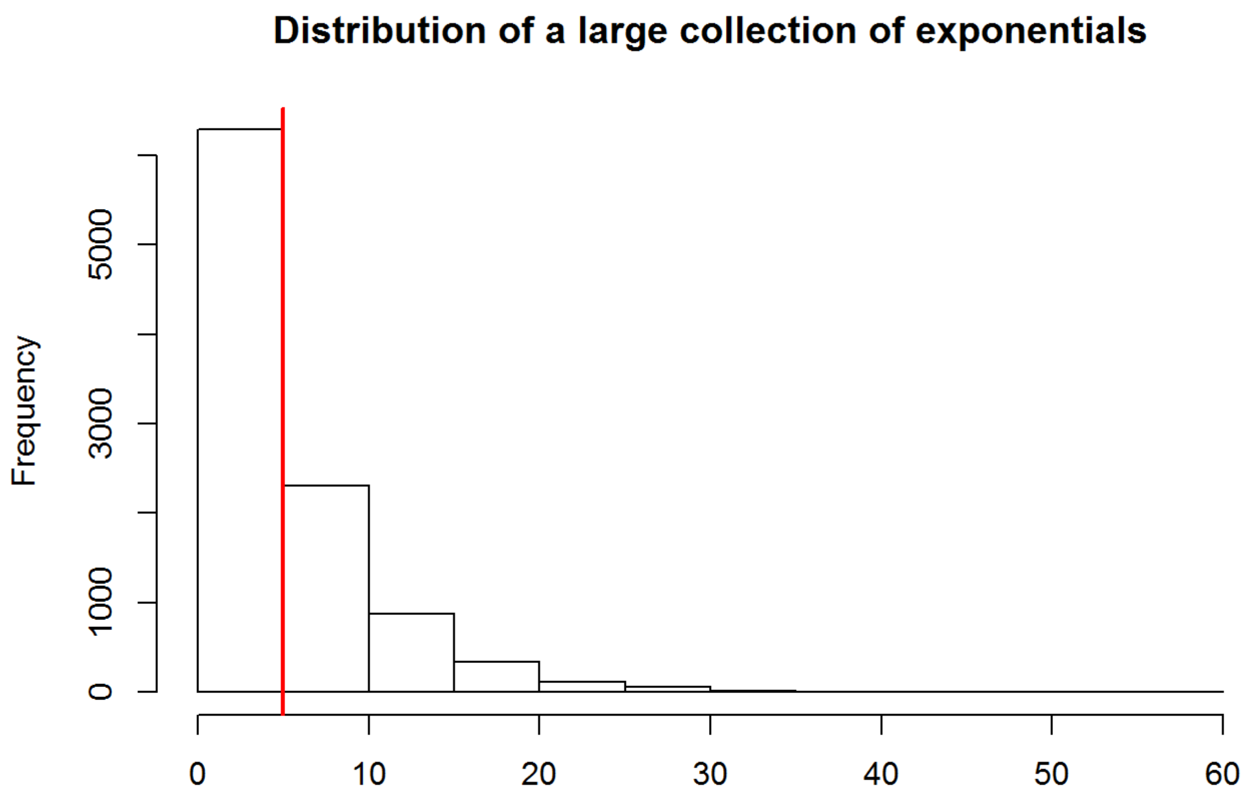
In this part we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. We will investigate the distribution of averages of 40 exponentials. Note that we will need to do a thousand simulations.

Set parameters of theoretical distribution

```
lambda<-0.2

#The theoretical mean is 1/lambda, which is 5
theory_mn<-1/lambda

#Standard deviation is also 1/lambda
sd<-1/lambda
```

Let’s build a theoretical distribution made of a large collection of random exponentials, e.g. 10000 exponentials and plot it with a histogram to see what the distribution looks like.

```
theory_distrib<-rexp(10000,lambda)

hist(theory_distrib,main = "Distribution of a large collection of exponentials",xlab="")
#Mean of the theoretical distribution was calculated earlier and we know is 5
abline(v=theory_mn, col="red",lwd=2)
```



Perform simulations

To simulate the distribution of averages of 40 exponentials we use a for loop. Note the loop interates 1000 times, each time calculating the mean of 40 exponentials. The result of each simulation is attached into a character vector named “sim_mns”, which starts empty and end up containing 1000 values (each one is a mean)

```
set.seed(23)

sim_mns = NULL
for (i in 1 : 1000) sim_mns = c(sim_mns, mean(rexp(40,lambda)))
```

Compare sample mean with theoretical mean

We already know that the theoretical mean is 1/lambda, that is equal to 5.

```
#As we can see the sample mean (the mean of the 1000 simulations we performed earlier) is aro
und 4.96. Hence very close to the actual mean.The difference in absolute terms is very low, a
bout 0.035
sample_mn<-mean(sim_mns)
sample_mn
```

```
## [1] 5.01425
```

Compare sample variance with theoretical variance

As stated above, the theoretical standard deviation is 1/lambda. To calculate its variance, we need to square the standard deviation and divide by the sample size which is 40.

Again, we can notice that the sample variance is very close to the theoretical variance (the true value).

```
theory_var<-sd^2/40
theory_var
```

```
## [1] 0.625
```

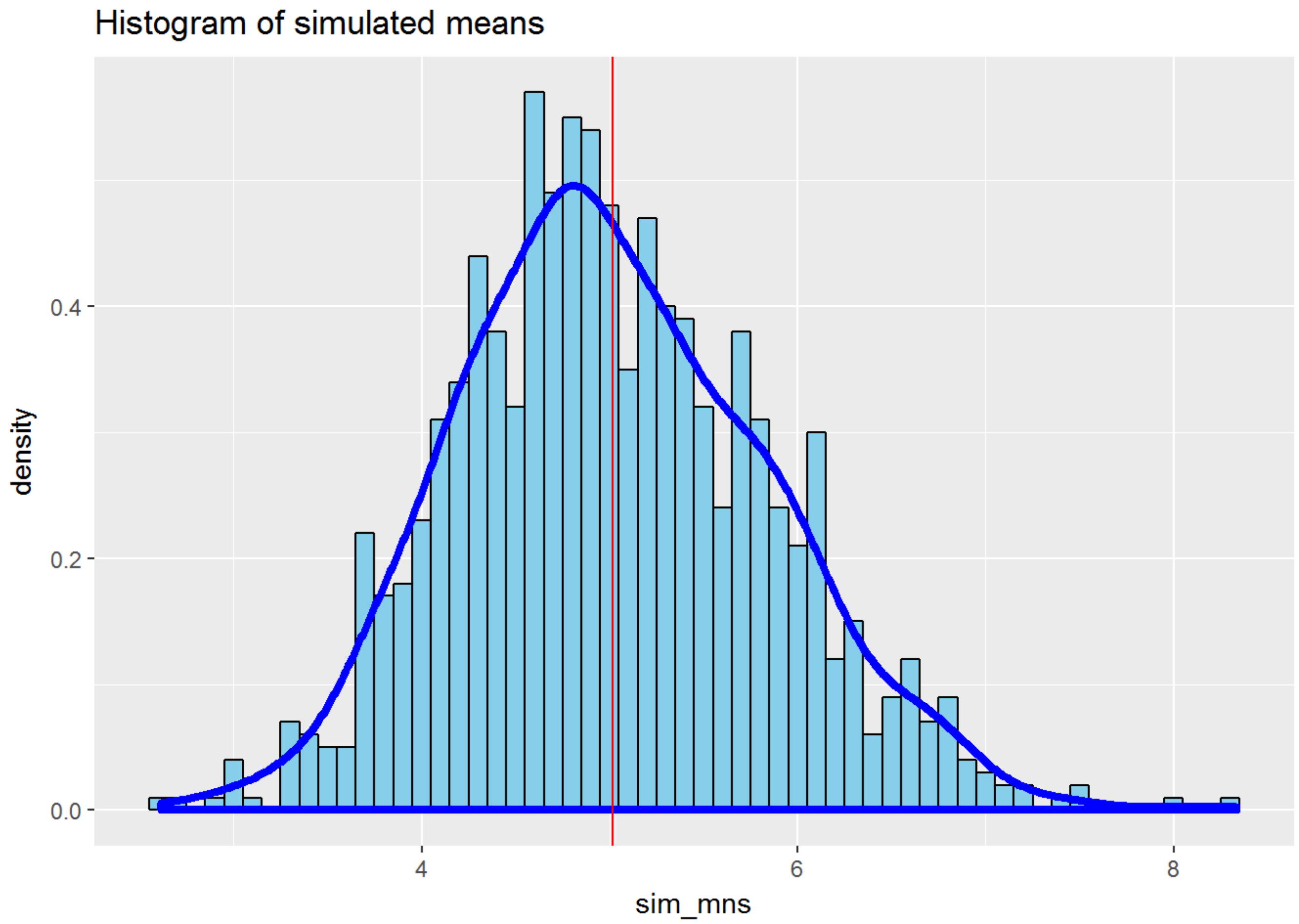
```
sample_var<-var(sim_mns)
sample_var
```

```
## [1] 0.6862199
```

Show that the distribution is approximately normal

Now let visualize instead the simulated distribution and compare it with the theoretical one plotted earlier. According to the **CLT (Central Limit Theorem)**, the distribution of the simulated means should be approximately normal. The result of the code below indeed generates a **bell curve** which resembles the normal distribution.

```
library(ggplot2)
library(magrittr)
data <- data.frame(sim_mns)
m <- ggplot(data, aes(x =sim_mns)) +
      geom_histogram(aes(y=..density..), colour="black", fill = "sky blue",binwidth = 0.1)+
      ggtitle("Histogram of simulated means")
m + geom_density(colour="blue", size=1.5)+geom_vline(xintercept=sample_mn,color="red")
```



We can conclude that by repeating numerous simulations we are able to get consistent estimates of the actual means and variance of the exponential distribution.

Part 2: Basic Inferential Data Analysis

In this second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

Load the ToothGrowth data and perform some basic exploratory data analyses

As per dicumentation `?ToothGrowth` : "The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid)."

```
data("ToothGrowth")
```

Provide a basic summary of the data.

We can see it's data frame with 60 observations on 3 variables.

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
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 ...
```

```
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  
```

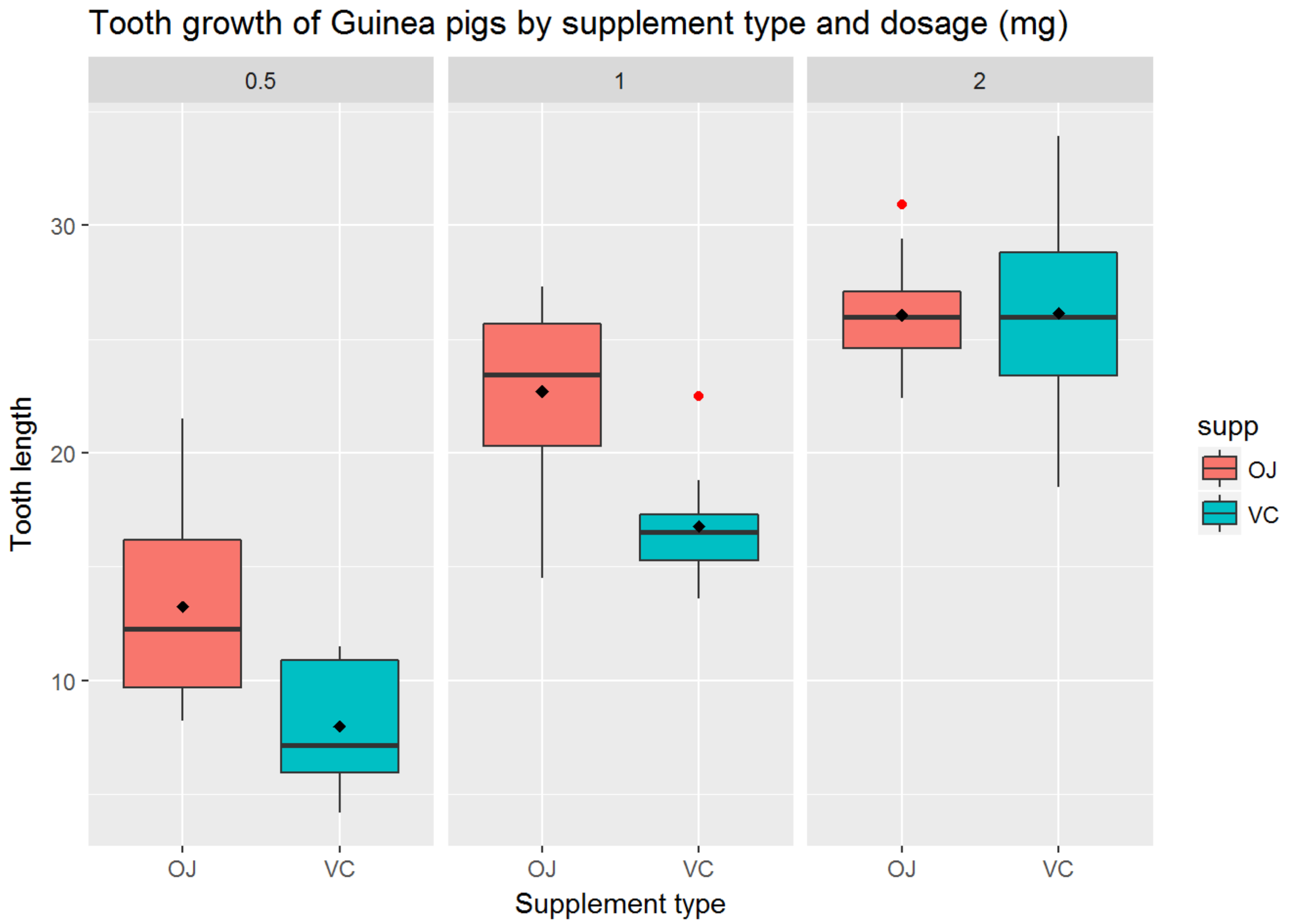
```
sd(ToothGrowth$len) #check the overall variability of response variable
```

```
## [1] 7.649315
```

We can also perform some **exploratory analysis in order to formulate hypothesis** later. Using a **boxplot** we are able to relate tooth length to both supplement type and dose to see which which inputs (or combination of inputs) have more impact on tooth growth.

Boxplots allow us to visualize distributions, hence understanding mean and variance (the box contains the middle 50% of observations and the horizontal line is the median; note we also plotted the mean by adding a black point in the box; the red points located at the extremes are outliers).

```
library(ggplot2)
b<- ggplot(aes(x = supp, y = len), data = ToothGrowth, main="Tooth growth of guinea pigs by s
upplement type and dosage (mg)",xlab="Supplement type", ylab="Tooth length")

b+geom_boxplot(aes(fill =supp),outlier.colour = "red") +
  facet_wrap(~ dose)+
  stat_summary(fun.y=mean, geom="point", shape=18, size=2) +
  ggtitle("Tooth growth of Guinea pigs by supplement type and dosage (mg)") +
  labs(x="Supplement type", y="Tooth length")
```



It looks as, generally there is a positive linear relationship between dosage and tooth growth: the more the dosage and the more teeths grow, using each of the two supplement types. The increase looks particular evident going from 0.5mg to 1mg dosage.

What is not so clear is which supplement type works best. It looks like the OJ type is better for low dosages but then when the dosage increases to 2mg, OJ and VC distributions overlaps.

Confidence intervals and hypothesis testing can help to understand whether the difference between tooth length means of different groups (dosage and supplements) is statistical significant.

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. For instance ANOVA could be a good approach in this case).

Lets first test relationship between dosage and tooth growth

Since there are 3 different levels of dosage (0.5,1,2) in the data, we will start performing 2 separate **t-tests**, comparing difference in means between:

- 0.5mg and 1mg dosage
- 1mg and 2mg dosage

Test 0.5mg vs 1mg dosage

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
test1<-filter(ToothGrowth, dose %in% c(0.5,1))
#in the t.test function we can set the paired parameter equal FALSe because pigs that receive
d 0.5 dosage were a separate set of pigs than those that received 1mg dosage. The same assump
tion will be made in subsequent tests.
t.test(len~dose,paired=FALSE,var.equal=TRUE,data=test1)
```

```
##
##  Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 38, p-value = 1.266e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983748  -6.276252
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

The test looks statistically significant given an alpha level of 5%. The p-values is much lower than 0.05. So, it should be very unlikely to get this t-statistic if the null hypothesis (there is not difference in length means between the two dosage levels) was actually true. In other words it looks **there is a statistically significant difference** between tooth lenght means of pigs who were treated with a dosage level of 1mg instead of 0.5mg.

Test 1mg vs 1.5mg dosage

```
test2<-filter(ToothGrowth, dose %in% c(1,2))
t.test(len~dose,paired=FALSE,var.equal=TRUE,data=test2)
```

```
##
##  Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 38, p-value = 1.811e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.994387 -3.735613
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

Again, p-value is much lower than our alpha level of 0.05, so it seems it’s unlikely that the higher tooth growth obtained with a dosage of 2mg would have happened by chance under the null hypothesis. In conclusion, **we can reject the null hypothesis**.

Given our results as well as looking at previous boxplots, we reckon it’s not necessary to test difference in means between the 0.5 and 2mg groups.

Finally test relationship between supplement type and tooth growth

We wil perform only one test since there are two supplemt types available in the dataset.

Test OJ vs VC supp type

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

```
## 
##  Two Sample t-test
## 
## data:  len by supp
## t = 1.9153, df = 58, p-value = 0.06039
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1670064  7.5670064
## sample estimates:
## mean in group OJ mean in group VC 
##         20.66333         16.96333
```

In this case the p-value is higher than our alpha threeshold of 0.05. Even if there might be some relationship between one specific supplement and the response variable, it is not statistically significant. Hence **we can’t reject the null Hypothesis** stating there is no difference between OJ tooth growth mean of VC tooth growth mean.

To undestand further the relationship between all variable, it might make sense to **understand the effect of the interaction between input variables** on the response variable. In particular, as showed in above the boxplot, it would be interesting to see if the OJ supplement is more effective for low dosages (reject null hypothesis) while there is no evidence of difference in means for higher dosages like 2mg. However we won’t perform these analysis.