Statistical Inference Course Project

Martin J Page

28/05/2020

Part One: Simulation Exercise

Overview

In this project we will investigate the exponential distribution by running two simulations. In the first simulation, we generate one distribution of 1000 random numbers from the exponential distribution. In the second simulation, we generate 1000 distributions with 40 random numbers from the exponential distribution. With the second simulation, we take the averages of the distributions to create one distribution of 1000 averages. Finally, we compare the single distribution to the distribution of averages to demonstate the principle of the Central Limit Theorem: when the sample size increases, the distribution of the averages of a random variable is approximated by the normal distribution.

Simulations of the Exponential Distribution

We will run two simulation protocols for the exponential distibution. The exponential distibution is simulated with `rexp(n, lambda)`, with lambda representing the rate. The first simulation generates 1000 random numbers from the exponential distibution and stores them in a vector `one_sim`. The second simulation generates 40 random numbers and repeats this 1000 times, storing the 1000 distrbutions of 40 exponentials in a 1000 x 40 matrix `forty_sims`. The formulae of the theortical mean and standard deviation of an exponential distibution are also defined, which are both $\frac{1}{lambda}$.

```
#Simulation Parameters
lambda <- 0.2 #rate parameter of the exponential  distribution
B <- 1000 #number of observations to generate for each distribution
n <- 40 #number of distributions to generate

#Simulation 1: One distribution of 1000 observations
set.seed(28052020)
one_sim <- rexp(B, lambda)

#Simulation 2: 1000 distributions of 40 observations
set.seed(28052020)
forty_sims <- matrix(rexp(n*B, lambda) , nrow = B, ncol = n)

#Theoretical Mean and SD
tMean <- function(lam) 1/lam
tSD <- function(lam) 1/lam
```

Q1: Sample Mean versus Theoretical Mean

```
sample_ave <- round(mean(one_sim),2) #calculate sample mean from the data
theortical_avg <- tMean(lambda) #calculate theortical mean with the formula

print(data.frame(Sample = sample_ave, Theoretical = theortical_avg), row.names = FALSE)
```

```
 Sample Theoretical
   5.22           5
```

The sample mean calculated from the simulation of 1000 random numbers from an exponential distribution is 5.22. This is close to the theortical mean calculated from the formula $\frac{1}{lambda}$, which is 5.

Q2: Sample Variance versus Theoretical Variance

```
sample_var <- round(var(one_sim),2) #calculate sample variance from the data
theortical_var <- tSD(lambda)^2 #calculate theoretical variance with the formula

print(data.frame(Sample = sample_var, Theoretical = theortical_var), row.names = FALSE)
```

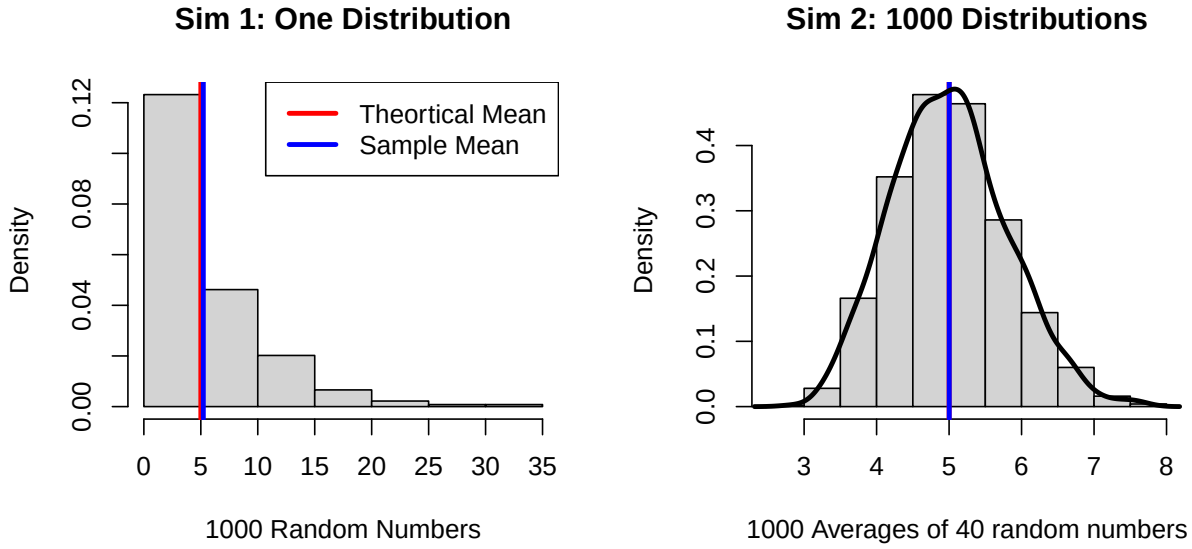
```
 Sample Theoretical
  25.67          25
```

Similarly, the sample variance calculated from the simulation of 1000 random numbers from an exponential distribution is 25.67. This is close to the theortical mean calculated from the formula $\frac{1}{lambda}^2$, which is 25.

Q3: Distribution

```
#Calculate the means of the 1000 simulations of 40 exponentials to create a vector
#of 1000 averages, i.e. the distribution of averages of 40 exponentials
distribition_avgs <- apply(forty_sims, 1, mean)

#Calculate the mean of the distribution of averages
dist_avg <- mean(distribition_avgs)

#Plotting the distributions
par(mfrow = c(1,2))

#Plot a histogram of the 1000 random numbers from Simulation 1
hist(one_sim, main = "Sim 1: One Distribution",
     xlab = "1000 Random Numbers", prob = TRUE) #plot as a density function
abline(v = theortical_avg, col = "red", lwd = 3)
                #add a red line at the theortical mean (from formula)
abline(v = sample_ave, col = "blue", lwd = 3)
              #add a blue line at the sample mean (from data)
legend("topright", c("Theortical Mean", "Sample Mean"), col = c("red", "blue"), lwd = 3)
```

```r
#Plot a histogram of the distribution of averages of 40 exponentials (1000 averages)
hist(distribition_avgs, main = "Sim 2: 1000 Distributions",
     xlab = "1000 Averages of 40 random numbers", prob = TRUE) #plot as a density function
abline(v = theortical_avg, col = "red", lwd = 3)
              #add a red line at the theortical mean (from formula)
abline(v = dist_avg, col = "blue", lwd = 3)
            #add a blue line at the sample mean (from data of averages)
lines(density(distribition_avgs), lwd = 3) #overlay a density distribution
```



The distribution of Simulation 1 (of one distribution with 1000 observations) is clearly not normally distributed. The theortical mean (red) and sample mean (blue) do not exactly align. Moreover, the data are not symetrically distrubted around the mean. However, when multiple distrubtions are averaged as in Simulation 2 (of 1000 distribution with 40 observations) the theortical and sample means align and the data are close to being symetrically distrubted around the mean, as established by the Central Limit Theorem.

Part 2: Basic Inferential Data Analysis

Overview

Q1: Exploratory Data Analyses

From `help(ToothGrowth)` we know that the `ToothGrowth` dataframe looks at the response is the length, `len`, of odontoblasts in 60 guinea pigs. The animals received vitamin C in one of three doses, `dose`, and by one of two delivery methods, `supp`.

```r
data("ToothGrowth")
ToothGrowth$dose <- factor(ToothGrowth$dose)
str(ToothGrowth)
```

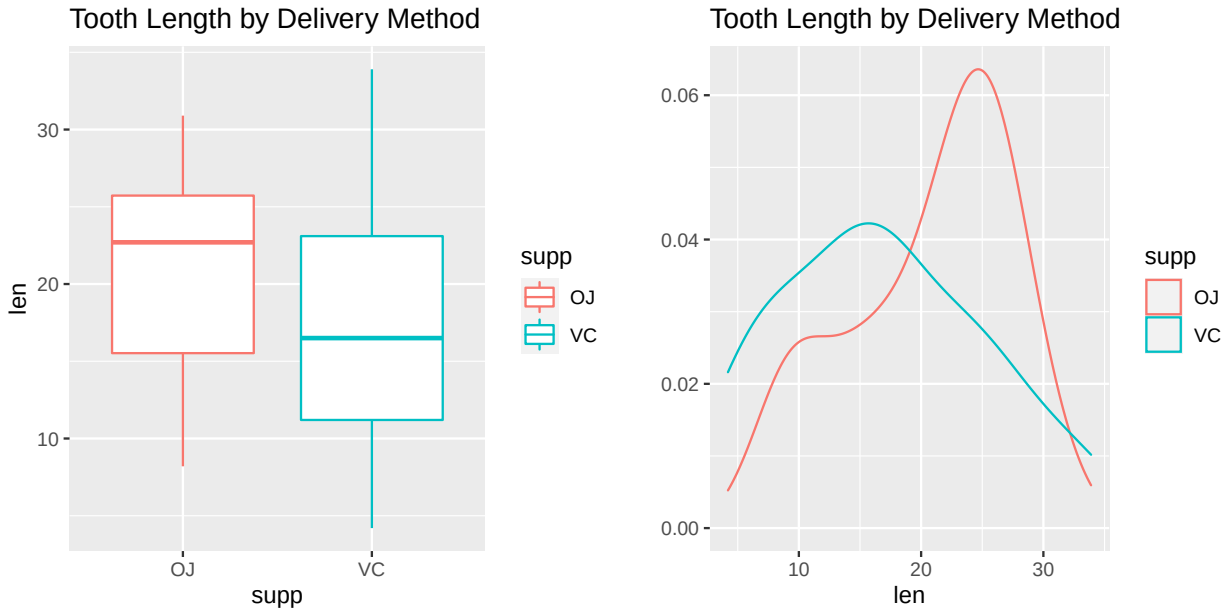
```
'data.frame':   60 obs. of  3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
```

```
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

Let's use the boxplot to see the distribution of the data and get an idea if the groups might be different/separated from each other. Let's also plot the data as a density function to see the distribution in another way.

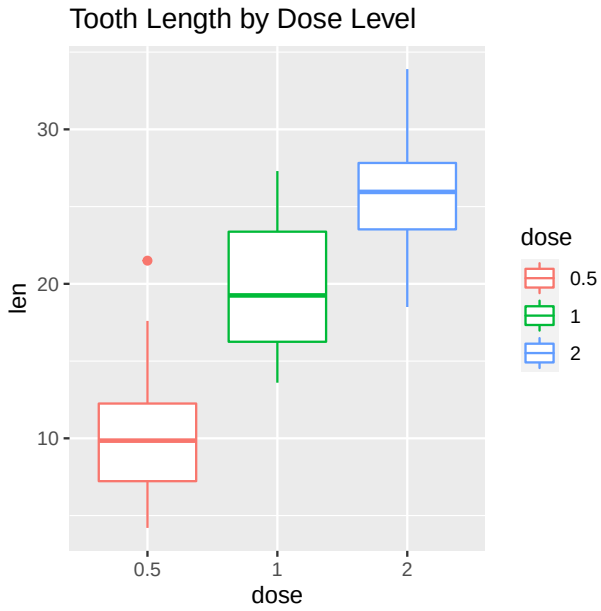
Exploring Delivery Method (supp)

```
library(ggplot2)

qp1 <- qplot(supp, len, data = ToothGrowth, colour = supp,
      main = "Tooth Length by Delivery Method", geom = "boxplot")
qp2 <- qplot(len, group = supp, data = ToothGrowth, col = supp,
      main = "Tooth Length by Delivery Method", geom = "density")

gridExtra::grid.arrange(qp1, qp2, ncol = 2)
```



Exploring Doseage (dose)

```
qp3 <- qplot(dose, len, data = ToothGrowth, colour = dose,
      main = "Tooth Length by Dose Level", geom = "boxplot")
qp4 <- qplot(len, group = dose, data = ToothGrowth, col = dose,
      main = "Tooth Length by Dose Level", geom = "density")

gridExtra::grid.arrange(qp3, qp4, ncol = 2)
```



Q2: Summary of the data

Let's look at the variance of `len` when we group by `supp` and `dose` to see if they are the same

Summarising Delivery Method (`supp`)

```
tapply(ToothGrowth$len, ToothGrowth$supp, var)
```

```
      OJ       VC
43.63344 68.32723
```

Summarising Doseage (`dose`)

```
tapply(ToothGrowth$len, ToothGrowth$dose, var)
```

```
     0.5        1        2
20.24787 19.49608 14.24421
```

Q3: Comparison of Tooth Growth by Supp and Dose

Test if there are differences between the Delivery Methods (`supp`)

```
test_supp <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE , data = ToothGrowth)
test_supp$conf.int
```

```
[1] -0.1710156  7.5710156
attr(,"conf.level")
[1] 0.95
```

Test if there are differences between the lowest and highest Doseages (`dose`)

```
library(dplyr)
sub_tooth <- ToothGrowth %>% filter(dose %in% c("0.5", "2"))
test_dose <- t.test(len ~ I(droplevels(dose)), paired = FALSE, var.equal = TRUE, data = sub_tooth)
test_dose$conf.int
```

```
[1] -18.15352 -12.83648
attr(,"conf.level")
[1] 0.95
```

```
?round
```

Q4: Conclusions and Assumptions

In these experiments different guinea pigs were given different treatments in terms of dosage and delivery method of vitamin C. This means that the groups are UNPAIRED. Groups may potentially be bigger or smaller than each other, so we use a two-sided test.

Delivery Methods

We assume that when the observations are grouped by the delivery method (`supp`) that they follow a t-distrubution, with degrees of freedom 55 and UNEQUAL variance. The t-test comparing the two groups, delivery with organge juice and with ascorbic acid, has a null hypothesis that there is no difference between the two groups. Our 95% confidence interval include the value 0 (and the p-value is 0.0606345), so we CANNOT REJECT the hypothesis that there is no difference between the two groups at an alpha level of 0.05.

Dose Levels

We compare the lowest dose with the highest dose of vitamin C. We assume that the when the observations are grouped by the dose level (`dose`) that they follow a t-distrubution, with degrees of freedom 38 (considering only the two levels: 0.5 mg/day and 2 mg/day) and have EQUAL variance. The t-test comparing the two groups has a null hypothesis that there is no difference between the two groups. Our 95% confidence interval does not include 0. Indeed, the t-statistic (-11.8) is much smaller than 0 and the p-value is very small at 2.8×10^{-14}. We thus REJECT the null hypothesis and conclude that the is a significant difference between the groups. High doseage of vitamin C has a greater value of tooth length than low dose of vitamin C.