Coursera - Statistical Inference

Ali Magzari

8/28/2021

Description

This project consists of two parts: Simulation and basic inference exercises.

https://www.coursera.org/learn/statistical-inference/peer/3k8j5/statistical-inference-course-project

Part 1: Simulation exercise

In this part, the exponential distribution will be investigated and compared to the Central Limit Theorem.

Set seed to recreate random experiments

```
set.seed(500)
```

Set simulation parameters (lambda, sample size and number of replications)

```
lambda <- 0.2
n <- 40
rep <- 1000
```

Create the random exponential distribution matrix

```
sim <- replicate(rep, rexp(n, lambda))
sim_matrix <- matrix(sim, rep, n)
```

Create the vector of means, compare theoretical and sample mean and variance

```
means <- rowMeans(sim_matrix)

sam_mean <- mean(means)
theo_mean <- 1/lambda
error_mean <- (theo_mean - sam_mean)/theo_mean

sam_var <- var(means)
theo_var <- (1/lambda)^2/n
error_var <- (theo_var - sam_var)/theo_var

message(sprintf(
  "Sample mean: %s; Theoretical mean: %s with a mere error of %s\n\n",
  round(sam_mean, 3), theo_mean, round(error_mean, 3)))
```

```
## Sample mean: 5.011; Theoretical mean: 5 with a mere error of -0.002
```

```
message(sprintf(
  "Sample variance: %s; Theoretical variance: %s with a mere error of %s\n\n",  round(sam_var, 3), theo_
```

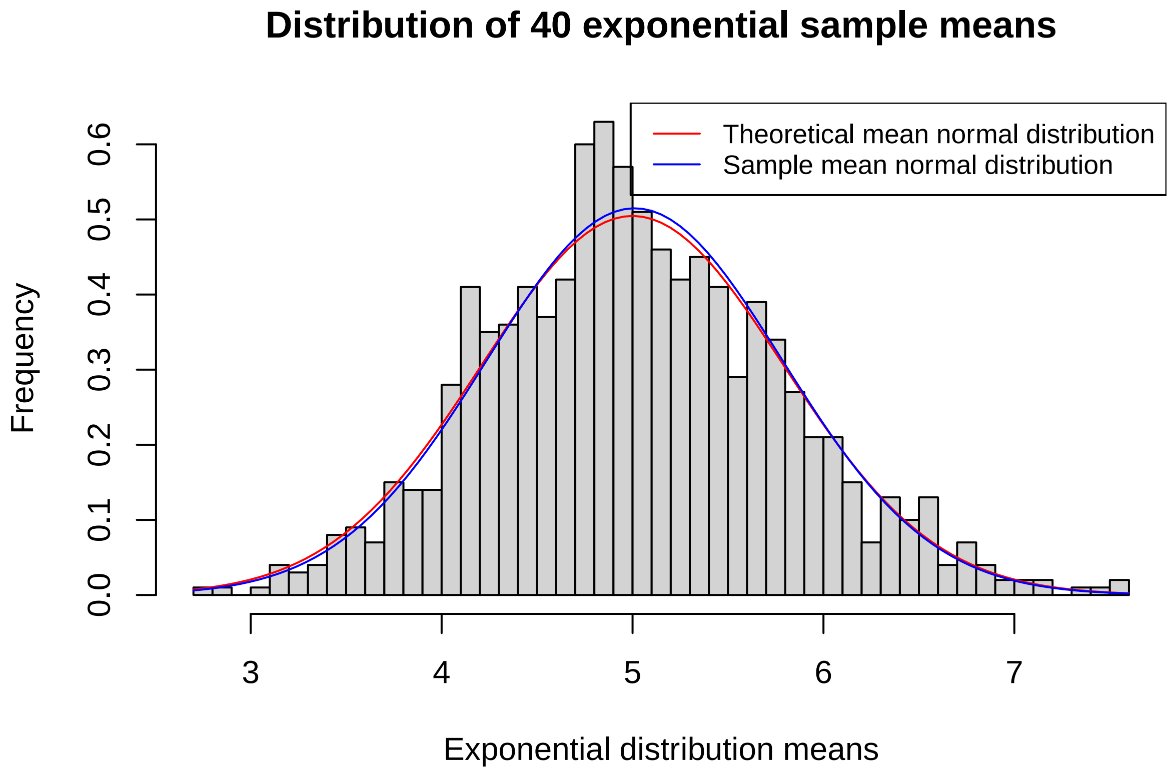
```
## Sample variance: 0.6; Theoretical variance: 0.625 with a mere error of 0.039
```

Plot the histogram of exponential distribution means and compare it to the curves of theoretical and sample mean normal distributions

```
hist(means,  xlab="Exponential distribution means", ylab = "Frequency",
     main="Distribution of 40 exponential sample means", breaks=n, prob=TRUE)

curve(dnorm(x, theo_mean, sqrt(theo_var)), col="red", add=TRUE)
curve(dnorm(x, sam_mean, sqrt(sam_var)), col="blue", add=TRUE)

legend(x = "topright",
       legend = c("Theoretical mean normal distribution", "Sample mean normal distribution"),
       lty = c(1, 1),
       col = c("red", "blue"),
       cex = 0.8)
```



Part 2: Basic Inferential Data Analysis

The data set used is named “ToothGrowth”. It contains 60 observations on 60 guinea pigs related to tooth length, vitamin C delivery method (orange juice or ascorbic acid), and the dose level (0.5, 1, 2).

This part consists of developing a hypothesis test on whether the delivery method or the dose level have any impact on tooth length.

Let’s start by loading the ToothGrowth data and explore its structure

```
data(ToothGrowth)
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 ...
```

```
table(data$dose)
```

```
##
## 0.5   1   2
##  20  20  20
```

```
table(data$supp)
```

```
##
## OJ VC
## 30 30
```

```
table(data$len)
```

```
##
##  4.2  5.2  5.8  6.4    7  7.3  8.2  9.4  9.7   10 11.2 11.5 13.6 14.5 15.2 15.5
##    1    1    1    1    1    1    1    1    2    2    2    1    1    3    2    1
## 16.5 17.3 17.6 18.5 18.8 19.7   20 21.2 21.5 22.4 22.5   23 23.3 23.6 24.5 24.8
##    3    2    1    1    1    1    1    1    2    1    1    1    2    2    1    1
## 25.2 25.5 25.8 26.4 26.7 27.3 29.4 29.5 30.9 32.5 33.9
##    1    2    1    4    1    2    1    1    1    1    1
```

Provide a basic summary of the data

```
summary(data)
```

```
##       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
```

Perform a t-test to test if vitamin C delivery method affects tooth length mean

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

```
##
##  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 between group OJ and group VC 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
```

The fact that the p-value is large enough (0.06063) and that 0 is enclosed within the confidence interval leads us to not reject the null hypothesis, and therefore state that vitamin C delivery method has no effect on tooth length.

Perform three t-tests to test whether or not odontoblast length is affected by changing the dose level

```
t.test(len ~ dose, data[data$dose == 0.5 | data$dose == 1, ])
```

```
##
##  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 between group 0.5 and group 1 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
```

```
t.test(len ~ dose, data[data$dose == 0.5 | data$dose == 2, ])
```

```
##
##  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 between group 0.5 and group 2 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
```

```
t.test(len ~ dose, data[data$dose == 1 | data$dose == 2, ])
```

```
##
##  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 between group 1 and group 2 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
```

Based on the fact that the p-value is negligible in all three tests above, we are allowed to reject the null hypothesis, and state that vitamin C dose has indeed an effect on tooth length.