Statistical Inference - Peer Graded Assignment

Pete Petersen III

6/20/2020

# Part 1: Simulation Exercise

#### Overview:

In this paper we demonstrate that the Central Limit Therom is in congruence with theoretical normal distributions. In order to accomplish this we will generate a 1000 samples of lenf 40, We will then take the means of each of these 1000 samples. This will show that the density of the means is very sumilar to the theoretic density. The exponential distribution will be simulated in R with rexp(n,lambda) where lambda is the rate parameter. rexp uses Ahrens, J. H. and Dieter, U. (1972). Computer methods for sampling from the exponential and normal distributions. Communications of the ACM, 15, 873–882.

#### Simulations:

The mean of exponential distribution and the standard deviation are both 1/lambda where lambda = 0.2, and distribution of averages of 40 exponentials and will perform 1000 simulations.

#### The Exponential Distribution:

Density, distribution function, quantile function and random generation for the exponential distribution with rate (i.e., mean 1/rate).  
  
 Usage  
 dexp(x, rate = 1, log = FALSE)  
 pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)  
 qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)  
 rexp(n, rate = 1)  
   
 Arguments:  
   
 x, q - vector of quantiles.  
 p - vector of probabilities.  
 n - number of observations. If length(n) > 1, the length is taken to be the number required.  
 rate - vector of rates.  
 log, log.p - logical; if TRUE, probabilities p are given as log(p).  
 lower.tail - logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x].

# Set the initial simulations paremeters wher lamda is the the rate parameter in exponential distrubution.  
 lambda <- 0.2  
  
# Set the seed for the random number generator for reproducabkility  
 set.seed(11)  
  
 # Set the number of observations. If length(n) > 1, the length is taken to be the number required  
 n <- 40  
  
 # Iterate a thousand times of the exponential distribution where length = 40 and lamda = .2 THis will result in a 40 row by 1000 column array.  
 simulation <- replicate(1000, rexp(n, lambda))  
   
# we use the apply to take means over the simulation distribution columns and arrive ar 1 mean per colimn.  
 mean\_simulation <- apply(simulation, 2, mean)

Sample Mean versus Theoretical Mean: Include figures with titles. In the figures, highlight the means you are comparing. Include text that explains the figures and what is shown on them, and provides appropriate numbers.

sampleMean <- mean(mean\_simulation)  
 simulationmean <- mean(simulation)  
 theoretical\_mean <- (1/lambda)

The mean of the simulations is 4.9871567  
The theoretical mean of:

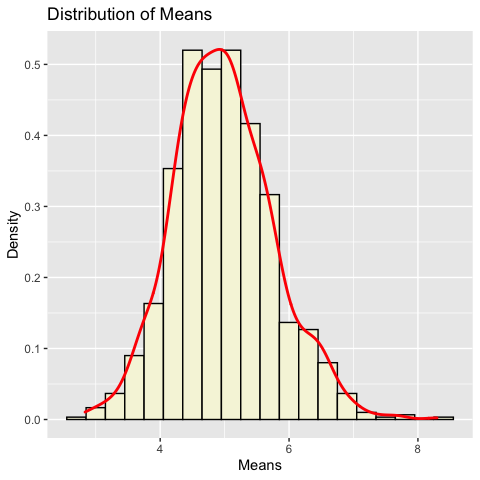


Fig. 1 Simulation Distribution snd Density

Sample Variance versus Theoretical Variance: Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances.

# sample deviation & variance  
sample\_sd <- sd(mean\_simulation)  
sample\_var <- sample\_sd^2  
simulationmean <- mean(mean\_simulation)  
# theoretical deviation & variance  
theoretical\_sd <- (1/lambda)/sqrt(n)  
theoretical\_var <- ((1/lambda)\*(1/sqrt(n)))^2  
theoretical\_mean <- 1/lambda

We now compare data from the simulation to the values the CTL. CTL predicts that the distribution of the simulation should be very close to normal distribution.

##### Defined As

##### where

##### and

Simulated vs Theoretical Values

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Std | Variance |
| Theoretical | 5.000 | 0.791 | 0.625 |
| Simulation | 4.987 | 0.775 | 0.601 |

#### Distribution:

In the figure below, we plot the histogram of the means and the density of the means in order to note the appearant normalcy. We also note that the distribution appears centered around.5 and is symetric. Further, the similarity of the simulation density in blue to the theoretical density in red highlights the proximity to normalcy of the simulation means which exemplifies the CLT’s correctneess.

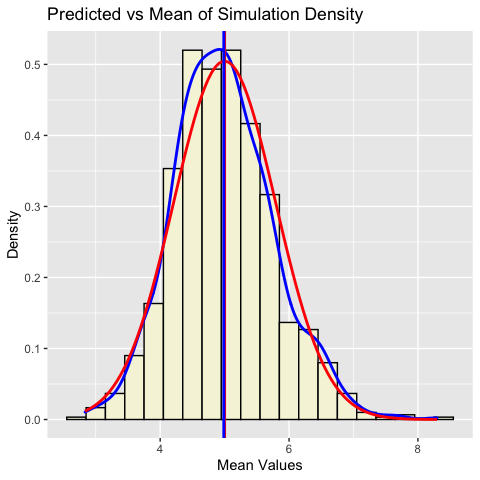


Fig. 2 Distribution Analysis

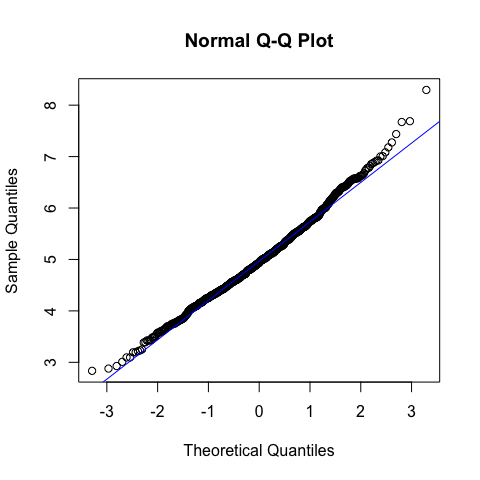


Fig. 3 Distribution Analysis

Q-Q Normal Plot also indicates the normal distribution. qqnorm is a generic function the default method of which produces a normal QQ plot of the values in y. qqline adds a line to a “theoretical”, by default normal, quantile-quantile plot which passes through the probs quantiles, by default the first and third quartiles. The proximity to the xy plot of the simulation pairs demonstrated very litte deviation from normalcy.

## Part 2: Basic Inferential Data Analysis Instructions

Load the ToothGrowth data and perform some basic exploratory data analyses

kable(head(ToothGrowth))

|  |  |  |
| --- | --- | --- |
| len | supp | dose |
| 4.2 | VC | 0.5 |
| 11.5 | VC | 0.5 |
| 7.3 | VC | 0.5 |
| 5.8 | VC | 0.5 |
| 6.4 | VC | 0.5 |
| 10.0 | VC | 0.5 |

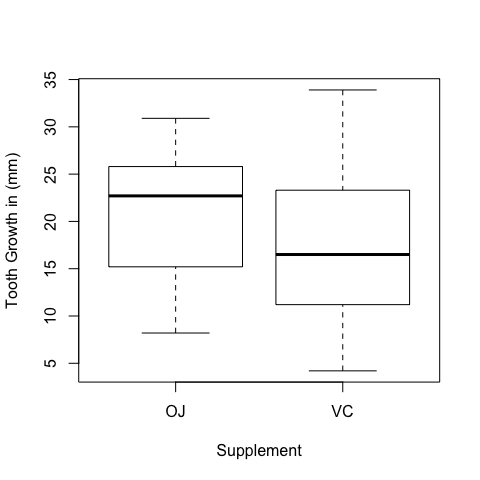


Fig. 4 Tooth Length

The Box Plot above seems to indicate additional tooth growth in the OJ supplement over the VC supplement.

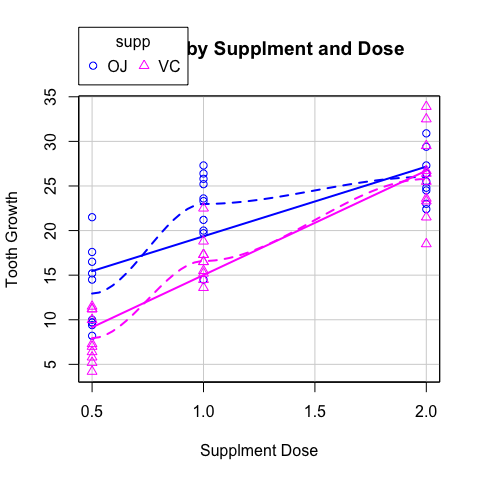


Fig. 5 Growth by Supplement and Dose

The scatter plot show a higher level of tooth growth at lower dosages as compared to the vitamin C. However, at the higher dosage (2mg) the supplments seem equally effective.

## Data 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))

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 |

## Confidence Intervals

# Subset the data for confints  
ToothGrowth\_oj <- subset(ToothGrowth, supp=='OJ')  
ToothGrowth\_vc <- subset(ToothGrowth, supp=='VC')  
ConfIntTable <- rbind(  
mean(ToothGrowth$len) + c(-1, 1) \* 1.96 \* sd(ToothGrowth$len)/sqrt(nrow(ToothGrowth)),  
mean(ToothGrowth\_oj$len) + c(-1, 1) \* 1.96 \* sd(ToothGrowth\_oj$len)/sqrt(nrow(ToothGrowth\_oj)),  
mean(ToothGrowth\_vc$len) + c(-1, 1) \* 1.96 \* sd(ToothGrowth\_vc$len)/sqrt(nrow(ToothGrowth\_vc)))  
  
row.names(ConfIntTable) = c('Overall', 'OJ', 'VC')  
df\_conf <- as.data.frame.matrix(ConfIntTable)   
df\_conf %>%   
 rename(  
 'ConfInt-Low' = 'V1',  
 'ConfInt-High' = 'V2')

## ConfInt-Low ConfInt-High  
## Overall 16.87779 20.74888  
## OJ 18.29956 23.02710  
## VC 14.00537 19.92129

## Hypothesis Testing and Sample Analysis

##### All Data

#Run Welch Test on entire Sample  
 model1 <- t.test(len ~ supp, data = ToothGrowth)  
 model1

##   
## 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

# Run power tests to determine power properties of our sample  
 all\_delta <- power.t.test(n = nrow(ToothGrowth), power = .90, sd = sd(ToothGrowth$len))$delta  
 all\_size <- power.t.test( power = .90, delta=3, sd = sd(ToothGrowth$len))$n  
 all\_power <- power.t.test(n = nrow(ToothGrowth), delta = 3, sd = sd(ToothGrowth$len))$power

The p value is somewhat high for all dosages. Therefore we accept the null H0 hypthesis that there is no difference in the means of the supplements and also accept that there maybe some significance in general on supplement type accross all dosages.

Power testing indicates that we could only determine a delta of 4.56 mm with our current sample size with .9 power. If we wanted to increase the length sensitivity to a delta of 3mm we would need a sample size of 138. At our current sample size of 60, we can only detect a 3mm growth delta with 0.57 power.

##### Analysis at .5mg Doses

five <- subset(ToothGrowth, dose == "0.5")  
model2 <- t.test(len ~ supp, data = five)  
model2

##   
## Welch Two Sample t-test  
##   
## data: len by supp  
## t = 3.1697, df = 14.969, p-value = 0.006359  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.719057 8.780943  
## sample estimates:  
## mean in group OJ mean in group VC   
## 13.23 7.98

five\_delta <- power.t.test(n = nrow(five), power = .90, sd = sd(five$len))$delta  
five\_size <- power.t.test(delta = 3, power = .90, sd = sd(five$len))$n  
five\_power <- power.t.test(delta = 3,n = nrow(five) , sd = sd(five$len))$power

The p value is low at the .5 mg dose level. Therefore we reject the H0 hypthesis that there is no difference in the means of the supplements and we accept that there is some signifcance to the impact on toothgrowth at this dose level.

Power testing indicates that we could only determine a delta of 4.73 mm with our current sample size with .9 power. If we wanted to increase the length sensitivity to a delta of 3mm we would need a sample size of 48. At our current sample size of 20, we can only detect a 3mm growth delta with 0.54 power.

##### Analysis at 1mg Doses

one <- subset(ToothGrowth, dose == "1")  
model3 <- t.test(len ~ supp, data = one)  
model3

##   
## Welch Two Sample t-test  
##   
## data: len by supp  
## t = 4.0328, df = 15.358, p-value = 0.001038  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 2.802148 9.057852  
## sample estimates:  
## mean in group OJ mean in group VC   
## 22.70 16.77

one\_delta <- power.t.test(n = nrow(one), power = .90, sd = sd(one$len))$delta  
one\_size <- power.t.test(delta = 3, power = .90, sd = sd(one$len))$n  
one\_power <- power.t.test(delta = 3, n = nrow(one), sd = sd(one$len))$power

The p value is low at the 1 mg dose level. Therefore we reject the H0 hypthesis that there is no difference in the means of the supplements and we accept that there is some signifcance to the impact on toothgrowth at this dose level.

Power testing indicates that we could only determine a delta of 4.65 mm with our current sample size with .9 power. If we wanted to increase the length sensitivity to a delta of 3mm we would need a sample size of 47. At our current sample size of 20, we can only detect a 3mm growth delta with 0.55 power.

##### Analysis at 2mg Doses

two <- subset(ToothGrowth, dose == "2")  
model4 <- t.test(len ~ supp, data = two)  
model4

##   
## Welch Two Sample t-test  
##   
## data: len by supp  
## t = -0.046136, df = 14.04, p-value = 0.9639  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.79807 3.63807  
## sample estimates:  
## mean in group OJ mean in group VC   
## 26.06 26.14

two\_delta <- power.t.test(n = nrow(two), power = .90, sd = sd(two$len))$delta  
two\_size <- power.t.test(delta = 3, power = .90, sd = sd(two$len))$n  
two\_power <- power.t.test(delta = 3,n = nrow(two), sd = sd(two$len))$power

Zero is in the confidence interval and the p value is high. Therefore we can not reject the null hypthesis that the means of the two supplememts are equal at the 2mg dose for tooth growth.

Power testing indicates that we could only determine a delta of 3.97 mm with our current sample size with .9 power. If we wanted to increase the length sensitivity to a delta of 3mm we would need a sample size of 34. At our current sample size of 20, we can only detect a 3mm growth delta with 0.69 power.

## Conclusion

The exploratory phase of the analysis indicated that there was some possibility of correlation of tooth growth to supplement type and dose. Both Confidence Intervals and Welch Two Sample hypothesis testing confirmed that there is significance at the .5 mg and 1mg dose levels and no attributable significance to the dose level by supplement type at the 2 mg level. However, we must careful when we interpret the dosages as the power t.test indicates that diffences in the means is just within the required sample size. We reccomend that the analysis be conducted with higher sample sizes in the future in order to confirm these findings.