Statistical Inferential Data Analysis

Trina Dey

11 July 2020

Overview

This report provides overview of two statistical analysis we are going to perform.
1. Analysing Exponential Distribution and comparing how the distribution of mean of samples appears close to Normal distribution.
2. Analysing ToothGrowth data to compare impact of tooth growth by supplement and dosage of Vitamin C on Guinea Pigs.

Exponential Distribution Analysis

Data Simulation and Calculation

We know that the exponential distribution has 2 parameters - n the population size and lambda the rate parameter. We also know that the mean and standard deviation of exponential distribution is 1/lambda. In our simulation we are going to set n = 40 and lambda = 0.2 and calculate means of 1000 samples derived from exponential distribution and plot it on a graph.

Here is the data for generating 1000 samples of means of 40 exponential data.

```
set.seed(512)
mean_data <- NULL
for( i in 1:1000)
  mean_data <- c(mean_data , mean(rexp(40,0.2)))
```

Now we will compute mean and standard deviation of the sampling distribution.

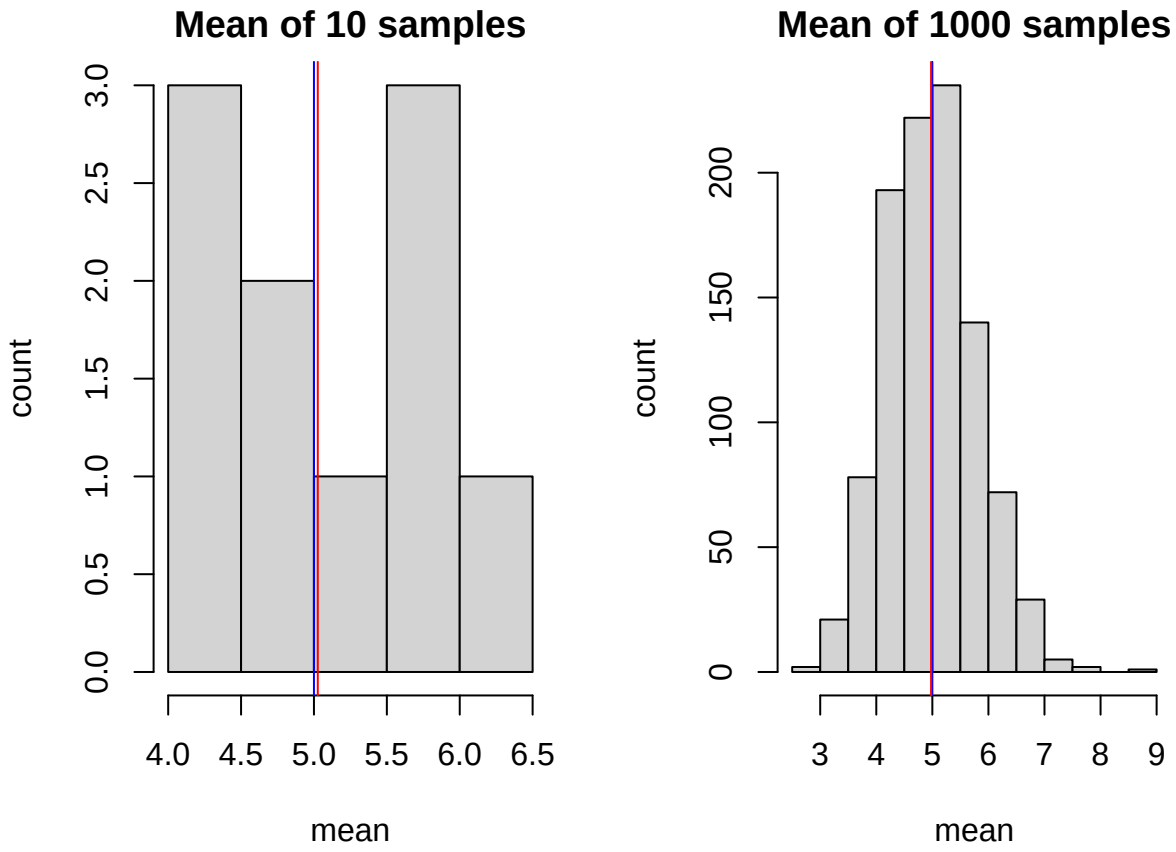
```
sample_mean <- mean(mean_data)
sample_sd <- sd(mean_data)
c(sample_mean,sample_sd)
```

```
## [1] 4.9777902 0.7905796
```

Sample Mean versus Theoretical Mean

The sample mean estimates the population mean. Population mean is 1/lambda which is equal to 5.0. So our sample mean should be same or close to our population mean. We will show this by plotting the sampling distribution. To show the difference of sample size and to verify the Central Limit Theorem, we will plot data for two different sample size say 10 and 1000.

```
par(mfrow = c(1,2) , mar = c(4,4,2,2))
hist(mean_data[1:10] ,xlab= "mean", ylab="count", main="Mean of 10 samples", border = "black")
abline(v=5.0, col="blue")
abline(v=mean(mean_data[1:10]), col="red")
hist(mean_data , xlab="mean", ylab="count", main = "Mean of 1000 samples",border = "black")
abline(v=5.0 , col="blue")
abline(v=sample_mean , col ="red")
```



The blue line shows the population mean which is 1/lambda i.e. 5.0 in our case and the red line shows sample mean which is closer to population mean. When the sample size is large enough, sampling distribution appears normally distributed.

Sample Standard Deviation versus Theoretical Standard Deviation

The standard deviation of sampling distribution or the standard error is defined as population standard deviation divided by square root of population size.

```
lambda <- 0.2
population_sd <- 1/lambda
populationsize <- 40
theoretical_sample_sd <- population_sd/ sqrt(populationsize)
c(sample_sd,theoretical_sample_sd)
```

```
## [1] 0.7905796 0.7905694
```

So our calculated sample standard devation turns out to be 0.7905796 compared to theoretical value 0.7905694 which seems pretty close.

Conclusion

As the central limit theorem predicts that if we take sufficiently large samples from the population with replacement, the sampling distribution ie central limits of the sample (mean or median) will be approximately normally distributed. The plots clearly show how the distribution seems approximately normal when the sample size taken is sufficiently large (say 1000) vs not so normally distributed when the sample size is small (say 10). Also, with sample size 10, we can see a difference between the blue line of population mean and the red line of sample mean, however as the sample size grows large, the sample mean clearly approaches population mean.

Basic Inferential Data Analysis of Tooth Growth data

We will analyse the ToothGrowth data which is the effect of Vitamin C on Tooth Growth in Guinea Pigs and compare the growth by different supplement type of Vitamin C (Orange Juice - OJ and Ascorbic Acid - VC) and dose (0.5, 1.0, 2.0 milligrams). We ran the command ?ToothGrowth on terminal to find the details of the dataset.

ToothGrowth is a data frame with 60 observations on 3 variables.

[,1] len numeric Tooth length
[,2] supp factor Supplement type (VC or OJ)
[,3] dose numeric Dose in milligrams/day

Here is the summary for the data.

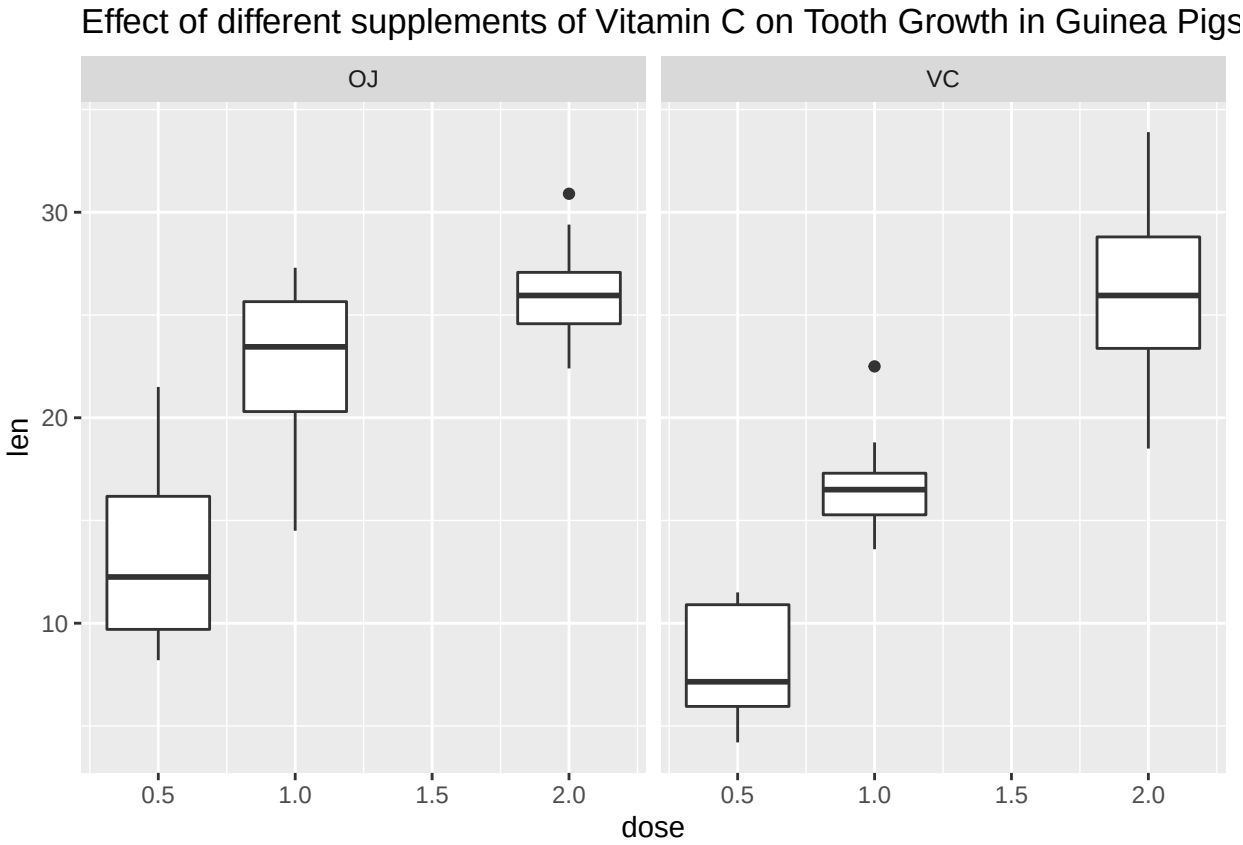
```
summary(ToothGrowth)
```

```
##       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
```

Exploratory Analysis

We will plot a box plot showing the growth of tooth with respect to different supplements and the dosage. This will help us understand how the two supplements are different and provide a visual representation of the variability in the data.

```
library(ggplot2)
ggplot(data = ToothGrowth, aes(dose, len ) ) +
    geom_boxplot(mapping=aes(group=dose) ) +
    facet_wrap(~ supp, nrow=1) +
    labs(title="Effect of different supplements of Vitamin C on Tooth Growth in Guinea Pigs")
```



From the plot it is very clear that the higher dosage has better effect on length of tooth in both the supplements of Vitamin C. However, we would like to see which supplement has better effect on the tooth growth.

Hypothesis Testing

To see which supplement has better effect on the tooth growth, we will create below hypotheis.
- H0: Null Hypothesis : Both the supplements has equal effect on the growth
- HA: Alternative Hypothesis: There is a measurable difference between the two supplements.

Assumption : For our Hypothesis Testing we will run T test and will assume the significance level at 5%. Any pvalue with less than 0.05 will make us reject the null hypotheis and accept the alternative hypothesis.

Since we are comparing for inequality, we will perform two sided T tests on difference of OJ and VC data dosage wise.

```
library(dplyr)
oj_0_5 <- filter(ToothGrowth ,supp=="OJ" , dose==0.5)$len
vc_0_5 <-filter(ToothGrowth ,supp=="VC" , dose==0.5)$len
t.test(oj_0_5 - vc_0_5 , paired = FALSE)
```

```
##
##  One Sample t-test
##
## data:  oj_0_5 - vc_0_5
## t = 2.9791, df = 9, p-value = 0.01547
## alternative hypothesis: true mean is not equal to 0
```

```
## 95 percent confidence interval:
##  1.263458 9.236542
## sample estimates:
## mean of x
##      5.25
```

We see that the mean of the difference is more than 0 for different supplements at 0.5 mg. We will repeat the same tests on doses 1mg and 2mg as well.

Test for 1 mg data

```
oj_1 <- filter(ToothGrowth ,supp=="OJ" , dose==1)$len
vc_1 <-filter(ToothGrowth ,supp=="VC" , dose==1)$len
t.test(oj_1 - vc_1 , paired = FALSE)
```

```
##
##  One Sample t-test
##
## data:  oj_1 - vc_1
## t = 3.3721, df = 9, p-value = 0.008229
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.951911 9.908089
## sample estimates:
## mean of x
##      5.93
```

Test for 2 mg data

```
oj_2 <- filter(ToothGrowth ,supp=="OJ" , dose==2)$len
vc_2 <-filter(ToothGrowth ,supp=="VC" , dose==2)$len
t.test(oj_2 - vc_2 , paired = FALSE)
```

```
##
##  One Sample t-test
##
## data:  oj_2 - vc_2
## t = -0.042592, df = 9, p-value = 0.967
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of x
##     -0.08
```

Conclusion

Since our assumption was to set significance level at 0.05, we see that the two sided T tests for 0.5, 1, 2 milligrams dose are 0.01547, 0.008229, 0.967. So we conclude -

1. With less dosages 0.05 mg and 0.01 mg, our pvalue is less than our significance value and we reject our null hypothesis and accpet our alternative hypothesis and since our mean is a positive one we assume that Orange Juice performs better than Ascorbic Acid.

2. However at higher dosage of 2mg, both Orange Juice and Ascorbic Acid has similar effect on tooth growth.

Also, the box plots clearly confirms our conclusions, especially for 0.5mg and 1.0mg dosage and shows that the variability is more in VC at 2mg than OJ at 2mg.