Statistical_Inference_Simulation_Project- Part1&2

Venkat.T

8/19/2019

Introduction

The project consists of two parts:

1. A simulation exercise - This is investgation of exponential distribution for the data generated using rexp().
2. Basic inferential data analysis - This is analysis of the Tooth growth data and inference of the results.

Part 1: Simulation Exercise with exponential distribution

Overview

In this project you 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. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

You should

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.
4. In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Variables and Data Simulation

Based on above requirements, we are defining the variables and also using rexp() generate sample data and using sapply simulate it 1000 times.

```
#set seed for reproducibility
set.seed(99999989)
n <- 40   #tcount of numbers in each simulation
lambda <- 0.2       #set lambda for each simulation
simulations <- 1:1000   # number of simulations
#Generating Data for Exponential Distribution using rexp() of rate 0.2.
avg <- sapply(simulations, function(x) {mean(rexp(n, lambda))})

# now we have averaged out values of the 40 numbers that are simulated 1000 times.
length(avg)
```

```
## [1] 1000
```

1. Compare Sample Mean versus Theoretical Mean

- For the **simulated data**, Mean of distribution : `4.988287`
- While for the **theoretical data**, mean of distribution is `5`

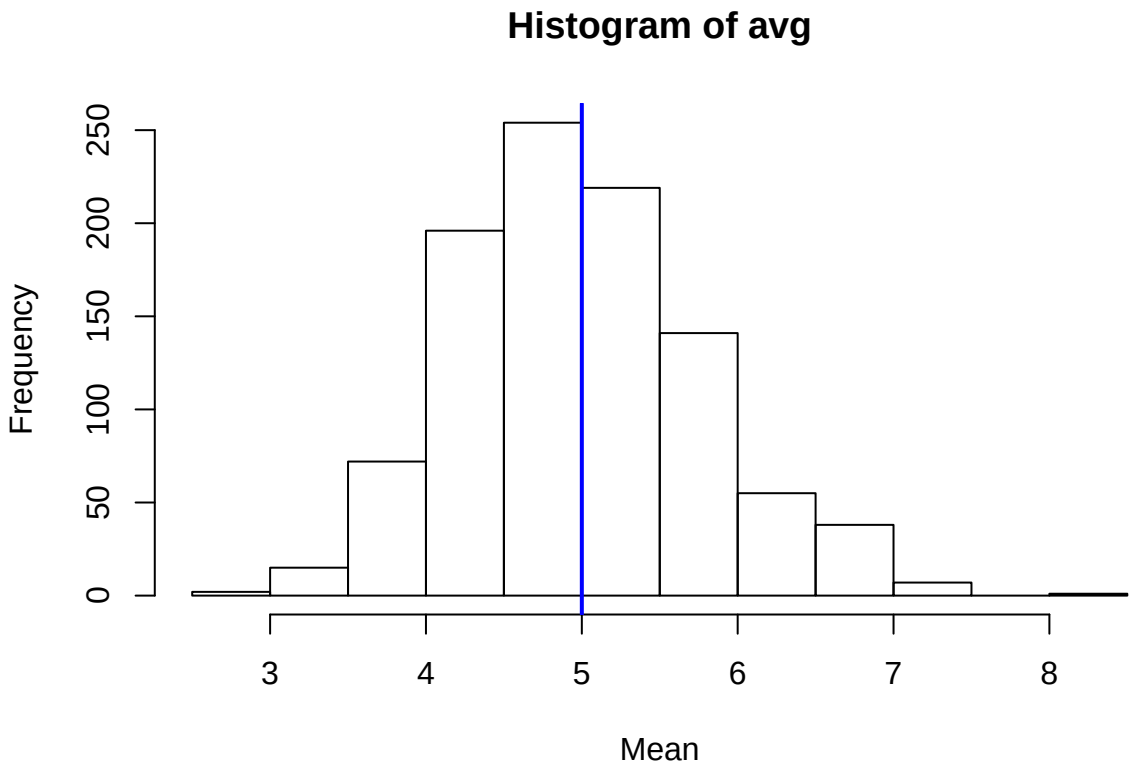
Difference between the 2 distributions: `0.011713`

2. Sample Variance versus Theoretical Variance

- For the **generated data**, the standard deviation: `0.7886625` & the **variance is :** `0.6219886`
- For **theoretical data**, The standard deviation: `0.7905694` & the **variance is:** `0.625`

3. Distribution

```
## Now plot the data as histogram
hist(avg, xlab="Mean", ylab="Frequency")
abline(v=1/lambda , col="blue", lwd=2)
```



Based on output of the histogram, it follows the bell curve, indicating the data is **distributed normally** and most of it is concentrated around the mean.

4. Confidence Interval

Based on the samples, the best estimate would be the sample mean, i.e.here it is 5. But if we have to tell how uncertain we are of this estimate, we need capture the confidence interval.

We can calculate a 95% confidence interval for a sample mean by adding and subtracting 1.96 standard errors to the point estimate.

$$\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$$

.

```
low = mean(avg) - 1.96*(sd(avg)/sqrt(n))
high= mean(avg) + 1.96*(sd(avg)/sqrt(n))
print(c(low, high))
```

```
## [1] 4.743878 5.232696
```

Based on the output we can say **with 95% confdence that the true mean lies between `4.743878, 5.2326961`**

Conclusion for Simulation

- The bell curve of the hostogram indicate this is Normal distribution.
- We see that the normal distribution indeed closely matches the barplot of the means.
- We can tell with 95% confidence interval that true mean lies between `4.743878, 5.2326961`

Part 2: Basic inferential data analysis

Introduction

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

We examine the ToothGrowth dataset in the R datasets package and analyzes the growth by supp and dose.The dataset records 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 (a form of vitamin C and coded as VC).

Sammary

Load data and analyze data

```
library(ggplot2)
library(datasets)
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
```

```
## Now load tg=he toonth growth data
data("ToothGrowth")

# now evalaute and verify 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
```

```r
# check the data types
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 ...
```

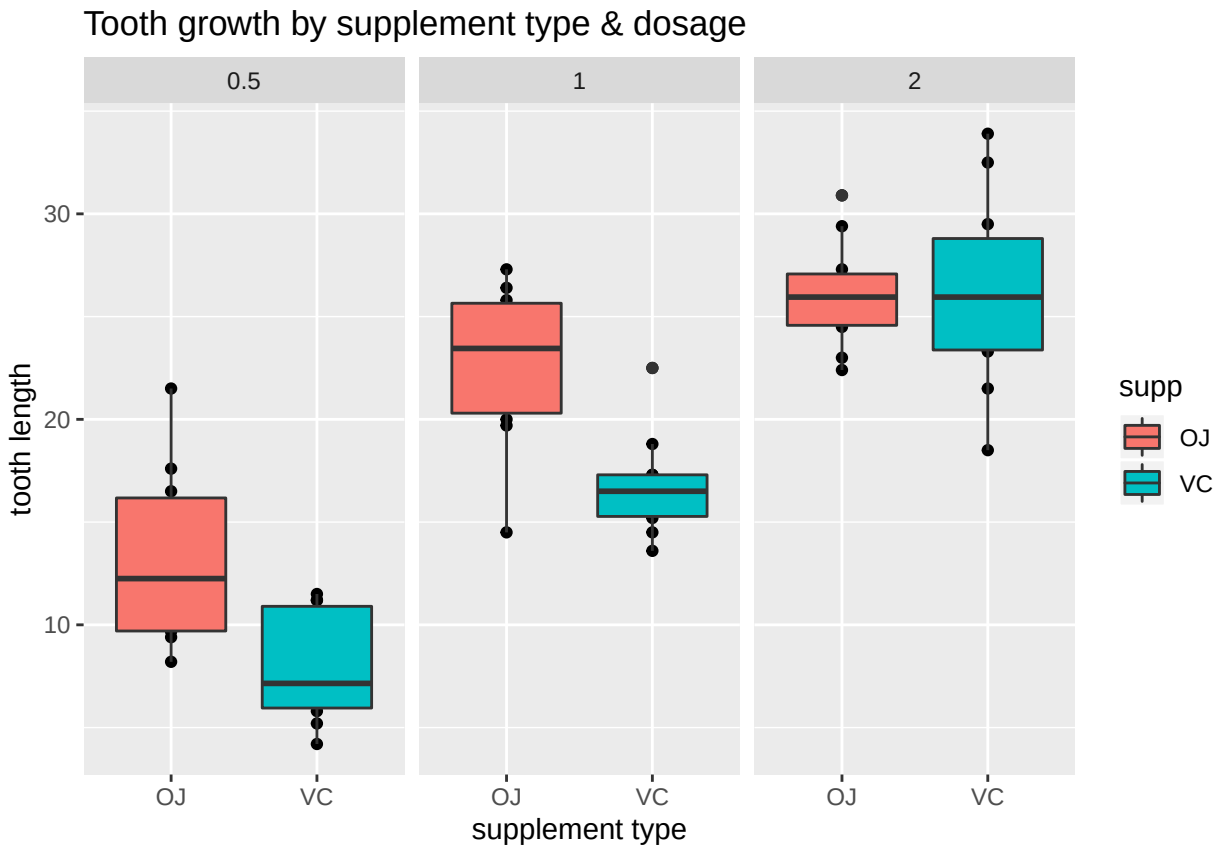
```r
head(ToothGrowth)
```

```
##    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
```

Based on the data summary, we see that data cintauns 3 variables - (Len, Supp, dose) and 60 observations. The str() function shows that the variables - length & dose are numeric and suppliment(supp) is a string with values OJ(orange Juice) and VC(vitamin C).

First, we evaluate the tooth growth length, Suppliment and Dosage relation.

```r
qplot(x=supp,y=len,data=ToothGrowth, facets=~dose, main="Tooth growth by supplement type & dosage",xlab=
```



Inferential Statistics

To test whether the dosage method has a statistically significant effect on tooth length, we perform a t-test against the two groups. For this let's start by splitting the data by the dosage, so that we can perform t-test on each dosage size group.

```
dose_0.5 <- filter(ToothGrowth, dose == 0.5)
dose_1.0 <- filter(ToothGrowth, dose == 1.0)
dose_2.0 <- filter(ToothGrowth, dose == 2.0)
```

This will help us test if the OJ or VC supp methods have any statistical difference in lean length.

Perform p-value test for dosage = 0.5.

```
t_test_dose_0.5 <- t.test(len ~ supp, data = dose_0.5, paired = FALSE)
print(t_test_dose_0.5)
```

```
##
##  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
```

We can see that with a P-value of 0.006, the results are not statistically significant at the 95% confidence level and that the confidence interval contains zero. Thus, we cannot reject the null hypothesis that the dosage method has no effect on tooth length.

Now perform p-value test for dosage 1.0:

```
t_test_dose_1.0 <- t.test(len ~ supp, data = dose_1.0, paired = FALSE)
print(t_test_dose_1.0)
```

```
##
##  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
```

In this instance, the P-value is 0.001 and we can reject the null hypothesis at the 95% confidence level. We can also see that the 95% confidence interval excludes 0.

Now perform p-value test for dosage 2.0:

```
t_test_dose_2.0 <- t.test(len ~ supp, data = dose_2.0, paired = FALSE)
print(t_test_dose_2.0)
```

```
##
##  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
```

In this instance, the P-value is 0.9 and we can *NOT* reject the null hypothesis at the 95% confidence level. We can also see that the 95% confidence interval includes 0.

Assumptions and Conclusions

Assuming that the sample is a simple random sample and that the data follows a normal probability distribution. From the tests we can reject the null hypothesis that Vitamin C dosage method at the dosage levels 0.5 and 1.0 mg/day has no statistically significant effect on tooth growth at the 95% confidence level, except when the dosage is high - 2.0 mg/ml.