Statistical Inference Course Project

yhuai

2018/4/21

Overview

This is the Coursera Johns Hopkins University Data Science Specialization – Statistical Inference Course Project

The project consists of two parts:

1.A simulation exercise. 2.Basic inferential data analysis.

Part 1: Simulation Exercise Instructions

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.

```
set.seed(19930319)
lambda <- 0.2; n <- 40;
means = NULL
for (i in 1:1000) means = c(means, mean(rexp(n,lambda)))
hist(means,breaks=50,xlim=c(2,8))

mean(means)
```

```
## [1] 4.999559
```

```
abline(v=mean(means),col="red",lwd="2")
```

Show the sample mean and compare it to the theoretical mean of the distribution.

```
a1 <- lambda^-1
theoretical_mean <- a1
b1 <- mean(means)
sample_mean <- b1
data.frame(theoretical_mean,sample_mean)
```

Show how variable it is and compare it to the theoretical variance of the distribution.

```
a2 <- (1/lambda)^2/n
theoretical_var <- a2
b2 <- var(means)
sample_var <- b2
data.frame(theoretical_var,sample_var)
```

Based on Central Limit Theorem, our number of smaples is big(1000), the simulated exponential distribution is close to normal distribution.

```
x <- seq(2,8,length=2*n)
y <- dnorm(x,mean(means),sd(means))
plot(x,y,type = "l",col="red")
```

Part 2: Basic Inferential Data Analysis Instructions

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

```
## '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 ...
```

```
##       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  
```

From the summary function and the plot, we can find that the tooth growth length is increasing as dosage increasing. The two kind of supplements-OJ and VC, looks have similar effects when dosage is 2.

```
g <- ggplot(data,aes(x=dose, y=len))
g <- g + facet_grid(.~ data$supp)
g <- g + geom_bar(stat = "identity",aes(fill=supp))
g
```

Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

We assume that OJ and VC have the same effect on tooth growth.

```r
#Hypothesis
h1 <- t.test(len ~ supp, data=subset(data,dose==0.5))
h1
```

```
##
##  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
```

```r
p1 <- h1$p.value
h2 <- t.test(len ~ supp, data=subset(data,dose==1.0))
h2
```

```
##
##  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
```

```r
p2 <- h2$p.value
h3 <- t.test(len ~ supp, data=subset(data,dose==2.0))
h3
```

```
##
##  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
```

```r
p3 <- h3$p.value
data.frame(p1,p2,p3)
```

Conclustions

As the hypothesis part show, the confidence interval is 95%, and only when the dosage is 2, the p-value is greater than 0.05 threshold. So we can say, the OJ and VC have the same effect only when dosage is 2.0