Assignment3\_stat359

Koki Itagaki

2023-02-15

#1. The following data represent the running times of films produced by #two motion-picture companies: #Test the hypothesis that the average running time of films produced by company #2 exceeds the average running time of films produced by company 1 by 10 #minutes against the one-sided alternative that the difference is less than 10 #minutes. Use a 0.1 level of significance. Please consider carefully assumptions #made on the data.

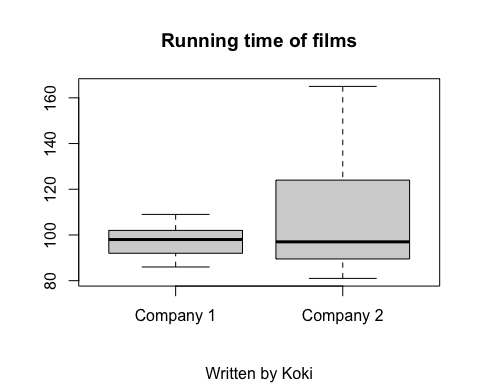
#Since both of the sample sizes are pretty small, and we do not know sigmas,  
#I use t distribution.  
#Test the hypothesis that the average running time of films produced by company  
#2 exceeds the average running time of films produced by company 1 by 10  
#minutes against the one-sided alternative that the difference is less than 10  
#minutes. Use a 0.1 level of significance. Please consider carefully assumptions  
#made on the data.  
c1<-c(102, 86, 98, 109, 92)  
c2<-c(81, 165, 97, 134, 92, 87, 114)  
summary(c1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 86.0 92.0 98.0 97.4 102.0 109.0

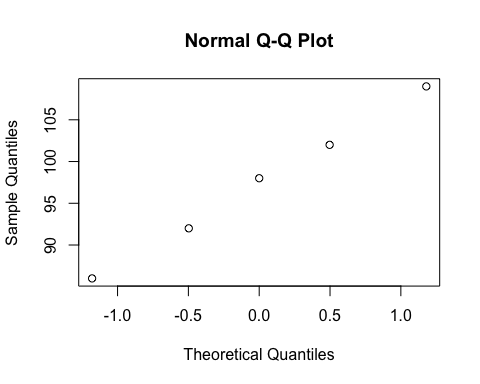
summary(c2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 81.0 89.5 97.0 110.0 124.0 165.0

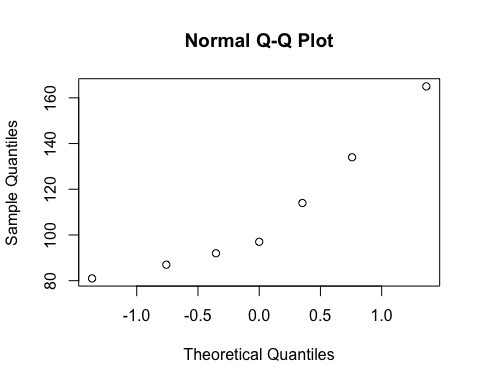
boxplot(c1,c2, names = c("Company 1", "Company 2"),  
 main = "Running time of films",sub = "Written by Koki")



#According to the boxplots, means are the almost the same,but the range of the   
#data from Company 2 is much larger than the data from company 1  
qqnorm(c1)



qqnorm(c2)



#There is a streighrt line in both of qq plots which means these data are   
#Normally distributed  
#To decide which t-test I will use, I need to know if the sample   
#variance of 2 different data sets are the same or not   
var.test(c1,c2,alternative = "two.sided",conf.level = 0.90)

##   
## F test to compare two variances  
##   
## data: c1 and c2  
## F = 0.086277, num df = 4, denom df = 6, p-value = 0.03298  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 90 percent confidence interval:  
## 0.01903033 0.53173886  
## sample estimates:  
## ratio of variances   
## 0.08627737

#Since P value = 0,003298 <= a = 0.1, this is a significant evidence against Ho.  
#Therefore, the variance is different.  
#Since the variance is different, I use Welch's t-test  
  
t.test(c1,c2,alternative = "less",mu = 10, var.equal = FALSE,conf.level = 0.90)

##   
## Welch Two Sample t-test  
##   
## data: c1 and c2  
## t = -1.8689, df = 7.3756, p-value = 0.05085  
## alternative hypothesis: true difference in means is less than 10  
## 90 percent confidence interval:  
## -Inf 4.420568  
## sample estimates:  
## mean of x mean of y   
## 97.4 110.0

#Since the p-value = 0.05085 <= a = 0.1, we reject Ho.  
#There is a significant evidence that the average running time of films produced  
#by company 2 exceeds the average running time of films produced by company 1  
#by 10 minutes against the one-sided alternative that the difference is less  
#than 10 minutes

#Question 2 #Six different machines are being considered for use in manufacturing rubber #seals. The machines are being compared with respect to tensile strength of the #product. A random sample of four seals from each machine is used to #determine whether the mean tensile strength varies from machine to #machine. The following are the tensile-strength measurements in kilograms #per square centimeter 10E-01.

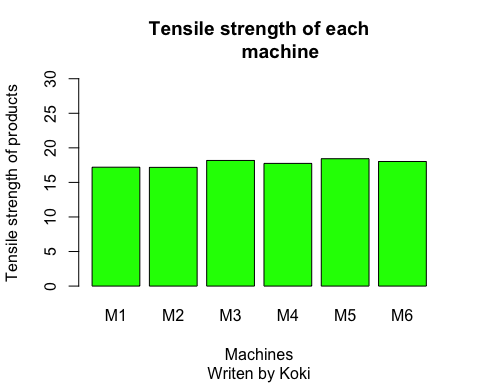
#k = 6, n = 4, N =24, a = 0.05  
  
#an analysis of variance at the 0.05 significance level.  
#The hypothesis is Ho: u1 = u2 = ...= u6 Ha: at least one u is different  
dataframe<-data.frame(strength = c(17.5, 16.9,15.8,18.6,16.4,19.2,17.7,15.4,  
20.3,15.7,17.8,18.9,14.6,16.7,20.8,18.9,17.5,19.2,  
16.5,20.5,18.3,16.2,17.5,20.1),  
Machines= c("M1","M1","M1","M1","M2","M2","M2","M2","M3","M3","M3","M3","M4",  
 "M4","M4","M4","M5","M5","M5","M5","M6","M6","M6","M6"))  
dataframe

## strength Machines  
## 1 17.5 M1  
## 2 16.9 M1  
## 3 15.8 M1  
## 4 18.6 M1  
## 5 16.4 M2  
## 6 19.2 M2  
## 7 17.7 M2  
## 8 15.4 M2  
## 9 20.3 M3  
## 10 15.7 M3  
## 11 17.8 M3  
## 12 18.9 M3  
## 13 14.6 M4  
## 14 16.7 M4  
## 15 20.8 M4  
## 16 18.9 M4  
## 17 17.5 M5  
## 18 19.2 M5  
## 19 16.5 M5  
## 20 20.5 M5  
## 21 18.3 M6  
## 22 16.2 M6  
## 23 17.5 M6  
## 24 20.1 M6

attach(dataframe)  
Strength<-tapply(strength,Machines,mean)  
Strength

## M1 M2 M3 M4 M5 M6   
## 17.200 17.175 18.175 17.750 18.425 18.025

#par(mfrow=c(2,3))  
barplot(Strength,col = "Green", ylim = c(0,30),main = "Tensile strength of each  
 machine",xlab = "Machines",  
 ylab = "Tensile strength of products", sub = "Writen by Koki")



#The means of tensile strength of products for the 4 different   
#machines are the almost same  
   
   
error.bars<-function(y,z){  
 x<-barplot(y, plot=F)  
 n<-length(y)  
 for (i in 1:n)  
 {  
 arrows(x[i],y[i]-z[i],x[i],y[i]+z[i],code=3,angle=90,length=0.15)  
 }  
}  
sigma.hat<-summary.lm(aov(strength~Machines))$sigma  
sigma.hat

## [1] 1.865476

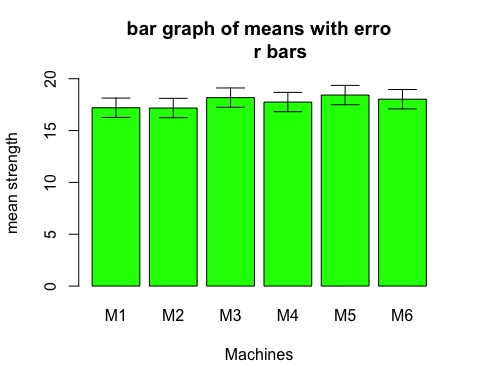
table(Machines)

## Machines  
## M1 M2 M3 M4 M5 M6   
## 4 4 4 4 4 4

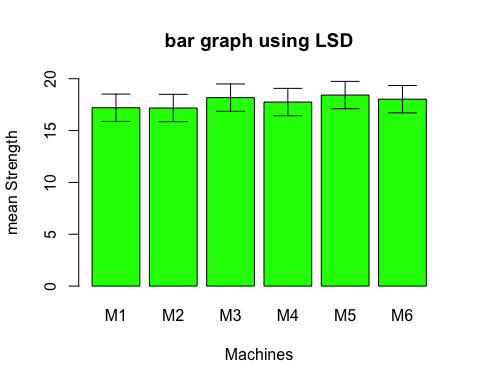
se.mean<-sigma.hat/sqrt(4)  
se.mean

## [1] 0.9327379

barplot(Strength, col="green", ylim=c(0,20),main = "bar graph of means with erro  
 r bars",ylab="mean strength", xlab = "Machines")  
bar.half.width<-rep(se.mean,6)  
error.bars(Strength,bar.half.width)



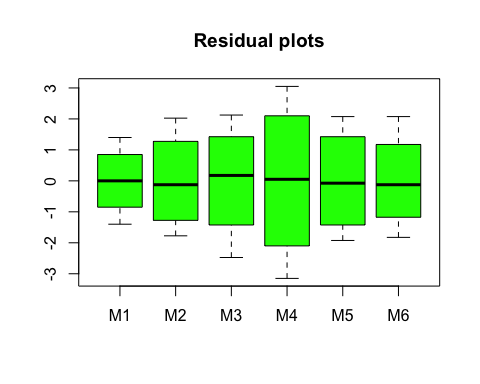
#According to the barplot with error bars, it is clear that the error bars are  
#overlaped. This means all means seem to be the same.  
  
#For certainly, I also use the least significant difference method.  
  
LSD<-2\*sqrt(2)\*se.mean  
LSD.bars<-rep(LSD,6)/2  
barplot(Strength, col="green", ylim=c(0,20),main = "bar graph using LSD",ylab=  
 "mean Strength", xlab ="Machines")  
error.bars(Strength,LSD.bars)



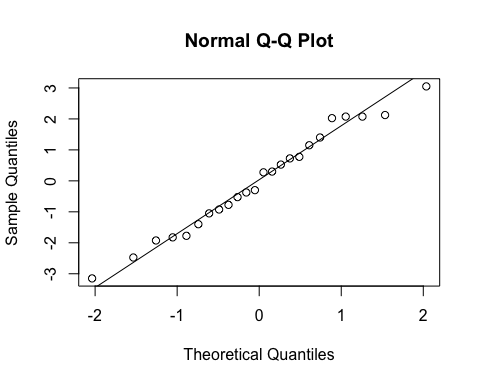
#From this graph using LSD, I can see that the error bars are overlaped as well.  
#so There is not significant difference among means of the strength  
#from different Machines  
  
  
summary(aov(strength~Machines))

## Df Sum Sq Mean Sq F value Pr(>F)  
## Machines 5 5.34 1.068 0.307 0.902  
## Residuals 18 62.64 3.480

#According to the ANOVA table, the p-value is 0.902.  
#Since the p-value >= 0.05 = a, we fail to reject Ho.  
#There is a insignificant evidence that at least one mean of 6 machines are  
#different.  
  
  
resid.plant<-resid(aov(strength~Machines))  
boxplot(resid.plant[Machines=="M1"],resid.plant[Machines=="M2"],  
 resid.plant[Machines=="M3"],resid.plant[Machines=="M4"],  
 resid.plant[Machines=="M5"],resid.plant[Machines=="M6"],  
 main = "Residual plots",  
 names=c('M1','M2','M3','M4','M5', 'M6'),  
 col="green")



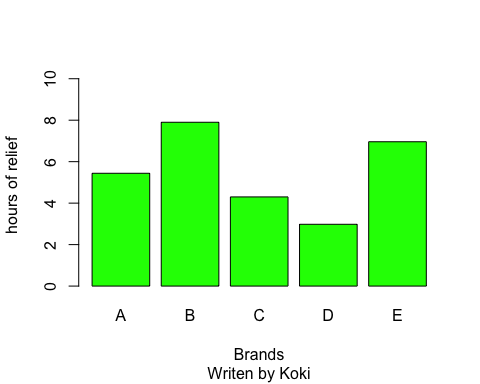
qqnorm(resid.plant)  
qqline(resid.plant)



#According to the qq plot of residuals, there is a clear streight line on the   
#graph. This means the residuals are normally distributed.

#Q3 #The data in the following table represent the number of hours of relief #provided by five different brands of headache tablets administered to 25 #ubjects experiencing fevers of 38 degrees Celsius or more. Perform the #analysis of variance and test the hypothesis at the 0.05 level of significance #that the mean number of hours of relief provided by the tablets is the same for #all five brands. Discuss the results.

data\_hours<-data.frame(Hours = c(5.2,4.7,8.1,6.2,3.0,9.1,7.1,8.2,6.0,9.1,  
 3.2,5.8,2.2,3.1,7.2,2.4,3.4,4.1,1.0,4.0,  
 7.1,6.6,9.3,4.2,7.6),  
 Brands= c("A","A","A","A","A","B","B","B","B","B","C","C","C",  
 "C","C","D","D","D","D","D","E","E","E","E","E"))  
  
  
  
attach(data\_hours)  
mhours<-tapply(Hours,Brands,mean)  
#par(mfrow=c(2,3))  
barplot(mhours,col = "Green", ylim = c(0,10),xlab = "Brands",  
 ylab = "hours of relief ", sub = "Writen by Koki")



#From the bar graph, it is clear that the means of brand c and d are much   
#smaller than others.  
  
error.bars<-function(y,z){  
 x<-barplot(y, plot=F)  
 n<-length(y)  
 for (i in 1:n)  
 {  
 arrows(x[i],y[i]-z[i],x[i],y[i]+z[i],code=3,angle=90,length=0.15)  
 }  
}  
sigma.hat<-summary.lm(aov(Hours~Brands))$sigma  
sigma.hat

## [1] 1.725283

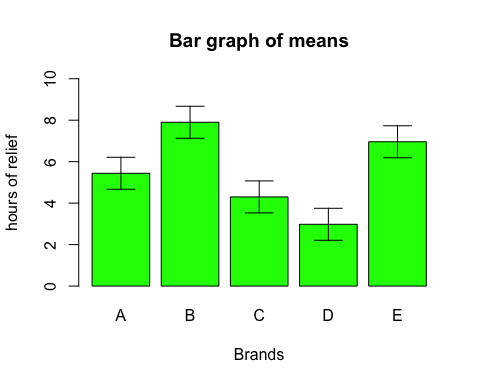
table(Brands)

## Brands  
## A B C D E   
## 5 5 5 5 5

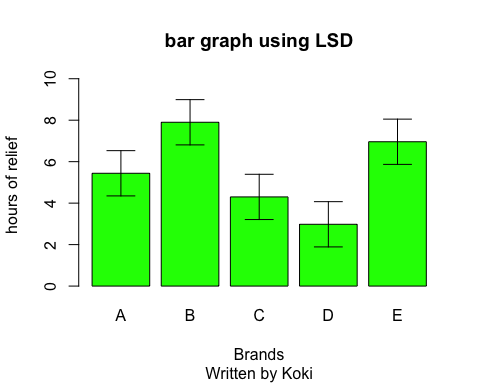
se.mean<-sigma.hat/sqrt(5)  
se.mean

## [1] 0.7715698

barplot(mhours, col="green", ylim=c(0,10),main = "Bar graph of means"  
 ,ylab="hours of relief", xlab = "Brands")  
bar.half.width<-rep(se.mean,6)  
error.bars(mhours,bar.half.width)



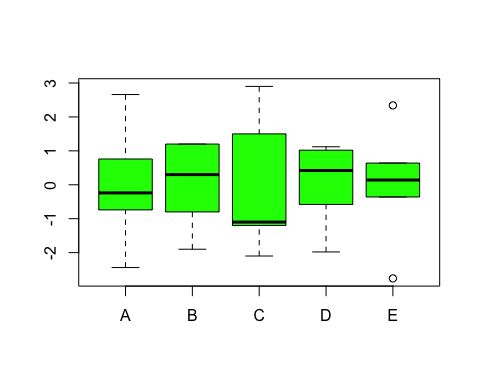
#According to the barplot with error bars, it is clear that the error bars of C  
#and D are quite small and do not overlap with the other error bars.  
#This means all means seem not to be the same.  
  
#For certainly, I also use the least significant difference method.  
  
LSD<-2\*sqrt(2)\*se.mean  
LSD.bars<-rep(LSD,6)/2  
barplot(mhours, col="green", ylim=c(0,10),main = "bar graph using LSD",  
 ylab="hours of relief ", xlab =  
 "Brands", sub = "Written by Koki")  
error.bars(mhours,LSD.bars)



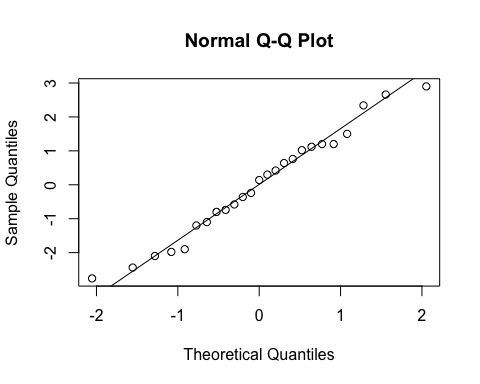
#From this graph, I can see that the all error bars are not overlaped as well.  
#The means from brands B and E are pretty large compared to the means from   
#brands C and D. This means means of C and D are not the same as the means   
#of B and E.  
  
  
summary(aov(Hours~Brands))

## Df Sum Sq Mean Sq F value Pr(>F)   
## Brands 4 78.42 19.605 6.587 0.0015 \*\*  
## Residuals 20 59.53 2.977   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#According to the ANOVA table, the p-value is 0.0015.  
#Since the p-value >= 0.05 = a, we reject Ho.  
#There is a significant evidence that at least one mean that the number of hours  
#of relief provided by five different brands of headache tablets is different  
  
  
resid.plant<-resid(aov(Hours~Brands))  
boxplot(resid.plant[Brands=="A"],resid.plant[Brands=="B"],  
 resid.plant[Brands=="C"],resid.plant[Brands=="D"],  
 resid.plant[Brands=="E"],  
 names=c('A','B','C','D','E'),  
 col="green")



#From the residual plots, we can see that mean of c is pretty low.  
#Also there is outliers in graph E.  
  
  
qqnorm(resid.plant)  
qqline(resid.plant)



#From the qq plot, the data make the stright line.  
#This means the data of residuals is normally distributed.