Assignment-5.R

Purbasha Chatterjee

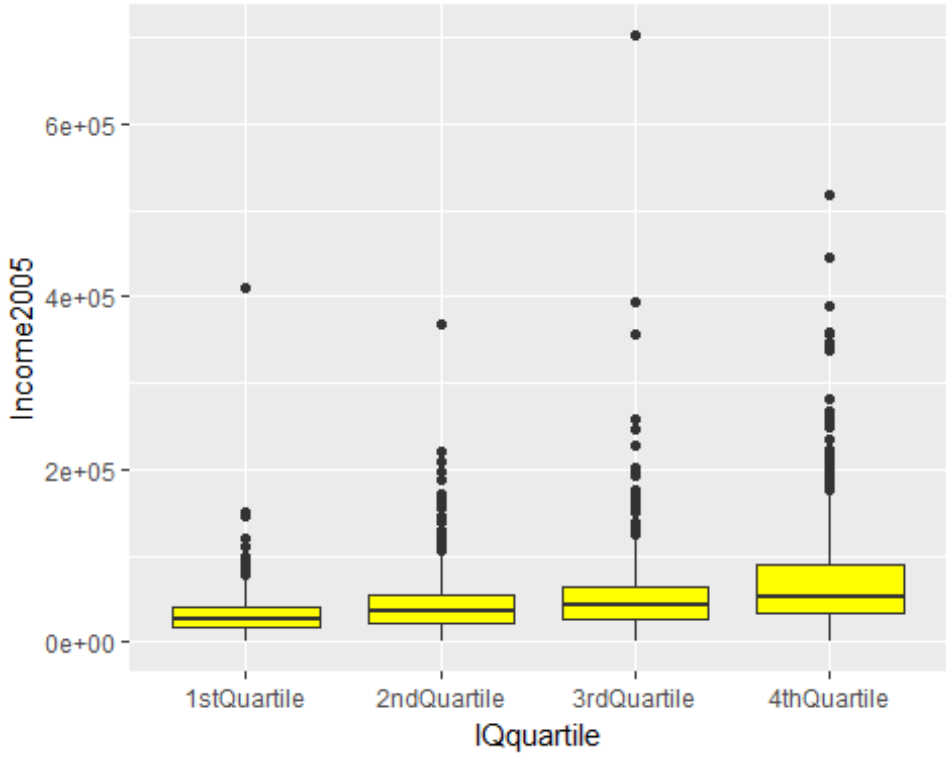
Fri Nov 03 20:26:13 2017

```
 Q1 - Ex:5.17
 We have total df = 31 , given within groups df = 24, hence dfof between grou
ps = 31-24 = 7
 Total sum sqaure = 70907 and within groups sum of sqaure = 35088 , thus betw
een groups sum of sqaure = 70907 - 35088 = 35819
 Mean sqaure = (sum of sq)/df
 between groups mean sq = 35819/7 = 5117
 within groups mean sq = 35088/24 = 1462
 F = (between groups mean sq)/(within groups mean sq) = 5117/1462 = 3.5
 Using F-statistics p-value = 0.009942

 Df_BG = Df_Tot - Df_WG = 31-24 = 7
 The number of groups are,
 Df_BG = k - 1 ,
 Hence, 7 = k - 1
 k = 7+1 = 8 groups

 Since the p-value is less than 0.05 so we can reject the null hypothesis and
state that there exist atleast one group mean value which is different
```

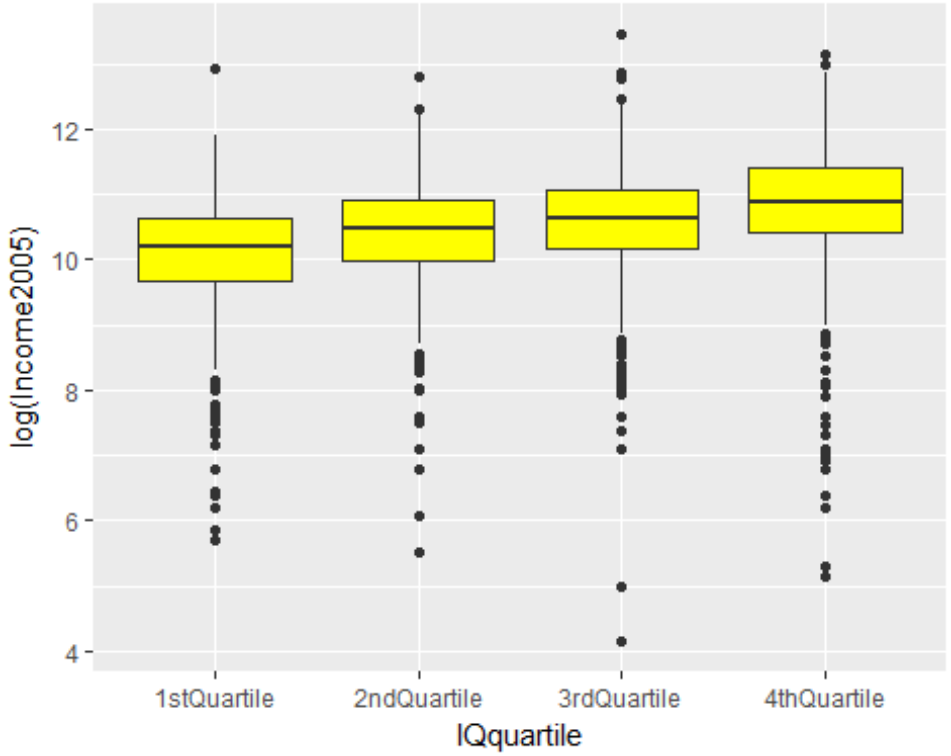
Source	df	Sum of sq	Mean sq	F	p-value
Between groups	7	35819	5117	3.5	0.009942
Within groups	24	35088	1462		
Total	31	70907			

```
 Q2 - Ex:5.20
 Assuming all 10 means having the same mean (mu) and the same variance (sigma
2).
 The sample variance of the sample means therefore estimates (sigma^2/24)
 Hence s^2 = sigma^2/24
 mu = (13.6 + 15.4 + 14.7 + 12.4 + 9.2 + 13.7 + 10.3 + 7 + 9.5 + 9.5)/10 = 11
.53
 s^2 = [sum (x-mu)^2]/(N-1)
 s^2 = [(4.2849)+(14.9769)+(10.0489)+(0.7569)+(5.4289)+(4.7089)+(1.5129) +(20
.5209)+(4.1209)+(4.1209)]/(10-1)
 s^2 = 70.481/9 = 7.8312
 Since s^2 = sigma^2/24 ; thus 7.8312 = sigma^2/24
```

```
 sigma^2 = 7.8312 * 24 = 187.949
 sigma = 13.71

Q3 - Ex:5.24
library(Sleuth3)

## Warning: package 'Sleuth3' was built under R version 3.4.2

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.2

step1 <- aov(Income2005 ~ IQquartile, data = ex0524)
summary(step1)

##               Df    Sum Sq   Mean Sq F value Pr(>F)
## IQquartile     3 4.934e+11 1.645e+11   82.44 <2e-16 ***
## Residuals   2580 5.147e+12 1.995e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

step1b <- aov(log(Income2005) ~ IQquartile, data = ex0524)
summary(step1b)

##               Df Sum Sq Mean Sq F value Pr(>F)
## IQquartile     3  184.8   61.60   70.17 <2e-16 ***
## Residuals   2580 2265.0    0.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ggplot(ex0524, aes(IQquartile, Income2005)) +
  geom_boxplot(fill = "yellow") + ylab("Income2005") + xlab("IQquartile")
```




```
ggplot(ex0524, aes(IQquartile, log(Income2005))) +
  geom_boxplot(fill = "yellow") + ylab("log(Income2005)") + xlab("IQquartile"
)
```



```r
quartile <- ex0524
quartile$high_low = with(ex0524, ifelse(IQquartile == "4thQuartile",1,0))
quartile <- quartile[!(quartile$IQquartile == "3rdQuartile" | quartile$IQquar
tile == "2ndQuartile"),]

step2 <- aov(Income2005 ~ IQquartile, data = quartile)
summary(step2)

##               Df    Sum Sq   Mean Sq F value Pr(>F)
## IQquartile     1 4.638e+11 4.638e+11   205.2 <2e-16 ***
## Residuals   1290 2.915e+12 2.260e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

step2b <- aov(log(Income2005) ~ IQquartile, data = quartile)
summary(step2b)

##               Df Sum Sq Mean Sq F value Pr(>F)
## IQquartile     1  174.3  174.35   185.1 <2e-16 ***
## Residuals   1290 1215.0    0.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(step2b <- t.test(Income2005 ~ IQquartile, data = quartile, var.equal = TRUE)
)

##
##  Two Sample t-test
##
## data:  Income2005 by IQquartile
## t = -14.326, df = 1290, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -43081.55 -32703.55
## sample estimates:
## mean in group 1stQuartile mean in group 4thQuartile
##                  32035.85                  69928.40

(step2c <- t.test(log(Income2005) ~ IQquartile, data = quartile, var.equal =
TRUE))

##
##  Two Sample t-test
##
## data:  log(Income2005) by IQquartile
## t = -13.605, df = 1290, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8406283 -0.6287509
## sample estimates:
```

```
## mean in group 1stQuartile mean in group 4thQuartile
##                   10.05009                   10.78478

## Thus the mean difference of the group is (10.78478 - 10.05009) = 0.73649

step2c$statistic^2

##        t
## 185.1016

anova(step2b)$`F value`[1]

## [1] 185.1016

kruskal.test(Income2005 ~ IQquartile, data = quartile)

##
##  Kruskal-Wallis rank sum test
##
## data:  Income2005 by IQquartile
## Kruskal-Wallis chi-squared = 232.62, df = 1, p-value < 2.2e-16
```

Statistical Summary:

- Without applying log transformation, it can be observed from the box plot that the income values are very low and the plot is skewed. Hence, log transformation was applied.
- After transformation, when ANOVA was applied, the p-value appeared to be less than 0.05 giving strong evidence of existence of difference in mean of at least one group and leading to rejection of the null hypothesis. The degree of freedom for residuals is 2580 whereas df for the group is 3. Moreover, mean square is 0.88 and 61.6 respectively. The F-statistics was observed to 70.17.
- The 1st and 4th quartile were measured by applying ANOVA which again rejected the null hypothesis with p-value less than 0.05. The degree of freedom for residuals is 1290 whereas df for the group is 1. Additionally, mean square is 0.94 and 174.35 respectively. The F-statistics was observed to 185.1.
- From two sample t-test , the df of the residuals is same. We observe that in this case p-value is same that is less than 0.05. The 95% CI was observed to be 0.629 to 0.841.
- The t-statistic (13.605) is observed to be the square root of the F-statistic (185.096) value.
- The two mean location of the distribution is observed to be `10.78478 - 10.05009 = 0.73649. Thus, the percentage by which the 4th quartile exceed in terms of mean location` distribution `is 6.8%`