BACS HW7 - 109006234

Credit: 109006278

April 9th 2023

Additional Functions

```
norm_qq_plot <-function(values, main) {
    probs1000 <- seq(0, 1, 0.001)
    q_vals <- quantile(values, probs=probs1000)
    q_norm <- qnorm(probs1000,mean=mean(values), sd = sd(values))
    plot(q_norm, q_vals,
        xlab = "normal quantiles", ylab = "values quantiles",
        main = main)
    abline(a=0, b=1, col="red", lwd = 2)
}
```

Loading The Datas

```
data1 <- read.csv("G:/My Drive/111_2_BACS/HW7/pls-media1.csv")$INTEND.0
data2 <- read.csv("G:/My Drive/111_2_BACS/HW7/pls-media2.csv")$INTEND.0
data3 <- read.csv("G:/My Drive/111_2_BACS/HW7/pls-media3.csv")$INTEND.0
data4 <- read.csv("G:/My Drive/111_2_BACS/HW7/pls-media4.csv")$INTEND.0
dataset <- list(data1, data2, data3, data4)
```

Problem 1

(a) What are the means of viewers' intentions to share (INTEND.0) on each of the four media types?

```
data1_mean <- mean(data1)
data2_mean <- mean(data2)
data3_mean <- mean(data3)
data4_mean <- mean(data4)
```

```
## [1] "The mean of pls-media1.csv:  4.80952380952381"
```

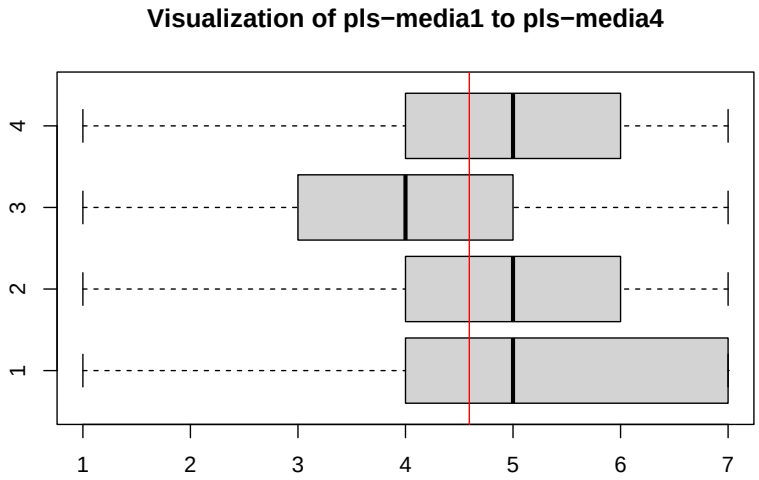
```
## [1] "The mean of pls-media2.csv:  3.94736842105263"
```

```
## [1] "The mean of pls-media3.csv:  4.725"
```

```
## [1] "The mean of pls-media4.csv:  4.89130434782609"
```

(b) Visualize the distribution and mean of intention to share, across all four media.

```r
boxplot(rev(dataset), horizontal=TRUE, main="Visualization of pls-media1 to pls-media4")
abline(v=mean(sapply(dataset, mean)), col="red")
```



(c) From the visualization alone, do you feel that media type makes a difference on intention to share?
From the boxplot, we can determine that pls-media2 has the smallest mean compared to the other datasets.

Problem 2

(a) State the null and alternative hypotheses when comparing INTEND.0 across four groups in ANOVA

```r
#H0: The means of the medias are the same
#H1: The means of the medias are not the same
```

(b) Let's compute the F-statistic ourselves
(i) Show the code and results of computing MSTR, MSE, and F

```r
sstr <- sum(sapply(dataset, length)*(sapply(dataset, mean)-
    mean(sapply(dataset, mean)))^2)
df_mstr <- 4-1
mstr <- sstr/df_mstr
```

```
## [1] "The MSTR value:  7.53238956378754"
```

```r
sse <- sum((sapply(dataset, length)-1)*sapply(dataset, var))
df_mse <- sum(sapply(dataset, length)) - 4
mse <- sse/df_mse
```

```
## [1] "The MSE value:  2.86915092010757"
```

```r
f_value <- mstr/mse
```

```
## [1] "The F-value:  2.62530266741951"
```

(ii) Compute the p-value of F, from the null F-distribution; is the F-value significant? If so, state your conclusion for the hypotheses.

```r
qf(p=0.95, df1=df_mstr, df2=df_mse)
```

```
## [1] 2.660406
```

```r
p_value <- pf(f_value, df_mstr, df_mse, lower.tail=FALSE)
```

```
## [1] "The P-value:  0.0523068574612358"
```

Since the F-value is less than the critical F-value, we fail to reject the null hypothesis. We can also observe that the P-value is not significant because it is less than the significant level specified in the question.

(c) Conduct the same one-way ANOVA using the aov() function in R – confirm that you got similar results.

```r
anova_data <- melt(dataset, id.vars = NULL,
    variable.name = "MediaType",
    value.name = "Intend")
anova_model <- aov(anova_data$Intend ~ factor(anova_data$L1))
summary(anova_model)
```

```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## factor(anova_data$L1)    3   22.5   7.508   2.617 0.0529 .
## Residuals              162  464.8   2.869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing the results, we can conclude that both calculations share similar results.

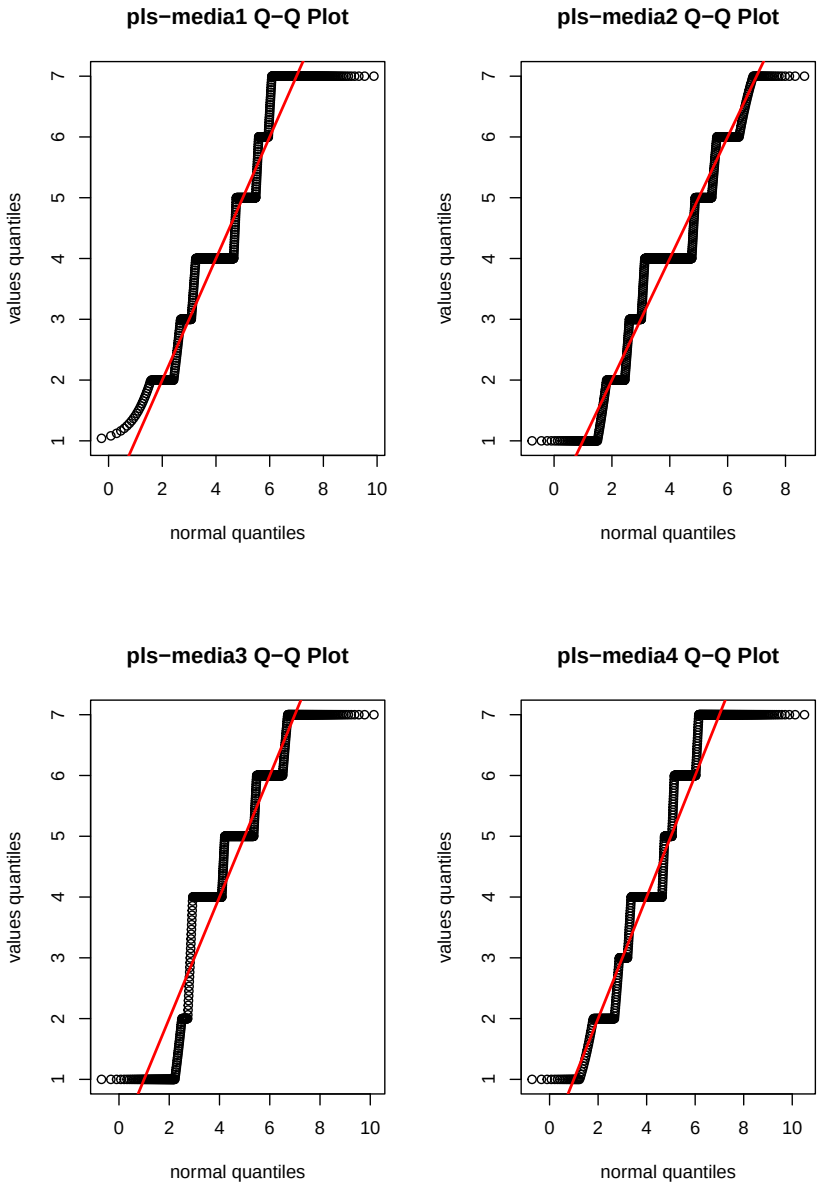
(d) Regardless of your conclusions, conduct a post-hoc Tukey test to see if any pairs of media have significantly different means – what do you find?

```r
TukeyHSD(anova_model, conf.level = 0.05)
```

```
##   Tukey multiple comparisons of means
##     5% family-wise confidence level
##
## Fit: aov(formula = anova_data$Intend ~ factor(anova_data$L1))
##
## $`factor(anova_data$L1)`
##             diff         lwr        upr     p adj
## 2-1 -0.86215539 -1.06562977 -0.6586810 0.1085727
## 3-1 -0.08452381 -0.28530983  0.1162622 0.9959223
## 4-1  0.08178054 -0.11218249  0.2757436 0.9959032
## 3-2  0.77763158  0.57175512  0.9835080 0.1825044
## 4-2  0.94393593  0.74470805  1.1431638 0.0573229
## 4-3  0.16630435 -0.03017708  0.3627858 0.9687417
```

From the p-adj value, we can observe that it is much greater than the specified alpha, which is 0.05. Hence, we fail to reject the null hypothesis. In addition to that, we also observe that each difference of the media data is not significant.

(e) Do you feel the classic requirements of one-way ANOVA were met?
Each treatment/population's response variable is normally distributed



As can be seen from the Q-Q plot, all media datas are not normally distributed. Hence, the first assumption fails.

The variance of the response variables is the same for all treatments/populations

```
bartlett.test(dataset)
```

```
##
##  Bartlett test of homogeneity of variances
```

```
##
## data:  dataset
## Bartlett's K-squared = 1.3958, df = 3, p-value = 0.7065
```

From this result, we can observe that the p-value is greater than the specified alpha. Hence, we can assume that the variances are equal, which proves the second assumption.

The observations are independent
We assume that the observation are independent.

Problem 3

(a) State the null and alternative hypotheses

```
#H0: The medians of the medias are the same
#H1: The medians of the medias are not the same
```

(b) Let's compute (an approximate) Kruskal Wallis H ourselves

```
anova_data$ranks <- rank(anova_data$Intend)
sum_ranks <- tapply(anova_data$ranks, anova_data$L1, sum) ; sum_ranks
```

```
##      1      2      3      4
## 3693.5 2421.0 3556.0 4190.5
```

```
R <- c(sum_ranks[[1]]^2 / length(data1),
    sum_ranks[[2]]^2 / length(data2),
    sum_ranks[[3]]^2 / length(data3),
    sum_ranks[[4]]^2 / length(data4))
N <- sum(length(data1), length(data2), length(data3), length(data4))
```

(i) Show the code and results of computing H

```
H <- (12/(N*(N+1))) * sum(R) - 3*(N+1)
```

```
## [1] "H:  8.45465979544389"
```

(ii) Compute the p-value of H, from the null chi-square distribution; is the H value significant? If so, state your conclusion of the hypotheses.

```
kruskal_p <- 1 - pchisq(H, df=4-1)
```

```
## [1] "Kruskal's P-value:  0.037492918119218"
```

From this result, we observe that the Kruskal's p-value is smaller than the significant value that is 0.05. Hence, we can reject the null hypothesis.

(c) Conduct the same test using the kruskal.wallis() function in R – confirm that you got similar results.

```r
kruskal.test(Intend~L1, data=anova_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Intend by L1
## Kruskal-Wallis chi-squared = 8.8283, df = 3, p-value = 0.03166
```

From the results above, we can conclude that using both ways, the results are pretty much the same.

(d) Regardless of your conclusions, conduct a post-hoc Dunn test to see if the values of any pairs of media are significantly different – what are your conclusions?

```r
dunnTest(Intend~L1, data=anova_data, method = "bonferroni")
```

```
## Warning: L1 was coerced to a factor.
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison
```

```
##   p-values adjusted with the Bonferroni method.
```

```
##   Comparison           Z     P.unadj      P.adj
## 1      1 - 2  2.30087819 0.021398517 0.12839110
## 2      1 - 3 -0.09233644 0.926430736 1.00000000
## 3      2 - 3 -2.36408588 0.018074622 0.10844773
## 4      1 - 4 -0.31452459 0.753122646 1.00000000
## 5      2 - 4 -2.65613380 0.007904225 0.04742535
## 6      3 - 4 -0.21613379 0.828883460 1.00000000
```

From the results above, we can see that pls-media2 and pls-media4 pair has a smaller p-value and adj p-value compared to the significant value that is 0.05. Hene, we can say that there are significant difference between the two groups.