HW7

110077443

3/28/2022

Please note that all code in this document is presented in a grey box and the output reflected below each box

- The below code allows lengthy lines of code to display neatly within the grey box (wrapping it)

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

1 Importing and develop some intuition about the data and results

```
media1 <- read.csv("pls-media1.csv", header = T)
media2 <- read.csv("pls-media2.csv", header = T)
media3 <- read.csv("pls-media3.csv", header = T)
media4 <- read.csv("pls-media4.csv", header = T)
```

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

```
mean(media1$INTEND.0)  # 4.809524
```

```
## [1] 4.809524
```

```
mean(media2$INTEND.0)  # 3.947368
```

```
## [1] 3.947368
```

```
mean(media3$INTEND.0)  # 4.75
```

```
## [1] 4.725
```

```
mean(media4$INTEND.0)  # 4.891304
```

```
## [1] 4.891304
```

```r
media_list <- list(media1$INTEND.0, media2$INTEND.0, media3$INTEND.0,
    media4$INTEND.0)  # Converting data to a list
```

b) Visualize the distrrevibution and mean of intention to share, across all four media

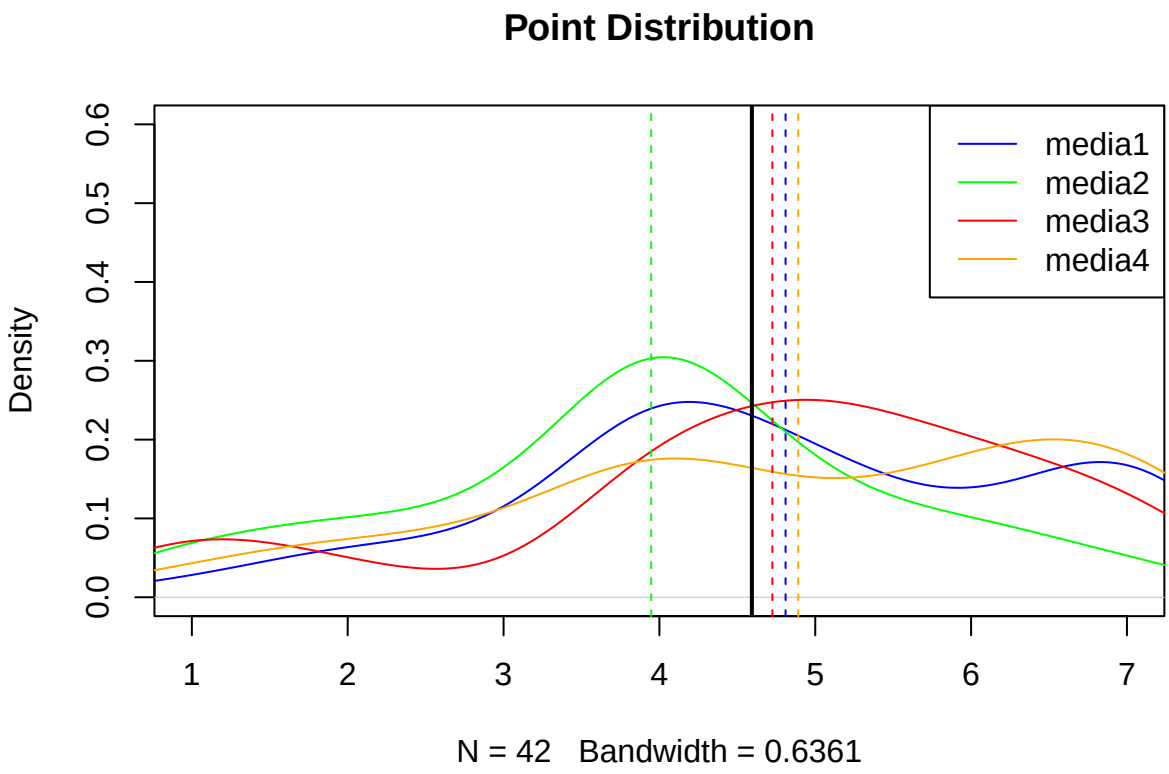
```r
# First calculate grand mean
all_means <- sapply(media_list, mean)  # All four means
all_means  # Print
```

```
## [1] 4.809524 3.947368 4.725000 4.891304
```

```r
total_mean <- mean(all_means)  # Grand mean
total_mean  # Print
```

```
## [1] 4.593299
```

```r
# Creating a function
media_plots <- function() {
    plot(density(media1$INTEND.0), xlim = c(1, 7), ylim = c(0,
        0.6), col = "blue", main = "Point Distribution")
    lines(density(media2$INTEND.0), col = "green")
    lines(density(media3$INTEND.0), col = "red")
    lines(density(media4$INTEND.0), col = "orange")
    abline(v = mean(media1$INTEND.0), col = "blue", lty = "dashed")
    abline(v = mean(media2$INTEND.0), col = "green", lty = "dashed")
    abline(v = mean(media3$INTEND.0), col = "red", lty = "dashed")
    abline(v = mean(media4$INTEND.0), col = "orange", lty = "dashed")
    abline(v = total_mean, col = "black", lwd = 2)  # Grand Mean
    legend(x = "topright", lty = 1, c("media1", "media2", "media3",
        "media4"), col = c("blue", "green", "red", "orange"))
}
media_plots()  # Print
```



c) From the visualization alone, do you feel that media type makes a difference on intention to share?

- From the visualization, we can see that the grand mean crosses through all four media types.
- The density plot indicates that the data for all four media types behave similar with slight variation between means.
- Thus, the media type **would not** really make a difference on intention to share

2) Traditional one-way ANOVA

a) State the null and alternative hypotheses

- H_null: mu1 = mu2 = mu3 = mu4
- H_alt: The means are NOT the same

b) Compute the F-statistic

i) Show the code and results of computing MSTR, MSE, and F

```
# Calculating MSTR
sstr <- sum(sapply(media_list, length) * ((all_means) - (total_mean))^2)  #22.59717
sstr  # Print
```

```
## [1] 22.59717
```

```
df_mstr <- 4 - 1
mstr <- sstr/df_mstr  # 7.53239
mstr  # Print
```

```
## [1] 7.53239
```

```
# Calculating MSE
sse <- sum((sapply(media_list, length) - 1) * sapply(media_list,
    var))  # 464.8024
sse  # Print
```

```
## [1] 464.8024
```

```
df_mse <- sum(sapply(media_list, length)) - 4
df_mse  #162
```

```
## [1] 162
```

```
mse <- sse/df_mse  # 2.869151
mse  # Print
```

```
## [1] 2.869151
```

```
# Calculating F-value
f_value <- mstr/mse  # 2.625303
f_value  # Print
```

```
## [1] 2.625303
```

ii) Compute the p-value of F, from the null F-distribution

```
# cut-off at 95%
cut_off <- qf(p = 0.95, df1 = df_mstr, df2 = df_mse)  # 2.660406
cut_off  # Print
```

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

```
# P-value
p_value <- pf(f_value, df_mstr, df_mse, lower.tail = FALSE)  # 0.05230686
p_value  # Print
```

```
## [1] 0.05230686
```

Conclusion of Hypothesis:

- The F-value of **2.625** is smaller than the cut_off of **2.6604** at 0.05% significance
- The p-value of **0.0523** is greater than the 95% significance level
- Thus, we **do not** reject the null hypothesis.
- This means the intention to share will not make a difference as the means may be the same.

c) Conduct the same one-way ANOVA using the aov() function in R

- I will first convert and clean data before testing

```
# First converting list to dataframe
media_frame <- as.data.frame(sapply(media_list, "[", seq(max(lengths(media_list)))))
colnames(media_frame) <- c("media1", "media2", "media3", "media4")
str(media_frame)  # Checking data frame
```

```
## 'data.frame':    46 obs. of  4 variables:
##  $ media1: int  3 5 4 5 5 4 4 5 4 1 ...
##  $ media2: int  4 6 4 4 5 4 4 4 4 7 ...
##  $ media3: int  1 4 1 5 6 6 5 7 5 5 ...
##  $ media4: int  3 4 4 2 7 7 5 7 5 6 ...
```

```
# Converting data to row-wise format
require(reshape2)
media_long <- melt(media_frame, na.rm = T, id.vars = NULL, variable.name = "media_type",
    value.name = "value")
str(media_long)  # Checking no.of observations match
```

```
## 'data.frame':    166 obs. of  2 variables:
##  $ media_type: Factor w/ 4 levels "media1","media2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ value     : int  3 5 4 5 5 4 4 5 4 1 ...
```

- Run one-way test:

```
anova1 <- oneway.test(media_long$value ~ factor(media_long$media_type),
    var.equal = T)
anova1  # Print
```

```
##
##  One-way analysis of means
##
## data:  media_long$value and factor(media_long$media_type)
## F = 2.6167, num df = 3, denom df = 162, p-value = 0.05289
```

- Using **aov** function

```
summary(aov(media_long$value ~ factor(media_long$media_type)))  # Print
```

```
##                                Df Sum Sq Mean Sq F value Pr(>F)
## factor(media_long$media_type)   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
```

- We can confirm that the results **are similar** as aov F-value = 2.6167, while the manual coding F-value = 2.625303.
- Thus, the conclusion will not change.

d) Conduct a post-hoc Tukey test

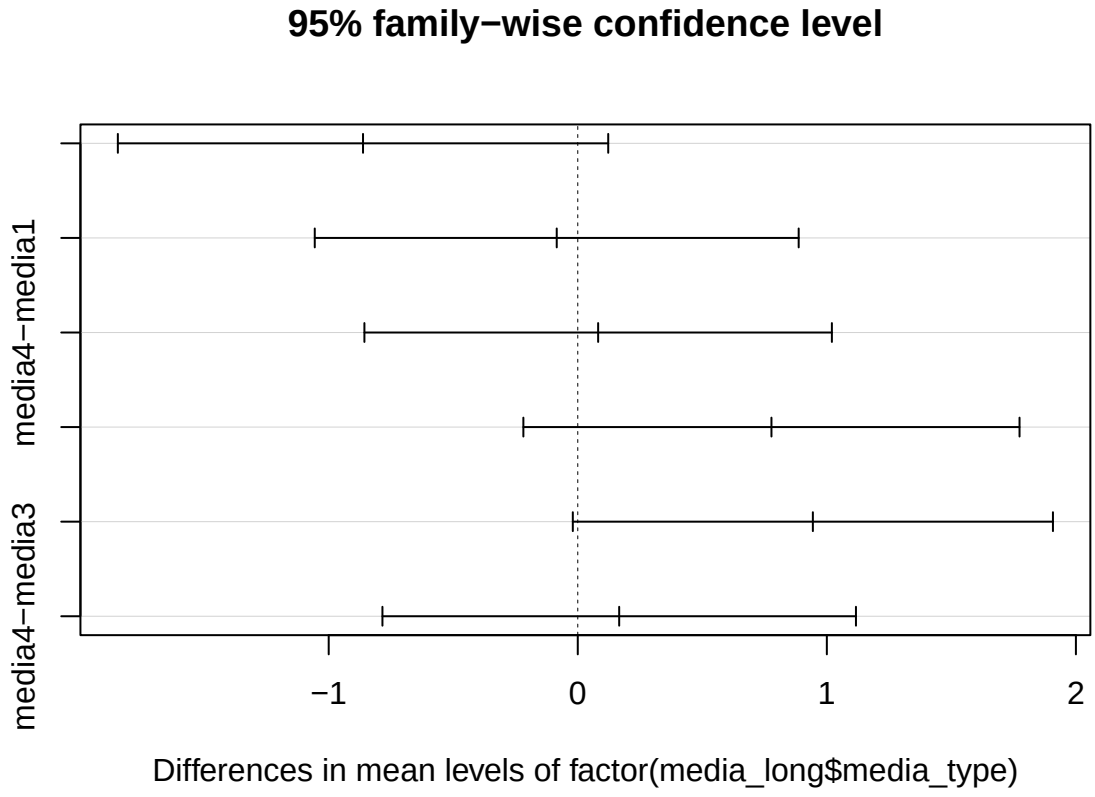
```
# Putting data in appropriate format for the Tukey Test
media_lm <- lm(media_long$value ~ factor(media_long$media_type),
    data = media_long)
media_aov <- aov(media_lm)
summary(media_aov)
```

```
##                                Df Sum Sq Mean Sq F value Pr(>F)
## factor(media_long$media_type)   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
```

```
# Tukey Test
tukey_test <- TukeyHSD(media_aov)
tukey_test  # Print
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = media_lm)
##
## $`factor(media_long$media_type)`
##                       diff         lwr       upr     p adj
## media2-media1 -0.86215539 -1.84660332 0.1222925 0.1085727
## media3-media1 -0.08452381 -1.05596494 0.8869173 0.9959223
## media4-media1  0.08178054 -0.85664966 1.0202107 0.9959032
## media3-media2  0.77763158 -0.21843807 1.7737012 0.1825044
## media4-media2  0.94393593 -0.01996662 1.9078385 0.0573229
## media4-media3  0.16630435 -0.78431033 1.1169190 0.9687417
```

```
plot(tukey_test)
```



What do we find from the Tukey Test?

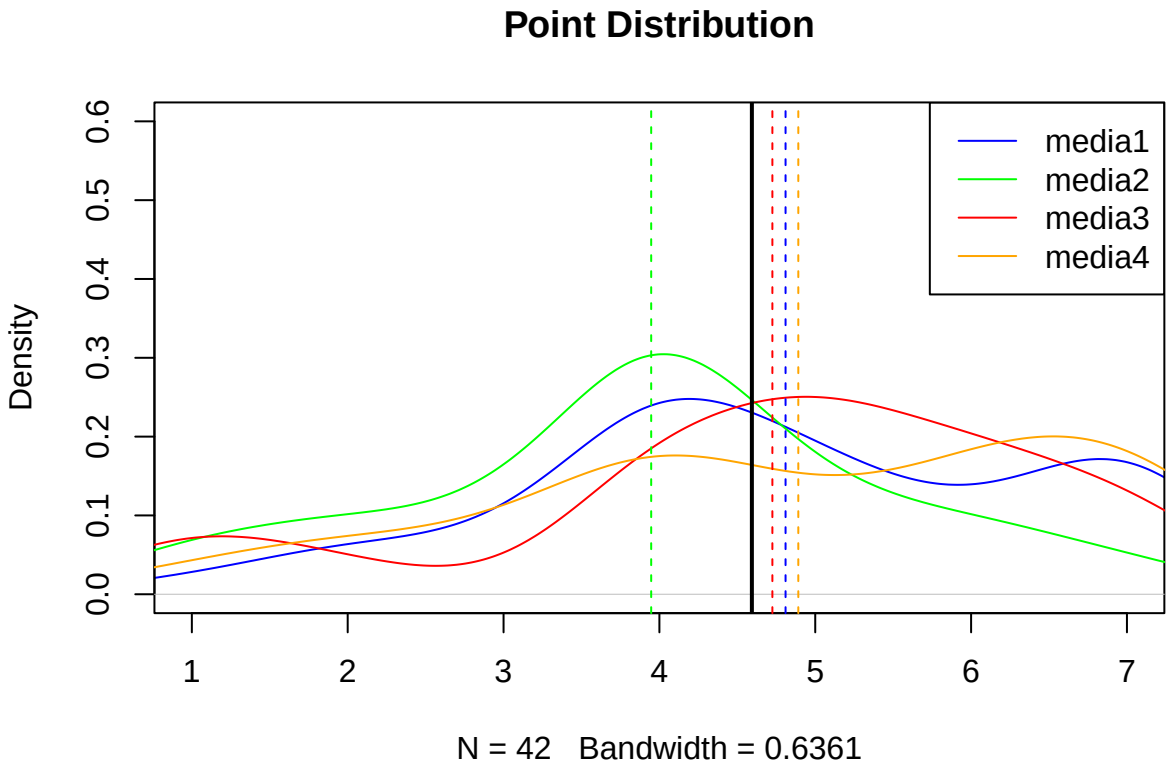
- As evident by the plot, the confidence levels and p-values show no significance between-group difference for all media pairs.
- Note all the pairs contain 0 in the confidence intervals and thus, have no significant difference (although **media pair 4 and 2** come close)

e) Do you feel the classic requirements of one-way ANOVA were met?

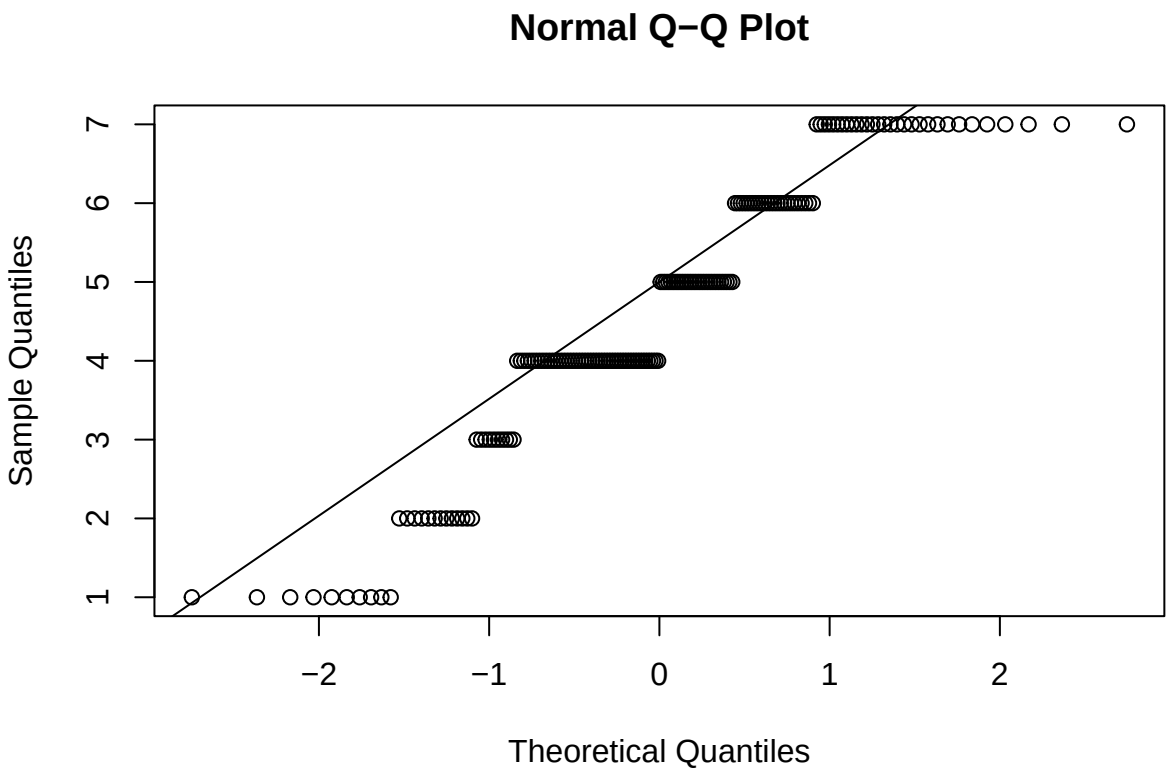
The classic requirements of one-way Anova are not met. Below I test the assumptions:

- **Assumption 1** not met:
  - Response variable should be normally distributed, but is not.

```
media_plots()  # Density Plots
```




```
# Departures from the line suggest violations of normality
# in the Q-Q plot:
attach(media_long)
qqnorm(value)
qqline(value)
```




```
# We can also use the Shapiro test to test for normality
require(dplyr)
shapiro.test(media_long$value)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  media_long$value
## W = 0.92051, p-value = 6.886e-08
```

- We can see that each sample group is not normally distributed from the above plots.
- From the Shapiro Test, the p-value is less than 0.05, which indicates that the sample groups are different from normal distribution.
- **Assumption 2** is met:
  - The variance of the response variables should be the same.
- Run a Bartlett test to verify if assumption is true
  - H_null: The variance among each group is equal
  - H_alt: At least one group has a variance that is not equal to the rest

```r
bartlett.test(value ~ media_type, media_long)
```

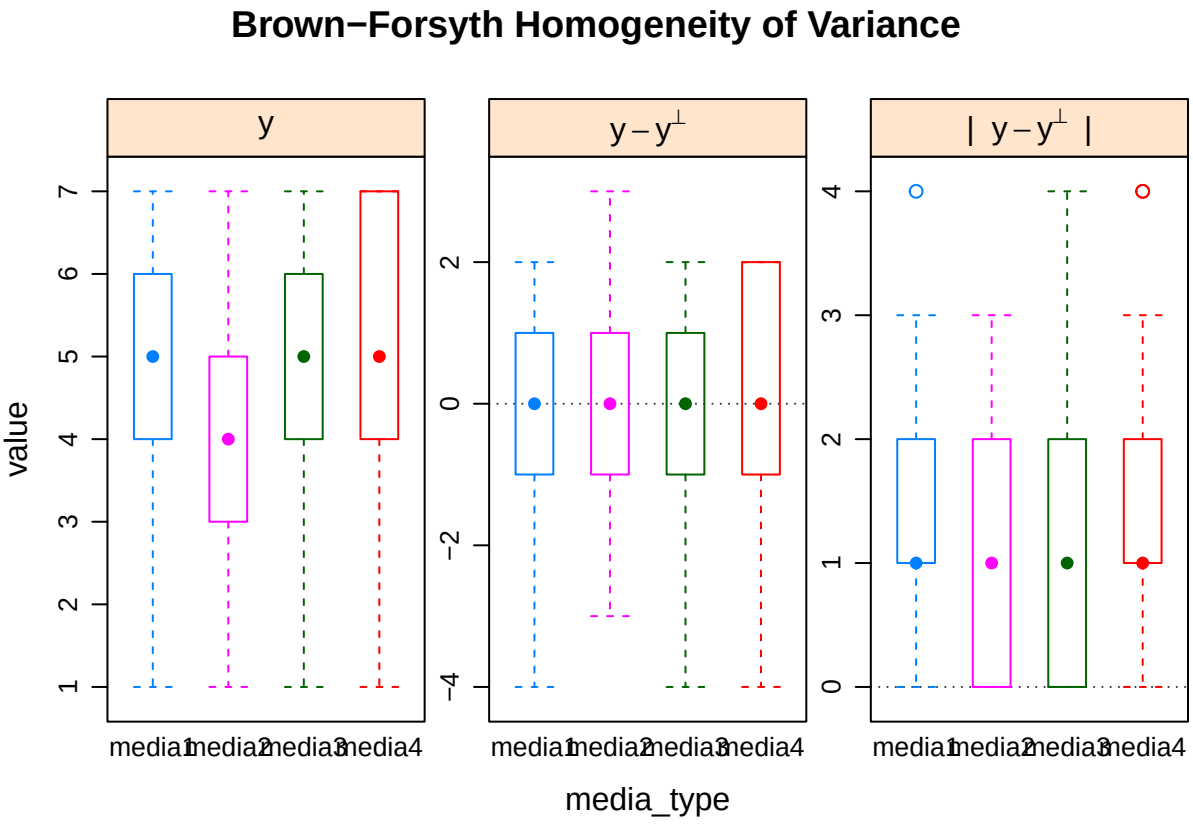
```
##
##  Bartlett test of homogeneity of variances
##
## data:  value by media_type
## Bartlett's K-squared = 1.3958, df = 3, p-value = 0.7065
```

- The **p-value is 0.7065** which is greater than the significance level of 0.05 and we can therefore **not reject** the null hypotheses.
- Thus, we don't have sufficient evidence to say that the four groups have different variances

```r
# Visualize homogeneity of Variance Plot
require(HH)
hovBF(value ~ media_type, media_long, na.rm = TRUE)
```

```
##
##  hov: Brown-Forsyth
##
## data:  value
## F = 1.5403, df:media_type = 3, df:Residuals = 162, p-value = 0.2061
## alternative hypothesis: variances are not identical
```

```r
hovplot <- hovplotBF(value ~ media_type, media_long, na.rm = TRUE,
    main = "Brown-Forsyth Homogeneity of Variance", plotmath = TRUE)
hovplot  # Print
```



- Since the p-value of **0.2061** is greater than the 0.05 significance level in the Brown-Forsyth test, we can therefore **not reject** the null that the variances among each group is different.
- Thus, we don't have sufficient evidence to say that the four groups have different variances

Conclusion:

- The first assumption is not met, while the second assumption is met.
- However, classic requirements should all be met.
- Thus, we can conclude that the classic requirements of one-way ANOVA were **not met**.

3) The non-parametric Kruskal Wallis test

a) State the null and alternative hypotheses (in terms of distribution or difference of mean ranks)

We are looking for similarity in the values in the different groups

- H_null: The ranks of the groups are the same
- H_alt: The ranks of the groups are NOT the same

b) compute (an approximate) Kruskal Wallis H

i) Show the code and results of computing H

```
# Rank all the combined values across groups
media_long$value_rank <- rank(media_long$value)
head(media_long, 5)  # Checking column was added
```

```
##   media_type value value_rank
## 1     media1     3       28.5
## 2     media1     5       97.5
## 3     media1     4       58.5
## 4     media1     5       97.5
## 5     media1     5       97.5
```

```
# Group the ranks into original groups
group_ranks <- split(x = media_long$value_rank, f = media_long$media_type)

# Group and Sum the ranks for each group
group_rank_sums <- media_long %>%
    group_by(media_type) %>%
    summarise(Freq = sum(value_rank))
group_rank_sums  # Print
```

```
## # A tibble: 4 x 2
##   media_type  Freq
##   <fct>      <dbl>
## 1 media1     3694.
## 2 media2     2421
## 3 media3     3556
## 4 media4     4190.
```

```
# Creating variables for calculation and ease of reading
N <- length(media_long$value_rank)  #166

n1 <- length(group_ranks$media1)  # 42
n2 <- length(group_ranks$media2)  # 38
n3 <- length(group_ranks$media3)  # 40
n4 <- length(group_ranks$media4)  # 46

R1 <- (group_rank_sums[1, 2])^2  # 13641942
R2 <- (group_rank_sums[2, 2])^2  # 5861241
R3 <- (group_rank_sums[3, 2])^2  # 12645136
R4 <- (group_rank_sums[4, 2])^2  # 17560290

# Apply the Kruskal Wallis formula to sum the squared ranks
H <- (12/(N * (N + 1))) * sum((R1/n1) + (R2/n2) + (R3/n3) + (R4/n4)) -
    3 * (N + 1)  # 8.45466
H  # Print
```

```
## [1] 8.45466
```

ii) Compute the p-value of H, from the null chi-square distribution (0.05% significance level)

```
# Find p-value of H from <U+03C7>2 distribution
kw_p <- 1 - pchisq(H, df = 4 - 1)  # 0.03749292
kw_p  # Print
```

```
## [1] 0.03749292
```

```
# Chi-Square Critical Value at 0.05 significance level
cut_off <- qchisq(0.05, 3, lower.tail = F)  # 7.814728
cut_off  # Print
```

```
## [1] 7.814728
```

- The approximate H-value is **significant** because it is greater than the critical value at a 0.05 significance level ($8.45466 > 7.814728$).

Conclusion of the hypotheses:

- As the p-value of **0.03749292** is less than the significance level of 0.05, thus, we can conclude that there may be significant differences between the groups.
- Thus, we **reject** the null hypothesis

c) Conduct the same test using the kruskal.wallis() function

```
Kruskal_test <- kruskal.test(value ~ media_type, data = media_long)
Kruskal_test  # Print
```

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

- We can confirm that the results **are similar** as H-value = 8.82851, while the manual coding H-value = 8.45466.
- Since both values are greater than the cut-off, and both p-values are less than a 0.05 significance level, the conclusion will not change.

d) Conduct a post-hoc Dunn test to see if any pairs of media are significantly different

```
require(FSA)
dunn_test <- dunnTest(value ~ media_type, data = media_long,
    method = "bonferroni")
dunn_test  # Print
```

```
##        Comparison           Z     P.unadj      P.adj
## 1 media1 - media2  2.30087819 0.021398517 0.12839110
## 2 media1 - media3 -0.09233644 0.926430736 1.00000000
## 3 media2 - media3 -2.36408588 0.018074622 0.10844773
## 4 media1 - media4 -0.31452459 0.753122646 1.00000000
## 5 media2 - media4 -2.65613380 0.007904225 0.04742535
## 6 media3 - media4 -0.21613379 0.828883460 1.00000000
```

What do we find from the Dunn Test?

- Based off the P.adj, we can see that only **media pairs 2 and 4** are below a 0.05 significance level, which indicates that this pair is significantly different from the other media pairs.