# BACS HW (Week 7)

108020024

due on 04/9 (Sun) Helped by 108020033

**Question 1) Let's explore and describe the data and develop some early intuitive thoughts:**

```
pls_media1 <- read.csv("pls-media1.csv")
pls_media2 <- read.csv("pls-media2.csv")
pls_media3 <- read.csv("pls-media3.csv")
pls_media4 <- read.csv("pls-media4.csv")
```

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

```
mean(pls_media1$INTEND.0)
```

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

```
mean(pls_media2$INTEND.0)
```

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

```
mean(pls_media3$INTEND.0)
```

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

```
mean(pls_media4$INTEND.0)
```

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

The means for each of the four media types:

pls_media1 = 4.809524

pls_media2 = 3.947368

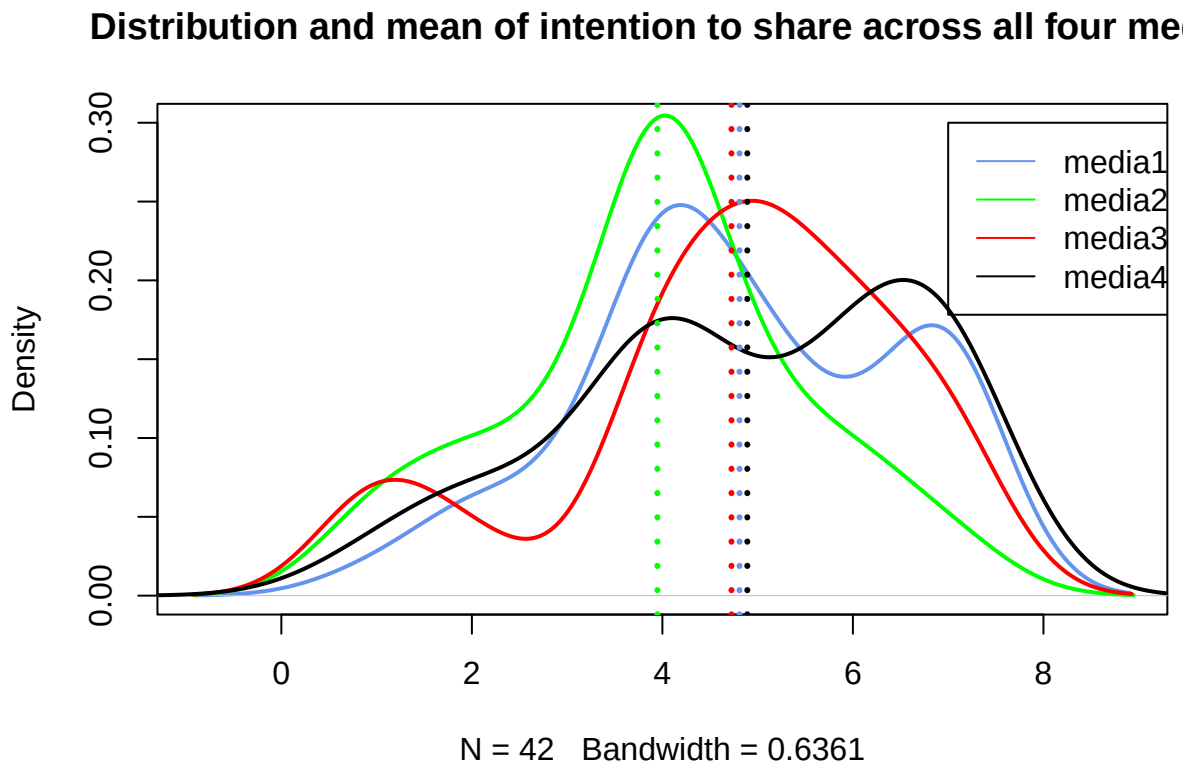
pls_media3 = 4.725

pls_media4 = 4.891304

**b) Visualize the distribution and mean of intention to share, across all four media.(Your choice of data visualization; Try to put them all on the same plot and make it look sensible)**

```
plot(density(pls_media1$INTEND.0),
     col="cornflowerblue",
     lwd=2,
     ylim = c(0,0.3),
     main = "Distribution and mean of intention to share across all four media")
lines(density(pls_media2$INTEND.0), col="green", lwd=2)
lines(density(pls_media3$INTEND.0), col="red", lwd=2)
lines(density(pls_media4$INTEND.0), col="black", lwd=2)
abline(v = mean(pls_media1$INTEND.0), col="cornflowerblue", lwd=3, lty=3)
abline(v = mean(pls_media2$INTEND.0), col="green", lwd=3, lty=3)
abline(v = mean(pls_media3$INTEND.0), col="red", lwd=3, lty=3)
abline(v = mean(pls_media4$INTEND.0), col="black", lwd=3, lty=3)


legend(7, 0.3, lty=1, c("media1", "media2","media3","media4"), col=c("cornflowerblue", "green","red","bl
```




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

I'll say I feel that media type makes a difference on intention to share from the visualization alone.

The visualization for the four media data have many distracting features that make comparisons hard. The trend of the distribution are similar, however it is difficult for me to describe the difference details on intention to share between the four media data with a clear description, such as the peak difference between the four medias.

**Question 2) Let's try traditional one-way ANOVA:**

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

$H_0$ : The means of the INTEND.0 across four medias are the same.

$H_1$ : The means of the INTEND.0 across four 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

```
library(reshape2)

df = list(media1=pls_media1$INTEND.0,
          media2=pls_media2$INTEND.0,
          media3=pls_media3$INTEND.0,
          media4=pls_media4$INTEND.0)



media <- melt(df, id.vars = NULL,
variable.name = "L1",
value.name = "INTEND.0")


#MSTR
sstr <-
  (length(pls_media1$INTEND.0)*(mean(pls_media1$INTEND.0)-mean(media$INTEND.0))^2)+
  (length(pls_media2$INTEND.0)*(mean(pls_media2$INTEND.0)-mean(media$INTEND.0))^2)+
  (length(pls_media3$INTEND.0)*(mean(pls_media3$INTEND.0)-mean(media$INTEND.0))^2)+
  (length(pls_media4$INTEND.0)*(mean(pls_media4$INTEND.0)-mean(media$INTEND.0))^2)

df_mstr <- 4-1
mstr <- sstr/df_mstr

#MSE
sse <-
  (length(pls_media1$INTEND.0)-1)*var(pls_media1$INTEND.0)+
  (length(pls_media2$INTEND.0)-1)*var(pls_media2$INTEND.0)+
  (length(pls_media3$INTEND.0)-1)*var(pls_media3$INTEND.0)+
  (length(pls_media4$INTEND.0)-1)*var(pls_media4$INTEND.0)

df_mse <-
  length(pls_media1$INTEND.0)+
  length(pls_media2$INTEND.0)+
```

```
  length(pls_media3$INTEND.0)+
  length(pls_media4$INTEND.0)- 4

mse <- sse/df_mse


#F
f_value <- mstr/mse
```

The value for MSTR, MSE, and F are:

MSTR = 7.5076174

MSE = 2.8691509

F = 2.6166687

```
    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.
```

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

The p_value is 0.0528902 > 0.05, cannot reject the null hypothesis that the means of the INTEND.0 across four medias are the same at 5% significance.

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

```
anova_model <- aov( media$INTEND.0 ~ factor(media$L1))
summary(anova_model)
```

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

The p_value is 0.0529 > 0.05, which confirm that I got similar results from (b), cannot reject the null hypothesis that the means of the INTEND.0 across four medias are the same at 5% significance.

**d) Regardless of your conclusions, conduct a post-hoc Tukey test (feel free to use the TukeyHSD() function included in base R) to see if any pairs of media have significantly different means – what do you find?**

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

```
##   Tukey multiple comparisons of means
##     5% family-wise confidence level
##
```

```
## Fit: aov(formula = media$INTEND.0 ~ factor(media$L1))
##
## $`factor(media$L1)`
##                       diff         lwr        upr     p adj
## media2-media1 -0.86215539 -1.06562977 -0.6586810 0.1085727
## media3-media1 -0.08452381 -0.28530983  0.1162622 0.9959223
## media4-media1  0.08178054 -0.11218249  0.2757436 0.9959032
## media3-media2  0.77763158  0.57175512  0.9835080 0.1825044
## media4-media2  0.94393593  0.74470805  1.1431638 0.0573229
## media4-media3  0.16630435 -0.03017708  0.3627858 0.9687417
```

From the result of the post-hoc Tukey test, didn't find any pairs of media have significantly different means.

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

ANOVA requires some assumptions to be met:

1. Each treatment/population's response variable is normally distributed

Check the normally assumptions:

```
shapiro.test(pls_media1$INTEND.0)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pls_media1$INTEND.0
## W = 0.91279, p-value = 0.003557
```

```
shapiro.test(pls_media2$INTEND.0)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pls_media2$INTEND.0
## W = 0.92974, p-value = 0.01969
```

```
shapiro.test(pls_media3$INTEND.0)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pls_media3$INTEND.0
## W = 0.88247, p-value = 0.0006139
```

```
shapiro.test(pls_media4$INTEND.0)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pls_media4$INTEND.0
## W = 0.89611, p-value = 0.0006242
```

Under 5% significance level, INTEND.0 across four medias did not passed the normally test, so we cannot assume the normality.

2. The variance (s2) of the response variables is the same for all treatments/populations

Perform F-test to test whether two population variances are equal.

```
var.test(pls_media1$INTEND.0, pls_media2$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  pls_media1$INTEND.0 and pls_media2$INTEND.0
## F = 1.1607, num df = 41, denom df = 37, p-value = 0.6488
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.610132 2.184962
## sample estimates:
## ratio of variances
##             1.1607
```

```
var.test(pls_media1$INTEND.0, pls_media3$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  pls_media1$INTEND.0 and pls_media3$INTEND.0
## F = 0.87591, num df = 41, denom df = 39, p-value = 0.6752
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4658417 1.6387455
## sample estimates:
## ratio of variances
##          0.8759084
```

```
var.test(pls_media1$INTEND.0, pls_media4$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  pls_media1$INTEND.0 and pls_media4$INTEND.0
## F = 0.81677, num df = 41, denom df = 45, p-value = 0.5139
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4472891 1.5043838
## sample estimates:
## ratio of variances
##          0.8167668
```

```r
var.test(pls_media2$INTEND.0, pls_media3$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  pls_media2$INTEND.0 and pls_media3$INTEND.0
## F = 0.75464, num df = 37, denom df = 39, p-value = 0.3918
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.396723 1.443427
## sample estimates:
## ratio of variances
##           0.754638
```

```r
var.test(pls_media2$INTEND.0, pls_media4$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  pls_media2$INTEND.0 and pls_media4$INTEND.0
## F = 0.70368, num df = 37, denom df = 45, p-value = 0.2741
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3806992 1.3258587
## sample estimates:
## ratio of variances
##          0.7036846
```

```r
var.test(pls_media3$INTEND.0, pls_media4$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  pls_media3$INTEND.0 and pls_media4$INTEND.0
## F = 0.93248, num df = 39, denom df = 45, p-value = 0.8282
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5076909 1.7361918
## sample estimates:
## ratio of variances
##          0.9324797
```

Under 5% significance level, INTEND.0 across four medias passed the F-test, we can consider variance of the response variables doesn't have significance difference for all four medias.

3. The observations are independent: the response variables are not related between groups

Perform Chi-squared test to test whether the observations are independent.

```
max_length <- max(c(length(pls_media1$INTEND.0),
                    length(pls_media2$INTEND.0),
                    length(pls_media3$INTEND.0),
                    length(pls_media4$INTEND.0)))

mdf = data.frame(media1 = c(pls_media1$INTEND.0,
                            rep(NA, max_length - length(pls_media1$INTEND.0))),
                 media2 = c(pls_media2$INTEND.0,
                            rep(NA, max_length - length(pls_media2$INTEND.0))),
                 media3 = c(pls_media3$INTEND.0,
                            rep(NA, max_length - length(pls_media3$INTEND.0))),
                 media4 = c(pls_media4$INTEND.0,
                            rep(NA, max_length - length(pls_media4$INTEND.0)))
                 )
chisq.test(table(mdf$media1, mdf$media2))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(mdf$media1, mdf$media2)
## X-squared = 43.304, df = 36, p-value = 0.1878
```

```
chisq.test(table(mdf$media1, mdf$media3))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(mdf$media1, mdf$media3)
## X-squared = 26.921, df = 30, p-value = 0.6274
```

```
chisq.test(table(mdf$media1, mdf$media4))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(mdf$media1, mdf$media4)
## X-squared = 44.32, df = 36, p-value = 0.1608
```

```
chisq.test(table(mdf$media2, mdf$media3))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(mdf$media2, mdf$media3)
## X-squared = 31.244, df = 30, p-value = 0.4035
```

```
chisq.test(table(mdf$media2, mdf$media4))
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  table(mdf$media2, mdf$media4)
## X-squared = 33.668, df = 36, p-value = 0.58
```

```
chisq.test(table(mdf$media3, mdf$media4))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(mdf$media3, mdf$media4)
## X-squared = 30.859, df = 30, p-value = 0.4224
```

Under 5% significance level, the observations can be considered as independent.

So the data does not met all th classic requirements of one-way ANOVA.

**Question 3) Let's use the non-parametric Kruskal Wallis test:**

**a) State the null and alternative hypotheses**

$H_0$ : All groups would give you a similar value if randomly drawn from them

$H_1$ : At least one group would give you a larger value than another if randomly drawn

**b) Let's compute (an approximate) Kruskal Wallis H ourselves (use the formula we saw in class or another formula might have found at a reputable website/book):**

i) Show the code and results of computing H

```
media_ranks <- rank(media$INTEND.0)
group_ranks <- split(media_ranks,media$L1)

N <- length(media$INTEND.0)
#
H <- unname((12/(N*(N+1)))*
  ((((sapply(group_ranks, sum)[1])^2)/length(pls_media1$INTEND.0))+
  (((sapply(group_ranks, sum)[2])^2)/length(pls_media2$INTEND.0))+
  (((sapply(group_ranks, sum)[3])^2)/length(pls_media3$INTEND.0))+
  (((sapply(group_ranks, sum)[4])^2)/length(pls_media4$INTEND.0)))-(3*(N+1)))

H
```

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

The value H is 8.4546598 .

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.

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

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

The p_value of H is $0.0374929 < 0.05$, reject the null hypothesis that all groups would give you a similar value if randomly drawn from them at 5% significance.

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

```r
kruskal.test(INTEND.0 ~ L1, data = media)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  INTEND.0 by L1
## Kruskal-Wallis chi-squared = 8.8283, df = 3, p-value = 0.03166
```

The p_value of H is $0.03166 < 0.05$, got a similar result from (b), reject the null hypothesis that all groups would give you av

**d) Regardless of your conclusions, conduct a post-hoc Dunn test (feel free to use the dunnTest() function from the FSA package) to see if the values of any pairs of media are significantly different – what are your conclusions?**

```r
#install.packages("FSA")
library(FSA)
```

```
## ## FSA v0.9.4. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```r
dunnTest(INTEND.0 ~ L1, data = media, method = "bonferroni")
```

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

##   p-values adjusted with the Bonferroni method.

##        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
```

In the comparison of media2 - media4, the adjusted p-value is $0.04742535 < 0.05$, media2 and media4 are the only two media that are statistically significantly different from each other under 5% significance level.