BACS-hw07-107070004

Question 1)

Let's develop some intuition about the data and results:

```
mydir = "pls-media"
myfiles = list.files(path=mydir, pattern="*.csv", full.names=TRUE)
df_total = vector()
for (i in c(1:length(myfiles))) {
  df_dict <- read.csv(myfiles[i])
  df = data.frame(df_dict$INTEND.0)
  colnames(df) <- paste("media",i)
  df_total <- c(df_total,df)
}
```

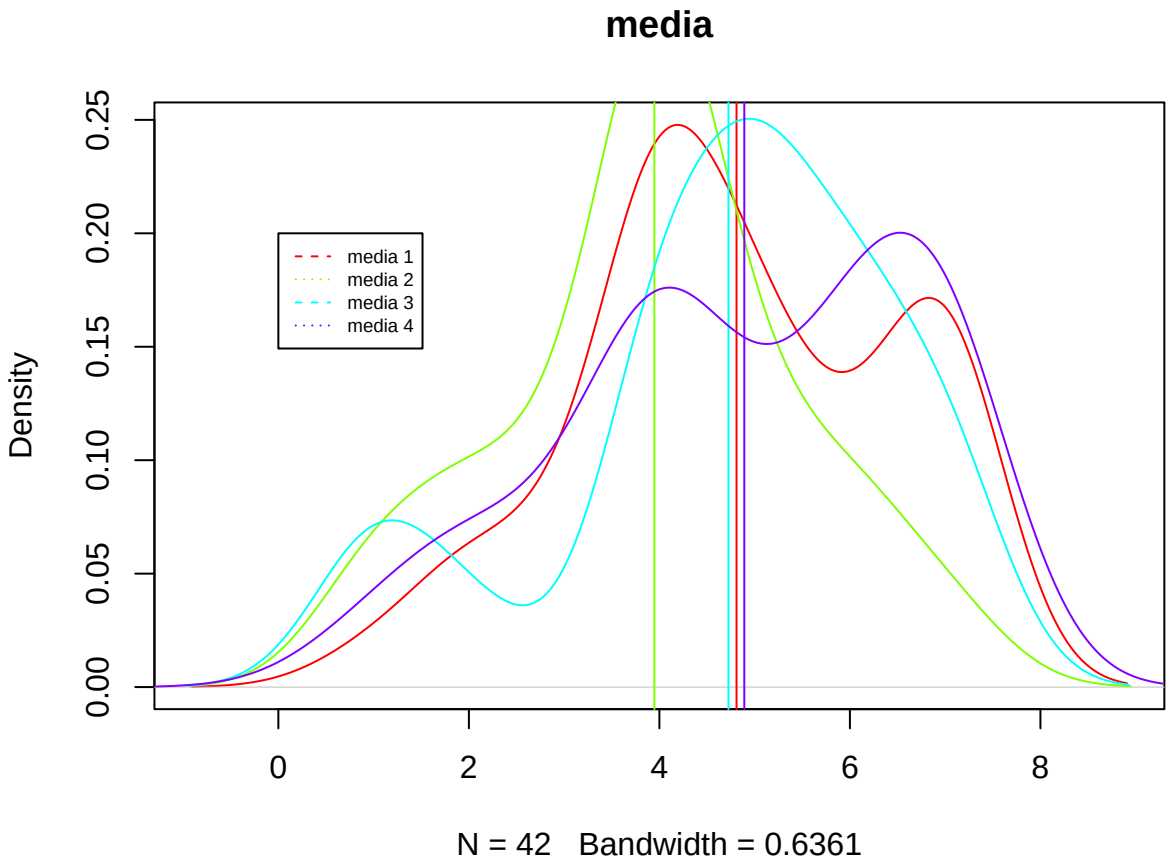
(a) What are the means of viewers' intentions to share (INTEND.0) on each of the four media types?

```
all_mean <- sapply(df_total, mean)
all_mean
```

```
##  media 1  media 2  media 3  media 4
## 4.809524 3.947368 4.725000 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)

```
i <- 1
cl <- rainbow(4)
for (df in df_total) {
  if(i == 1) plot(density(df),col = cl[i], main="media")
  else lines(density(df), col = cl[i])
  abline(v=mean(df), col = cl[i])
  i <- i+1
}
legend(0, 0.2, legend = c("media 1", "media 2", "media 3", "media 4"),
       col = cl, lty = 2:3, cex = 0.6)
```



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

Yes.

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

$H_{null} : \mu_1 = \mu_2 = \mu_3 = \mu_4$ $H_{alt} : \mu_1 \neq \mu_2; \mu_1 \neq \mu_3; \mu_1 \neq \mu_4; \mu_2 \neq \mu_3; \mu_2 \neq \mu_4; \mu_3 \neq \mu_4$

(b) Let's compute the F-statistic ourselves:

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

- MSTR

```
k <- length(df_total)
SSTR <- function(df) {
  n <- length(df)
  sstr <- n*sum((mean(df)-mean(all_mean))^2)
}
sstr_total <- sum(sapply(df_total, SSTR))
df_mstr <- k-1
df_mstr
```

```
## [1] 3
```

```
mstr <- sstr_total/df_mstr
mstr
```

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

- MSE

```
all_num <- 0
MSE <- function(df) {
  n <- length(df)
  all_num <<- all_num + length(df)
  sse <- sum((n-1)*(sd(df)^2))
}
sse_total <- sum(sapply(df_total, MSE))
df_mse <- all_num - k
mse <- sse_total/df_mse
mse
```

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

- F

```
F <- mstr/mse
F
```

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

(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 <- pf(F, df_mstr, df_mse, lower.tail=FALSE)
p
```

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

The p-value of the test is 0.05230686, which is more than the alpha level of 0.05. meaning that there was not sufficiently strong evidence to reject the null hypothesis and conclude that the groups are different.

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

```
ads <-melt(df_total,id.vars= NULL,variable.name= "L1",value.name= "value")
oneway.test(ads$value~factor(ads$L1),var.equal=TRUE)
```

```
##
##  One-way analysis of means
##
## data:  ads$value and factor(ads$L1)
## F = 2.6167, num df = 3, denom df = 162, p-value = 0.05289
```

```
summary(aov(ads$value~factor(ads$L1)))
```

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

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

```
TukeyHSD(aov(ads$value~factor(ads$L1)),conf.level= 0.01)
```

```
##   Tukey multiple comparisons of means
##     1% family-wise confidence level
##
## Fit: aov(formula = ads$value ~ factor(ads$L1))
##
## $`factor(ads$L1)`
##                          diff         lwr         upr     p adj
## media 2-media 1 -0.86215539 -0.97829137 -0.74601940 0.1085727
## media 3-media 1 -0.08452381 -0.19912537  0.03007775 0.9959223
## media 4-media 1  0.08178054 -0.02892670  0.19248778 0.9959032
## media 3-media 2  0.77763158  0.66012457  0.89513859 0.1825044
## media 4-media 2  0.94393593  0.83022369  1.05764816 0.0573229
## media 4-media 3  0.16630435  0.05415969  0.27844900 0.9687417
```

- All of P adj. is > 0.05 -There is no significant difference between all groups.

(e) Do you feel the classic requirements of one-way ANOVA were met? (Feel free to use any combination of methods we saw in class or any analysis we haven't covered)

1. Normality – Each sample was drawn from a normally distributed population.

2. Equal Variances – The variances of the populations that the samples come from are equal.
3. Independence – The observations in each group are independent of each other and the observations within groups were obtained by a random sample.

If these assumptions aren't met, then the results of our one-way ANOVA could be unreliable.

Question 3)

One of the assumptions of some classical statistical tests is that our population data should be roughly normal. Let's explore one way of visualizing whether a sample of data is normally distributed.

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

$H_{null} : \mu_1 = \mu_2 = \mu_3 = \mu_4$ $H_{alt} : \mu_1 \neq \mu_2; \mu_1 \neq \mu_3; \mu_1 \neq \mu_4; \mu_2 \neq \mu_3; \mu_2 \neq \mu_4; \mu_3 \neq \mu_4$

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

(i) Show the code and results of computing H

```
ranks <- rank(ads$value)
ranks
```

```
##   [1]  28.5  97.5  58.5  97.5  97.5  58.5  58.5  97.5  58.5   5.5 151.5  28.5
##  [13]  58.5 151.5 151.5  97.5  58.5 124.0  58.5 124.0  58.5  17.0 151.5 151.5
##  [25] 124.0  58.5  17.0  28.5  97.5 151.5  58.5  97.5  58.5  58.5 151.5 151.5
##  [37]  58.5  17.0  97.5 124.0 151.5 151.5  58.5 124.0  58.5  58.5  97.5  58.5
##  [49]  58.5  58.5  58.5 151.5  58.5  58.5  17.0  58.5  58.5  97.5  58.5   5.5
##  [61] 151.5  28.5  58.5  28.5 124.0  58.5  17.0  97.5  97.5  28.5   5.5  17.0
##  [73]  97.5  17.0 124.0  28.5   5.5  58.5  58.5 124.0   5.5  58.5   5.5  97.5
##  [85] 124.0 124.0  97.5 151.5  97.5  97.5 151.5  97.5  58.5  17.0  97.5  58.5
##  [97] 124.0  97.5  58.5 151.5  97.5  58.5  97.5 124.0 124.0   5.5  58.5  17.0
## [109] 151.5  58.5  58.5   5.5  58.5 151.5 124.0 124.0  97.5 124.0  97.5 151.5
## [121]  28.5  58.5  58.5  17.0 151.5 151.5  97.5 151.5  97.5 124.0  58.5 151.5
## [133] 124.0  58.5  58.5   5.5  17.0   5.5 124.0 151.5  17.0 151.5  58.5 124.0
## [145]  97.5  97.5 124.0  58.5  28.5  17.0 124.0  58.5 124.0 124.0  58.5 151.5
## [157] 124.0 151.5  28.5  58.5 151.5  58.5 151.5 151.5 151.5  58.5
```

```
group_ranks <- split(ranks, ads$L1)
group_ranks
```

```
## $`media 1`
##  [1]  28.5  97.5  58.5  97.5  97.5  58.5  58.5  97.5  58.5   5.5 151.5  28.5
## [13]  58.5 151.5 151.5  97.5  58.5 124.0  58.5 124.0  58.5  17.0 151.5 151.5
## [25] 124.0  58.5  17.0  28.5  97.5 151.5  58.5  97.5  58.5  58.5 151.5 151.5
## [37]  58.5  17.0  97.5 124.0 151.5 151.5
##
## $`media 2`
```

```
##  [1]  58.5 124.0  58.5  58.5  97.5  58.5  58.5  58.5  58.5 151.5  58.5  58.5
## [13]  17.0  58.5  58.5  97.5  58.5   5.5 151.5  28.5  58.5  28.5 124.0  58.5
## [25]  17.0  97.5  97.5  28.5   5.5  17.0  97.5  17.0 124.0  28.5   5.5  58.5
## [37]  58.5 124.0
##
## $`media 3`
##  [1]   5.5  58.5   5.5  97.5 124.0 124.0  97.5 151.5  97.5  97.5 151.5  97.5
## [13]  58.5  17.0  97.5  58.5 124.0  97.5  58.5 151.5  97.5  58.5  97.5 124.0
## [25] 124.0   5.5  58.5  17.0 151.5  58.5  58.5   5.5  58.5 151.5 124.0 124.0
## [37]  97.5 124.0  97.5 151.5
##
## $`media 4`
##  [1]  28.5  58.5  58.5  17.0 151.5 151.5  97.5 151.5  97.5 124.0  58.5 151.5
## [13] 124.0  58.5  58.5   5.5  17.0   5.5 124.0 151.5  17.0 151.5  58.5 124.0
## [25]  97.5  97.5 124.0  58.5  28.5  17.0 124.0  58.5 124.0 124.0  58.5 151.5
## [37] 124.0 151.5  28.5  58.5 151.5  58.5 151.5 151.5 151.5  58.5
```

```
group_ranksums <- sapply(group_ranks, sum)
group_ranksums
```

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

```
group_ranknum <- sapply(group_ranks, length)
group_ranknum
```

```
## media 1 media 2 media 3 media 4
##      42      38      40      46
```

```
N <- sum(group_ranknum)
H <- (12/N*(N+1))*sum((group_ranksums^2)/group_ranknum)-3*(N+1)
H
```

```
## [1] 14207680
```

(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.

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

```
## [1] 0
```

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

```r
kruskal.test(value ~ L1, data = ads)
```

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

(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 any pairs of media are significantly different – what do you find?

```r
dunnTest(value ~ L1, data = ads, 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 media 1 - media 2  2.30087819 0.021398517 0.12839110
## 2 media 1 - media 3 -0.09233644 0.926430736 1.00000000
## 3 media 2 - media 3 -2.36408588 0.018074622 0.10844773
## 4 media 1 - media 4 -0.31452459 0.753122646 1.00000000
## 5 media 2 - media 4 -2.65613380 0.007904225 0.04742535
## 6 media 3 - media 4 -0.21613379 0.828883460 1.00000000
```

- media 2 - media 4 is the only one which P adj. is < 0.05, and its Z value of -2.65613380 is less than 0, indicating that the latter group "media 4" is significantly larger than the former group "media 2".
- There was no significant difference between other groups.