BACS - HW7

109006241

Load Data

```
media1 = read.csv("D:/Users/User/Documents/R/BACS/Homeworks/HW7/pls-media1.csv")
media2 = read.csv("D:/Users/User/Documents/R/BACS/Homeworks/HW7/pls-media2.csv")
media3 = read.csv("D:/Users/User/Documents/R/BACS/Homeworks/HW7/pls-media3.csv")
media4 = read.csv("D:/Users/User/Documents/R/BACS/Homeworks/HW7/pls-media4.csv")
```

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

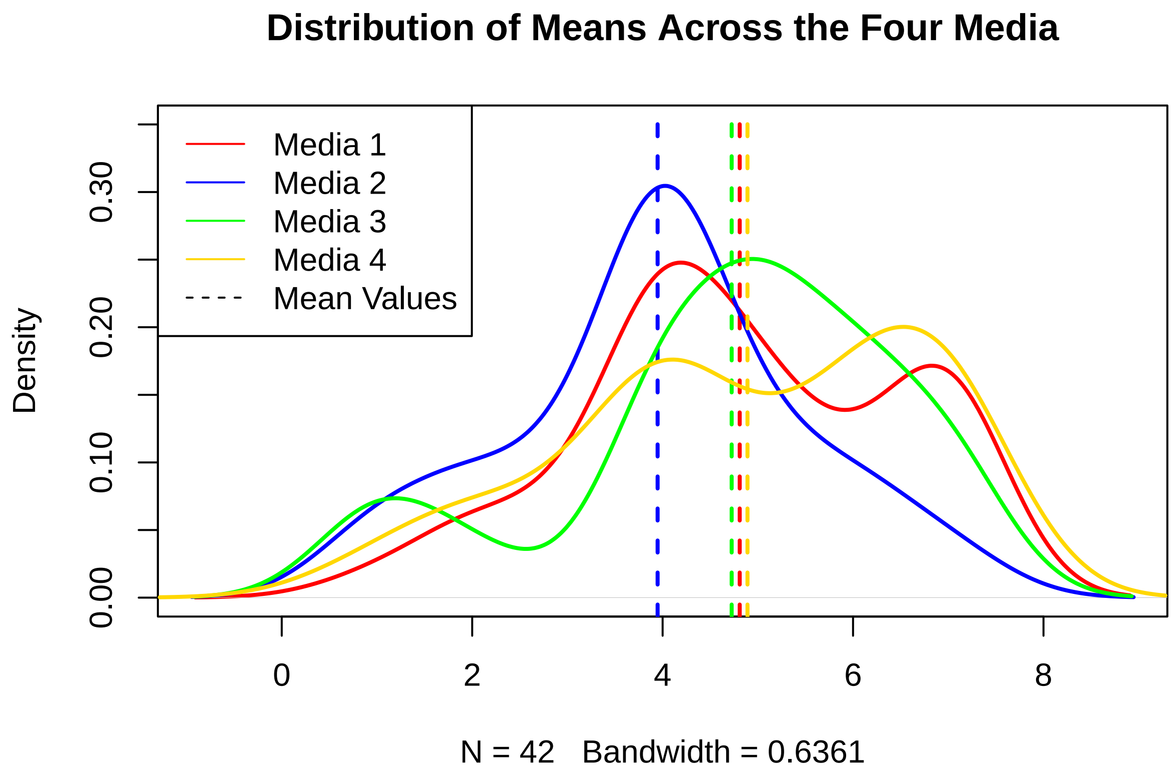
a. What are the means of viewers' intentions to share (INTEND.0) on each of the four media types?

```
mean1 = mean(media1$INTEND.0)
mean2 = mean(media2$INTEND.0)
mean3 = mean(media3$INTEND.0)
mean4 = mean(media4$INTEND.0)
cat(" Mean of Media 1 =", mean1, "\n",
     "Mean of Media 2 =", mean2, "\n",
     "Mean of Media 3 =", mean3, "\n",
     "Mean of Media 4 =", mean4)
```

```
##  Mean of Media 1 = 4.809524
##  Mean of Media 2 = 3.947368
##  Mean of Media 3 = 4.725
##  Mean of Media 4 = 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(media1$INTEND.0), col="red", lty="solid", lwd=2, main="Distribution of Means Across the Fo
abline(v=mean1, col="red", lty="dashed", lwd=2)
lines(density(media2$INTEND.0), col="blue", lty="solid", lwd=2)
abline(v=mean2, col="blue", lty="dashed", lwd=2)
lines(density(media3$INTEND.0), col="green", lty="solid", lwd=2)
abline(v=mean3, col="green", lty="dashed", lwd=2)
lines(density(media4$INTEND.0), col="gold", lty="solid", lwd=2)
abline(v=mean4, col="gold", lty="dashed", lwd=2)
legend("topleft", legend=c("Media 1","Media 2","Media 3","Media 4","Mean Values"), col=c("red","blue","
```



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

   Yes, because the shape of the distributions of the four medias have a significant difference from each other.

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

   H0: The mean values of INTEND.0 across the four groups are the same

   H1: The mean values of INTEND.0 across the four groups are not the same

b. Let's compute the F-statistic ourselves:

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

```
media1234 = list("media1"=media1$INTEND.0,
                 "media2"=media2$INTEND.0,
                 "media3"=media3$INTEND.0,
                 "media4"=media4$INTEND.0)
```

```r
all_mean = sapply(media1234, mean)

# MSTR
k = length(media1234)
SSTR = function(df) {
  n = length(df)
  sstr = n * sum((mean(df) - mean(all_mean))^2)
}
sstr_tot = sum(sapply(media1234, SSTR))
df_mstr = k - 1
MSTR = sstr_tot / df_mstr

# MSE
all_num = 0
SSE = function(df) {
  n = length(df)
  all_num <<- all_num + length(df)
  sse = sum((n - 1) * (sd(df)^2))
}
sse_tot = sum(sapply(media1234, SSE))
df_mse = all_num - k
MSE = sse_tot / df_mse

# F-statistic
F = MSTR / MSE

cat(" MSTR =", MSTR, "\n",
     "MSE =", MSE, "\n",
     "F-statistic =", F)
```

```
##  MSTR = 7.53239
##  MSE = 2.869151
##  F-statistic = 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.

```r
f_val = pf(F, df1=df_mstr, df2=df_mse, lower.tail=FALSE)
cat("F-value =", f_val)
```

```
## F-value = 0.05230686
```

> The p-value is larger than alpha = 0.05, meaning that we don't have enough evidence to say that the mean values of INTEND.0 across the four groups are the same.

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

```r
media1234_long = melt(media1234, id.vars=NULL, variable.name="media", value.name="value")

anova_model = aov(media1234_long$value ~ factor(media1234_long$L1))
summary(anova_model)
```

```
##                            Df Sum Sq Mean Sq F value Pr(>F)
## factor(media1234_long$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
```

Yes, we got the same results indeed, as can be seen from the summary above.

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=0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = media1234_long$value ~ factor(media1234_long$L1))
##
## $`factor(media1234_long$L1)`
##                      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
```

Because all the p-values are above the alpha level, which is 0.05, we can conclude that there is a strong evidence that the mean values of INTEND.0 across the four groups are the same.

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)

```
shapiro.test(media1$INTEND.0)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  media1$INTEND.0
## W = 0.91279, p-value = 0.003557
```

```
shapiro.test(media2$INTEND.0)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  media2$INTEND.0
## W = 0.92974, p-value = 0.01969
```

```
shapiro.test(media3$INTEND.0)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  media3$INTEND.0
## W = 0.88247, p-value = 0.0006139
```

```
shapiro.test(media4$INTEND.0)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  media4$INTEND.0
## W = 0.89611, p-value = 0.0006242
```

Using shapiro.test() to check for normality, we can see from the p-values that the distributions of the four groups are not normally distributed, so we can say that **the classic requirements of a one-way ANOVA test are not met**.

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

a. State the null and alternative hypotheses

H0: The four groups of media will give us a similar value if we randomly draw from them (mean values are the same)

H1: At least one group will give us a larger value than another group if we randomly draw from them (mean values are different)

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

```
value_ranks = rank(media1234_long$value)
group_ranks = split(value_ranks, media1234_long$L1)

group_ranks_R = sapply(group_ranks, sum)
group_ranks_n = sapply(group_ranks, length)
N = sum(group_ranks_n)

H = (12 / (N * (N+1))) * sum((group_ranks_R^2) / group_ranks_n) - 3 * (N + 1)

cat("H-statistic =", H)
```

```
## H-statistic = 8.45466
```

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.

```
k = length(media1234)
kw_p = 1 - pchisq(H, df=k-1)
cat("P-value =", kw_p)
```

```
## P-value = 0.03749292
```

The p-value is smaller than alpha = 0.05, meaning that we can reject the null hypothesis. Thus, we can also say that there is a strong evidence that the mean values of INTEND.0 across the four groups are not the same.

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

```
kruskal.test(value ~ L1, media1234_long)
```

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

Yes, we also got similar results.

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?

```
dunnTest(value ~ L1, media1234_long, 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 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
```

The p-value for media2 vs media4 is smaller than alpha = 0.05, so we can conclude that there is a strong evidence that the mean values for media2 and media4 are different.