HW07

107070008

Question 1) Let's develop some intuition about the data and results:

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

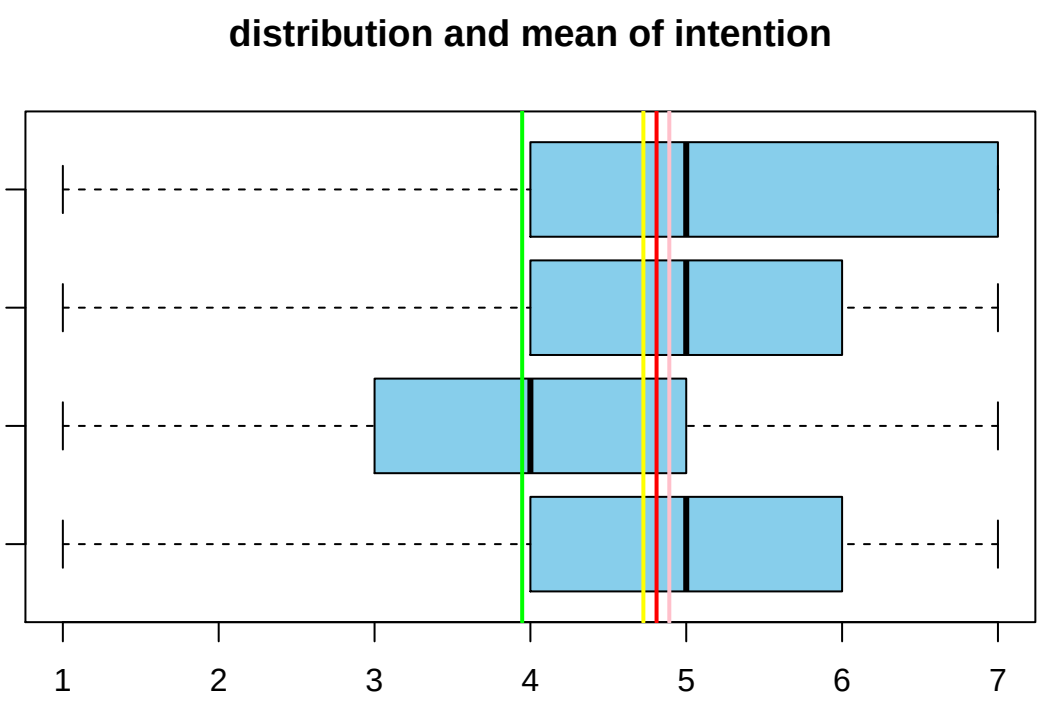
```
data1 <- read.csv("./pls-media/pls-media1.csv")
data2 <- read.csv("./pls-media/pls-media2.csv")
data3 <- read.csv("./pls-media/pls-media3.csv")
data4 <- read.csv("./pls-media/pls-media4.csv")

max.len <- max(length(data1$INTEND.0), length(data2$INTEND.0),
               length(data3$INTEND.0), length(data4$INTEND.0))
datas <- data.frame(
  d1 = c(data1$INTEND.0, rep(NA, max.len - length(data1$INTEND.0))),
  d2 = c(data2$INTEND.0, rep(NA, max.len - length(data2$INTEND.0))),
  d3 = c(data3$INTEND.0, rep(NA, max.len - length(data3$INTEND.0))),
  d4 = c(data4$INTEND.0, rep(NA, max.len - length(data4$INTEND.0)))
)
means <- sapply(datas, mean, na.rm=TRUE)
means
```

```
##       d1       d2       d3       d4
## 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)

```
boxplot(data1$INTEND.0, data2$INTEND.0, data3$INTEND.0, data4$INTEND.0,
        horizontal = TRUE, col = "skyblue", main = "distribution and mean of intention")
mean1 <- mean(data1$INTEND.0)
mean2 <- mean(data2$INTEND.0)
mean3 <- mean(data3$INTEND.0)
mean4 <- mean(data4$INTEND.0)
abline(v = mean1, col = "red", lwd = 2)
abline(v = mean2, col = "green", lwd = 2)
abline(v = mean3, col = "yellow", lwd = 2)
abline(v = mean4, col = "pink", lwd = 2)
```



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

```
Yes, people prefer to share text with each other, and they are a little bit less inclined
to share pictures and audio.
```

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

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.

```
means <- c(mean1, mean2, mean3, mean4)
sstr <- (length(data1$INTEND.0)*((mean1 - mean(means))^2)+length(data2$INTEND.0)
         *((mean2 - mean(means))^2)+length(data3$INTEND.0)*((mean3 - mean(means))^2)
         + length(data4$INTEND.0)*((mean4 - mean(means))^2))
df_mstr <- 4 - 1
mstr <- sstr/df_mstr
mstr
```

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

```
var1 <- sd(data1$INTEND.0)^2
var2 <- sd(data2$INTEND.0)^2
var3 <- sd(data3$INTEND.0)^2
var4 <- sd(data4$INTEND.0)^2
vars <- c(unname(sapply(datas, var, na.rm=TRUE)))
sse<- (length(data1$INTEND.0)-1)*vars[1]+(length(data2$INTEND.0)-1)*vars[2]+
  (length(data3$INTEND.0)-1)*vars[3]+(length(data4$INTEND.0)-1)*vars[4]
df_mse <- length(data1$INTEND.0) + length(data2$INTEND.0) + length(data3$INTEND.0) +
  length(data4$INTEND.0)- 4
mse <- sse/df_mse
mse
```

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

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

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

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

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

```
f-value > p-value in lower tail, so we should reject our null hypothesis.
```

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

```
library(reshape2)
test <- melt(datas, na.rm = TRUE, id.vars = NULL, variable.name = "media",
             value.name = "intend")
oneway.test(test$intend~factor(test$media), var.equal=TRUE)
```

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

```
summary(aov(test$intend~factor(test$media)))
```

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

```
anova_model <- aov(test$intend~factor(test$media))
TukeyHSD(anova_model, conf.level= 0.01)
```

```
##   Tukey multiple comparisons of means
##     1% family-wise confidence level
##
## Fit: aov(formula = test$intend ~ factor(test$media))
##
## $`factor(test$media)`
##               diff         lwr         upr     p adj
## d2-d1 -0.86215539 -0.97829137 -0.74601940 0.1085727
## d3-d1 -0.08452381 -0.19912537  0.03007775 0.9959223
## d4-d1  0.08178054 -0.02892670  0.19248778 0.9959032
## d3-d2  0.77763158  0.66012457  0.89513859 0.1825044
## d4-d2  0.94393593  0.83022369  1.05764816 0.0573229
## d4-d3  0.16630435  0.05415969  0.27844900 0.9687417
```

There is no significant different in those 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)

Not Really. Some of the groups are not normall distribution, so we need to use Kruskal Wallis test.

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

(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:

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.

```
sales_ranks <- rank(test$intend, na.last = NA)

group_ranks <- split(sales_ranks, test$media)

group_ranksums <- sapply(group_ranks, sum) #, na.rm=TRUE)))

group_length <- sapply(group_ranks, length)

N <- length(test$intend)
H <- (12/N*(N+1))*sum((group_ranksums^2)/group_length)-3*(N+1)
H
```

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

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

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

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

```
kruskal.test(intend ~ media, data = test)
```

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

```
kw_p = 0, p-value of kruskal-Wallis rank sum test is 0.03166. They are similar.
```

(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?

```
library(FSA)
```

```
## ## FSA v0.9.3. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
dunnTest(intend ~ media, data = test, method = "bonferroni")
```

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

##   p-values adjusted with the Bonferroni method.

##   Comparison           Z     P.unadj      P.adj
## 1    d1 - d2  2.30087819 0.021398517 0.12839110
## 2    d1 - d3 -0.09233644 0.926430736 1.00000000
## 3    d2 - d3 -2.36408588 0.018074622 0.10844773
## 4    d1 - d4 -0.31452459 0.753122646 1.00000000
## 5    d2 - d4 -2.65613380 0.007904225 0.04742535
## 6    d3 - d4 -0.21613379 0.828883460 1.00000000
```

d1-d4 is significant different. Its P.adj is less than 0.05, and its abs(z value) is the greatest one among other datas.