BACS HW 7

109090046

2023-03-31

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?

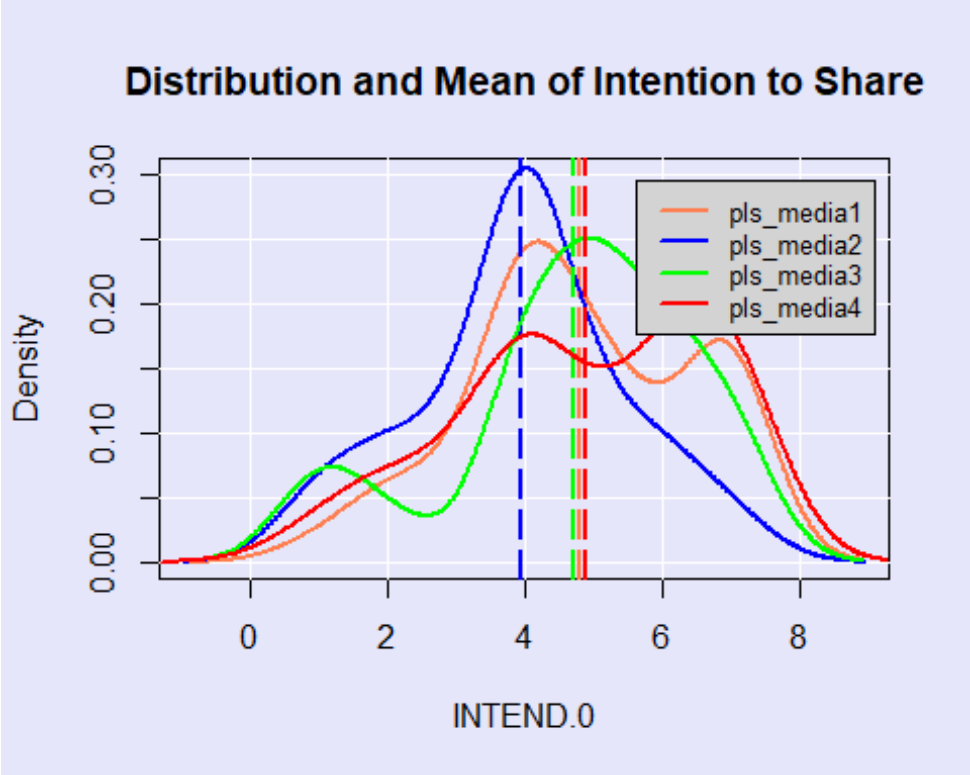
```
pls_media <- c("pls_media1", "pls_media2", "pls_media3", "pls_media4")

for (i in pls_media) {
  cat("Mean of (INTEND.0) on", i, "is", mean(get(i)$INTEND.0, na.rm=TRU
E), "\n")
}
```

```
## Mean of (INTEND.0) on pls_media1 is 4.809524
## Mean of (INTEND.0) on pls_media2 is 3.947368
## Mean of (INTEND.0) on pls_media3 is 4.725
## Mean of (INTEND.0) on pls_media4 is 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)

```
color <- c("coral", "blue", "green", "red")

par(bg="lavender")
plot(density(pls_media1$INTEND.0), ylim=c(0,0.3), xlab="INTEND.0", main
="Distribution and Mean of Intention to Share")
grid(col="white", lty=1)

for (i in 1:4) {
  lines(density(get(pls_media[i])$INTEND.0), lwd=2, col=color[i])
  abline(v=mean(get(pls_media[i])$INTEND.0), lwd=2, lty="longdash", col
=color[i])
}

legend("topright", legend = pls_media, col = color, inset = c(0.02, 0.0
5), lwd = 2, cex=0.8, bg="lightgrey")
```



c.

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

ANS:

Yes, the media type did make a difference on intention to share. For example, `pls_media2`(blue line) have different mean and different distribution from others. Although the other three types have close mean value, their distribution are a lot different from others.

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

ANS:

The null hypothesis (H0) would be that there is no difference in the mean intention to share information (INTEND.0) across the four groups.

The alternative hypothesis (Ha) is that at least one of the population means is different from the others.

H0: μ1 = μ2 = μ3 = μ4 Ha: At least one of the population means is different from the others.

b.

Let's compute the F-statistic ourselves:

i.

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

```
# Create a data frame containing the response variable 'INTEND.0' and t
he grouping variable 'group'
pls_media1_df <- data.frame(Group = rep("pls_media1", nrow(pls_media1)),
 INTEND.0=pls_media1$INTEND.0)
pls_media2_df <- data.frame(Group = rep("pls_media2", nrow(pls_media2)),
 INTEND.0=pls_media2$INTEND.0)
pls_media3_df <- data.frame(Group = rep("pls_media3", nrow(pls_media3)),
 INTEND.0=pls_media3$INTEND.0)
pls_media4_df <- data.frame(Group = rep("pls_media4", nrow(pls_media4)),
 INTEND.0=pls_media4$INTEND.0)

combined_df <- bind_rows(pls_media1_df, pls_media2_df, pls_media3_df, p
ls_media4_df)

# First, calculate the group means and the grand mean
group_means <- tapply(combined_df$INTEND.0, combined_df$Group, mean)
grand_mean <- mean(combined_df$INTEND.0)

# Calculate the degrees of freedom for between groups and within groups
df_between <- length(group_means) - 1
df_within <- length(combined_df$INTEND.0) - length(group_means)

# Calculate the sum of squares between groups (SSBG)
SSBG <- sum((group_means - grand_mean)^2) * length(combined_df$INTEND.0)
 / length(group_means)

# Calculate the sum of squares within groups (SSWG)
SSWG <- sum((combined_df$INTEND.0 - group_means[combined_df$Group])^2)

# Calculate MSTR and MSE
MSTR <- SSBG / df_between
MSE <- SSWG / df_within

# Calculate the F-statistic
F <- MSTR / MSE
```

```
## MSTR: 7.911596

## MSE: 2.869151

## F: 2.757469
```

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 <- 1 - pf(F, df_between, df_within)

## p-value: 0.04413113
```

Based on the p-value of 0.04413113 that we calculated, it appears that the result is significant at the 0.05 significance level. This means that we can reject the null hypothesis (H0) and conclude that there is a significant difference in the mean intention to share information (INTEND.0) between at least two of the four groups.

c.

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

```
# Fit the ANOVA model
fit <- aov(INTEND.0 ~ Group, data = combined_df)

summary(fit)

##              Df Sum Sq Mean Sq F value Pr(>F)
## Group         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
```

Based on the output, MSTR is 7.508, MSE is 2.869, F is 2.617 and p-value is 0.0529. In the previous question, we calculated that MSTR is 7.912, MSE is 2.869, F is 2.757 and p-value is 0.0441. The answer are pretty close, there could be slightly different when calculated using different methods because of rounding errors or differences in the way the calculations are performed.

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?

```
tukey <- TukeyHSD(fit)
tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = INTEND.0 ~ Group, data = combined_df)
##
## $Group
##                                diff         lwr       upr     p adj
## pls_media2-pls_media1 -0.86215539 -1.84660332 0.1222925 0.1085727
## pls_media3-pls_media1 -0.08452381 -1.05596494 0.8869173 0.9959223
## pls_media4-pls_media1  0.08178054 -0.85664966 1.0202107 0.9959032
## pls_media3-pls_media2  0.77763158 -0.21843807 1.7737012 0.1825044
## pls_media4-pls_media2  0.94393593 -0.01996662 1.9078385 0.0573229
## pls_media4-pls_media3  0.16630435 -0.78431033 1.1169190 0.9687417
```

By the `p adj` (adjusted p-value), we can see that every p-values are larger than the significant value of 0.05. It means that the difference between the every two sample mean being compared is not statistically significant. Therefore, we cannot reject the null hypothesis and each two sample means are equal.

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)

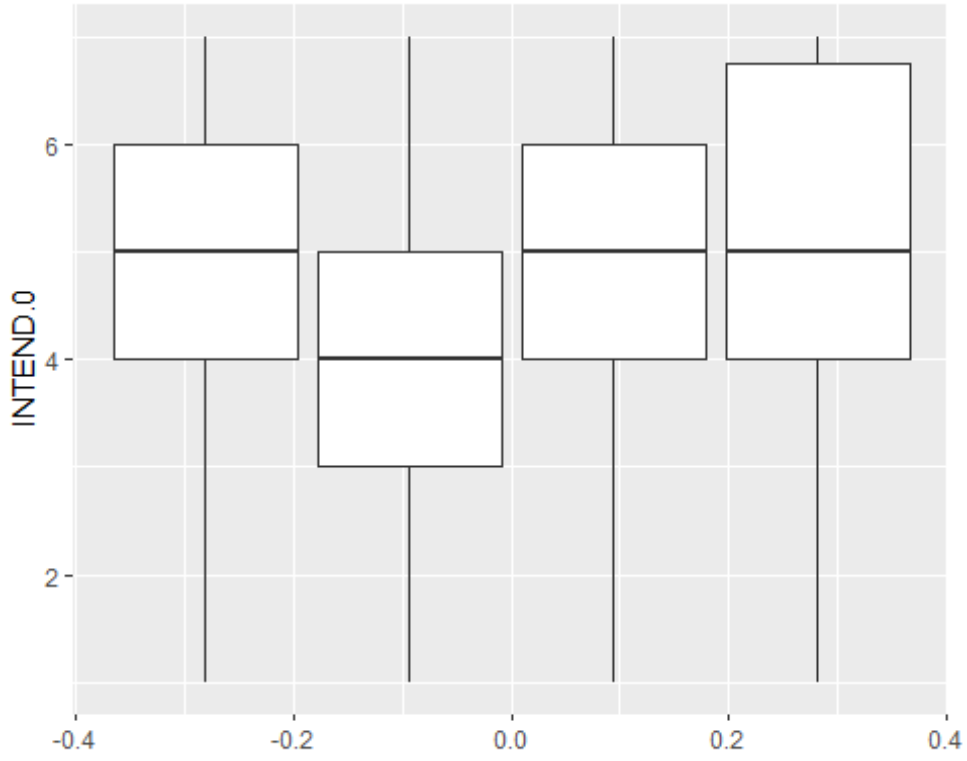
```
# Shapiro-Wilk test for normality within each group
shapiro.test(pls_media1$INTEND.0)

##
##  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
```

One-way ANOVA has several assumptions that must be met for the test to be valid. These assumptions include independence of observations, normality of the data within each group, and homogeneity of variance across groups.

The Shapiro-Wilk test tests the null hypothesis that a sample of data comes from a normally distributed population. Since all the p-values are less than significant level of 0.05, we can reject the null hypothesis and conclude that the data does not come from a normally distributed population. In other words, all the data don't meet the classic requirements of one-way ANOVA with normality.

```
pls_media <- bind_rows(pls_media1, pls_media2, pls_media3, pls_media4)
# Boxplot for homogeneity of variance across groups
ggplot(pls_media, aes(group = media, y = INTEND.0)) +
  geom_boxplot()
```



A boxplot can be used to visually compare the spread of the data across groups. If the boxes have approximately the same length, then it suggests that the variances of the groups are equal. But, the result negative. Therefore, in this case, homogeneity of variance across groups doesn't meet the classic requirements of one-way ANOVA.

The conclusion is that the dataset maybe not met the classic requirements of one-way ANOVA.

Question 3)

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

a.

State the null and alternative hypotheses

ANS:

The null hypothesis for the Kruskal-Wallis test is that the medians of the dependent variable are equal across all groups. The alternative hypothesis is that at least one group median is different from the others.

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

```
# calculate ranks
ranks <- rank(pls_media$INTEND.0)

# calculate sum of ranks for each group
sum_of_ranks <- tapply(ranks, pls_media$media, sum)

# calculate sample sizes for each group
n <- tapply(pls_media$INTEND.0, pls_media$media, length)

# calculate H
H <- (12 / (sum(n) * (sum(n) + 1))) * sum(sum_of_ranks^2 / n) - 3 * (su
m(n) + 1)
```

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.

```
p_value <- 1 - pchisq(H, length(n) - 1)

## The p-value of H is: 0.03749292
```

The p-value is less than 0.05, we can reject the null hypothesis and conclude that there is a statistically significant difference in the median intention to share between at least two of the media formats.

c.

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

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

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

The resulting p-values are similar. By both p-values, we can reject the null hypothesis and conclude that there is a statistically significant difference in the median intention to share between at least two of the media formats.

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?

```
# Change media into factor
pls_media$media <- as.factor(pls_media$media)

library(FSA)
```

```
## Warning: 套件 'FSA' 是用 R 版本 4.2.3 來建造的
```

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

```
dunnTest(INTEND.0 ~ media, data = pls_media)
```

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

##   p-values adjusted with the Holm method.

##   Comparison           Z     P.unadj      P.adj
## 1      1 - 2  2.30087819 0.021398517 0.08559407
## 2      1 - 3 -0.09233644 0.926430736 0.92643074
## 3      2 - 3 -2.36408588 0.018074622 0.09037311
## 4      1 - 4 -0.31452459 0.753122646 1.00000000
## 5      2 - 4 -2.65613380 0.007904225 0.04742535
## 6      3 - 4 -0.21613379 0.828883460 1.00000000
```

The Dunn test is a post-hoc test that can be used after a Kruskal-Wallis test to determine which pairs of groups have significantly different medians. The test returns adjusted p-values (p adj) for each pairwise comparison to account for multiple testing.

If the adjusted p-value for a pairwise comparison is greater than 0.05, it means that the difference between the medians of those two groups is not statistically significant at the 0.05 level. In other words, we cannot reject the null hypothesis that the medians of those two groups are equal.

The adjusted p-values of all pairs of groups are larger than 0.05 except for 2-4, so we can reject the null hypothesis and conclude that there is a statistically significant difference in the median intention to share between at least two of the media formats (which may be 2 and 4).

The result is same with the previous Kruskal-Wallis test.