HW6

113078506

2025-03-29

- Gen AI Usage: I use Gen AI to refine my grammar and verify my reasoning.
- Students who helped: 113078505 (Helped with checking the plot type of Q1_b), 113078502 (suggest using test to validate the classic requirements of ANOVA)

Question 1

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

```r
# read files
files <- paste0("pls-media", 1:4, ".csv")
data_list <- lapply(files, read.csv)

# calculate means
mean_values <- sapply(data_list, function(x) mean(x$INTEND.0))

# paste the name in the mean_values vector
names(mean_values) <- paste0("mean_pls-media", 1:4)

print(mean_values)
```

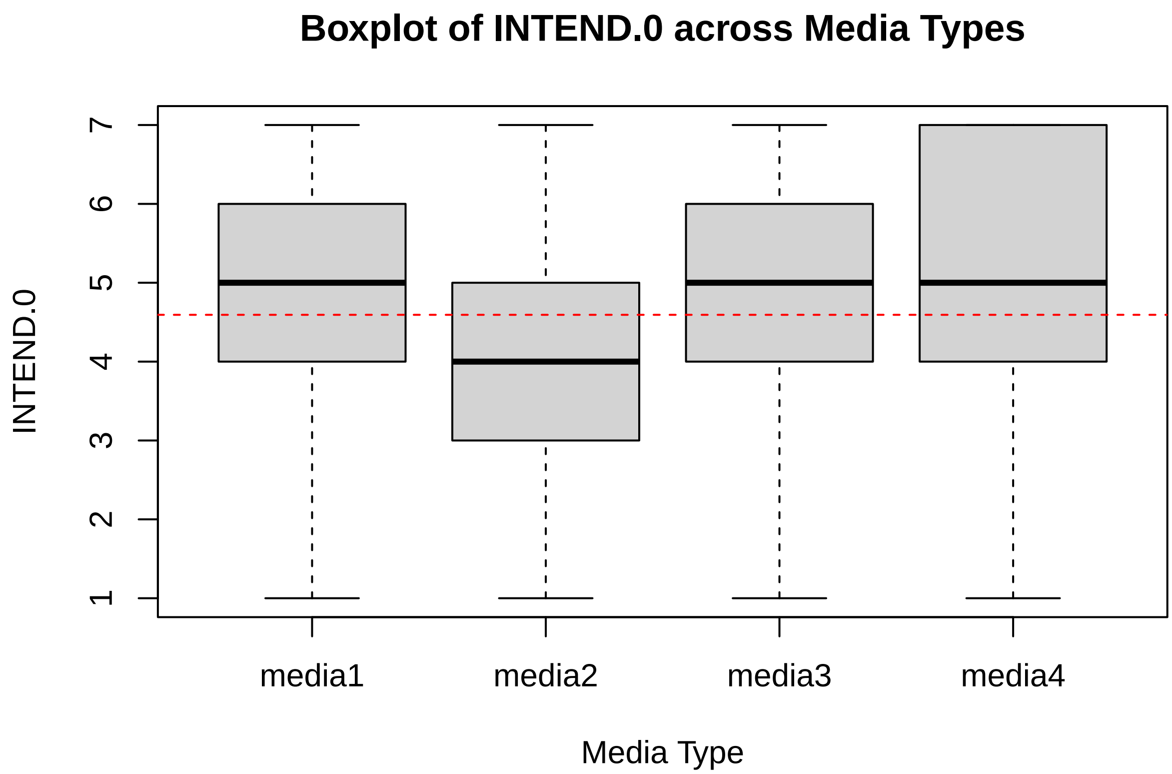
```
## mean_pls-media1 mean_pls-media2 mean_pls-media3 mean_pls-media4
##        4.809524        3.947368        4.725000        4.891304
```

b. Visualize the distribution and mean of intention to share, across all four media.

```r
# Combine data from all media types into one data frame
combined_data <- do.call(rbind, lapply(1:4, function(i) {
  data.frame(media = paste0("media", i), INTEND.0 = data_list[[i]]$INTEND.0)
}))

# Plot boxplots
boxplot(INTEND.0 ~ media, data = combined_data,
        main = "Boxplot of INTEND.0 across Media Types",
        xlab = "Media Type",
        ylab = "INTEND.0"
        )
```

```r
# add abline
abline(h = mean(mean_values), col = "red", lty = 2)
```



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

- Yes, the mean of pls-media2 (green box) is smaller than grand mean while other medias are all greater than the grand mean, I think that provide a evidence of media types makes a difference on intention to share.

Question 2

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

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_1: Not all means are equal

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

```r
# MSTR
grand_mean <- mean(mean_values)

sstr <- sum(sapply(data_list, function(i) {
  length(i$INTEND.0)*(mean(i$INTEND.0) - grand_mean)^2}))

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

# MSE
sse <- sum(sapply(data_list, function(group) {
  (length(group$INTEND.0) - 1) * var(group$INTEND.0)
}))
n_T <- sum(sapply(data_list, function(group) length(group$INTEND.0)))
df_mse <- n_T - 4
mse <- sse/df_mse

# F Statistic
f = mstr/mse

cat("MSTR:", mstr, "MSE:", mse, "F:", f)
```

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

```
## MSTR: 7.53239 MSE: 2.869151 F: 2.625303
```

```r
p_value <- pf(f, df_mstr, df_mse, lower.tail=FALSE)
cat("p-value:", p_value)
```

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: 0.05230686
```

- p-value is slightly greater than 0.05, therefore we do not have strong evidence to reject H_0. Resulting to the conclusion that the average intention of sharing among all media types are the same.

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

```r
combined_data$media <- as.factor(combined_data$media)

anova_model <- aov(INTEND.0 ~ media, data = combined_data)
summary(anova_model)
```

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

- The p-value is approximately the same as the result computed in (b).

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?

```r
TukeyHSD(anova_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = INTEND.0 ~ media, data = combined_data)
##
## $media
##                    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
```

- None of the media pairs show a statistically significant difference at the 95% confidence level. All adjusted p-values are greater than 0.05, indicating that we fail to reject the null hypothesis for each pair. This validates our conclusion of accepting H_0.

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

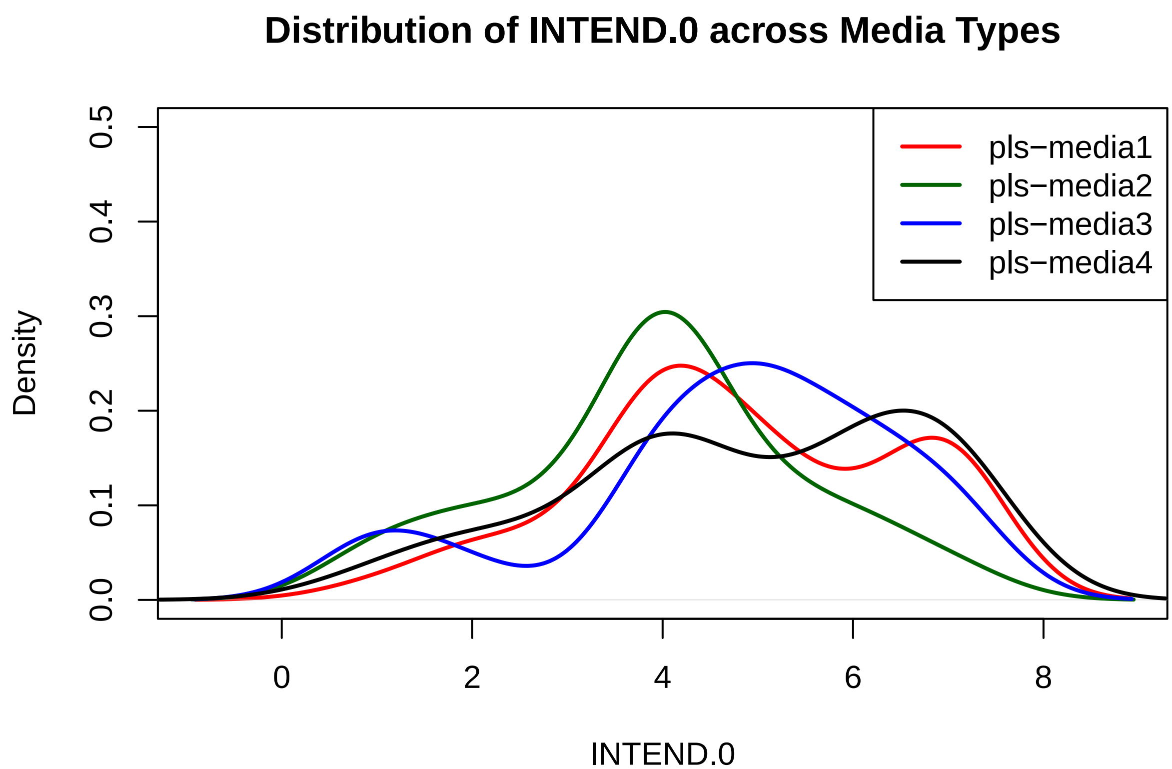
- The classic requirement including 1) each population is normally distributed, 2) The variance among the group should be the same, and 3) the observation should be independent. (Source: Course Slide: 06 Multigroup Comparisons p.18)
- Normally distributed: The density plot of 4 different groups didn't all show like normal distribution, only pls-media2 (green line) shows like a symmetrical distribution.

```r
# Plot the first media type with a density line
plot(density(data_list[[1]]$INTEND.0),
     main = "Distribution of INTEND.0 across Media Types",
     xlab = "INTEND.0",
     lwd = 2,
     ylim = c(0,0.5),
     col = "red")

# Add density lines for the other media types
colors <- c('darkgreen', 'blue', 'black')
lapply(2:4, function(i) {
  lines(density(data_list[[i]]$INTEND.0), col = colors[i - 1], lwd = 2)
})
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
```

```r
legend("topright", legend = paste0("pls-media", 1:4), col = c("red", colors), lwd = 2)
```



- Variances of different group should be the same: conduct Bartlett's Test. The p-value is greater than 0.05, so the variances of different group are the same at 95% confidence level.

```r
# Bartlett's Test
bartlett.test(INTEND.0 ~ media, data = combined_data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  INTEND.0 by media
## Bartlett's K-squared = 1.3958, df = 3, p-value = 0.7065
```

- The observation should be independent: I don't think there's a mathematical test can validate the independent of different observation. However we can make this assumption according the description of how the experiment is conducted: "Each of these four alternative media is shown to a different group of randomly assigned people."

Question 3

a. State the null and alternative hypotheses

- H_0 : All media types would give similar values if randomly drawn
- H_1 : At least media type would give larger values 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):

```
#rank
combined_data$rank <- rank(combined_data$INTEND.0)

#rank sum
rank_sums <- aggregate(rank ~ media, combined_data, sum)
n_i <- table(combined_data$media)

# Compute H statistic
N <- sum(n_i)
H <- (12 / (N * (N + 1))) * sum((rank_sums$rank^2 / n_i)) - 3 * (N + 1)
H
```

i. Show the code and results of computing H

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

```
kw_p <- 1 - pchisq(H, df = 4-1)
cat("p-value:",kw_p)
```

ii. Compute the p-value of H, from the null chi-square distribution; is the H value significant at 5% significance? If so, state your conclusion of the hypotheses.

```
## p-value: 0.03749292
```

- The p-value is smaller than critical value 0.05, we reject H_0 at 95% confidence level, showing a strong evidence that at least one media type would give larger values than another if randomly drawn.

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

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

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

- The result calculated by the built-in function in R shows the similar result as shown in (b).

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
library("FSA")
```

```
## ## FSA v0.9.6. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```r
dunnTest(INTEND.0 ~ media, data = combined_data, 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
```

- The adjusted p-value of media2 - media4 pair is smaller than 0.05, which mean this pair is significantly different at confidence level 95%. In contrary, the adjusted p-value of other pairs are all above 0.05, meaning that there's no significant difference among these pairs.