HW_Week7_108020033

Che-Wei, Chang

2023-04-07 helped by 108020024

The data in this assignment is from an actual study, but the scenario is fictional:

A health researcher is investigating how health information spreads through word-of-mouth across different media (video, pictures, text, etc.). She has

prepared some informative content about avoiding stomach aches. She is curious which media format to use, or avoid, in order to encourage people to share

such health-related information. She wants to conduct her tests at 95% confidence.

So she prepares the following four alternative media that share the same informational content:

(1) video [animation + audio]: A fully animated video with audio narration

(2) video [pictures + audio]: Video of sequence of still pictures (not-animated) with audio narration

(3) webpage [pictures + text]: Static web page with still pictures (no video) and accompanying text narration (no audio)

(4) webpage [text only]: Static web page of text narration (no audio) but no pictures

Our researcher runs an experiment where each of these four alternative media is shown to a different group of randomly assigned people. Afterwards, viewers

were surveyed about their thoughts, including a question (labeled INTEND.0) about their intention to share what they had seen with others:

(answered on 7 point scale: 1=strongly disagree; 4=neutral; 7=strongly agree)

You may find the researcher's data in a ZIP file containing four CSV files named: pls-media[1-4].csv

(note: the number in the filename corresponds to the type of media listed above)

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?

```
# load the data
media1 <- read.csv("pls-media1.csv")
media2 <- read.csv("pls-media2.csv")
media3 <- read.csv("pls-media3.csv")
media4 <- read.csv("pls-media4.csv")
```

```r
# find the means of INTEND.0 for each media type
mean(media1$INTEND.0)
```

```
## [1] 4.809524
```

```r
mean(media2$INTEND.0)
```

```
## [1] 3.947368
```

```r
mean(media3$INTEND.0)
```

```
## [1] 4.725
```

```r
mean(media4$INTEND.0)
```

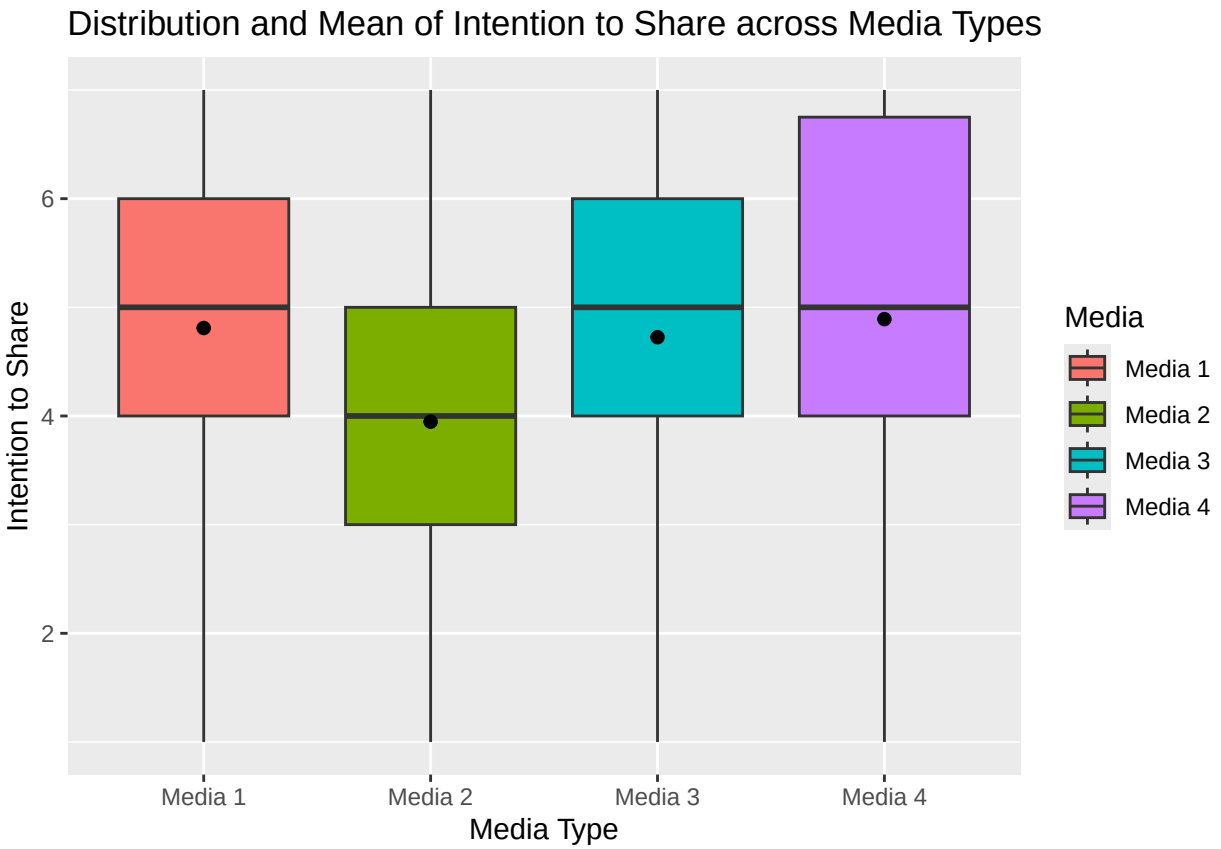
```
## [1] 4.891304
```

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)

```r
library(ggplot2)

# Combine the data from all four media types into one data frame
all_media <- data.frame(
  Media = c(rep("Media 1", nrow(media1)), rep("Media 2", nrow(media2)),
            rep("Media 3", nrow(media3)), rep("Media 4", nrow(media4))),
  INTEND = c(media1$INTEND.0, media2$INTEND.0, media3$INTEND.0, media4$INTEND.0)
)

# Create a box plot
ggplot(all_media, aes(x = Media, y = INTEND, fill = Media)) +
  geom_boxplot() +
  labs(title = "Distribution and Mean of Intention to Share across Media Types",
       x = "Media Type", y = "Intention to Share") +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 3, color = "black", fill = "white")
```



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

```
# We can see that there is a little difference from the box-plot.
# Therefore, I think that media type makes a difference on intention to share.
```

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

```
# Null hypotheses: The means of INTEND.0 are equal across all four media types.
# Alternative hypotheses: The means of INTEND.0 are not equal across at least one pair of media types.
```

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

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

```
# Declare variables to store mean of INTEND0
mean_all <- mean(all_media$INTEND)
mean1 <- mean(media1$INTEND.0)
mean2 <- mean(media2$INTEND.0)
mean3 <- mean(media3$INTEND.0)
mean4 <- mean(media4$INTEND.0)
```

```r
# Declare variables to store length of data
len1  <- length(media1$INTEND.0)
len2  <- length(media2$INTEND.0)
len3  <- length(media3$INTEND.0)
len4  <- length(media4$INTEND.0)

# Calculate MSTR and show the result
sstr <- len1 * ((mean1 - mean_all) ^2) + len2 * ((mean2 - mean_all) ^2) + len3 * ((mean3 - mean_all) ^2
df_mstr <- 4 - 1
mstr <- sstr / df_mstr
cat("MSTR: ", mstr)
```

```
## MSTR:  7.507617
```

```r
# Declare variable to store variance of each data
var1 <- var(media1$INTEND.0)
var2 <- var(media2$INTEND.0)
var3 <- var(media3$INTEND.0)
var4 <- var(media4$INTEND.0)

# Calculate MSE and show the result
sse <- (len1 - 1) * var1 + (len2 - 1) * var2 + (len3 - 1) * var3 + (len4 - 1) * var4
df_mse <- len1 + len2 + len3 + len4 - 4
mse <- sse / df_mse
cat("MSE: ", mse)
```

```
## MSE:  2.869151
```

```r
# Calculate F and show the result
f_value <- mstr / mse
cat("F: ", f_value)
```

```
## F:  2.616669
```

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
# Use function to calculate the p-value and show the result
p_value <- pf(f_value, df_mstr, df_mse, lower.tail = FALSE)
cat("p-value: ", p_value)
```

```
## p-value:  0.05289015
```

```r
# Suppose we consider the situation at 95% significance level
# Since the p-value = 0.05289015 > 0.05, we can't reject the null hypotheses.
# Therefore, F-value is not significant.
```

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

```r
oneway.test(all_media$INTEND ~ factor(all_media$Media), var.equal = TRUE)
```

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

```r
anova_model <- aov(all_media$INTEND ~ factor(all_media$Media))
summary(anova_model)
```

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

```r
# We can see that F value and P-value is same as the answer of part b
# Therefore, we got similar results
```

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
Post_Hoc_Tukey_model <- TukeyHSD(anova_model, conf.level = 0.05)
Post_Hoc_Tukey_model
```

```
##   Tukey multiple comparisons of means
##     5% family-wise confidence level
##
## Fit: aov(formula = all_media$INTEND ~ factor(all_media$Media))
##
## $`factor(all_media$Media)`
##                          diff         lwr        upr     p adj
## Media 2-Media 1 -0.86215539 -1.06562977 -0.6586810 0.1085727
## Media 3-Media 1 -0.08452381 -0.28530983  0.1162622 0.9959223
## Media 4-Media 1  0.08178054 -0.11218249  0.2757436 0.9959032
## Media 3-Media 2  0.77763158  0.57175512  0.9835080 0.1825044
## Media 4-Media 2  0.94393593  0.74470805  1.1431638 0.0573229
## Media 4-Media 3  0.16630435 -0.03017708  0.3627858 0.9687417
```

```r
# We can see that p adj of all pairs are larger than 0.05.
# Therefore, all pairs have no significant difference.
```

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)

```r
# We check three assumptions of ANOVA
# 1. Each treatment/population's response variable is normally distributed
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
```

```
# From four tables, all p-values are smaller than 0.05.
# Therefore, we can conclude that these four data are not normally distributed.
```

```
# 2. The variance (s2) of the response variables is the same for all treatments/populations
var.test(media1$INTEND.0, media2$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  media1$INTEND.0 and media2$INTEND.0
## F = 1.1607, num df = 41, denom df = 37, p-value = 0.6488
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.610132 2.184962
## sample estimates:
## ratio of variances
##             1.1607
```

```
var.test(media1$INTEND.0, media3$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  media1$INTEND.0 and media3$INTEND.0
## F = 0.87591, num df = 41, denom df = 39, p-value = 0.6752
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4658417 1.6387455
## sample estimates:
## ratio of variances
##          0.8759084
```

```
var.test(media1$INTEND.0, media4$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  media1$INTEND.0 and media4$INTEND.0
## F = 0.81677, num df = 41, denom df = 45, p-value = 0.5139
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4472891 1.5043838
## sample estimates:
## ratio of variances
##          0.8167668
```

```
var.test(media2$INTEND.0, media3$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  media2$INTEND.0 and media3$INTEND.0
## F = 0.75464, num df = 37, denom df = 39, p-value = 0.3918
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.396723 1.443427
## sample estimates:
## ratio of variances
##           0.754638
```

```
var.test(media2$INTEND.0, media4$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  media2$INTEND.0 and media4$INTEND.0
## F = 0.70368, num df = 37, denom df = 45, p-value = 0.2741
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3806992 1.3258587
## sample estimates:
## ratio of variances
##          0.7036846
```

```r
var.test(media3$INTEND.0, media4$INTEND.0, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  media3$INTEND.0 and media4$INTEND.0
## F = 0.93248, num df = 39, denom df = 45, p-value = 0.8282
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5076909 1.7361918
## sample estimates:
## ratio of variances
##          0.9324797
```

```r
# We do six variance test. All the p-values are larger than 0.05.
# Therefore, we can conclude that the variance is equal.
```

```r
# 3. The observations are independent: the response variables are not related between groups
new_media1_intend0 <- c(media1$INTEND.0, NA, NA, NA, NA)
new_media2_intend0 <- c(media2$INTEND.0, NA, NA, NA, NA, NA, NA, NA, NA)
new_media3_intend0 <- c(media3$INTEND.0, NA, NA, NA, NA, NA, NA)
new_media4_intend0 <- c(media4$INTEND.0)
chisq.test(new_media1_intend0, new_media2_intend0)
```

```
## Warning in chisq.test(new_media1_intend0, new_media2_intend0): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_media1_intend0 and new_media2_intend0
## X-squared = 43.304, df = 36, p-value = 0.1878
```

```r
chisq.test(new_media1_intend0, new_media3_intend0)
```

```
## Warning in chisq.test(new_media1_intend0, new_media3_intend0): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_media1_intend0 and new_media3_intend0
## X-squared = 26.921, df = 30, p-value = 0.6274
```

```r
chisq.test(new_media1_intend0, new_media4_intend0)
```

```
## Warning in chisq.test(new_media1_intend0, new_media4_intend0): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_media1_intend0 and new_media4_intend0
## X-squared = 44.32, df = 36, p-value = 0.1608
```

```
chisq.test(new_media2_intend0, new_media3_intend0)
```

```
## Warning in chisq.test(new_media2_intend0, new_media3_intend0): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_media2_intend0 and new_media3_intend0
## X-squared = 31.244, df = 30, p-value = 0.4035
```

```
chisq.test(new_media2_intend0, new_media4_intend0)
```

```
## Warning in chisq.test(new_media2_intend0, new_media4_intend0): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_media2_intend0 and new_media4_intend0
## X-squared = 33.668, df = 36, p-value = 0.58
```

```
chisq.test(new_media3_intend0, new_media4_intend0)
```

```
## Warning in chisq.test(new_media3_intend0, new_media4_intend0): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_media3_intend0 and new_media4_intend0
## X-squared = 30.859, df = 30, p-value = 0.4224
```

```
# We do six chi-square test. All the p-values are larger than 0.05.
# Therefore, we can't conclude that these observations are independent.
# Since we can't satisfy these three assumptions, we can't use anova table to analyze these data.
```

Question 3)

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

a. State the null and alternative hypotheses

```
# NULL hypotheses: All groups would give you similar a value if randomly drawn from them
# Alternative hypotheses: At least one group would give you a larger value 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):

i. Show the code and results of computing H

```
# Declare a variable to store the length
n = length(all_media$INTEND)

# Create a vector to represent the groups
groups <- rep(1:4, c(len1, len2, len3, len4))

# Calculate the H statistic manually
ranks <- rank(all_media$INTEND)
group_ranks <- split(ranks, all_media$Media)
R <- sapply(group_ranks, sum)
n_i <- tapply(all_media$INTEND, groups, length)
H <- (12 / (n * (n + 1))) * sum(R ^ 2 / n_i) - 3 * (n + 1)

# Print the H statistic
cat("H: ", H)
```

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

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

```
# Since the p-value = 0.03749292 < 0.05, we reject null hypotheses.
# Therefore, the H value is significant.
```

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

```
kruskal.test(all_media$INTEND ~ all_media$Media, data = all_media)
```

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

```
# We can see that H value and P-value is similar to the answer of part b
# Therefore, we 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?

```
# import the library
library(FSA)
```

```
## ## FSA v0.9.5. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
# Use dunnTest function to show the result
dunnTest(all_media$INTEND ~ all_media$Media, data = all_media, method = "bonferroni")
```

```
## Warning: all_media$Media 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
```

```
# We can see that only at p adj of Media 2 - Media 4 is smaller than 0.05.
# Therefore, only Media 2 - Media 4 has significant difference.
```