HW7

108048110

3/31/2022

BACS HW - Week 7

A health researcher is investigating how health information spreads through word-mouth across different media. She is curious which media format to use, or avoid. She wants to conduct her tests at 95% confidence.

Four alternative media formats:

Table 1: Media formats

Index	Media Formats	Information content
1	Animation + Audio	A fully animated video with audio narration.
2	Pictures + Audio	Video of sequence of still pictures with audio narration.
3	Pictures + Text	Static web page with still pictures and accompanying text narration.
4	Text only	Static web page of text narration but no pictures.

Viewers were surveyed about their thoughts, including a question about their intention to share what they had seen with others.[1]

Download packages, Load 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')

media1 <- data.frame(media1$media, media1$INTEND.0)
colnames(media1) <- c('Media1', 'Intend1')
media2 <- data.frame(media2$media, media2$INTEND.0)
colnames(media2) <- c('Media2', 'Intend2')
media3 <- data.frame(media3$media, media3$INTEND.0)
colnames(media3) <- c('Media3', 'Intend3')
media4 <- data.frame(media4$media, media4$INTEND.0)
colnames(media4) <- c('Media4', 'Intend4')
```

[1] answered on 7 point scale: 1=strongly disagree; 4=neutral; 7=strongly agree

```
knitr::kable(head(media1))
```

Media1	Intend1
1	3
1	5
1	4
1	5
1	5
1	4

Question 1)

- **a.** What are the means of viewers' intentions to share on the four media types?

```
mean(media1$Intend1)
```

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

```
mean(media2$Intend2)
```

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

```
mean(media3$Intend3)
```

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

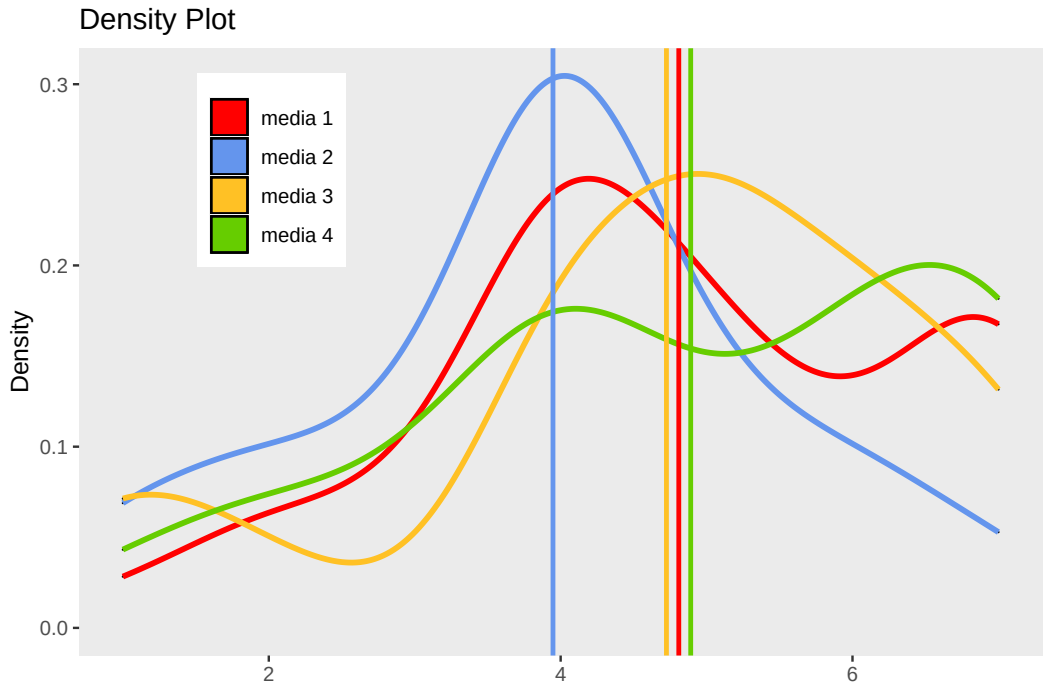
```
mean(media4$Intend4)
```

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

- **b.** Visualize the distribution and mean of intend to share, across all 4 media.

```
label <- c( "media 1"='red',
            "media 2"='cornflowerblue',
            "media 3"='goldenrod1',
            "media 4"='chartreuse3')
plt <- ggplot() +
  ggtitle('Density Plot')+
  geom_density(aes(media1$Intend1)) +
  stat_density(aes(media1$Intend1),
               color="red",
               geom="line",
               position="identity",
               lwd=1.2)+
  geom_density(aes(media2$Intend2)) +
  stat_density(aes(media2$Intend2),
               color="cornflowerblue",
               geom="line",
               position="identity",
               lwd=1.2) +
```

```
  geom_density(aes(media3$Intend3)) +
  stat_density(aes(media3$Intend3),
               color="goldenrod1",
               geom="line",
               position="identity",
               lwd=1.2) +
  geom_density(aes(media4$Intend4)) +
  stat_density(aes(media4$Intend4),
               color='chartreuse3',
               geom="line",
               position="identity",
               lwd=1.2)

plt+geom_vline(xintercept = mean(media1$Intend1),
               color='red',
               lwd=1)+
  geom_vline(xintercept = mean(media2$Intend2),
             color='cornflowerblue',
             lwd=1)+
  geom_vline(xintercept = mean(media3$Intend3),
             color='goldenrod1',
             lwd=1)+
  geom_vline(xintercept = mean(media4$Intend4),
             color='chartreuse3',
             lwd=1)+
  labs(x='', y='Density')+
  scale_fill_manual(values=label, aesthetics = c("colour", "fill"))+
  theme(
    legend.position=c(0.2,0.8),
    legend.title = element_blank(),
    panel.grid = element_blank(),
    panel.border = element_blank()
  )
```

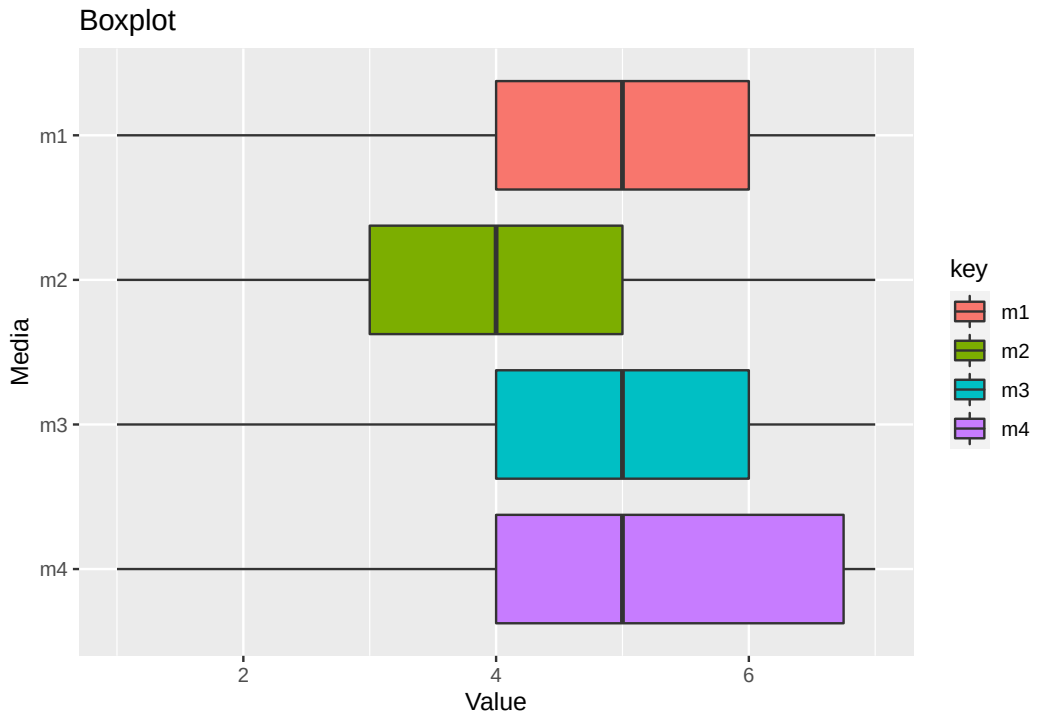



```
# transfer into long data
long_media1 <- gather(as.data.frame(media1$Intend1))
long_media2 <- gather(as.data.frame(media2$Intend2))
long_media3 <- gather(as.data.frame(media3$Intend3))
long_media4 <- gather(as.data.frame(media4$Intend4))

long_media1$key <- 'm1'
long_media2$key <- 'm2'
long_media3$key <- 'm3'
long_media4$key <- 'm4'

all_media <- rbind(long_media1, long_media2, long_media3, long_media4)

knitr::kable(head(all_media))
```

key	value
m1	3
m1	5
m1	4
m1	5
m1	5
m1	4

```
all_media$key <- as.factor(all_media$key)
plt <- ggplot(all_media, aes(x=key, y=value, fill=key))+
  geom_boxplot()+
  coord_flip()+
  scale_x_discrete(limits=rev(levels(all_media$key)))
```

```
plt+
  ggtitle('Boxplot')+
  labs(x='Media', y='Value')+
  theme(
    legend.position = 'right',
    legend.text = element_text()
  )+
  guides(fill=guide_legend(reverse=FALSE))
```



- Do you feel that media type makes a difference on intention to share?

  Ans. By observing from the plots presented above, I will infer that the mean of media 2 and the values of media 4 may be slightly different from the others.

Question 2)

Traditional one-way ANOVA

- **a.** State the null and alternative hypotheses when comparing intention to share across 4 groups in ANOVA.

  Ans.

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ $H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$
- **b.** Compute the F-statistic ourselves.
  - ***i.*** Show the code & results of computing MSTR, MSE and F.

```r
media_values <- split(all_media$value, f=all_media$key)
lengths <- sapply(media_values, length)
lengths
```

```
## m1 m2 m3 m4
## 42 38 40 46
```

```r
MSTR <- function(all_values){
  sstr=0

  grandmean <- mean(all_media$value)
  df_mstr <- length(all_values)-1

  for(i in c(1:(df_mstr+1))){
    n = length(all_values[[i]])
    xtilde = mean(all_values[[i]])
    sstr = sstr+n*(xtilde-grandmean)^2
    #print(sstr)
  }
  return(data.frame('df_mstr' = df_mstr,
                    'mstr' = round(sstr/df_mstr, 2)))
}

MSE <- function(all_values){
  sse=0

  nT <- length(all_media$value)
  k <- length(all_values)
  df_mse <- nT-k

  for(i in c(1:k)){
    n = length(all_values[[i]])
    #print(n)
    Sj = var(all_values[[i]])
    sse = sse+(n-1)*Sj
  }

  return(data.frame('df_mse' = df_mse,
                    'mse' = round(sse/df_mse, 2)))
}

mstr_value <- MSTR(media_values); mstr_value
```

```
##   df_mstr mstr
## 1       3 7.51
```

```r
mse_value <- MSE(media_values); mse_value
```

```
##   df_mse  mse
## 1    162 2.87
```

```r
Fvalue <- mstr_value$mstr/mse_value$mse; round(Fvalue, 4)
```

```
## [1] 2.6167
```

– ***ii.*** Compute the p-value of F. Is the F-value significant? State your conclusion for the hypothesis.

```
qf(p=0.95, df1 <- mstr_value$df_mstr, df2 <- mse_value$df_mse)
```

```
## [1] 2.660406
```

```
Pvalue <- pf(Fvalue,
             mstr_value$df_mstr,
             mse_value$df_mse,
             lower.tail = FALSE)
Pvalue > 1-0.95
```

```
## [1] TRUE
```

```
round(Pvalue, 4)
```

```
## [1] 0.0529
```

  - ***Statement:*** Since the calculated p-value is slightly larger than the significance level, I may state that we do not have enough confidence to reject the null hypothesis.

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

```
oneway.test(all_media$value~factor(all_media$key), var.equal=TRUE)
```

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

```
summary(aov(all_media$value~factor(all_media$key)))
```

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

  - ***Statement:*** As we observe from the data presented above, our calculation match the results.

- **d.** Regardless of your conclusions, conduct a **post-hoc Tukey Test** to see if any pairs of media have significantly different means. What do you find?

```
anova_model <- aov(all_media$value~factor(all_media$key))
TukeyHSD(anova_model, conf.level = 0.05)
```

```
##   Tukey multiple comparisons of means
##     5% family-wise confidence level
##
## Fit: aov(formula = all_media$value ~ factor(all_media$key))
##
## $`factor(all_media$key)`
##              diff         lwr        upr     p adj
## m2-m1 -0.86215539 -1.06562977 -0.6586810 0.1085727
## m3-m1 -0.08452381 -0.28530983  0.1162622 0.9959223
## m4-m1  0.08178054 -0.11218249  0.2757436 0.9959032
## m3-m2  0.77763158  0.57175512  0.9835080 0.1825044
## m4-m2  0.94393593  0.74470805  1.1431638 0.0573229
## m4-m3  0.16630435 -0.03017708  0.3627858 0.9687417
```

– ***Statement:*** As we can observe from the summary, the lower different values are between the means, the higher the p-values. All pairs of media have insignificant different means, especially between m1-m3, m1-m4, and m3-m4, which indicates that these pairs of media might not be far different from each other. In conclusion, it seems that we do not have enough evidence to reject the null hypothesis.

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

Table 4: Requirements for ANOVA

Index	Requirements
1	Each treatments or populations response variable is normally distributed.
2	The variance (s^2) of the response variables is the same for all treatments or populations.
3	The observations are independent.

Ans. ANOVA requires some assumption to be met. So I decide to verify if the classic requirements of one-way ANOVA were met by conducting two tests, Bartlett test and Dunn Test.

Bartlett test *(proving assumption 2)*

H_0 : The variability in all_media is equal for all categories.

H_1 : The variability in all_media is not equal for all categories.

```
bartlett.test(all_media$value~all_media$key)
```

```
##
##    Bartlett test of homogeneity of variances
##
## data:  all_media$value by all_media$key
## Bartlett's K-squared = 1.3958, df = 3, p-value = 0.7065
```

Dunn Test *(non-parametric test)*

```
dunnTest(all_media$value~factor(all_media$key),
         data=all_media,
         method='bonferroni')
```

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

##   p-values adjusted with the Bonferroni method.

##   Comparison           Z     P.unadj      P.adj
## 1    m1 - m2  2.30087819 0.021398517 0.12839110
## 2    m1 - m3 -0.09233644 0.926430736 1.00000000
## 3    m2 - m3 -2.36408588 0.018074622 0.10844773
## 4    m1 - m4 -0.31452459 0.753122646 1.00000000
## 5    m2 - m4 -2.65613380 0.007904225 0.04742535
## 6    m3 - m4 -0.21613379 0.828883460 1.00000000
```

– ***Statement:*** When there is no assumptions of variance homogeneity and no assumptions of equal group sizes, we can easily observe from the table that almost no pairs of media have significant different means except for m2 and m4. Their p-value lies slightly below the significance level: 0.05, this result verify that the classic requirements of one-way ANOVA are met.

Question 3)

Non parametric Kruskal Wallis Test

None parametric means that the test does not assume your data comes from a particular distribution. The H test is used when the assumptions for ANOVA aren't met.

- **a.** State the null and alternative hypotheses. [2]
  Ans.
  $$H_0 : \frac{\sum_{i=1}^{n_1} rank(V_1)}{n_1} = \frac{\sum_{i=1}^{n_2} rank(V_2)}{n_2} = \frac{\sum_{i=1}^{n_3} rank(V_3)}{n_3} = \frac{\sum_{i=1}^{n_4} rank(V_4)}{n_4}$$
  $$H_1 : \frac{\sum_{i=1}^{n_1} rank(V_1)}{n_1} \neq \frac{\sum_{i=1}^{n_2} rank(V_2)}{n_2} \neq \frac{\sum_{i=1}^{n_3} rank(V_3)}{n_3} \neq \frac{\sum_{i=1}^{n_4} rank(V_4)}{n_4}$$
- **b.** Compute H value ourselves.
  - ***i.*** Show the code and results of computing **H**.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

```
all_media$rank <- rank(all_media$value)

knitr::kable(head(all_media))
```

key	value	rank
m1	3	28.5
m1	5	97.5
m1	4	58.5
m1	5	97.5
m1	5	97.5
m1	4	58.5

```
group_rank <- split(all_media$rank, f=all_media$key)
sapply(group_rank, sum)
```

```
##     m1     m2     m3     m4
## 3693.5 2421.0 3556.0 4190.5
```

```
H <- function(all_values){
  h=0
  N = length(all_media$value)
  k = length(all_values)
  for(i in 1:k){
    R = sum(group_rank[[i]])
    n = length(group_rank[[i]])
    h = h+R^2/n
  }
  return(12/(N*(N+1))*h-3*(N+1))
}

H_value <- H(media_values); H_value
```

[2] V_i : values of group i.

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

- ***ii.*** Compute the p-value of H. Is the **H** value significant? State your conclusion of the hypotheses.

```
kw_p <- round(1-pchisq(H_value, df=3), 5); kw_p<1-0.95; kw_p
```

```
## [1] TRUE
## [1] 0.03749
```

- ***Statement:*** Based on the result we get, H value is significant, hence we can conclude that we have enough evidence to reject the null hypothesis, not all the mean ranks of the groups are the same; in other words, the medians of the groups are different .

- **c.** Conduct the same test using the **`kruskal.wallis()`** function - confirm that you got similar results.

```
kruskal.test(all_media$value~factor(all_media$key), data=all_media)
```

```
##
##    Kruskal-Wallis rank sum test
##
## data:  all_media$value by factor(all_media$key)
## Kruskal-Wallis chi-squared = 8.8283, df = 3, p-value = 0.03166
```

- ***Statement:*** There will be some difference in the calculation of H because **kurskal.test()** accounts for ties in rank. However, we still got the similar results, not all the mean ranks of the groups are the same.

- **d.** Regardless of your conclusion, conduct a **post-hoc Dunn Test** to see if any pairs of media are significantly different. What do you find?

```
dunnTest(all_media$value~factor(all_media$key),
         data=all_media,
         method='bonferroni')
```

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

##   p-values adjusted with the Bonferroni method.

##   Comparison           Z     P.unadj      P.adj
## 1    m1 - m2  2.30087819 0.021398517 0.12839110
## 2    m1 - m3 -0.09233644 0.926430736 1.00000000
## 3    m2 - m3 -2.36408588 0.018074622 0.10844773
## 4    m1 - m4 -0.31452459 0.753122646 1.00000000
## 5    m2 - m4 -2.65613380 0.007904225 0.04742535
## 6    m3 - m4 -0.21613379 0.828883460 1.00000000
```

Ans. We can conclude that nearly all 4 media formats have similar results except for m2 and m4.