HW7

110078509 this guy helps me: 109065707

20220402

< Preface >

I merge pls-media{1~4}.csv 's column "INTEND.0" into a dataframe, called df as below.

In the process, I pad each vector into the same length with NA.

Then I hide the code for better review quality.

```
head(df, 3)
```

```
##   df1_intend0 df2_intend0 df3_intend0 df4_intend0
## 1           3           4           1           3
## 2           5           6           4           4
## 3           4           4           1           4
```

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?

```
sapply(df, mean)
```

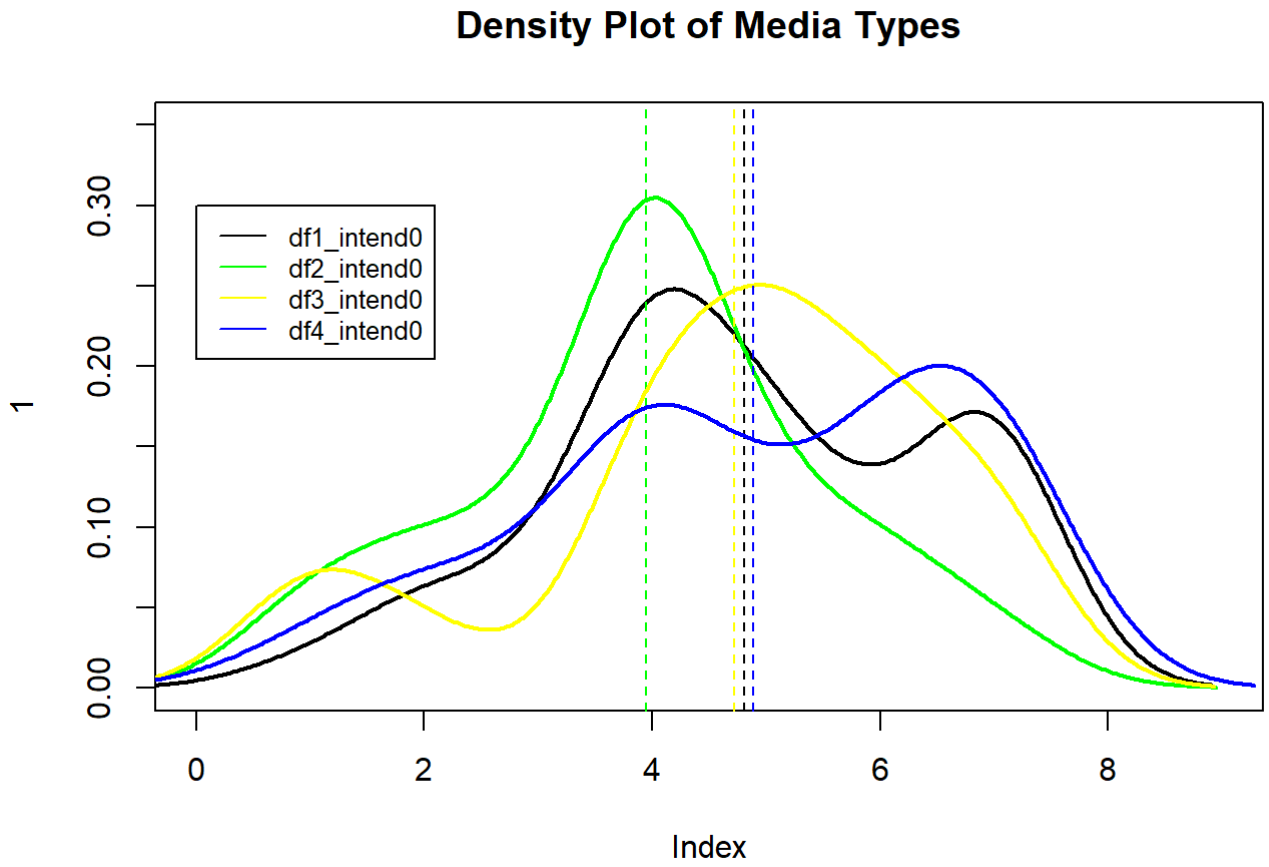
```
## df1_intend0 df2_intend0 df3_intend0 df4_intend0
##    4.809524    3.947368    4.725000    4.891304
```

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

- Note: Even though there's lots of repeated code as below,

I did not write it as a function. Because we don't need to reuse these function.

```
# Set the color
color_set = c('black','green', 'yellow', 'blue' )
# empty plot
plot(1,main='Density Plot of Media Types',lwd=2, xlim = c(0, 9) ,ylim=c(0,0.35) , col=color_s
et[1])
# Loop for plotting process
for(i in 1:length(df)) {
    lines(density(df[[i]]),lwd=2,col= color_set[i])
    abline(v=mean(df[[i]]),lty=2,col= color_set[i] ,lwd=1)
}
legend(x = 0, y = 0.3, legend=colnames(df),col= color_set ,  lty=c(1,1,1,1), cex=0.8)
```




```
#----------
```

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

- Ans:

Each of them have a different distribution, however, their means are pretty close. So, it would be hard to tell the media makes a difference on intention to share.

Question 2) Let's try traditional one-way ANOVA:

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

H0:The means of the four treatment populations are the same

Event: <μ1 =μ2 =μ3 = μ4>

H1:The means of the four treatments populations are not the same

Event:<μ1 ≠μ2 ; μ1 ≠μ3 ;μ1 ≠μ4 ; μ2 ≠μ3 , μ2 ≠μ4, μ3 ≠μ4 >

(b). Complete the F-statistic ourselves:

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

- Ans:

MSTR = SSTR / df_MSTR

MSE = SSE / df_MSE

F_score = mstr/ mse

```
# Function definition for mu, n (total data size)
mu<- mean(sapply(df, mean))
valid_n <- sum(length(df[,1]),length(df[,2]),length(df[,3]) ,length(df[,4]))
# Function definition for SSTR
sstr_count <- function(x){length(x)*(( mean(x) - mu)^2)}
sstr <- as.numeric(sum(sapply( df , sstr_count)))

# Unbiased degree of freedom of MSTR = 4-1
df_mstr= 4-1

# Function definition for mstr
mstr= sstr/df_mstr;
sprintf("mstr: %f ",mstr)
```

```
## [1] "mstr: 7.532390 "
```

```
# Function definition for SSE
sse_count <- function(x){sum(( length(x) - 1 ) * var(x))}
sse <- sum(sapply(df, sse_count))

# Unbiased degree of freedom of MSE = 166 - 4
df_mse <- valid_n - length(df) #166-4
# MSE
mse <- sse / df_mse ;
sprintf("MSE: %f ",mse)
```

```
## [1] "MSE: 2.869151 "
```

```
# F
f_score <-mstr/mse
sprintf("F: %f ",f_score)
```

```
## [1] "F: 2.625303 "
```

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.

```
f_cut <- qf(p=0.95, df1=df_mstr, df2=df_mse)

# Use pf() CDF with df seting 3 & 163
p_value <- pf(f_score, df_mstr, df_mse, lower.tail=FALSE)
cat("F:",f_score,"\n","F Cut:",f_cut, "\n","p value:",p_value )
```

```
## F: 2.625303
##  F Cut: 2.660406
##  p value: 0.05230686
```

```
# Note that the argument "lower.tail"
#logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x]
# In other words, we set lower.tail=FALSE to get the right part area of the X
```

- Ans:

No,it's not significant.

p_value = 0.05230686 > 0.05 under 95% confident level. Therefore, we can not reject the H0, which means that the means of the four treatment populations are the same.

---

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

- *Ans:*

It seems that the variance of each group is 0, they're almost the same.

Hence, we can set var.equal=TRUE as following:

As the question asking us to use aov(),

the question assume that it reach the assumption of the test itself. Therefore, we skipped the independent, nomlity, and the test of Homogeneity of variance in this question.

```
# It weird that melt function has strange number inside it
# long_df <-melt(df, id.vars= NULL, variable.name= "Set",value.name= "value")
m1 <- data.frame(set=rep("m1",length(df$df1_intend0)), value=df$df1_intend0)
m2 <- data.frame(set=rep("m2",length(df$df2_intend0)), value=df$df2_intend0)
m3 <- data.frame(set=rep("m3",length(df$df3_intend0)), value=df$df3_intend0)
m4 <- data.frame(set=rep("m4",length(df$df4_intend0)), value=df$df4_intend0)
long_df <- rbind(m1, m2, m3, m4)

# Run oneway.test()function for one-way ANOVA
# var.equal=TRUE
# one-way ANOVA model Using aov()
model <- aov(value~factor(set), data=long_df); model
```

```
## Call:
##    aov(formula = value ~ factor(set), data = long_df)
##
## Terms:
##                 factor(set) Residuals
## Sum of Squares      22.5229  464.8024
## Deg. of Freedom           3       162
##
## Residual standard error: 1.693857
## Estimated effects may be unbalanced
```

```
summary(model)
```

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

- *Ans:*

No, it's not significant.the p-value of F value is 0.0529 > 0.05 under 95% confident level.

Hence, we can not reject H0.

And Yes, we get the same result as the question above.

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?

Well, the assumptions of TukeyHSD are:

1- All the observations set are independent

2- Sample from Normal distribution population ;

3- Equal variation across observations ;

4- equal sample sizes, and if it's not, use Tukey-Kramer Test.

As far as I can tell, the TukeyHSD function uses the Tukey-Kramer procedure. So we don't have to worry about it.

```
HSD <- TukeyHSD(model, conf.level = 0.01); HSD
```

```
##   Tukey multiple comparisons of means
##     1% family-wise confidence level
##
## Fit: aov(formula = value ~ factor(set), data = long_df)
##
## $`factor(set)`
##               diff         lwr         upr     p adj
## m2-m1 -0.86215539 -0.97829137 -0.74601940 0.1085727
## m3-m1 -0.08452381 -0.19912537  0.03007775 0.9959223
## m4-m1  0.08178054 -0.02892670  0.19248778 0.9959032
## m3-m2  0.77763158  0.66012457  0.89513859 0.1825044
## m4-m2  0.94393593  0.83022369  1.05764816 0.0573229
## m4-m3  0.16630435  0.05415969  0.27844900 0.9687417
```

Ans: Due to the p-value shown above, I think there is no significant value among any 2 pairs under the 95% confident level.

---

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

ANOVA requires some assumptions to be met:

Each group of data should pass:

- 1. Normality assumption (Use Shapiro-Wilk Test)
- 2. Homogeneity of variance assumption

(a.Bartlett’s test, b.Levene’s test)

- 3. The observations are independent
- e-1. Normality Test:

```
# Shapiro-Wilk常態性檢定
temp <- sapply(df,shapiro.test)
attributes(temp)
```

```
## $dim
## [1] 4 4
##
## $dimnames
## $dimnames[[1]]
## [1] "statistic" "p.value"   "method"    "data.name"
##
## $dimnames[[2]]
## [1] "df1_intend0" "df2_intend0" "df3_intend0" "df4_intend0"
```

- *Ans:*

Appearently, each of them are normal distribution according to their p-value . Hence, normality passed.

- e-2. Homogeneity of variance

Most of time, Bartlett test is for Normal Distribution datasets.In contrast, Levene test is mostly used for the condition of non-normal distribution situation.Here, I use Bartlett Test of Homogeneity of Variances due to the normality of them are passed.

(Caution: Not Bartlett of sphericity test)

```
# For parametric population
bartlett.test(value ~ set , data = long_df)
```

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

```
# ps: what's the different of of this concept:
# Test the sample variance of 4 group # sum(x - sample mean)/ n-1
# sample_variance_test <- function(x){
#     df_len <- length(x)-1;
#     sum(x - mean(x))/ df_len
# }
# sapply(df, sample_variance_test)
```

- *Ans:*

p-value < 2.2e-16, it indicates that each sets of data do not have the same variance.

Therefore, it doesn’t pass the Homogeneity of variance test. Therefore, it doesn’t net requirements of one-way ANOVA.

- e-3. The observations are independent

Each of these four alternative media is shown to a different panel of randomly assigned people.

Afterwards, viewers were surveyed for the grading. And there's no any "time" related variable is includes in the data provided. So we considered it as independent.

To Summary:

It does not pass the Bartlett test of homogeneity of variances. Therefore, it doesn't reach the requirements.

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)

- *Ans:*

Let mean rank of group i equal to ηi H0： η1 =n2 = η3 = η4 H1： η1≠ n2 | η1≠ n3 | η1≠ n4 | η2≠ n3 | η2≠ n4 | η3≠ n4

(ps: "|" represents "OR" instead of conditional probability given)

b. Let's compute (an approximate) Kruskal Wallis H ourselves:

- i. Show the code and results of computing H

```
# Rank all the combined values across groups
rank<- rank(long_df$value)

# combine rank into medias_intention
long_df_rk <- cbind(long_df, rank)

# split the same rank into same group
group_rank <- split(long_df_rk$rank, long_df_rk$set )
group_ranksum <- sapply(group_rank, sum)
group_rank_length <- sapply(group_rank, length)
N <- sum(group_rank_length)
H <- (12 / (N * (N + 1))) * sum((group_ranksum^2) / group_rank_length) - 3 * (N + 1)

paste("H value is", H)
```

```
## [1] "H value is 8.45465979544389"
```

- Ans:

H = 8.45465979544389

- ii. Compute the p-value of H, from the null chi-square distribution; is the H value significant?

```
# . Find p-value of H from chi-square distribution
k = 4
kw_p <- 1 - pchisq(H, df=k-1)
kw_p
```

```
## [1] 0.03749292
```

- Ans:

The p- value of Kruskal-Wallis chi-squared is 0.03749292 < 0.05. We reject null chi-square distribution, accept H1 that the mean rank among 4 groups are not the same.

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

```
kruskal.test(value ~ set , data = long_df)
```

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

- Ans:

Kruskal-Wallis chi-squared = 8.8283 ≈ 8.45465979544389(H value).

p-value = 0.03166 < 0.05, we reject H0, accept H1. Hence, we get similar results.

d. Regardless of your conclusions, conduct a post-hoc Dunn test . what do you find?

```
dunnTest(value ~ set, data = long_df, method = "bonferroni")
```

```
## Warning: set 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    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:

I figured out that m2-m4's p-value = 0.04742535 < 0.05. It represents that the pair m2 & m4 is significant different.