HW07

111078513

Help By 111078511, 111078505

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?

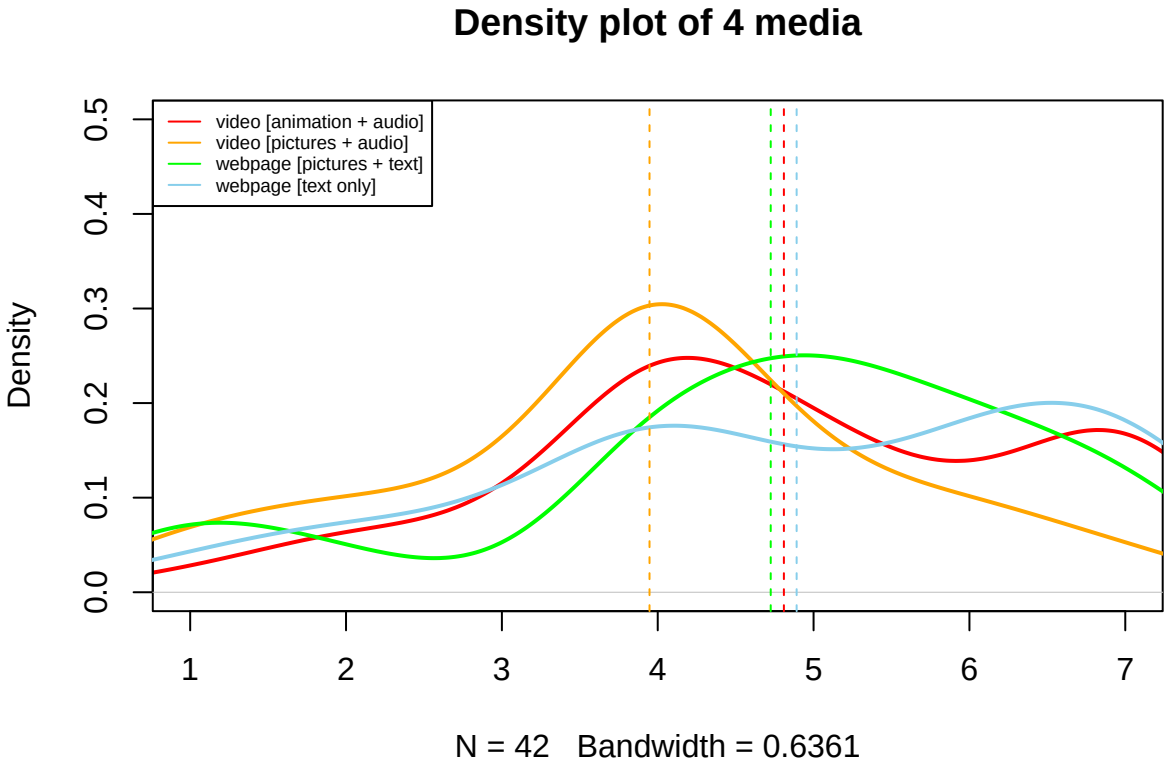
```
# install.packages("data.table")
library(data.table)
pls_media1 <- fread("pls-media1.csv")
pls_media2 <- fread("pls-media2.csv")
pls_media3 <- fread("pls-media3.csv")
pls_media4 <- fread("pls-media4.csv")

mean1 <- mean(pls_media1$INTEND.0)
mean2 <- mean(pls_media2$INTEND.0)
mean3 <- mean(pls_media3$INTEND.0)
mean4 <- mean(pls_media4$INTEND.0)
cat("mean1:",mean1,", mean2:",mean2, ", mean3:",mean3, ", mean4:",mean4)
```

```
## mean1: 4.809524 , mean2: 3.947368 , mean3: 4.725 , mean4: 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)

```
plot(density(pls_media1$INTEND.0), col="red", lwd=2, main="Density plot of 4 media", xlim=c(1, 7),ylim=
lines(density(pls_media2$INTEND.0), col="orange", lwd=2)
lines(density(pls_media3$INTEND.0), col="green", lwd=2)
lines(density(pls_media4$INTEND.0), col="skyblue", lwd=2)
abline(v=mean1,col="red",lwd=1,lty=2)
abline(v=mean2,col="orange",lwd=1,lty=2)
abline(v=mean3,col="green",lwd=1,lty=2)
abline(v=mean4,col="skyblue",lwd=1,lty=2)
legend('topleft',0.2, cex = 0.55, lty=1
       , c("video [animation + audio]", "video [pictures + audio]", "webpage [pictures + text]", "webpag
       , col = c("red", "orange", "green", "skyblue"))
```



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

Even though there are three types of media that would typically agree to share, and a 50-50 split for sharing on one type of media, I do not observe any significant differences among these four types of media. This could be due to various reasons, such as the video not being attractive enough for the participants. In conclusion, people tend to become neutral when deciding whether to share information with others, regardless of the type of media.

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

- H0:There is no difference among the four 4 types of media.
- H1:There is a difference among the four types of media.

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

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

```
# Combine the datasets into one data frame
data <- data.frame(value = c(pls_media1$INTEND.0, pls_media2$INTEND.0,
                             pls_media3$INTEND.0, pls_media4$INTEND.0),
                   group = rep(1:4, c(42, 38, 40, 46)))
# Calculate the overall mean
grand_mean <- mean(data$value)
# Calculate the sum of squares due to treatments (SSTR)
SSTR <- sum((tapply(data$value, data$group, mean) - grand_mean)^2 * c(42, 38, 40, 46))
```

```r
# Calculate the mean square due to treatments (MSTR)
k <- length(unique(data$group))
df_mstr <- k - 1
MSTR <- SSTR / df_mstr
# Calculate the sum of squares due to error (SSE)
SSE <- sum((tapply(data$value, data$group, sd)^2) * (c(42, 38, 40, 46) - 1))
# Calculate the mean square due to error (MSE)
nT <- length(data$value)
df_mse <- nT - k
MSE <- SSE / df_mse
# Calculate the F-statistic
F_stat <- MSTR / MSE
# Print the results
cat("MSTR:", MSTR, "\n")
```

```
## MSTR: 7.507617
```

(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
p_value <- pf(F_stat, df_mstr, df_mse, lower.tail=FALSE);p_value
```

```
## [1] 0.05289015
```

p-value is 0.05289015, is bigger than 0.05(we set confidence interval 95%), so we should reject H0, which means the mean of 'INTEND.0' in 4 media are not the same.

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

```r
anova_model <- aov( data$value ~ factor(data$group))
summary(anova_model)
```

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

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, conf.level = 0.05)
```

```
##   Tukey multiple comparisons of means
##     5% family-wise confidence level
##
## Fit: aov(formula = data$value ~ factor(data$group))
##
## $`factor(data$group)`
##            diff         lwr        upr     p adj
## 2-1 -0.86215539 -1.06562977 -0.6586810 0.1085727
## 3-1 -0.08452381 -0.28530983  0.1162622 0.9959223
```

```
## 4-1  0.08178054 -0.11218249  0.2757436 0.9959032
## 3-2  0.77763158  0.57175512  0.9835080 0.1825044
## 4-2  0.94393593  0.74470805  1.1431638 0.0573229
## 4-3  0.16630435 -0.03017708  0.3627858 0.9687417
```

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.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
```

The key assumption for conducting a valid one-way ANOVA is that the data within each group should follow a normal distribution. However, when we performed the Shapiro-Wilk test (a simple function to test for normality), the p-values for all the groups were found to be less than 0.05, indicating that the data is not normally distributed. As a result, since the assumption of normality has been violated, alternative methods such as nonparametric tests may need to be considered instead of ANOVA. # Question 3) Let's use the non-parametric Kruskal Wallis test:

a. State the null and alternative hypotheses

- H0: All groups would give you similar a value if randomly drawn from them.
- H1: At least one group would give you a larger value than another if randomly

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

```
media_ranks <- rank(data$value)
group_ranks <- split(media_ranks,data$group)
sapply(group_ranks, sum)
```

```
##      1      2      3      4
## 3693.5 2421.0 3556.0 4190.5
```

```
N <- length(data$value)
H <- 12/(N*(N+1)) *
  sum(tapply(media_ranks,data$group,sum)^2/
        tapply(data$value,data$group,FUN = length)) - 3*(N+1);H
```

```
## [1] 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.

```
kw_p <- 1 - pchisq(H, df=k-1);kw_p
```

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

The p-value of H is 0.0375, which is smaller than 0.05, so the H value is significant. Therefore, we should reject the null hypothesis and say that there is at least one group would give you a larger value than another if randomly drawn.

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

```
kruskal.test(value ~ group, data = data)
```

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

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?

```
#install.packages("FSA")
require(FSA)
```

```
## Loading required package: FSA
```

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

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

```
## Warning: group 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      1 - 2  2.30087819 0.021398517 0.12839110
## 2      1 - 3 -0.09233644 0.926430736 1.00000000
## 3      2 - 3 -2.36408588 0.018074622 0.10844773
## 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
```

Due to the post-hoc test, after viewing the adjustment of p-value, we can find that media2 has slightly significant difference from media4 at 95% confidence, but except of this, there is no significant difference between 4 types of media.