HW6

Answer are in red text!

Question 1) Let's describe and visualize the data:

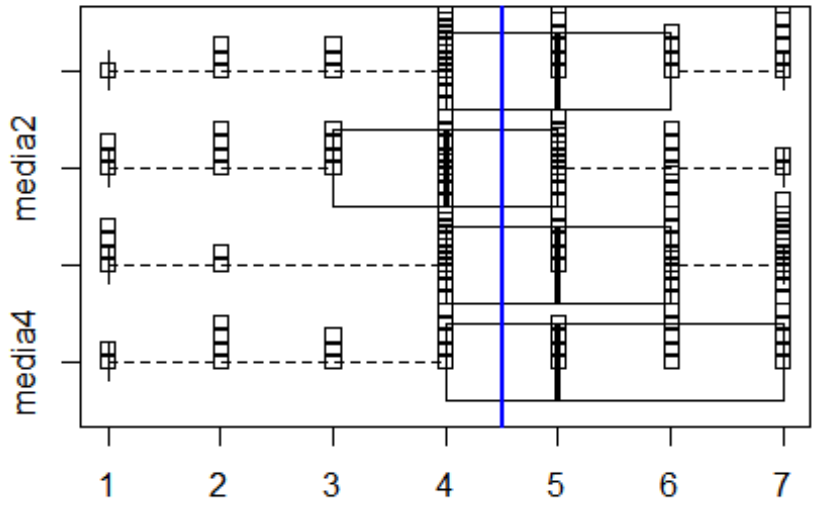
```
#load data
media1 <- read.csv("C:/Users/tsunh/Desktop/Schoolwork/BASM/health-media
1.csv")
media2 <- read.csv("C:/Users/tsunh/Desktop/Schoolwork/BASM/health-media
2.csv")
media3 <- read.csv("C:/Users/tsunh/Desktop/Schoolwork/BASM/health-media
3.csv")
media4 <- read.csv("C:/Users/tsunh/Desktop/Schoolwork/BASM/health-media
4.csv")
```

a. What are the means of viewers intentions to share (INTEND.0) for each media type? (report four means)

```
#report four mean
mean(media1$INTEND.0)

## [1] 4.809524

mean(media2$INTEND.0)

## [1] 3.947368

mean(media3$INTEND.0)

## [1] 4.725

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; Recommendation: conceptualize your visualization on paper, then search online for how to produce it)

```
data_list <- list(media1$INTEND.0,media2$INTEND.0,media3$INTEND.0,media
4$INTEND.0)
media_frame <- as.data.frame(sapply(data_list, '[', seq(max(lengths(dat
a_list)))))
colnames(media_frame) <- c("media1","media2","media3","media4")
boxplot(rev(media_frame), horizontal=TRUE)
stripchart(rev(media_frame),method="stack", add=TRUE)
abline(v=mean(sapply(na.omit(media_frame), mean)),col = "blue",lwd = 2)
```



c. Based on the visualization, do you feel that the type of media make a difference on intention to share?

From the above picture, we can find out that the blue line, which represents the total mean of four media channels, penetrates all of the box. Besides, the strip plot shows that the score from different medias behavior similarly. Therefore, I think the type of media might not make a difference on intention to share.

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

a. Satate the null and alternative hypotheses when comparing INTEND.0 across four groups using ANOVA

Let $\mu_1, \mu_2, \mu_3, \mu_4$ corresponding to each media

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_{alt}: the means are not same

b. Model and produce the F-statistic for our test

```
four_mean <- sapply(na.omit(media_frame), mean)
total_mean <- mean(sapply(na.omit(media_frame), mean))
sstr <- 0
for (media in colnames(media_frame)){
  sstr = sstr + dim(na.omit(media_frame[media]))[1]*((four_mean[media]-
total_mean)^2)
```

```
}
unname(sstr)

## [1] 17.39439

df_mstr<-4-1
mstr<-sstr/df_mstr
unname(mstr)

## [1] 5.79813

four_var <- sapply(na.omit(media_frame), var)
sse <- 0
N <- 0
for (media in colnames(media_frame)){
  sse = sse + (dim(na.omit(media_frame[media]))[1]-1)*(four_var[media])
  N = N+dim(na.omit(media_frame[media]))[1]
  }

df_mse<- N-4
mse<-sse/df_mse
unname(mse)

## [1] 2.882971

f_value = mstr/mse
cat("f-value",f_value)

## f-value 2.011165

cat("\ncut-off",qf(p=0.95, df1=df_mstr, df2=df_mse))

##
## cut-off 2.660406

p_value<-pf(f_value, df_mstr, df_mse, lower.tail=FALSE)
cat('\n',p_value)

##
##  0.1144646
```

c. What is the appropriate cut-off values of F for 95% and 99% confidence

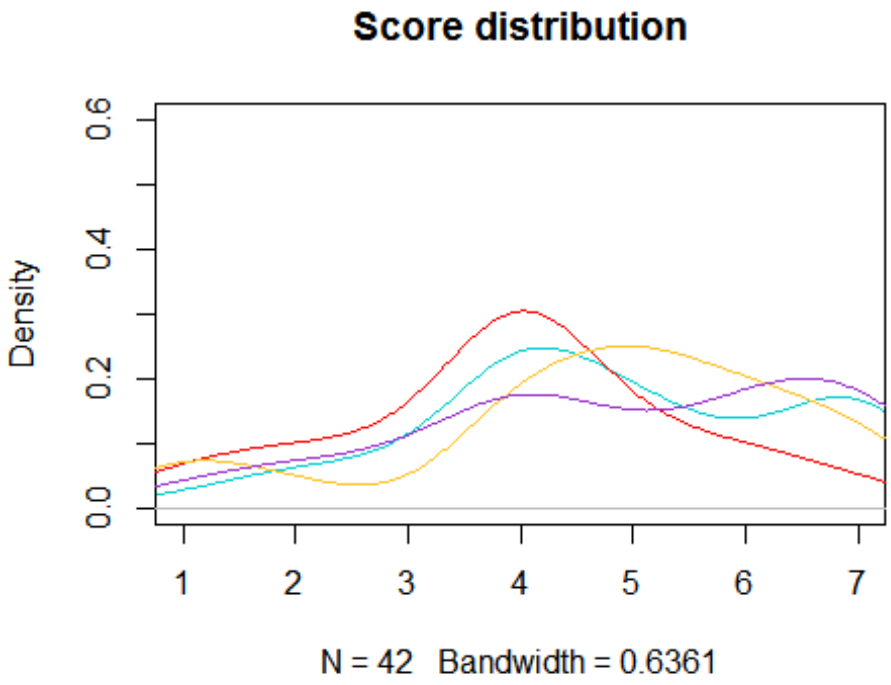
```
cat("\ncut-off 95%",qf(p=0.95, df1=df_mstr, df2=df_mse))

##
## cut-off 95% 2.660406

cat("\ncut-off 99%",qf(p=0.99, df1=df_mstr, df2=df_mse))

##
## cut-off 99% 3.904807
```

d. According to the traditional ANOVA, do the four types of media produce the same mean intention to share, at 95% confidence? How about at 99% confidence?

Since the F score could not reach the cut-off point at both 95% an 99% confidence interval, we should not reject the null hypothesis. Therefore, the four types of media might produce the same mean intention to share.

e. Are the classic requirements of one-way ANOVA met? Why or why not?

```
plot(density(media1$INTEND.0),xlim=c(1,7),ylim=c(0,0.6),col="darkturquoise",main="Score distribution")
lines(density(media2$INTEND.0),col="red")
lines(density(media3$INTEND.0),col="goldenrod1")
lines(density(media4$INTEND.0),col="darkorchid3")
```



I think the classic requirements did not be met in this dataset. First of all, each group sample should be drawn from a normally distributed population. However, from the above plot, it's clear that the distribution are not normal.

```
var(na.omit(media1$INTEND.0))

## [1] 2.694541

var(na.omit(media2$INTEND.0))

## [1] 2.321479
```

```r
var(na.omit(media3$INTEND.0))
```

```
## [1] 3.076282
```

```r
var(na.omit(media4$INTEND.0))
```

```
## [1] 3.299034
```

Secondly, all populations should have a common variance. But from the above caculation, the variances seem to be different. Therefore, the condition might not fit the requirement of tranditional ANOVA.

Question 3)

a. Bootstrap the null values of F and also the actual F-statistic.

```r
media_tidy1 <-data.frame(strategy=rep(1,length(na.omit(media_frame$media1))),
                 score=na.omit(media_frame$media1))
media_tidy2 <-data.frame(strategy=rep(2,length(na.omit(media_frame$media2))),
                 score=na.omit(media_frame$media2))
media_tidy3 <-data.frame(strategy=rep(3,length(na.omit(media_frame$media3))),
                 score=na.omit(media_frame$media3))
media_tidy4 <-data.frame(strategy=rep(4,length(na.omit(media_frame$media4))),
                 score=na.omit(media_frame$media4))
media_tidy <-rbind(media_tidy1, media_tidy2, media_tidy3, media_tidy4)
#define the boostrap function
boot_anova<-function(t1, t2, t3, t4, treat_nums) {
  size1 = length(t1)
  size2 = length(t2)
  size3 = length(t3)
  size4 = length(t4)
  null_grp1 = sample(t1 -mean(t1), size1, replace=TRUE)
  null_grp2 = sample(t2 -mean(t2), size2, replace=TRUE)
  null_grp3 = sample(t3 -mean(t3), size3, replace=TRUE)
  null_grp4 = sample(t4 -mean(t4), size4, replace=TRUE)
  null_values= c(null_grp1, null_grp2, null_grp3, null_grp4)
  alt_grp1 = sample(t1, size1, replace=TRUE)
  alt_grp2 = sample(t2, size2, replace=TRUE)
  alt_grp3 = sample(t3, size3, replace=TRUE)
  alt_grp4 = sample(t4, size4, replace=TRUE)
  alt_values= c(alt_grp1, alt_grp2, alt_grp3, alt_grp4)
  return(c(oneway.test(null_values~ treat_nums, var.equal=TRUE)$statistic,
           oneway.test(alt_values~ treat_nums, var.equal=TRUE)$statistic))
}

#compute the F-stastistic
```

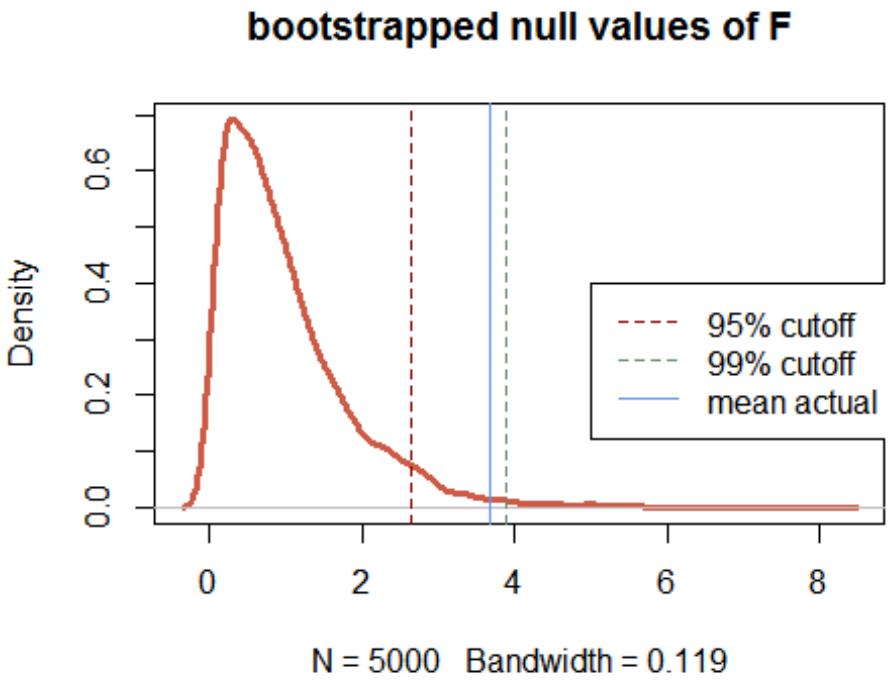
```
set.seed(42)
score1 = media_tidy$score[media_tidy$strategy==1]
score2 = media_tidy$score[media_tidy$strategy==2]
score3 = media_tidy$score[media_tidy$strategy==3]
score4 = media_tidy$score[media_tidy$strategy==4]
strategies = media_tidy$strategy
f_values<-replicate(5000, boot_anova(score1, score2, score3, score4,str
ategies))
f_nulls<-f_values[1,]
f_alts<-f_values[2,]

mean(f_nulls)

## [1] 0.99976

mean(f_alts)

## [1] 3.68306
```

b. According to the bootstrapped null values of F, What are the cutoff values for 95% and 99% confidence?

```
quantile(f_nulls, 0.95)

##      95%
## 2.647299

quantile(f_nulls, 0.99)

##      99%
## 3.886698
```

c. Show the distribution of bootstrapped null values of F, the 95% and 99% cutoff values of F (according to the bootstrap), and also the mean actual F-statistic.

```
plot(density(f_nulls), col='coral3', lwd=3,main="bootstrapped null valu
es of F")
abline(v=quantile(f_nulls, 0.95), col = "darkred",lty="dashed")
abline(v=quantile(f_nulls, 0.99), col="darkseagreen4",lty="dashed")
abline(v=mean(f_alts),col='cornflowerblue')
legend(5,0.4,c("95% cutoff","99% cutoff","mean actual F-statistic"),lty
 = c(2,2,1),
       col =c("darkred","darkseagreen4","cornflowerblue") )
```



d. According to the bootstrap, do the four types of media produce the same mean intention to share, at 95% confidence? How about at 99% confidence?

In terms of the F-statistic from the bootstrap, we could find out that the mean actual F-statistic is larger than 95% cutoff and smaller than 99% cutoff of F. Therefore, I think we could reject the null under 95% CI and don't reject under 99% CI.