HW7

106022113

```
media1 <- read.csv("pls-media/pls-media1.csv")
media2 <- read.csv("pls-media/pls-media2.csv")
media3 <- read.csv("pls-media/pls-media3.csv")
media4 <- read.csv("pls-media/pls-media4.csv")
```

Question 1 Describe and Visualize the Data

a. What are the means for each media type?

```
paste("Mean for media1: ",mean(media1$INTEND.0))
```

```
## [1] "Mean for media1:  4.80952380952381"
```

```
paste("Mean for media2: ",mean(media2$INTEND.0))
```

```
## [1] "Mean for media2:  3.94736842105263"
```

```
paste("Mean for media3: ",mean(media3$INTEND.0))
```

```
## [1] "Mean for media3:  4.725"
```

```
paste("Mean for media4: ",mean(media4$INTEND.0))
```

```
## [1] "Mean for media4:  4.89130434782609"
```

b. Visualize the distribution and mean

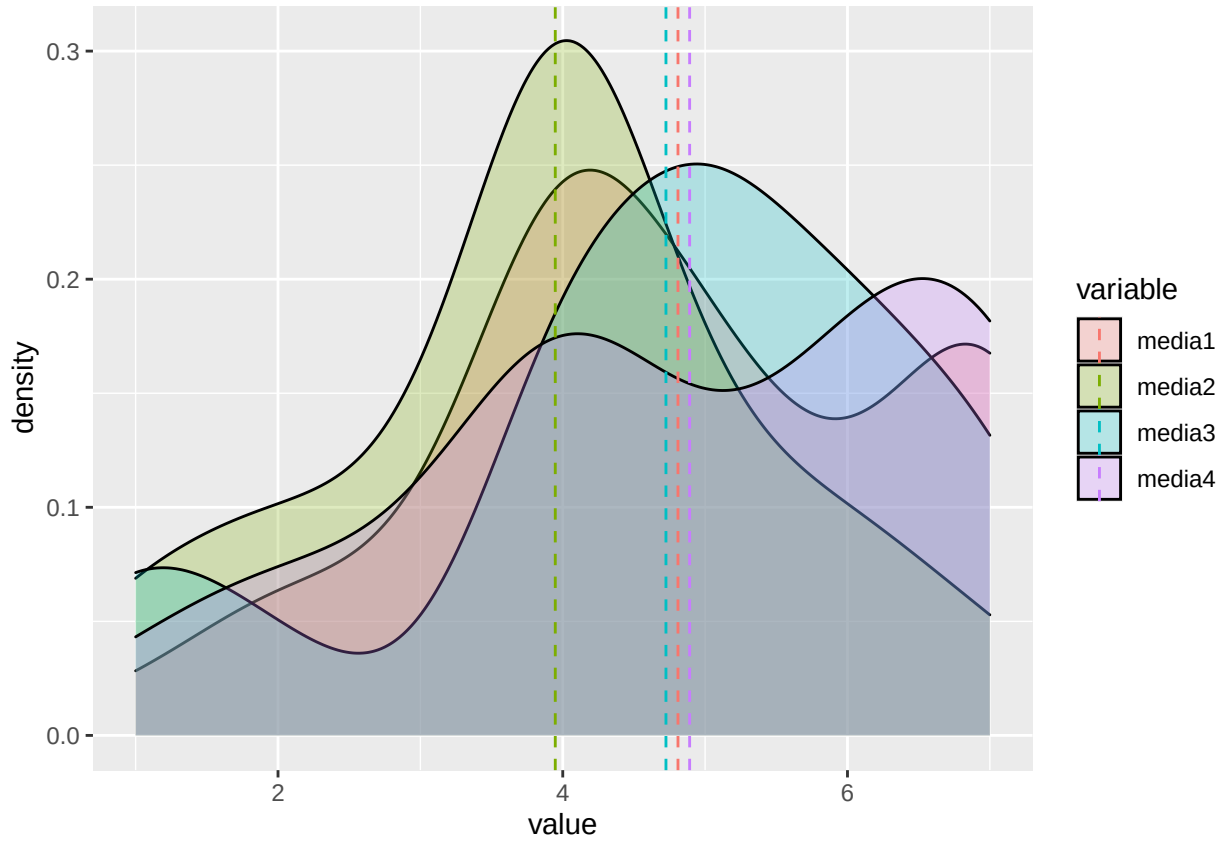
```
data_list <- list(media1$INTEND.0,media2$INTEND.0,media3$INTEND.0,media4$INTEND.0)
df <- as.data.frame(sapply(data_list, '[', seq(max(lengths(data_list)))))
colnames(df) <- c("media1","media2","media3","media4")
library(ggplot2);library(reshape2);library(plyr)
data <- melt(df)
```

Density Plots

```
## No id variables; using all as measure variables
```

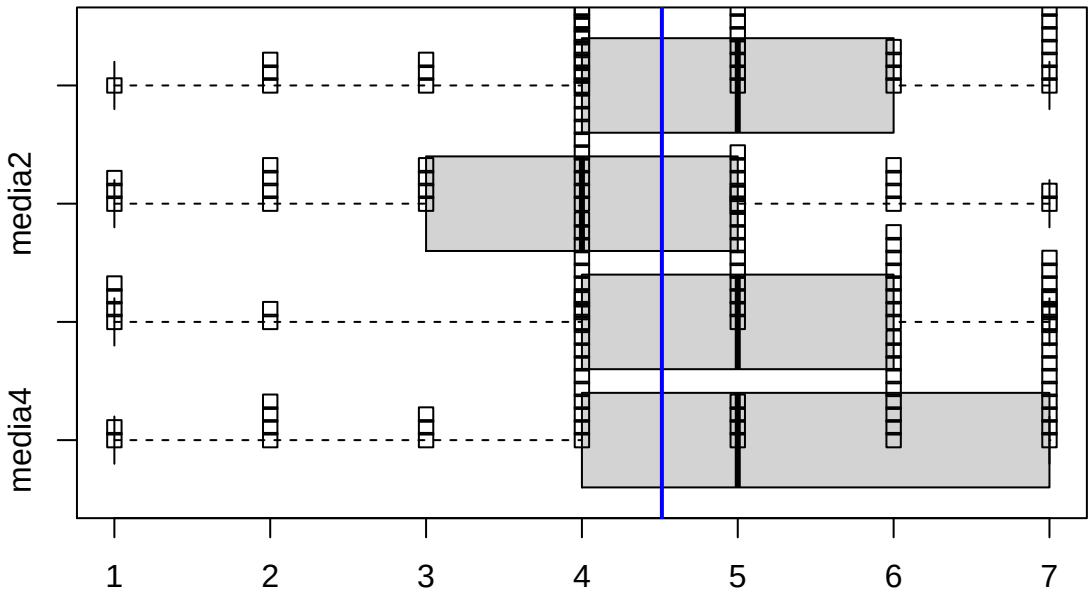
```
mu <- ddply(data, .(variable), summarise, mean=mean(value,na.rm = TRUE))
ggplot(data,aes(x=value,fill = variable)) + geom_density(alpha=0.25) + geom_vline(data = mu, aes(xinter
```

```
## Warning: Removed 18 rows containing non-finite values (stat_density).
```



Comments: Cannot get anything sensible using the form of plotting, lets try boxplots!

```
boxplot(rev(df), horizontal=TRUE)
stripchart(rev(df),method="stack", add=TRUE)
abline(v=mean(sapply(na.omit(df), mean)),col = "blue",lwd = 2)
```



Boxplots
Comments: Yes! There is actually something going on here, described in (c)!

c. Based on visualization, does different types of media differ in sharing?

ANSWER: Using boxplots to plot each media, we can see that the mean of all four media types, which is the blue line, penetrates through all the boxes! So it seems that the medias doesn't make a difference.

Question 2 Traditional one-way ANOVA

a. State null and alternative hypothesis across four groups in ANOVA

ANSWER: $H_{Null} : \mu_1 = \mu_2 = \mu_3 = \mu_4$(corresponding to different media) $H_{Alt} : means\ different$

b. Produce f-statistic

```
each_mean <- sapply(na.omit(df), mean)
t_mean <- mean(each_mean)
sstr <- 0
for (media in colnames(df)){
  sstr <- sstr + dim(na.omit(df[media]))[1]*((each_mean[media]-t_mean)^2)
}
mstr <- sstr/(4-1)
four_var <- sapply(na.omit(df), var)
```

```r
sse <- 0
N <- 0
for (media in colnames(df)){
  sse = sse + (dim(na.omit(df[media]))[1]-1)*(four_var[media])
  N = N+dim(na.omit(df[media]))[1]
}
mse <- sse/(N-4)
paste("F value: ", mstr/mse)
```

```
## [1] "F value:  2.01116515476246"
```

c. Cut-off values of F for 95% and 99% confidence according the the null distribution of F

```r
q_95 <- qf(p=0.95,df1 = 3, df2 = N-4)
paste("95% confidence: ",q_95)
```

```
## [1] "95% confidence:  2.66040556386687"
```

```r
q_99 <- qf(p=0.99,df1 = 3, df2 = N-4)
paste("99% confidence: ",q_99)
```

```
## [1] "99% confidence:  3.90480746374924"
```

d. Do the medias produce same mean at 95% and 99% confidence?

ANSWER : The f score is lower than both 95% and 99% confidence, hence we do not have enough evidence to reject the null hypothesis, the four types may produce the same mean!

e. Are the classic requirements of one-way ANOVA met?

First Requirement —> Each response variable should be normally distributed. By the density plots above, the distributions aren't normal like. We can illustrate the normality with QQ plot here. Requirement not satisfied.

```r
four_var
```

Second Requirement —> Variance of the response variables should be the same.

```
##   media1   media2   media3   media4
## 2.663585 2.321479 3.096017 3.359886
```

Second requirement not satisfied :(

Third Requirement —> Observations should be independent. However, these medias share some similiar features, so cannot be fully independent. —> Requirements not satisfied!!

Question 3 Bootstrapping ANOVA

a. Bootstrap the null and the alternative values of the F-statistic

```
set.seed(42)
boot_anova<-function(t1, t2, t3, t4, treat_nums) {
  null_grp1 = sample(t1 -mean(t1), length(t1), replace=TRUE)
  null_grp2 = sample(t2 -mean(t2), length(t2),replace=TRUE)
  null_grp3 = sample(t3 -mean(t3), length(t3),replace=TRUE)
  null_grp4 = sample(t4 -mean(t4), length(t4),replace=TRUE)
  null_values= c(null_grp1, null_grp2, null_grp3, null_grp4)
  alt_grp1 = sample(t1, replace=TRUE)
  alt_grp2 = sample(t2, replace=TRUE)
  alt_grp3 = sample(t3, replace=TRUE)
  alt_grp4 = sample(t4, 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))
}
med1 <- data.frame(strategy = rep(1, length(media1$INTEND.0)),score = media1$INTEND.0)
med2 <- data.frame(strategy = rep(2, length(media2$INTEND.0)), score = media2$INTEND.0)
med3 <- data.frame(strategy = rep(3, length(media3$INTEND.0)), score = media3$INTEND.0)
med4 <- data.frame(strategy = rep(4, length(media4$INTEND.0)), score = media4$INTEND.0)
meds <- rbind(med1,med2,med3,med4)
score1 <- meds$score[meds$strategy==1]
score2 <- meds$score[meds$strategy==2]
score3 <- meds$score[meds$strategy==3]
score4 <- meds$score[meds$strategy==4]
strategies <- meds$strategy
f_values <- replicate(5000,boot_anova(score1,score2,score3,score4,strategies))
f_nulls<-f_values[1,]
f_alts<-f_values[2,]
paste("F null values: ",mean(f_nulls))
```

```
## [1] "F null values:  1.01107337743971"
```

```
paste("F alternative values: ", mean(f_alts))
```

```
## [1] "F alternative values:  3.72545760669289"
```

b. Cut off values at 95% and 99% confidence ?

```
q1_95 <- quantile(f_nulls,0.95)
paste("Cutoff at 95%: ",q1_95)
```

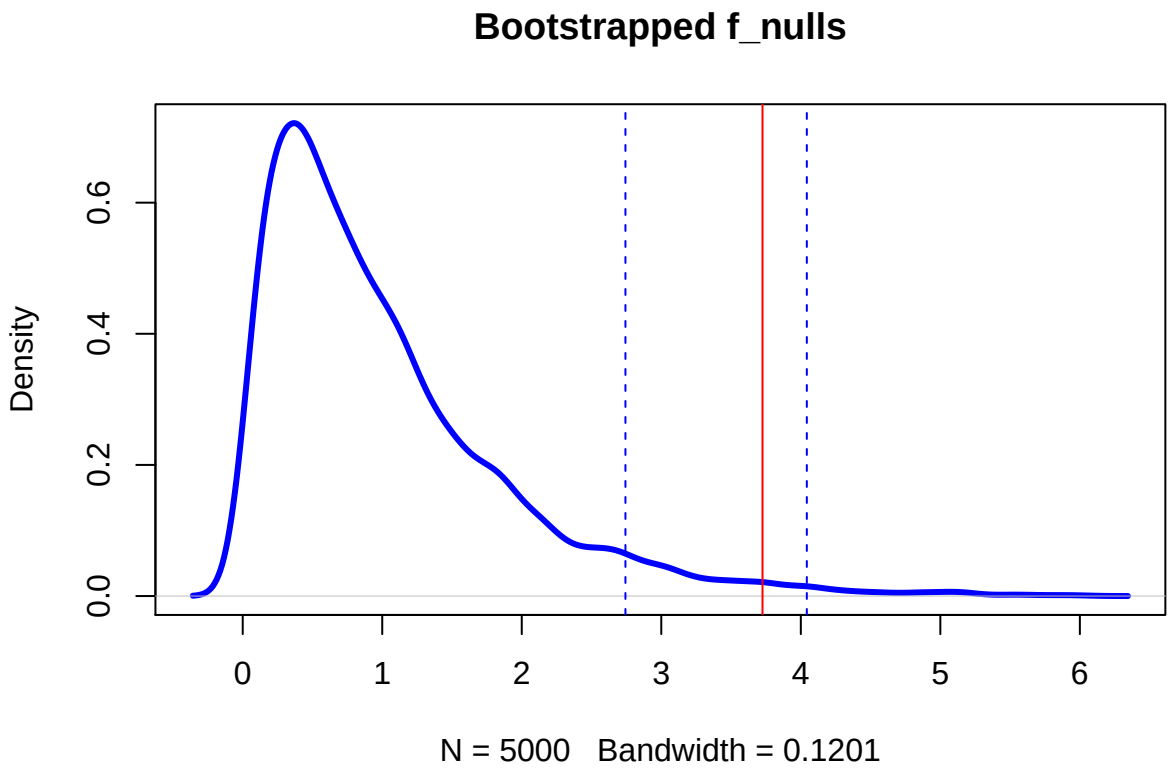
```
## [1] "Cutoff at 95%:  2.74306313151254"
```

```r
q1_99 <- quantile(f_nulls,0.99)
paste("Cutoff at 99%: ",q1_99)
```

```
## [1] "Cutoff at 99%:  4.0430658369394"
```

c. Show the distribution of f_nulls and the 95%, 99%, f_alternative stats

```r
plot(density(f_nulls), col = 'blue',lwd = 3, main = 'Bootstrapped f_nulls')
abline(v=q1_95,lty ='dashed', col = 'blue')
abline(v=q1_99,lty ='dashed', col = 'blue')
abline(v = mean(f_alts),col = 'red')
```



d. Do the four types of medias produce the intention to share at 95% and 99% confidence

ANSWER: The mean statistics of the f-value is larger than the 95% cutoff value but smaller than the 99% cutoff value. Hence, we have enough evidence to reject the null hypothesis below 95% confidence interval, but we are not able to acquire enought evidence to reject the null hypothesis at 99% confidence.