Analysis of Variance

Yu Han Huang

The researcher's data can be found in four CSV files named: pls-media[1-4].csv

> A health researcher, investigating how health information spreads through word-of-mouth, has prepared some informative about avoiding stomach aches. But she is unsure which media format to use, or avoid, in order to encourage people to share the information. So she prepares the same information in four alternative media formats: (1) video [animation + audio]
> (2) video [pictures + audio]
> (3) webpage [pictures + text]
> (4) webpage [text only] Our researcher runs an experiment where each of these four alternative media is shown to one of four different panels of randomly assigned people. After viewing their media material, the viewers were surveyed about their thoughts, including a question (labeled INTEND.0) about their intention to share what they had seen with others: [INTEND.0]: I intend to share the information I saw with others. (answered on 7 point scale: 1=strongly disagree; 4=neutral; 7=strongly agree)

Read in the files:

```
health1 <- read.csv(file.choose(),header = TRUE)
health2 <- read.csv(file.choose(),header = TRUE)
health3 <- read.csv(file.choose(),header = TRUE)
health4 <- read.csv(file.choose(),header = TRUE)
health <- rbind(health1[,1:2],health2[,1:2],health3[,1:2],health4[,1:2])
```

> Question 1) Describe and visualize the data:

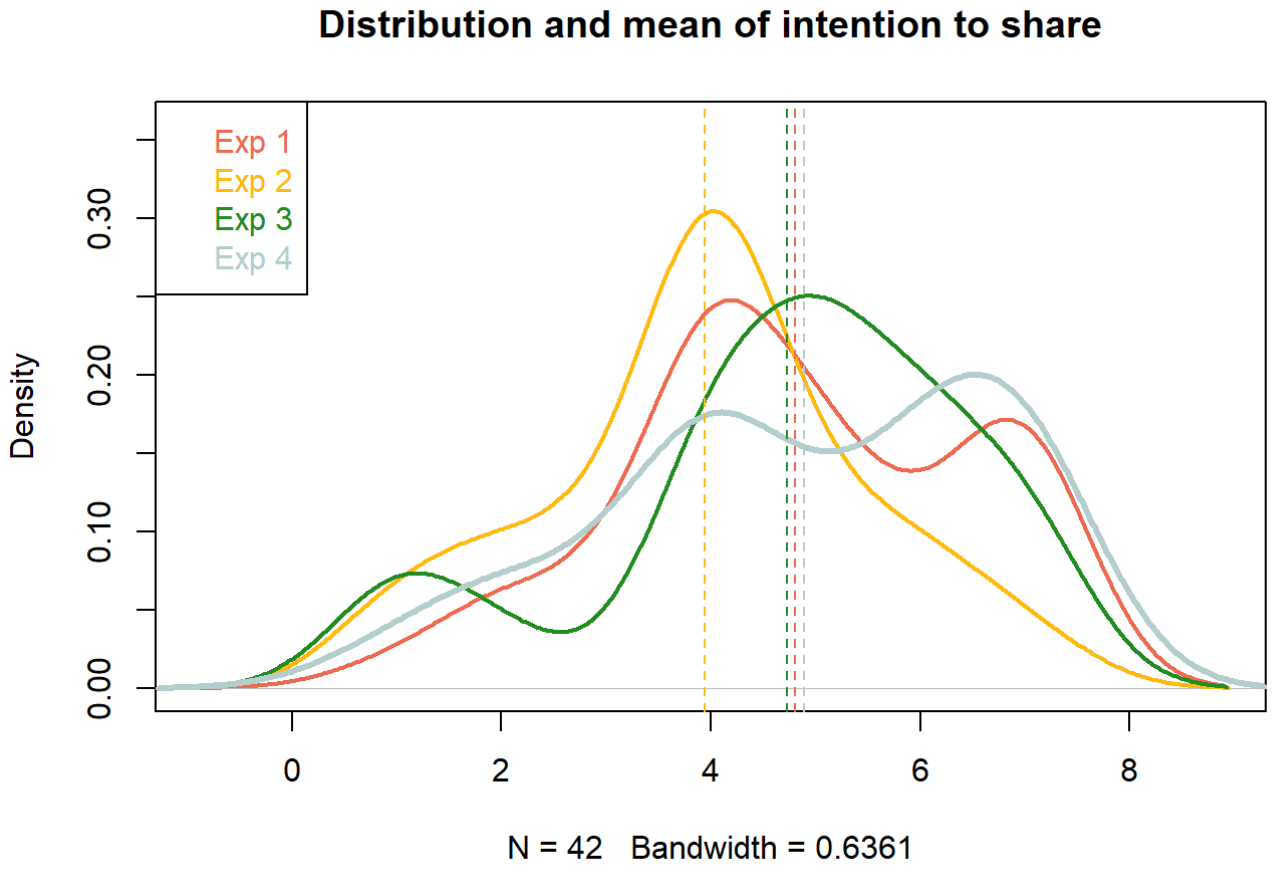
a. What are the means of viewers intentions to share (INTEND.0) for each media type? (report four means)

```
round(sapply(split(health$INTEND.0, health$media), mean),2)
```

```
##    1    2    3    4
## 4.81 3.95 4.72 4.89
```

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

```
plot(density(health1$INTEND.0), col="coral2",ylim=c(0,0.36), lwd=2, main="Distribution and mean of intention to s
hare")
lines(density(health2$INTEND.0), col="darkgoldenrod1",ylim=c(0,0.36), lwd=2)
lines(density(health3$INTEND.0), col="forestgreen",ylim=c(0,0.36), lwd=2)
lines(density(health4$INTEND.0), col="lightcyan3",ylim=c(0,0.36), lwd=3)
abline(v=mean(health1$INTEND.0), col="coral2",lty="dashed")
abline(v=mean(health2$INTEND.0), col="darkgoldenrod1",lty="dashed")
abline(v=mean(health3$INTEND.0), col="forestgreen",lty="dashed")
abline(v=mean(health4$INTEND.0), col="lightcyan3",lty="dashed")
legend("topleft", c("Exp 1", "Exp 2", "Exp 3", "Exp 4"), text.col = c("coral2", "darkgoldenrod1", "forestgreen",
"lightcyan3"))
```



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

```
"Though the mean is quite close to one another, there are some slight differences between medias, as from the dis
tribution, we may see that only health2 more closely approximated to normal distribution, while the others, for e
xample the fourth, is quite skewed. Also, only media 2's mean is further apart from the other threes."
```

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

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

```
"H0: The means of the three treatment populations are the same"
"H1: The means of the three treatments populations are not the same"
```

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

```
oneway.test(health$INTEND.0 ~ factor(health$media), var.equal=TRUE)$statistic
```

```
##        F
## 2.616669
```

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

```
qf(df1=length(unique(health$media))-1, df2=nrow(health)-1, p=c(0.95, 0.99))
```

```
## [1] 2.659384 3.902523
```

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?

```
"At 95% confidence, since our f-value exceeds cut-off value, we reject null hypothesis that the four media have t
he same mean.
However, at 99% confidence, since our f-value doesn't exceed cut-off value, we can't reject our null hypothesis t
hat the four media have the same mean."
```

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

```
round(sapply(split(health$INTEND.0, health$media), var),2)
```

```
##    1    2    3    4
## 2.69 2.32 3.08 3.30
```

```
"ANOVA requires the variance (s2) of the response variables is the same for all treatments/populations, but it is
n't met here."
```

Question 3) Let's try bootstrapping ANOVA:

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

```
#Bootstrapped ANOVA
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))
}
```

```
#Calculation
set.seed(42)
boot_f_values <- replicate(2000, boot_anova(health1$INTEND.0, health2$INTEND.0, health3$INTEND.0, health4$INTEND.
0, health$media))
boot_f_null <- boot_f_values[1,]
boot_f_alts <- boot_f_values[2,]
```

```
#Results
paste("Null F:",mean(boot_f_null),"and Alt F:",mean(boot_f_alts))
```

```
## [1] "Null F: 1.00150226688429 and Alt F: 3.69052865040187"
```

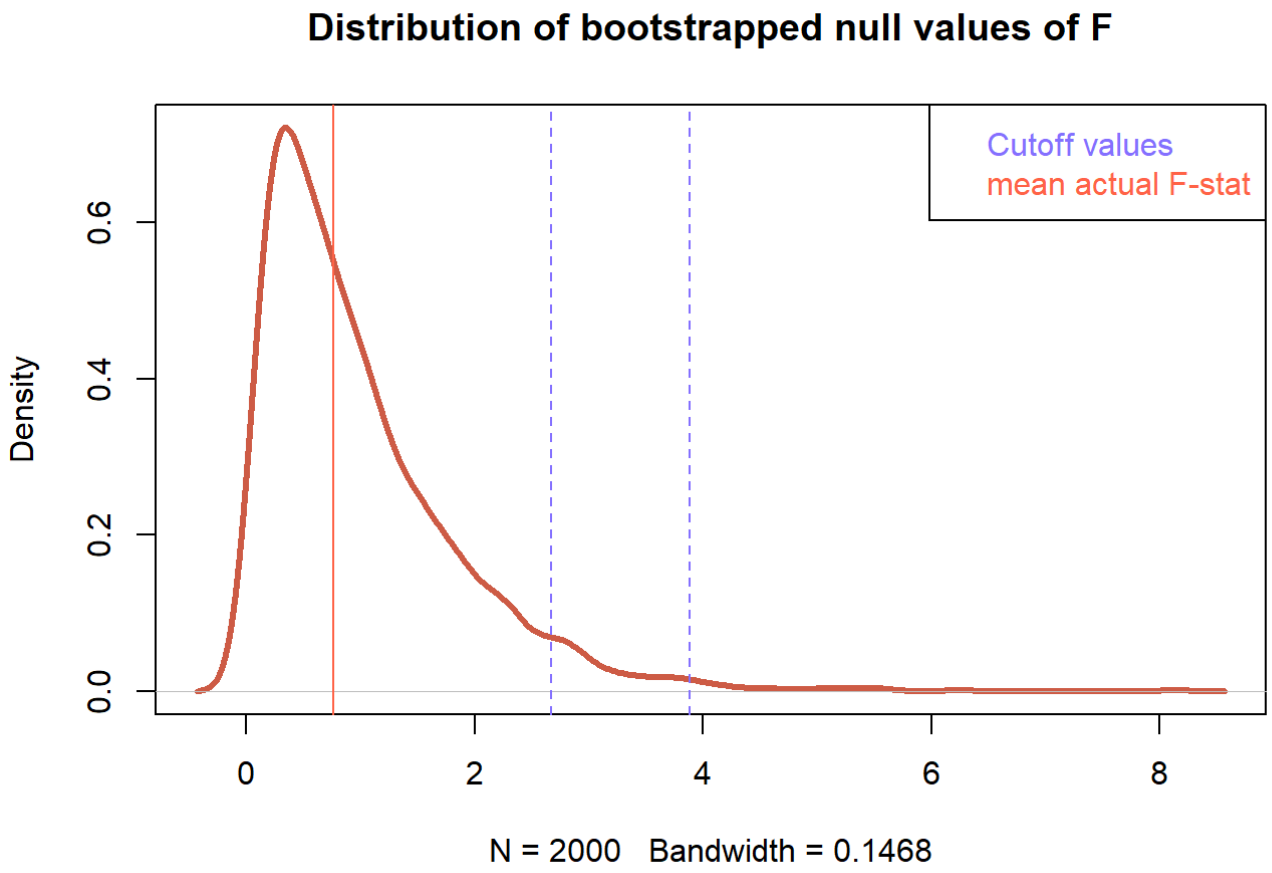
b. According to the bootstrapped null values of F, What are the cutoff values for 95% and 99% confidence?

```
paste("95% confidence:",quantile(boot_f_null, 0.95),"and 99% confidence:",quantile(boot_f_null, 0.99))
```

```
## [1] "95% confidence: 2.66927129458528 and 99% confidence: 3.88669817136418"
```

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(boot_f_null), col='coral3', lwd=3, main="Distribution of bootstrapped null values of F")
abline(v=quantile(boot_f_null, 0.95), col='lightslateblue',lty="dashed")
abline(v=quantile(boot_f_null, 0.99), col='lightslateblue',lty="dashed")
abline(v=mean(boot_f_null, 0.95), col='tomato1')
legend("topright", c("Cutoff values", "mean actual F-stat"), text.col = c("lightslateblue", "tomato1"))
```



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?

```
"They produce the same mean intention in either confidence since it lies beneath either cutoff values."
```