hw6

Bijon Setyawan Raya

4/17/2021

Data Preparation

```
# Import data sets in Linux
m1 <- read.csv("/home/johnbjohn/Documents/git_repos/bacs-hw/hw6/pls-media1.csv")
m2 <- read.csv("/home/johnbjohn/Documents/git_repos/bacs-hw/hw6/pls-media2.csv")
m3 <- read.csv("/home/johnbjohn/Documents/git_repos/bacs-hw/hw6/pls-media3.csv")
m4 <- read.csv("/home/johnbjohn/Documents/git_repos/bacs-hw/hw6/pls-media4.csv")

media <- rbind(m1,m2,m3,m4)
```

1. Let's describe and visualize the data:

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

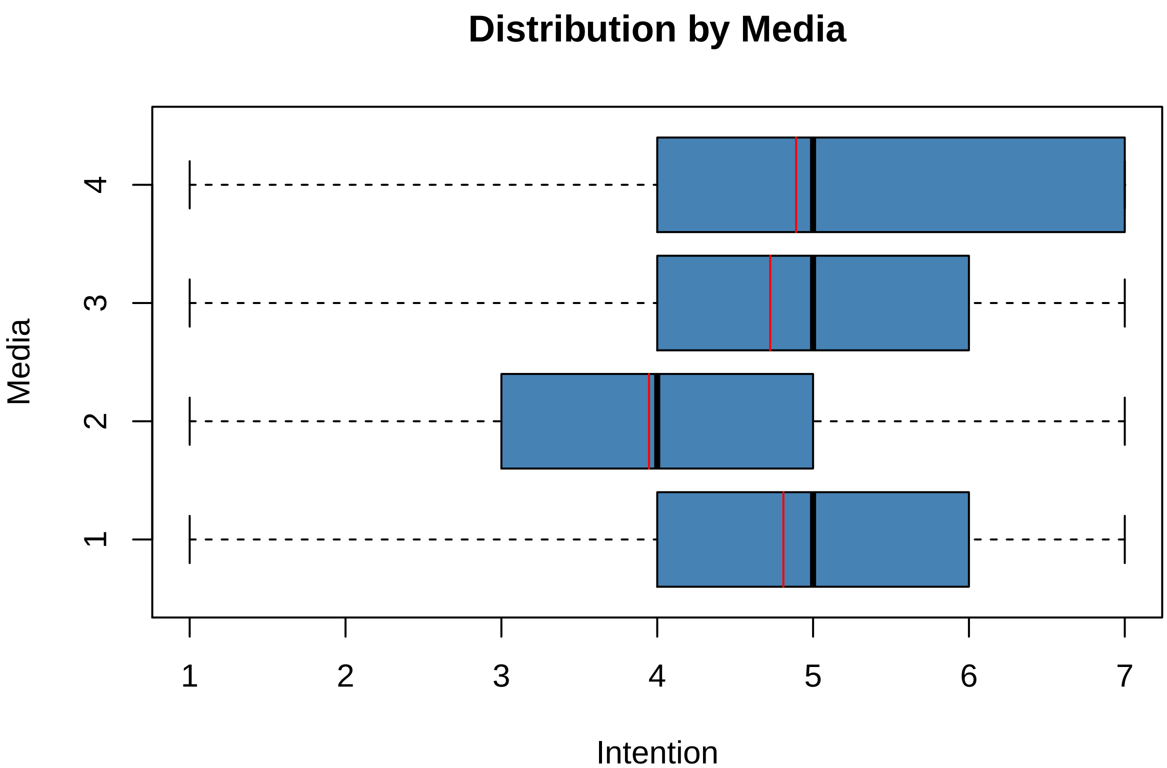
Here is the mean of `INTEND.0` in each data set: (starting from media1 all the way to media4)

1. `media1$INTEND.0` mean: 4.8095238
2. `media2$INTEND.0` mean: 3.9473684
3. `media3$INTEND.0` mean: 4.725
4. `media4$INTEND.0` mean: 4.8913043

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

```
boxplot(
  INTEND.0 ~ media,
  data=media,
  main="Distribution by Media",
  xlab="Intention",
  ylab="Media",
  col="steelblue",
  border="black",
  horizontal = TRUE
)
```

```r
x0s <- 1:4 - 0.4
x1s <- 1:4 + 0.4
y0s <- c(mean(m1$INTEND.0), mean(m2$INTEND.0), mean(m3$INTEND.0), mean(m4$INTEND.0))
segments(x0=y0s, y0=x0s, y1=x1s, col = "red")
```



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

In my opinion, the media won't affect the difference on intention to share since the `grand_mean` value is 4.6144578.

2. Let's try traditional one-way ANOVA:

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

Null hypothesis is the means of INTEND.0 in the four data sets are similar to each other. While the **Alternative hypothesis** is the means of INTEND.0 in the four data sets are not similar to each other.

b. Produce the traditional F-statistic for our test

```r
grand_mean <- mean(media$INTEND.0)
```

```r
intend0 <- tibble(
  media = c(1,2,3,4),
  instances = c(
    nrow(m1),
    nrow(m2),
    nrow(m3),
    nrow(m4)
  ),
  group_mean = c(
    mean(m1$INTEND.0),
    mean(m2$INTEND.0),
    mean(m3$INTEND.0),
    mean(m4$INTEND.0)
  ),
  var = c(
    var(m1$INTEND.0),
    var(m2$INTEND.0),
    var(m3$INTEND.0),
    var(m4$INTEND.0)
  ),
  sstr = c(
    nrow(m1) * (mean(m1$INTEND.0) - grand_mean)^2,
    nrow(m2) * (mean(m2$INTEND.0) - grand_mean)^2,
    nrow(m3) * (mean(m3$INTEND.0) - grand_mean)^2,
    nrow(m4) * (mean(m4$INTEND.0) - grand_mean)^2
  ),
  sse = c(
    (nrow(m1) - 1) * sd(m1$INTEND.0)^2,
    (nrow(m2) - 1) * sd(m2$INTEND.0)^2,
    (nrow(m3) - 1) * sd(m3$INTEND.0)^2,
    (nrow(m4) - 1) * sd(m4$INTEND.0)^2
  ),
)

intend0
```

```
## # A tibble: 4 x 6
##   media instances group_mean   var   sstr   sse
##   <dbl>     <int>      <dbl> <dbl>  <dbl> <dbl>
## 1     1        42       4.81  2.69  1.60  110.
## 2     2        38       3.95  2.32 16.9    85.9
## 3     3        40       4.72  3.08  0.489 120.
## 4     4        46       4.89  3.30  3.53  148.
```

```r
n <- 4
sstr <- sum(intend0$sstr)
mstr <- sstr/(n-1)
sse <- sum(intend0$sse)
mse <- sse/(nrow(media) - n)
f <- mstr / mse
f
```

```
## [1] 2.616669
```

The F-statistics value is 2.6166687.

c. What are the cut-off values of F for 95% and 99% confidence according the the null distribution of F?

The cut-off value of F for 95% is 2.6593838, and the cut-off value of F for 99% is 3.9025235.

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?

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

3. Let's try bootstrapping ANOVA

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

```
boot_anova <- function(t1, t2, t3, t4, treat_nums) {
  null_sample <- c(
    sample(t1 - mean(t1), length(t1), replace = TRUE),
    sample(t2 - mean(t2), length(t2), replace = TRUE),
    sample(t3 - mean(t3), length(t3), replace = TRUE),
    sample(t4 - mean(t4), length(t4), replace = TRUE)
  )

  alt_sample <- c(
    sample(t1, length(t1), replace = TRUE),
    sample(t2, length(t2), replace = TRUE),
    sample(t3, length(t3), replace = TRUE),
    sample(t4, length(t4), replace = TRUE)
  )

  c(
    oneway.test(null_sample ~ treat_nums, var.equal = TRUE)$statistic,
    oneway.test(alt_sample ~ treat_nums, var.equal = TRUE)$statistic
  )
}

f_values <- replicate(5000, boot_anova(m1$INTEND.0, m2$INTEND.0, m3$INTEND.0, m4$INTEND.0, media$media)

f_nulls <- f_values[1,]
f_alts <- f_values[2,]
```

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

```
quantile(f_nulls, 0.95)
```

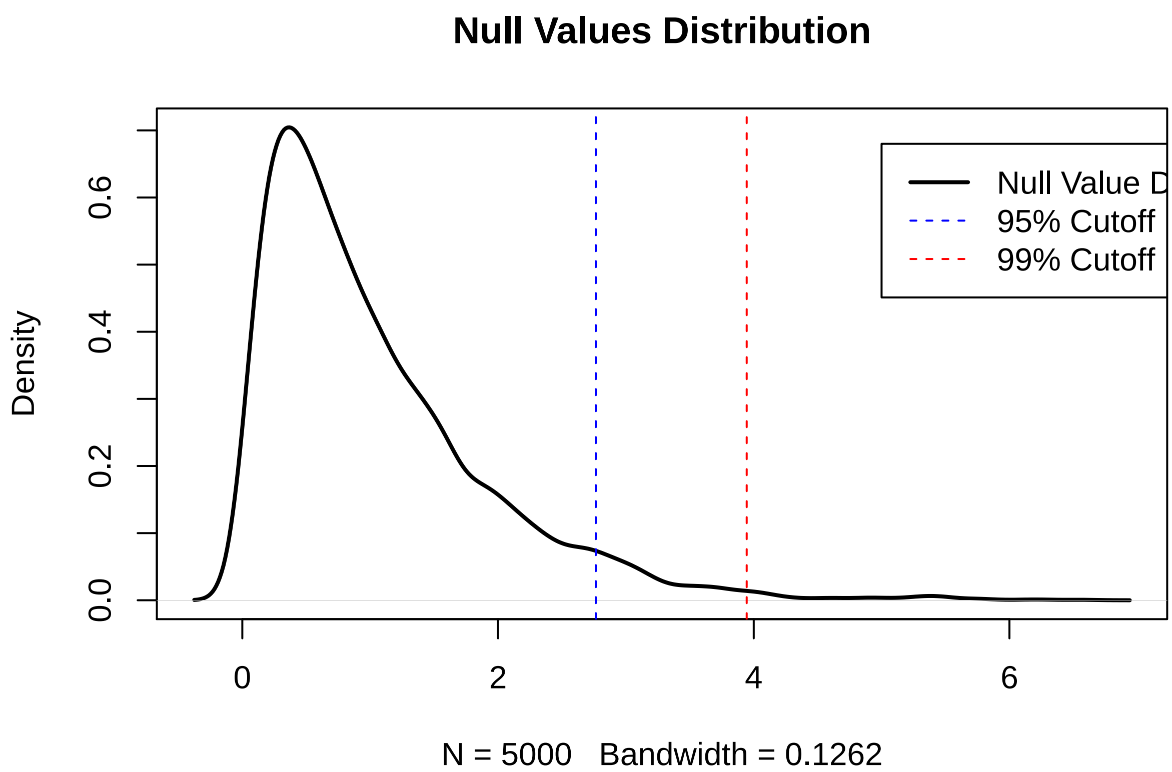
```
##      95%
## 2.76514
```

```r
quantile(f_nulls, 0.99)
```

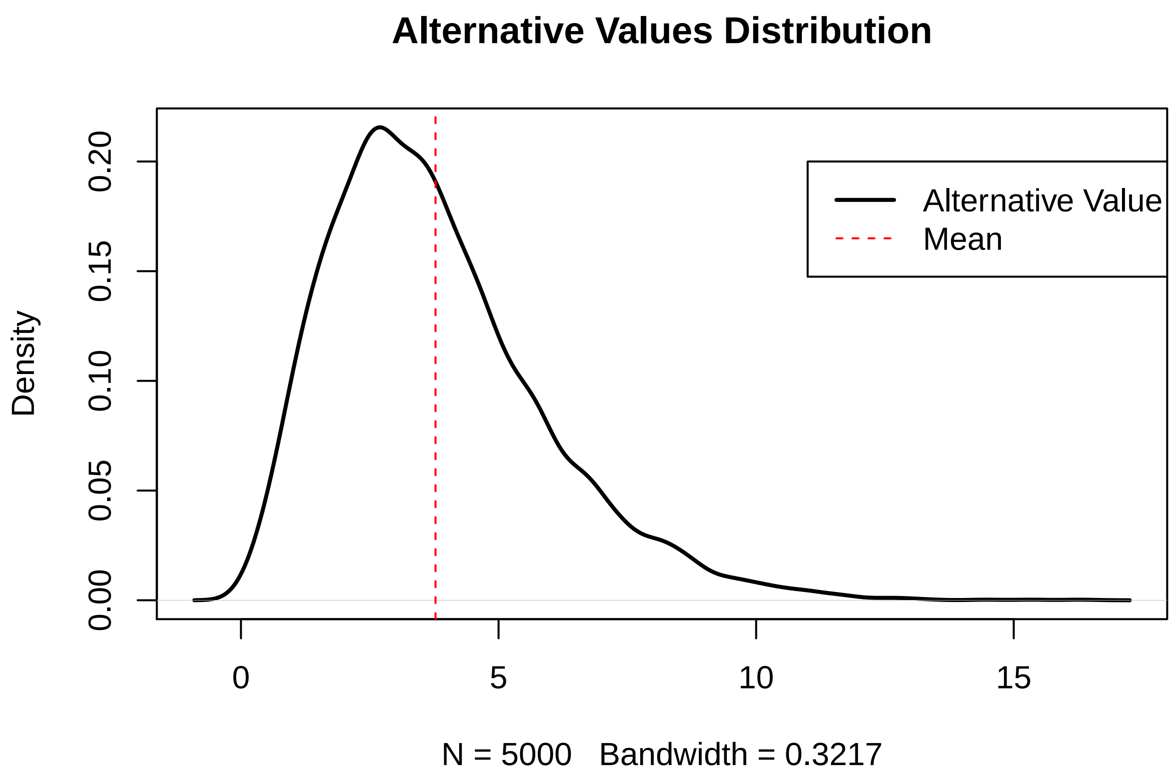
```
##      99%
## 3.944959
```

c. Visualize the distribution of bootstrapped null values of F, the 95% and 99% cutoff values of F, and also the original F-value from bootstrapped alternative values.

```r
plot(
  density(f_nulls),
  col = "black",
  lwd = 2,
  main = "Null Values Distribution"
)
abline(v = quantile(f_nulls, 0.95), col="blue", lty = "dashed")
abline(v = quantile(f_nulls, 0.99), col="red",lty = "dashed")
legend(
  5,
  0.68,
  c("Null Value Distribution", "95% Cutoff", "99% Cutoff"),
  lwd = c(2,1,1),
  lty = c("solid", "dashed", "dashed"),
  col = c("black", "blue", "red")
)
```




```r
plot(
  density(f_alts),
  col = "black",
  lwd = 2,
  main = "Alternative Values Distribution"
)
abline(v = mean(f_alts), col="red", lty = "dashed")
legend(
  11,
  0.20,
  c("Alternative Value", "Mean"),
  lwd = c(2,1),
  lty = c("solid", "dashed"),
  col = c("black", "red")
)
```


Alternative Values Distribution
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
Alternative Value
Mean
N = 5000 Bandwidth = 0.3217