Business Analytics using Statistical Modeling
Assignment 6

Load the researcher's data

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.2.5
```

```
health_media1 <- fread('../6-health-media1.csv')
health_media2 <- fread('../6-health-media2.csv')
health_media3 <- fread('../6-health-media3.csv')
health_media4 <- fread('../6-health-media4.csv')
```

Question 1

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

```
mean(health_media1$INTEND.0)
```

```
## [1] 4.809524
```

```
mean(health_media2$INTEND.0)
```

```
## [1] 3.947368
```

```
mean(health_media3$INTEND.0)
```

```
## [1] 4.725
```

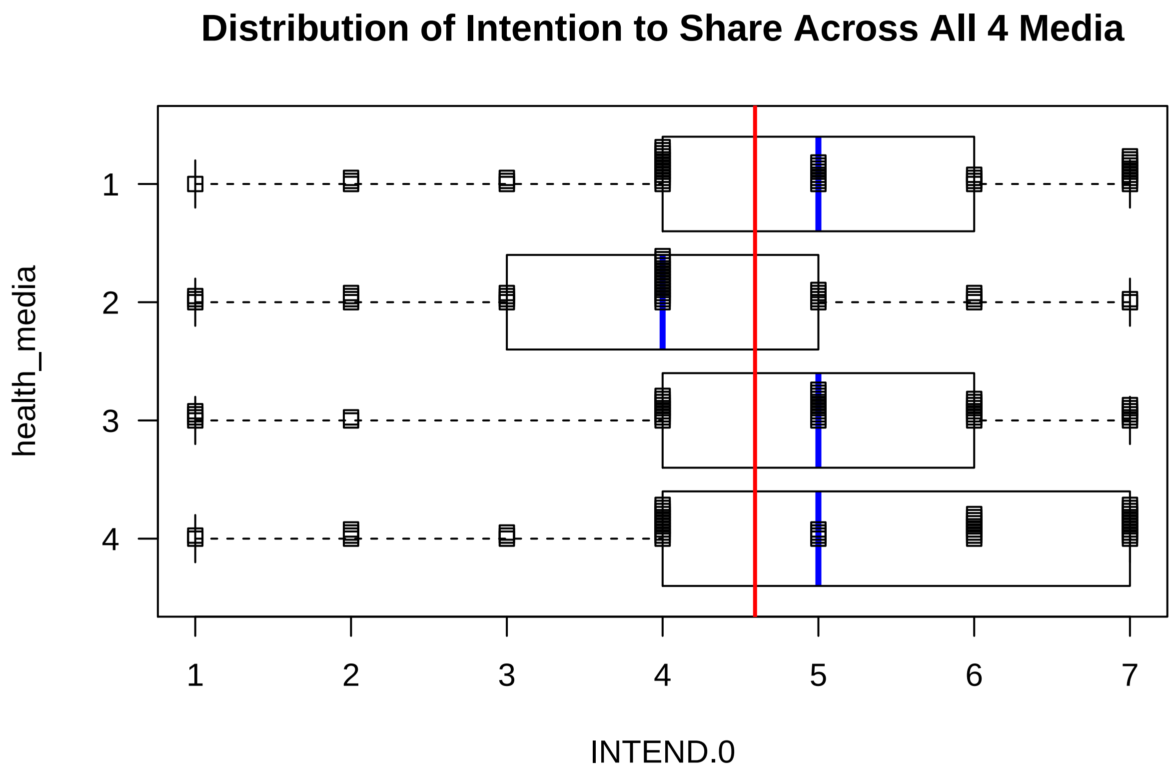
```
mean(health_media4$INTEND.0)
```

```
## [1] 4.891304
```

b. Visualize the distribution and mean of intention to share, across all 4 media.

```r
intend0 <- list(health_media1$INTEND.0, health_media2$INTEND.0, health_media3$INTEND.0,
                health_media4$INTEND.0)
intend0_nrow <- max(sapply(intend0, length))
intend0 <- sapply(intend0, function(x) c(x, rep(NA, intend0_nrow - length(x))))
intend0 <- as.data.frame(intend0)
colnames(intend0) <- c('health_media1', 'health_media2', 'health_media3', 'health_media4')

boxplot(rev(intend0), horizontal = TRUE, xlab = 'INTEND.0', ylab = 'health_media',
        las = 1, names = c('4', '3', '2', '1'), medcol = 'blue',
        main = 'Distribution of Intention to Share Across All 4 Media')
stripchart(rev(intend0), method = 'stack', add = TRUE, offset = 0.08)
abline(v = mean(sapply(intend0, mean, na.rm = TRUE)), col = 'red', lwd = 2)
```



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

The medians and boxplot distributions look very close to each other, so looks like the type of media doesn't make much difference on intention to share.

Question 2

a. State the null and alternative hypotheses when comparing INTEND.0 across 4 groups using ANOVA.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. H_a: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

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

```r
intend0_row_wise <- na.omit(melt(intend0))
```

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

```r
colnames(intend0_row_wise) <- c('health_media', 'INTEND.0')
oneway_test <- oneway.test(intend0_row_wise$INTEND.0 ~ intend0_row_wise$health_media,
                           var.equal = TRUE)
oneway_test
```

```
##
##  One-way analysis of means
##
## data:  intend0_row_wise$INTEND.0 and intend0_row_wise$health_media
## F = 2.6167, num df = 3, denom df = 162, p-value = 0.05289
```

```r
oneway_test$statistic
```

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

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

```r
cut_off_95 <- qf(p = 0.95, df1 = oneway_test$parameter[1], df2 = oneway_test$parameter[2])
cut_off_95
```

```
## [1] 2.660406
```

```r
cut_off_99 <- qf(p = 0.99, df1 = oneway_test$parameter[1], df2 = oneway_test$parameter[2])
cut_off_99
```

```
## [1] 3.904807
```

d. According to the traditional ANOVA, do the 4 types of media produce the same mean intention to share at 95% confidence? How about at 99% confidence?

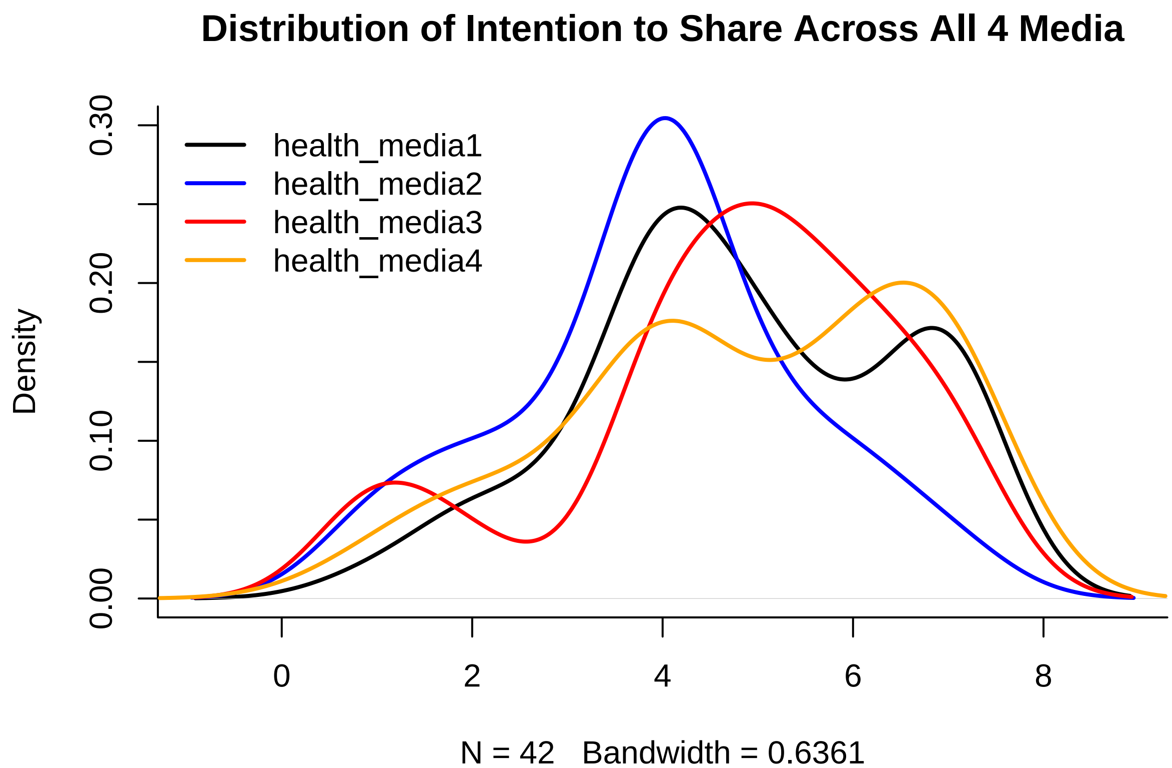
In both cases (95% and 99% confidence), the F-statistic is less than the corresponding cut-off values, so we do not reject the null hypothesis.

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

Requirements for ANOVA:

1. Each treatment/population's response variable is normally distributed

```
plot(density(intend0$health_media1, na.rm = TRUE), lwd = 2, bty = 'l', ylim = c(0, 0.3),
     main = 'Distribution of Intention to Share Across All 4 Media')
lines(density(intend0$health_media2, na.rm = TRUE), col = 'blue', lwd = 2)
lines(density(intend0$health_media3, na.rm = TRUE), col = 'red', lwd = 2)
lines(density(intend0$health_media4, na.rm = TRUE), col = 'orange', lwd = 2)
legend('topleft', lwd = c(2, 2, 2, 2), col = c('black', 'blue', 'red', 'orange'),
       bty = 'n',
       legend = c('health_media1', 'health_media2', 'health_media3', 'health_media4'))
```



They are not normally distributed.

.

2. The variance of the response variables is the same for all treatments / populations

```
sapply(intend0, var, na.rm = TRUE)
```

```
## health_media1 health_media2 health_media3 health_media4
##      2.694541      2.321479      3.076282      3.299034
```

They don't have the same variances.

3. The observations are independent: the response variable are not related

Those 4 alternative media is shown to one of four different panels of randomly assigned people. So we could say this one requirement is met.

Question 3

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

```
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))
}
intend0_1 <- intend0_row_wise$INTEND.0[intend0_row_wise$health_media == 'health_media1']
intend0_2 <- intend0_row_wise$INTEND.0[intend0_row_wise$health_media == 'health_media2']
intend0_3 <- intend0_row_wise$INTEND.0[intend0_row_wise$health_media == 'health_media3']
intend0_4 <- intend0_row_wise$INTEND.0[intend0_row_wise$health_media == 'health_media4']
health_medias <- intend0_row_wise$health_media

set.seed(1)
f_values <- replicate(10000, boot_anova(intend0_1, intend0_2, intend0_3, intend0_4,
                                        health_medias))
f_nulls <- f_values[1, ]
f_alts <- f_values[2, ]
```

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

```r
boot_cut_off_95 <- quantile(f_nulls, 0.95)
boot_cut_off_95
```

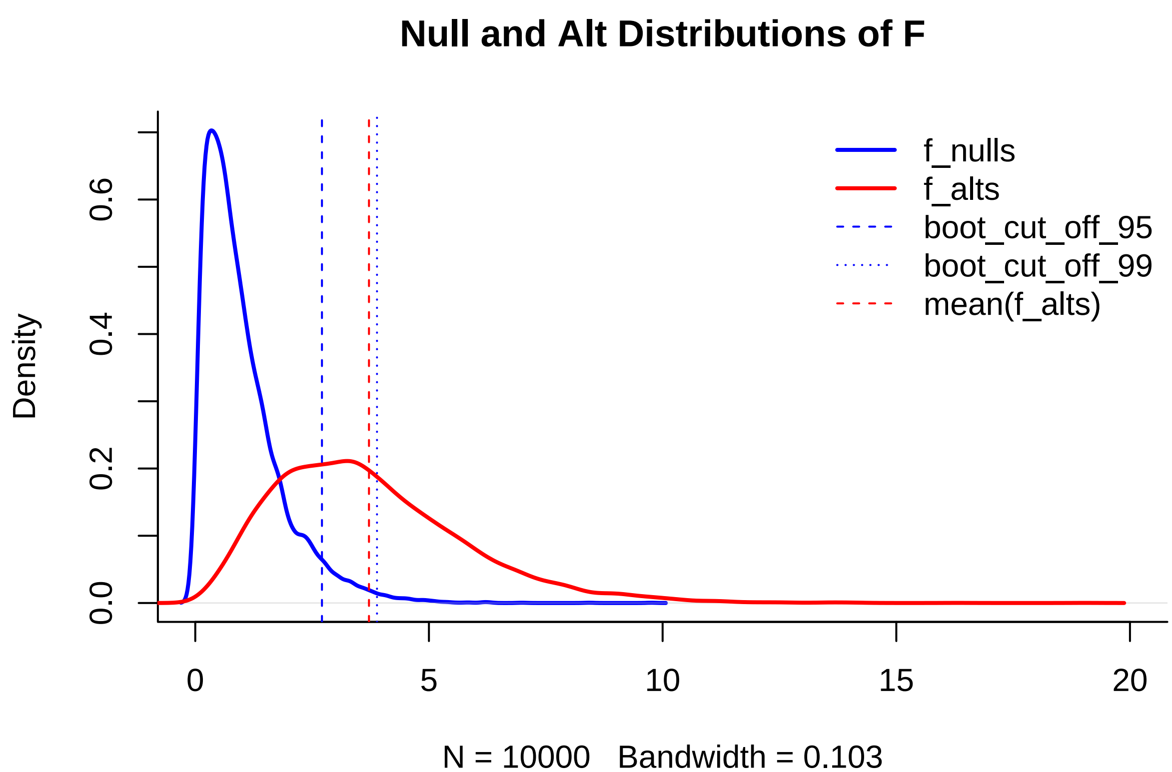
```
##      95%
## 2.710922
```

```r
boot_cut_off_99 <- quantile(f_nulls, 0.99)
boot_cut_off_99
```

```
##      99%
## 3.888907
```

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 = 'blue', lwd = 2, bty = 'l', xlim = c(0, 20),
     main = 'Null and Alt Distributions of F')
lines(density(f_alts), col = 'red', lwd = 2)
abline(v = boot_cut_off_95, lty = 2, col = 'blue')
abline(v = boot_cut_off_99, lty = 3, col = 'blue')
abline(v = mean(f_alts), lty = 2, col = 'red')
legend('topright', bty = 'n', lty = c(1, 1, 2, 3, 2), lwd = c(2, 2, 1, 1, 1),
       col = c('blue', 'red', 'blue', 'blue', 'red'),
       legend =
         c('f_nulls', 'f_alts', 'boot_cut_off_95', 'boot_cut_off_99', 'mean(f_alts)'))
```



d. According to the bootstrap, do the 4 types of media produce the same mean intention to share at 95% confidence? How about 99% confidence?

At 95% confidence, since the `mean(f_alts)` is greater than the `boot_cut_off_95`, we reject the null hypothesis; meaning that the 4 types of media do not produce the same mean intention to share.

At 99% confidence, since the `mean(f_alts)` is less than the `boot_cut_off_99`, we do not reject the null hypothesis.