HW Week8

106022103

2021/4/18

- Helped by
  - `106070038`: About the part of ANOVA.

Set up

import libary

```
library(ggplot2)
library(plyr)
require(qqplotr)
```

Random seed

```
SEED <- 1234
```

Read File

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

Background Description

> A health researcher, investigating how health information spreads through word-of-mouth, has prepared some informative content about avoiding stomach aches. She is curious which media format to use, or avoid, in order to encourage people to share such health related information.

So she prepares the similar information content in four alternative media formats: + (1) video [animation + audio]: A fully animated video with audio narration + (2) video [pictures + audio]: Video of sequence of still pictures (not-animated) with audio narration + (3) webpage [pictures + text]: Static webpage with still pictures (no video) and accompanying text narration (no audio) + (4) webpage [text only]: Static webpage of text narration (no audio) but no pictures

> You may find the researcher's data in a ZIP file containing four CSV files named: pls-media[1-4].csv (note: the number in the filename corresponds to the type of media listed above)

Question 1 - describe and visualize

(a)

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

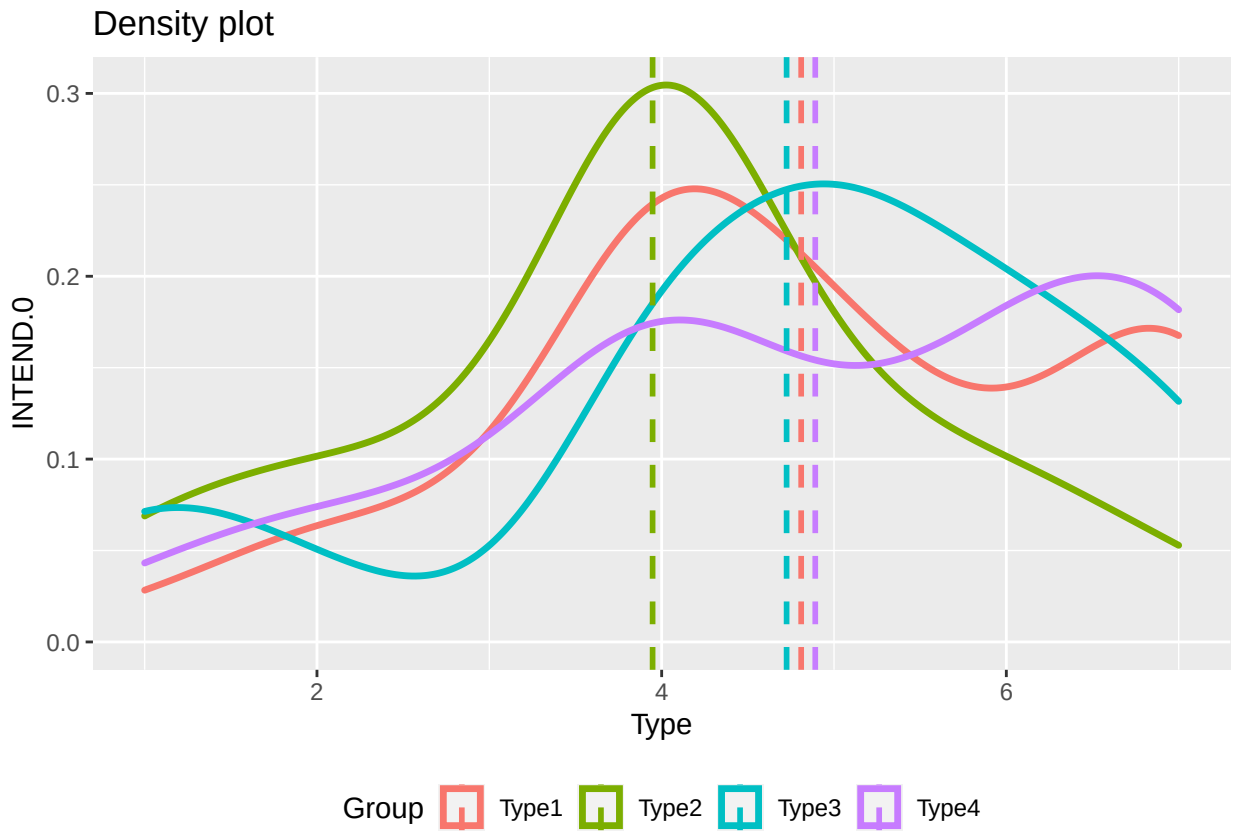
```r
df1 <- data.frame(value = media1$INTEND.0, Group = rep("Type1", length(media1$INTEND.0)))
df2 <- data.frame(value = media2$INTEND.0, Group = rep("Type2", length(media2$INTEND.0)))
df3 <- data.frame(value = media3$INTEND.0, Group = rep("Type3", length(media3$INTEND.0)))
df4 <- data.frame(value = media4$INTEND.0, Group = rep("Type4", length(media4$INTEND.0)))

sprintf("The mean of four types are: %.2f, %.2f, %.2f, %.2f",
        mean(df1$value), mean(df2$value), mean(df3$value), mean(df4$value))
```

```
## [1] "The mean of four types are: 4.81, 3.95, 4.72, 4.89"
```

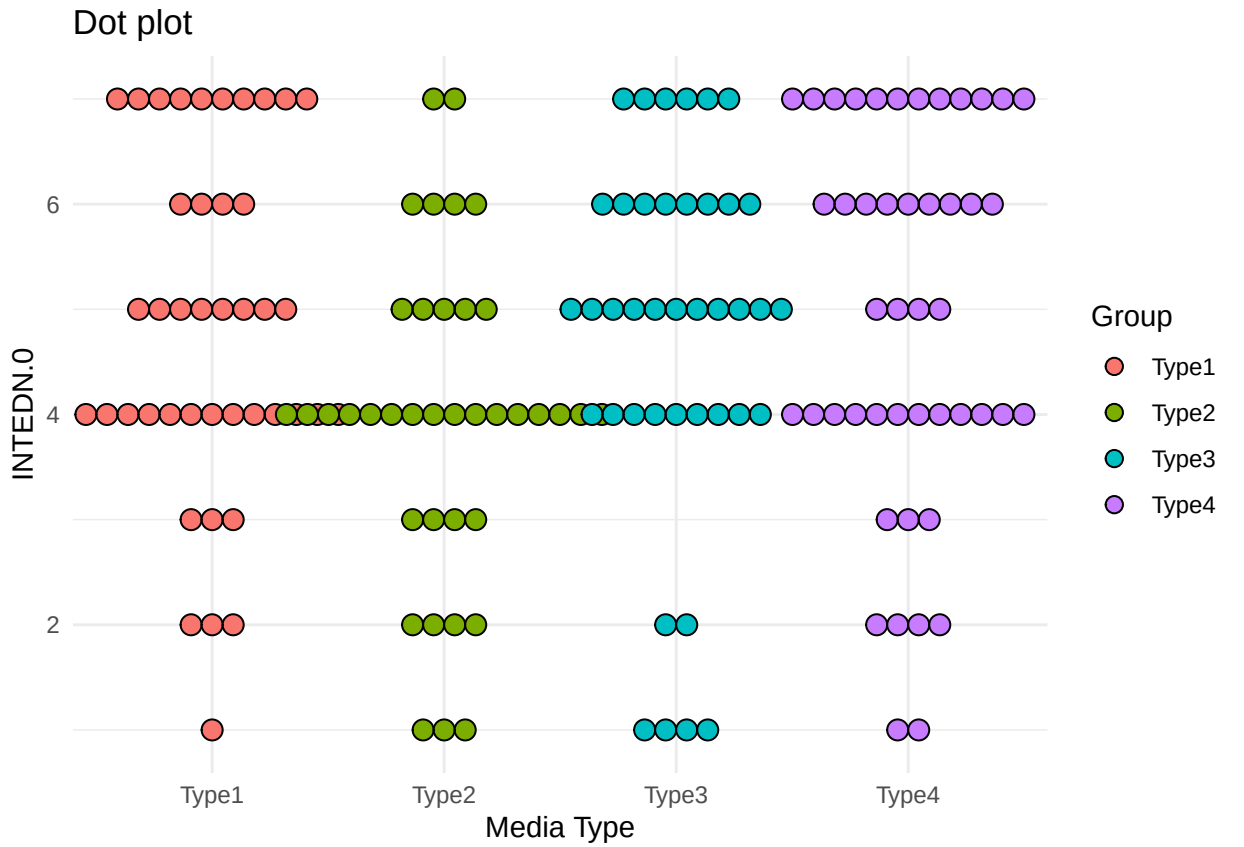
(b)

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)

```r
df <- rbind(df1,df2,df3,df4)
# get means of each group
mu <- ddply(df, "Group", summarise, grp.mean=mean(value))
# Visualize the density plot and add mean lines.
p <- ggplot(df, aes(x=value, color=Group)) +
  geom_density(lwd=1.2) +
  geom_vline(data=mu, aes(xintercept=grp.mean, color=Group),
             linetype="dashed",lwd=1) +
  labs(title="Density plot",x="Type", y = "INTEND.0") +
  theme(legend.position="bottom")
p
```



```
p <-ggplot(df, aes(x=Group, y=value, fill=Group)) +
  geom_dotplot(binaxis='y', stackdir='center')+
  labs(title="Dot plot",x="Media Type", y = "INTEDN.0")
p + theme_minimal()
```



(c)

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

ANSWER: From the results of the above graph, it can be considered that the `INTEND.0` in **Type2**(video:"pictures + audio") is less than the others.

Question 2 - traditional one-way ANOVA

(a)

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

ANSWER:

- H_{null}: The means of the four populations are the same.
  - $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_{alt}: The means of the four populations are not the same.

(b)

Produce the traditional F-statistic for our test (you may use any method or function you wish)

```r
var.test(df1$value, df2$value, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  df1$value and df2$value
## F = 1.1607, num df = 41, denom df = 37, p-value = 0.6488
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.610132 2.184962
## sample estimates:
## ratio of variances
##             1.1607
```

```r
var.test(df1$value, df3$value, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  df1$value and df3$value
## F = 0.87591, num df = 41, denom df = 39, p-value = 0.6752
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4658417 1.6387455
## sample estimates:
## ratio of variances
##          0.8759084
```

```r
var.test(df1$value, df4$value, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  df1$value and df4$value
## F = 0.81677, num df = 41, denom df = 45, p-value = 0.5139
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4472891 1.5043838
## sample estimates:
## ratio of variances
##          0.8167668
```

```r
var.test(df2$value, df3$value, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  df2$value and df3$value
## F = 0.75464, num df = 37, denom df = 39, p-value = 0.3918
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.396723 1.443427
## sample estimates:
## ratio of variances
##           0.754638
```

```r
var.test(df2$value, df4$value, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  df2$value and df4$value
## F = 0.70368, num df = 37, denom df = 45, p-value = 0.2741
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3806992 1.3258587
## sample estimates:
## ratio of variances
##          0.7036846
```

```r
var.test(df3$value, df4$value, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  df3$value and df4$value
## F = 0.93248, num df = 39, denom df = 45, p-value = 0.8282
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5076909 1.7361918
## sample estimates:
## ratio of variances
##          0.9324797
```

ANSWER:

- Not reject each $\sigma_i = \sigma_j$ $(i, j \in \{1, 2, 3, 4\})$
- So we can use ANOVA-test in below.

(c)

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

```r
qf(p=0.95, df1=3, df2=162)
```

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

```r
qf(p=0.99, df1=3, df2=162)
```

```
## [1] 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?

```r
summary(aov(df$value~df$Group))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## df$Group      3   22.5   7.508   2.617 0.0529 .
## Residuals   162  464.8   2.869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANSWER:

- F = 2.61
- p-value = 0.0529
- **Reject** H_{null} under the confidence level 95%.
- **NOT Reject** H_{null} under the confidence level 99%.

(e)

Do you feel the classic requirements of one-way ANOVA are met? (Use appropriate tools or logic)

ANSWER:

Requirements for ANOVA: + 1. Each treatment/population's response variable is normally distributed + Not met. + 2. The variance (s^2) of the response variables is the same for all treatments/populations + met. + 3. The observations are independent: the response variables are not related + met.

Question 3 - bootstrapping ANOVA

(a)

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

```
# lecture 8
boot_anova <- function(t1, t2, t3, t4, treat_nums) {
  null_grp1 = sample(t1 - mean(t1), replace=TRUE)
  null_grp2 = sample(t2 - mean(t2), replace=TRUE)
  null_grp3 = sample(t3 - mean(t3), replace=TRUE)
  null_grp4 = sample(t4 - mean(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)
  c(oneway.test(null_values ~ treat_nums, var.equal=TRUE)$statistic,
  oneway.test(alt_values ~ treat_nums, var.equal=TRUE)$statistic)
}
```

```
sales1 = df$value[df$Group=="Type1"]
sales2 = df$value[df$Group=="Type2"]
sales3 = df$value[df$Group=="Type3"]
sales4 = df$value[df$Group=="Type4"]
strategies = df$Group
```

```
f_values <- replicate(5000, boot_anova(sales1, sales2, sales3,sales4, df$Group))
f_nulls <- f_values[1,]
f_alts <- f_values[2,]
```

```
mean(f_nulls)
```

```
## [1] 0.9979202
```

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

```
##      95%
## 2.685553
```

```
mean(f_alts)
```

```
## [1] 3.752789
```

(b)

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

```
qf(p=0.95, df1=3, df2=162)
```

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

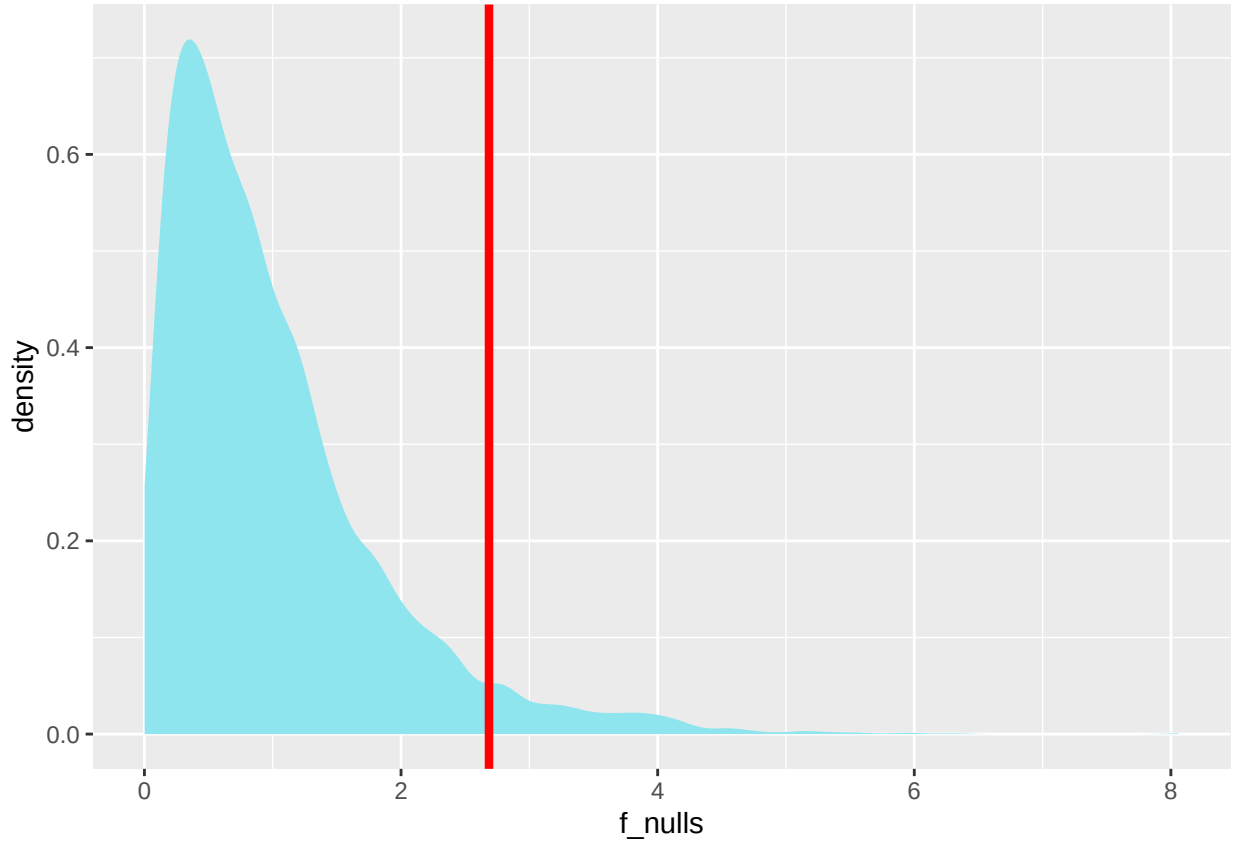
```
qf(p=0.99, df1=3, df2=162)
```

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

(c)

Visualize the distribution of bootstrapped null values of F, the 95% and 99% cutoff values of F (according to bootstrap), and also the original F-value from bootstrapped alternative values.

```
p <- ggplot(mapping = aes(f_nulls)) +
  geom_area(stat = "density", fill = "cadetblue2") +
  geom_vline(xintercept=quantile(f_nulls, 0.95), size=1.5, color="red")
p
```



(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?

ANSWER:

- **Reject** H_{null} under the confidence level 95%.
- **NOT Reject** H_{null} under the confidence level 99%.