Analysis of Variance

Two-way ANOVA

Jenine Harris
Brown School

Importing and cleaning the data

```
# load GSS rda file
load(file = "/Users/harrisj/Box/teaching/Teaching/Fall2020/data/gss2018.

# assign GSS to gss.2018
gss.2018 <- GSS
# remove GSS
rm(GSS)

# recode variables of interest to valid ranges
library(package = "tidyverse")
gss.2018.cleaned <- gss.2018 %>%
  select(HAPPY, SEX, DEGREE, USETECH, AGE) %>%
  mutate(USETECH = na_if(x = USETECH, y = -1)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 999)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 998)) %>%
  mutate(AGE = na_if(x = AGE, y = 98)) %>%
  mutate(AGE = na_if(x = AGE, y = 99)) %>%
  mutate(DEGREE = na_if(x = DEGREE, y = 8)) %>%
  mutate(DEGREE = na_if(x = DEGREE, y = 9)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 8)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 9)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 0)) %>%
  mutate(SEX = factor(x = SEX, labels = c("male","female"))) %>%
  mutate(DEGREE = factor(x = DEGREE, labels = c("< high school",
                                                "high school", "junior c
                                                "college", "grad school"
  mutate(HAPPY = factor(x = HAPPY, labels = c("very happy",
```

Understanding and conducting two-way ANOVA

- One-way ANOVA is useful when there is a single categorical variable (with 3+ categories) and the means of a continuous variable being compared across the categories.
- What happens when there there are two categorical variables that may both be useful in explaining a continuous outcome?
- For example, technology use varies by sex.
- Could ANOVA answer a research question that asked whether technology use varied by educational attainment AND sex?
- Two-way ANOVA is useful for situations where means are compared across the categories of two variables.

Exploratory data analysis for two-way ANOVA

- The boxplots for technology use by degree showed an increase in the percentage of time using technology for those with higher educational attainment.

- Examine the use of technology by sex with a boxplot.

The purpose of two-way ANOVA

- Two-way ANOVA can be used to determine if educational attainment and sex both have relationships with technology use by themselves and whether they **interact** to explain technology use.
  - That is, does technology use differ by educational attainment differently for males compared to females.

```
# graph usetech by degree & sex
gss.2018.cleaned %>%
ggplot(aes(y = USETECH, x = DEGREE)) +
  geom_boxplot(aes(fill = SEX), alpha = .4) +
  scale_fill_manual(values = c("gray70", "#7463AC")) +
  theme_minimal() +
  labs(x = "Educational attainment",
       y = "Percent of time spent using technology",
       title = "Distribution of percentage time using technology by educ
```

Technology use by educational attainment & sex

- There is a different pattern of technology use for males and females.
- Females with less than a high school degree were using technology a lower percentage of the time than males in this group.
- However, females use technology more of the time compared to the males in the high school group and for the junior college group.
- Males and females seem to have relatively equal time spent with technology once a bachelor or graduate degree is earned.
- This pattern of differences is consistent with an interaction.
- A traditional **means plot** is useful for visualizing the idea of an interaction.

Visualizing interaction with a means plot

```
# means plots graph
gss.2018.cleaned %>%
  ggplot(aes(y = USETECH, x = DEGREE, color = SEX)) +
  stat_summary(fun.y = mean, geom="point", size = 3) +
  stat_summary(fun.y = mean, geom="line", aes(group = SEX), size = 1) +
  scale_color_manual(values = c("gray70", "#7463AC")) +
  theme_minimal() +
  labs(x = "Educational attainment",
       y = "Percent of time spent using technology",
       title = "Means plot of technology use by educational attainment a
  ylim(0, 100)
```

Interpreting a means plot

- When the lines in means plots like this one are parallel, it indicates that there is no interaction between the two categorical variables.
- Parallel lines show that the mean of the continuous variable is consistently higher or lower for certain groups compared to others.
- When a means plot shows lines that cross or diverge, this indicates that there is an interaction between the categorical variables.
    - The mean of the continuous variable is different at different levels of one categorical variable depending on the value of the other categorical variable.
    - For example, mean technology use is lower for females compared to males for the lowest and highest educational attainment categories, but female technology use is higher than male technology use for the three other categories of educational attainment.
    - The two variables are working together to influence the value of technology use.

Descriptive stats for two-way ANOVA

- Given the interaction and the differences seen in technology use by `DEGREE` and by `SEX`, it seems likely that the two-way ANOVA would significant relationships for `DEGREE`, `SEX`, and the interaction between the two.
- Before conducting the ANOVA, examine the group means using `group_by()` with both grouping variables in the parentheses.

```
# means by degree and sex
use.stats.2 <- gss.2018.cleaned %>%
  group_by(DEGREE, SEX) %>%
  drop_na(USETECH) %>%
  summarize(m.techuse = mean(USETECH),
            sd.techuse = sd(USETECH))
use.stats.2
```

```
## # A tibble: 10 x 4
## # Groups:   DEGREE [5]
##    DEGREE         SEX    m.techuse sd.techuse
##    <fct>          <fct>      <dbl>      <dbl>
##  1 < high school  male        25.7       35.4
##  2 < high school  female      23.7       37.5
##  3 high school    male        43.5       37.8
##  4 high school    female      55.9       38.6
```

NHST Step 1: Write the null and alternate hypotheses

H0: The mean time using technology is the same across groups by degree, sex, and their interaction.

HA: The mean time using technology is not the same across the groups.

NHST Step 2: Compute the test statistic

- Include the interaction term in the ANOVA `aov()` by multiplying the two categorical variables.
- The terms for `DEGREE` and `SEX` are not needed in the `aov()` function if there is an interaction since `aov()` will include these terms, which are called **main effects**, for any variables included in an interaction.

```
# two-way ANOVA technology use by degree and sex
techuse.by.deg.sex <- aov(formula = USETECH ~ DEGREE * SEX, data = gss.2
summary(techuse.by.deg.sex)
```

```
##                Df  Sum Sq Mean Sq F value   Pr(>F)    
## DEGREE          4  221301   55325  44.209  < 2e-16 ***
## SEX             1   16473   16473  13.163 0.000296 ***
## DEGREE:SEX      4   26510    6627   5.296 0.000311 ***
## Residuals    1399 1750775    1251                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 936 observations deleted due to missingness
```

- There are three F-statistics for this ANOVA, one for each of the two individual variables, the *main effects*, and one for the interaction term.

NHST Step 3: Compute the probability for the test statistic (p-value)

- The p-values in this case were `< 2e-16`, 3\times 10^{-4}, and 3\times 10^{-4}.
- These are very tiny p-values and so the value of an F-statistic being as large or larger than the F-statistics for the two main effects and the interaction term happen a tiny percentage of the time when the null hypothesis is true.

NHST Steps 4 & 5: Interpret the probability and write a conclusion

- The mean time spent on technology use was significantly different across degree groups [$F(4,1399) = 44.21$; $p < .001$] indicating these groups likely came from populations with different mean time spent on technology use. Means were also statistically significant by sex [$F(1,1399) = 13.16$; $p < .001$] and there was a statistically significant interaction between degree and sex on technology use [$F(4,1399) = 5.3$; $p < .001$]. The highest mean was 72.1% of time used for technology for males with graduate degrees. The lowest mean was 23.7% of the time for females with less than a high school diploma. The interaction between degree and sex shows that time spent on technology use increases more quickly for females with both males and females eventually having high tech use in the top two educational attainment groups.

Post-hoc test for two-way ANOVA

- The Bonferroni post-hoc test was not available in R for two-way ANOVA, but the Tukey's HSD test still works.

```
# Tukey's HSD post-hoc test
TukeyHSD(techuse.by.deg.sex)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = USETECH ~ DEGREE * SEX, data = gss.2018.cleaned)
##
## $DEGREE
##                                   diff        lwr       upr     p adj
## high school-< high school    24.8247754 15.244768 34.404783 0.0000000
## junior college-< high school 37.6070312 25.329478 49.884584 0.0000000
## college-< high school        43.0859568 32.760484 53.411429 0.0000000
## grad school-< high school    43.9107249 32.376284 55.445165 0.0000000
## junior college-high school   12.7822558  3.459487 22.105024 0.0017563
## college-high school          18.2611813 11.719691 24.802671 0.0000000
## grad school-high school      19.0859494 10.766152 27.405746 0.0000000
## college-junior college        5.4789255 -4.608337 15.566188 0.5733923
## grad school-junior college    6.3036936 -5.018002 17.625389 0.5490670
## grad school-college           0.8247681 -8.343540  9.993076 0.9991960
```

Interpreting & reporting two-way post-hoc

- There are so many groups with significant differences that it would be more useful to just include the boxplot from the exploratory analysis or the means plot, a table of p-values for the comparisons, and a short paragraph about any interesting overall patterns in the comparisons.
- There were significant differences between males and females in the high school and junior college groups, but that males and females were not significantly different across the other educational groups.
- College groups spent significantly more time using technology than the two other groups, but were not statistically significantly different from each other.
- Overall it appeared that higher education groups spent more time using technology for males and females, and that high school and junior college educated females spent more time using technology than males with these same education levels.

Two-way ANOVA assumptions

- The assumptions of homogeneity of variances and normality were also applicable in two-way ANOVA.
- Normality would be a little trickier to test by looking at each group since there are five degree groups, two sex groups, and 10 degree-by-sex groups (e.g., male and < high school).
- Instead of checking normality one group at a time when there are a large number of groups in an ANOVA model, this assumption can be checked by examining the **residuals**.
- The residuals are the distances between the value of the outcome for each person and the value of the group mean for that person.
- When the residuals are normally distributed, this indicates that the values in each group are normally distributed around the group mean.

Testing the normality assumption

- The `techuse.by.deg.sex` object in the Environment pane shows the residuals.
- Use a Shapiro-Wilk test to check normality statistically and plot the residuals for a visual assessment:

```
# statistical test of normality for groups
shapiro.test(techuse.by.deg.sex$residuals)
```

```
##
##      Shapiro-Wilk normality test
##
## data:  techuse.by.deg.sex$residuals
## W = 0.95984, p-value < 2.2e-16
```

- The null hypthesis for the Shapiro-Wilk test is that the distribution is normal.
- By rejecting this null hypothesis with a tiny p-value, the assumption is failed.
- So, this test shows that the residuals fail the normality assumption.

Visualizing the residuals

- The `ggplot()` function does not work directly with the ANOVA object, so convert the residuals to a new data frame first and then graph them.

```
# make a data frame
tech.deg.sex <- data.frame(techuse.by.deg.sex$residuals)

# plot the residuals
tech.deg.sex %>%
ggplot(aes(x = techuse.by.deg.sex.residuals)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal() +
  labs(x = "Residuals",
       y = "Number of observations",
       title = "Distribution of residuals from ANOVA explaining tech use
```

Testing the homogeneity of variances assumption

- The `leveneTest()` function could be used to test the null hypothesis that the variances are equal:

```
# Levene test for ANOVA
car::leveneTest(y = USETECH ~ DEGREE*SEX, center = mean,
                data = gss.2018.cleaned)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##         Df F value    Pr(>F)
## group    9  8.5912 1.289e-12 ***
##       1399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The results were statistically significant so the null hypothesis was rejected.
- The equal variances assumption was not met.
- The two-way ANOVA has failed its assumptions.

Alternatives when two-way ANOVA assumptions fail

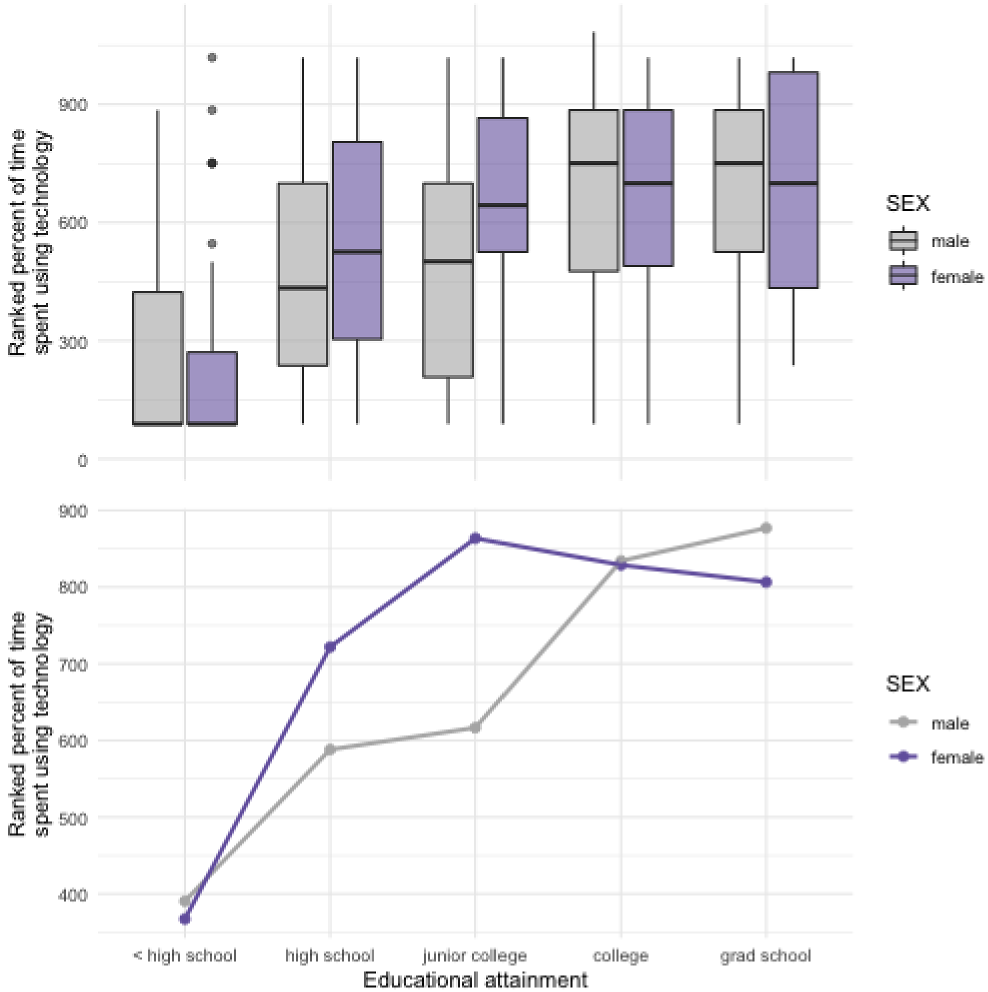
- One suggested method is to compute the ranks of the outcome and conduct the two-way ANOVA on the ranked outcome variable.

- Use the ranked values of `USETECH` from the Dunn's test earlier and the two-way ANOVA code with the transformed outcome variable.

```
# two-way ANOVA technology use by degree and sex
techuse.by.deg.sex.t <- aov(formula = usetech.rank ~ DEGREE * SEX,
                            data = gss.2018.cleaned)
summary(techuse.by.deg.sex.t)
```

```
##                Df    Sum Sq Mean Sq F value   Pr(>F)
## DEGREE          4  23270305 5817576   40.26  < 2e-16 ***
## SEX             1   1849104 1849104   12.80 0.000359 ***
## DEGREE:SEX      4   3120976  780244    5.40 0.000258 ***
## Residuals    1399 202148767  144495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 936 observations deleted due to missingness
```

Visualizing the transformed outcome



Interpreting the two-way ANOVA with ranks

- The plots showed difference from one educational attainment group to another and between males and females.

- Interpretation of the ranked outcome ANOVA:

    - A two-way ANOVA with ranked technology time use as the outcome found statistically significant main effects of degree and sex on technology use ($p < .05$) and a statistically significant interaction between degree and sex ($p < .05$). The overall pattern of results indicates that males and females with less than a high school education use technology the least, while those with a college degree use technology the most. Males and females differ a lot in use of technology for those with a junior college degree, with females having a junior college degree having the highest use of technology of all females.