Analysis of Variance

Assumptions

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",
```

Visualizing the groups

```
# graph usetech
gss.2018.cleaned %>%
  drop_na(USETECH) %>%
  ggplot(aes(y = USETECH, x = DEGREE)) +
  geom_jitter(aes(color = DEGREE), alpha = .6) +
  geom_boxplot(aes(fill = DEGREE), alpha = .4) +
  scale_fill_brewer(palette = "Spectral", guide = FALSE) +
  scale_color_brewer(palette = "Spectral", guide = FALSE) +
  theme_minimal() +
  labs(x = "Highest educational attainment",
       y = "Percent of time spent using technology",
       title = "Distribution of time spent using technology\nuse by educ
```

Group means

```
# mean and sd of age by group
use.stats <- gss.2018.cleaned %>%
  drop_na(USETECH) %>%
  group_by(DEGREE) %>%
  summarize(m.techuse = mean(USETECH),
            sd.techuse = sd(USETECH))
use.stats
```

```
## # A tibble: 5 x 3
##   DEGREE          m.techuse sd.techuse
##   <fct>               <dbl>      <dbl>
## 1 < high school        24.8       36.2
## 2 high school          49.6       38.6
## 3 junior college       62.4       35.2
## 4 college              67.9       32.1
## 5 grad school          68.7       30.2
```

ANOVA results

```
# conduct ANOVA for technology use by degree category with oneway.test
techuse.by.deg <- oneway.test(formula = USETECH ~ DEGREE,
                              data = gss.2018.cleaned,
                              var.equal = TRUE)
techuse.by.deg
```

```
##
##      One-way analysis of means
##
## data:  USETECH and DEGREE
## F = 43.304, num df = 4, denom df = 1404, p-value < 2.2e-16
```

```
# conduct ANOVA for technology use by degree category with aov
techuse.by.deg.aov <- aov(formula = USETECH ~ DEGREE,
            data = gss.2018.cleaned)
techuse.by.deg.aov
```

```
## Call:
##    aov(formula = USETECH ~ DEGREE, data = gss.2018.cleaned)
##
## Terms:
##                     DEGREE Residuals
## Sum of Squares   221300.6 1793757.2
## Deg. of Freedom         4      1404
##
## Residual standard error: 35.7436
```

Testing ANOVA assumptions

The assumptions of ANOVA are:

- continuous variable and three or more independent groups
- independent observations
- normal distribution in each group
- equal variances for each group

Testing normality

- There are many ways to test for normality, one way is with density plots:

```
#graph tech use by degree
gss.2018.cleaned %>%
  drop_na(USETECH) %>%
  ggplot(aes(x = USETECH)) +
  geom_density(aes(fill = DEGREE)) +
  facet_wrap(~ DEGREE, nrow = 2) +
  scale_fill_brewer(palette = "Spectral", guide = FALSE) +
  theme_minimal() +
  labs(x="Percent of time using tech",
       y="Probability density",
       title = "Distribution of technology use by educational attainment
```

Testing normality with Q-Q plots

- Based on the density plots, none of the groups looked normally distributed.
- Some Q-Q plots might confirm this:

```
#graph tech use by degree
gss.2018.cleaned %>%
  drop_na(USETECH) %>%
  ggplot(aes(sample = USETECH)) +
  geom_abline(aes(intercept = mean(USETECH), slope = sd(USETECH), linety
              color = "gray60", size = 1) +
  stat_qq(aes(color = DEGREE)) +
  scale_color_brewer(palette = "Spectral", guide = FALSE) +
  scale_linetype_manual(values = 1, name = "") +
  labs(x = "Theoretical normal distribution",
       y = "Observed values of percent time using tech",
       title = "Distribution of time spent on technology use by educatio
  theme_minimal() +
  facet_wrap(~ DEGREE, nrow = 2)
```

Shapiro-Wilk test of normality

- None of the groups appear to be normally distributed based on either type of plot.

- The floor and ceiling values appeared to be driving some of the non-normality.

- The Shapiro-Wilk test is not necessary given the big deviations from normality in the histograms and Q-Q plots, however, try it just to confirm.

- The Shapiro-Wilk test tests the null hypothesis that the data are normally distributed.

- The Shapiro-Wilk test is for a single group, but using `summarize()` after `group_by()` can compute it for each of the groups separately, then print the p-value each of the tests:

```
# statistical test of normality for groups
gss.2018.cleaned %>%
  drop_na(USETECH) %>%
  group_by(DEGREE) %>%
  summarize(shapiro.pval = shapiro.test(x = USETECH)$p.value)
```

```
## # A tibble: 5 x 2
##   DEGREE          shapiro.pval
##   <fct>                  <dbl>
## 1 < high school       1.83e-14
## 2 high school         5.99e-24
## 3 junior college      2.92e- 9
## 4 college             1.22e-16
## 5 grad school         4.34e-11
```

Homogeneity of variances assumption

- The second assumption for ANOVA is the assumption of *homogeneity of variances* or *equal variances across groups*.

- Levene's Test is widely used to test the assumption of equal variances.

- The null hypothesis is that *the variances are equal* while the alternate is that at least two of the variances are different.

- The `leveneTest()` function can be used to conduct the Levene's Test.

```
# equal variances for systolic by sex
car::leveneTest(y = USETECH ~ DEGREE, data = gss.2018.cleaned)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)    
## group    4   18.44 8.845e-15 ***
##       1404                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value for the Levene's test suggests rejecting the null hypothesis; the variances of `USETECH` are statistically significantly different across groups (p < .05).