Analysis of Variance

Alternate tests for failing 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
```

Calculating an alternate F-statistic for failing the homogeneity assumption

- The first options are for when the normality assumption is met but the homogeneity of variances assumption fails.
- In this situation, the standard approach is to use ANOVA but compute an alternate F-statistic that does not rely on equal variances.
- There are two alternate F-statistics that are widely used for this purpose:
- Brown-Forsythe
- Welch's

Brown-Forsythe F-statistic

- The Brown-Forsythe approach to calculating F starts with a transformation of the continuous variable from its measured values to values that represent the distance each observation is from the median of the variable.

$$t_{ij} = |y_{ij} - median_{y_j}|$$

- In this equation, y_{ij} is each observation i in group j, $median_{y_j}$ is the median of group j, and enclosing the equation in | is for absolute value.

- The alternate F-statistic is then computed as in Equation \@ref(eq:bftest2) using the same F formula but with the means computed from the transformed (t_{ij}) of the technology use variable rather than from the raw values of the continuous variable.

$$F_{BF} = \frac{\frac{\sum n_{i.}(\bar{t}_{i.} - \bar{t})^2}{k-1}}{\frac{\sum\sum(t_{ik} - \bar{t}_{i.})^2}{n-k}}$$

Computing Brown-Forsythe F stat

- While there are R packages that can be used to compute the Brown-Forsythe directly, another option is to transform the outcome variable and use the `aov()` command used for ANOVA.

- Create the transformed version of the `USETECH` variable using `abs()` to get the absolute value of the difference between each value of `USETECH` and the median of `USETECH`, making sure to remove the `NA` in the `median()` function so it works.

```
# add new variable to data management
gss.2018.cleaned <- gss.2018 %>%
  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(SEX, labels = c("male","female"))) %>%
  mutate(DEGREE = factor(x = DEGREE, labels = c("< high school",
                                          "high school", "junior c
                                          "bachelor", "graduate"))
```

Check the transformed variable

```
# check new variable
summary(object = gss.2018.cleaned$usetech.tran)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   20.00   35.00   34.08   45.00   60.00     936
```

- Use the `aov()` or `oneway.test()` from earlier, changing the outcome to use the `usetech.tran` transformed variable instead.

NHST Step 1: Write the null and alternate hypotheses

The null and alternate hypotheses would be:

H0: The mean value of the transformed technology use variable is the same across educational attainment groups.

HA: The mean value of the transformed technology use variable is not the same across educational attainment groups.

NHST Step 2: Compute the test statistic

```r
# brown-forsythe anova
usetech.t.by.degree <- oneway.test(formula = usetech.tran ~ DEGREE,
                                   data = gss.2018.cleaned)
usetech.t.by.degree
```

```
##
##     One-way analysis of means (not assuming equal variances)
##
## data:  usetech.tran and DEGREE
## F = 36.096, num df = 4.00, denom df = 393.53, p-value < 2.2e-16
```

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

The p-value in this case is 0.0000000000000000000000001054491. The value of an F_{BF}-statistic being this large or larger happens a tiny percentage of the time when the null hypothesis is true.

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

- The results show a statistically significant difference of the means of the transformed technology use variable by educational attainment group [F_{BF} (4, 393.5309584) = 36.1; p < .05].

Examine the transformed variable

- To better understand the results, compute descriptive statistics and examine a graph of the transformed variable.

```
# means of transformed variable
usetech.t.stats <- gss.2018.cleaned %>%
  drop_na(usetech.tran) %>%
  group_by(DEGREE) %>%
  summarise(m.usetech.tran = mean(usetech.tran),
            sd.usetech.tran = sd(usetech.tran))
usetech.t.stats
```

```
## # A tibble: 5 x 3
##   DEGREE         m.usetech.tran sd.usetech.tran
##   <fct>                   <dbl>           <dbl>
## 1 < high school            47.3            17.4
## 2 high school              36.2            17.0
## 3 junior college           30.6            17.5
## 4 bachelor                 29.5            14.8
## 5 graduate                 28.0            14.2
```

Interpreting the transformed variable

- The mean of the transformed `USETECH` variable, `usetech.tran`, which were differences between the original values and the median value of `USETECH`, was 47.3 for the < high school group and 36.18 for the high school group.

- The rest of the means were smaller.

- The transformation made the differences among the means somewhat smaller and the transformed means were higher in the lower education attainment groups.

```
# graph transformed USETECH variable
gss.2018.cleaned %>%
  drop_na(usetech.tran) %>%
  ggplot(aes(y = usetech.tran, 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 = "Educational attainment",
       y = "Distance to median of tech use time for group",
       title = "Transformed time using technology by educational attainm
```

Welch's F-statistic

- Rather than use a transformed outcome variable, the main idea behind the Welch's F-statistic is to use weights in the calculation of the group means and overall mean (also known as the grand mean).

- The weight is computed for each group to account for the different variances across groups:

$$w_k = \frac{n_k}{s_k^2}$$

- Where n_k is the sample size in group k and s_k^2 is the variance in group k.

Using the weight to compute means

- The grand/overall mean is then computed using the weight and the weighted mean for each of the groups:

$$\bar{y}_{welch} = \frac{\sum_{j=1}^{k} w_k \bar{y}_k}{\sum_{j=1}^{k} w_k}$$

- Where w_k is the weight for group k, $\bar{y}_k$ is the mean of the continuous variable for group k.

- The $\sum_{j=1}^{k}$ is the sum of each group from the first group, j = 1 to the last group, j = k.

- The grand mean for Welch's F-statistic is used in the final computation of the Welch's F-statistic.

NHST Step 1: Write the null and alternate hypotheses

The null and alternate hypotheses would be:

H0: Time spent using technology is the same across educational attainment groups.

HA: Time spent using technology is not the same across educational attainment groups.

NHST Step 2: Compute the test statistic

```r
# welch test for unequal variances
welch.usetech.by.degree <- oneway.test(formula = USETECH ~ DEGREE,
                data = gss.2018.cleaned,
                var.equal = FALSE)
welch.usetech.by.degree
```

```
##
##      One-way analysis of means (not assuming equal variances)
##
## data:  USETECH and DEGREE
## F = 46.06, num df = 4.00, denom df = 400.31, p-value < 2.2e-16
```

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

The p-value in this case is `< 2.2e-16`, which is much less than .05. The value of an F_W-statistic being this large or larger happens a tiny amount of the time when the null hypothesis is true.

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

- The results show a statistically significant difference in the mean of the `USETECH` variable by degree group [F_W (4, 400.3121225) = 46.06; p < .05].
- The F-statistic was a little larger and the degrees of freedom for the denominator was a smaller number.
- The weighting was also used in the calculation of the denominator degrees of freedom.

$$df_{denom} = \frac{1}{\frac{3\sum_{j=1}^{k}\left(1-\frac{w_k}{\sum_{j=1}^{k} w_k}\right)^2}{n_k-1}}$$

Degrees of freedom and the F distribution

- With fewer degrees of freedom, the F-statistic has to be a larger number to reach statistical significance.

Interpreting the distributions

- While it is true that there isn't much difference between the distributions, the area under the curves is what matters for the p-value cutoff.
- When the line is just slightly closer to the x-axis, this changes things quickly for the area under the curve.
- The thresholds for statistical significance ($p < .05$) for these three lines are 2.376377 for the 2000 degrees of freedom, 2.7587105 for the 25 degrees of freedom, and 3.4780497 for the 10 degrees of freedom.

The Kruskal-Wallis test for failing the normality assumption

- The Kruskal-Wallis test is used to compare three or more groups when the normal distribution assumption fails for ANOVA.
- Like several of the tests used when the the outcome is not normally distributed for a t-test, the Kruskal-Wallis test compares ranks among groups.
- Specifically, the observations are put in order by size and each is assigned a rank.
- The mean rank for each group is then computed and used to calculate the K-W test statistic, H.

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j(\overline{r}_j - \frac{n+1}{2})^2$$

- In this formula, n is the overall sample size, n_j is the sample size for group j and $\overline{r}_j$ is the mean rank for group j.

NHST Step 1: Write the null and alternate hypotheses

H0: The mean rank of technology use is the same across educational degree groups.

HA: The mean rank of technology use is not the same across educational degree groups.

NHST Step 2: Compute the test statistic

```r
# compare usetech by degree
kw.usetech.by.degree <- kruskal.test(formula = USETECH ~ DEGREE,
                                     data = gss.2018.cleaned)
kw.usetech.by.degree
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  USETECH by DEGREE
## Kruskal-Wallis chi-squared = 142.21, df = 4, p-value < 2.2e-16
```

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

The p-value is `< 2.2e-16`, which, as usual, is very tiny. The value of an H-statistic being this large or larger happens a tiny percentage of the time when the null hypothesis is true.

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

- The conclusion is that there is a difference in the mean rank for technology use by degree group ($H(4) = 4$; $p < .05$).
- Like the ANOVA results, the K-W test identifies if there is a difference somewhere among the means but does not identify which groups are different from one another.
- A post-hoc test like *Bonferroni* or *Tukey's HSD* could help.
- For K-W, *Dunn's test* of multiple comparisons is useful for identifying which groups are *statistically significantly* different from which other groups.

Dunn's post-hoc test for Kruskal-Wallis

- Dunn's test function, `dunn.test()` requires a method be selected for adjusting the p-value to account for the multiple comparisons.
- Bonferroni is one of the methods commonly used with ANOVA.
- The `dunn.test()` function takes three arguments, the `x =` argument is the continuous variable, the `g =` argument is for the groups, and the `method =` argument is the p-value adjustment method.

```
# post-hoc test for usetech by degree
dunn.usetech.by.degree <- dunn.test::dunn.test(x = gss.2018.cleaned$USET
                                               g = gss.2018.cleaned$DEGR
                                               method = "bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 142.2141, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                 (Bonferroni)
## Col Mean-|
```

Interpreting Dunn's test results

- The Dunn's test is a rank-sum test just like the Mann-Whitney U and can be interpreted in the same way.
- In this case it appears that there is *no difference* in technology use for *graduate vs. bachelor*, *junior college vs. bachelor*, or *junior college vs. graduate*.
- All other pairings have statistically significant differences between the mean ranks.
- The table shows a z-statistic for each pair computed from the sum of the ranks for the pair. Below the z-statistic is a p-value associated with the z-statistic.
- The p-value is adjusted using a Bonferroni adjustment, which means it was multiplied by the number of comparisons.
    - In this case, the number of comparisons was $\frac{5 \cdot (5-1)}{2} = 10$.

Visualizing the differences in rank

- The default for `rank()` is to give the `NA` values a rank; since it makes more sense to leave them as `NA`, add the "keep" option to the argument so that the code keeps `NA` as `NA`.

```
# add new variable to data management
gss.2018.cleaned <- gss.2018 %>%
  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(SEX, labels = c("male","female"))) %>%
  mutate(DEGREE = factor(x = DEGREE, labels = c("< high school",
                                               "high school", "junior c
                                               "bachelor", "graduate"))
  mutate(HAPPY = factor(x = HAPPY, labels = c("very happy",
                                             "pretty happy",
                                             "not too happy"))) %>%
  mutate(usetech.t = abs(x = USETECH - median(USETECH, na.rm = TRUE))) %
```

Examine new rank variable

```
# check new variable
summary(object = gss.2018.cleaned$usetech.rank)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    88.5   357.5   699.5   705.0  1019.0  1272.0     936
```

Graph the ranks

```
# graph the ranks
gss.2018.cleaned %>%
  ggplot(aes(y = usetech.rank, 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 = "Educational attainment",
       y = "Ranks of technology use time",
       title = "Ranks of percentage of time using technology use by educ
```

Effect size for Kruskal-Wallis

- eta-squared works for Kruskal-Wallis as an effect size:

$$\eta_H^2 = \frac{H - k + 1}{n - k}$$

- To use the eta-squared formula for the effect size of the Kruskal-Wallis test of technology use time by educational attainment, the H test statistic, k groups, and n number of observations are needed.

Computing effect size for Kruskal-Wallis

- The H is 142.21, there are 5 educational attainment groups, and there are 936 `NA` values out of the 2345 observations in the data frame, so n = 2345 - 936 = 1409.

$$\eta^2_H = \frac{142.21 - 5 + 1}{1409 - 5} = .098$$

The cutoff values are the same as for the omega-squared:

- $\eta^2 = .01$ to $\eta^2 < .06$ is a small effect
- $\eta^2 = .06$ to $\eta^2 < .14$ is a medium effect
- $\eta^2 \geq .14$ is a large effect
- In this case, consistent with the original ANOVA results, the eta-squared effect size for the Kruskal-Wallis test is medium.
- There is a medium strength relationship between educational attainment and percentage of time spent using technology.

Interpreting K-W with effect size

- A Kruskal-Wallis test found a statistically significant difference in technology use time across educational attainment groups ($H = 142.21$; $p < .05$). Based on a Dunn's post-hoc test, those with less than a high school education had statistically significantly lower mean ranked technology use time than all of the other groups ($p < .05$), people with a bachelor's degree, a master's degree, or a junior college degree had significantly higher mean ranks than those with a high school diploma. There were no statistically significant differences among the three college groups. There was a medium effect size for the relationship between educational attainment and ranked values of technology use time ($\eta^2 = .098$).