Analysis of Variance

Conducting ANOVA

Jenine Harris
Brown School

Exploring the data using graphics and descriptive statistics

```
# 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
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)) %>%
```

Understanding and conducting one-way Analysis of Variance (ANOVA)

- ANOVA is the statistical test used to compare means across 3 or more groups.
- The t-tests were great for comparing two means, could t-tests be used to compare all the means for the five `DEGREE` variable groups?
- Perhaps this could be done with one t-test for each pair of means?
- The problem with using the independent samples t-test is that each test comes with a p-value which represent the probability that the two group means were the same in the populations the samples came from.
- Usually the threshold for the p-value to indicate statistical significance is .05, which means that there can be up to a 5% probability that *rejecting the null hypothesis is an error*.
    - Rejecting the null hypothesis when it is true is called a **Type I error** (Box \@ref(ch7leslie)).
- With five groups in the degree variable, comparing each pair with a t-test (i.e., conducting **pairwise comparisons**) would result in 10 t-tests.

Familywise Type I error

- The formula for this probability of a Type I error when there are multiple comparisons is:.

$$\alpha_f = 1 - (1 - \alpha_i)^c$$

- Where α_f is the **familywise** Type I error rate, α_i is the individual alpha set as the statistical significance threshold, and c is the number of comparisons.

- The formula for computing c is $\frac{k \cdot (k-1)}{2}$, where k is the total number of groups.

Familywise error for 5 groups

- For a five-group `DEGREE` variable with α = .05 for each pairwise comparison, the familywise α would be .40.

$$\alpha_f = 1 - (1 - .05_i)^{10} = .40$$

- With 10 pairwise comparisons, the familywise α indicated there would be a 40% probability that a conclusion was wrong on at least one of the comparisons.
- To control this error rate, and for efficiency, use a single analysis of variance (ANOVA) test instead of 10 t-tests.
- ANOVA is useful for testing whether three or more means are equal.
- It can be used with two means, but the t-test is preferable because it is more straightforward.

The F test statistic for ANOVA

- To compare mean technology use time across the five degree categories, use the `oneway.test()` function.
- The `oneway.test()` function has several arguments.
- The first argument is `formula =`, where the formula for testing would be entered.
- The formula for `oneway.test()` includes the continuous variable first, then the tilde, then the categorical variable.
- The formula would be `continuous ~ categorical`. In this case, with the `USETECH` and `DEGREE` variables, the formula is `USETECH ~ DEGREE`.
- After the formula, the data frame name is entered for the `data =` argument and the final argument is `var.equal =` which refers to one of the assumptions of ANOVA.
- For now use `var.equal = TRUE`.

```
# mean tech use percent by degree groups
techuse.by.deg <- oneway.test(formula = USETECH ~ DEGREE,
                              data = gss.2018.cleaned,
                              var.equal = TRUE)
```

Examining R output for ANOVA

```
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
```

- The output the F-statistic, which is a ratio where the variation *between* the groups is compared to the variation *within* the groups.
- The *between group variation* is in the numerator to calculate F, while the *within group variation* is in the denominator.
- A plot can be useful for thinking about this.

Computing the F statistic

- Notice that the group mean for the junior college group was higher than the overall mean.
- That is, the mean percentage of time people use technology in this group is higher than the overall mean percentage of time people use technology.
- For each group, the group mean does a better job than the overall mean of explaining tech use *for that group*.
- The difference between the group mean and the overall mean is *how much better* the group mean is at representing the data in the group.
- This difference is used to compute the numerator of the F-statistic.

$$F = \frac{\frac{\sum n_{i.}(\bar{y}_{i.} - \bar{y})^2}{k-1}}{\frac{\sum\sum(y_{ik} - \bar{y}_{i.})^2}{n-k}}$$

- Where y represents the continuous variable, n represents number of observations, k represents the groups, and i stands for individual.
- In the numerator, $\bar{y}_{i.}$ is mean of the continuous variable for each group and $\bar{y}$ is the overall or **grand mean** of the continuous variable.
- $n_{i.}$ is the number of people in the group and divided and k is the number of groups.

The F statistic conceptually

- Altogether, the numerator quantifies the variation between the group means and the grand mean, while the denominator quantifies the variation between the individual values and the group means.

$$F = \frac{between - group - variability}{within - group - variability}$$

- Because the difference between the group means and the grand mean represents the variability that the group mean explains for the group, the numerator is also sometimes referred to as **explained variance**.
- The denominator sums the distances between the observations and their group mean, which is the variability that the group mean cannot explain. The denominator is sometimes called **unexplained variance**.
- The F-statistic, then, could be referred to as a ratio of explained to unexplained variance.
- That is, how much of the variability in the data does the ANOVA explain compared to how much it leaves unexplained.
- The larger the F-statistic, the more the ANOVA has explained compared to what it has left unexplained.

An alternate way to compute F

- True to the *analysis of variance* name, the F-statistic can also be represented as the ratio of the variance between the groups to the variance within the groups.

$$F = \frac{s_b^2}{s_w^2}$$

Using the F statistic

- Once the F-statistic is computed, the probability of finding an F-statistic at least as big as the one computed is determined by the F-distribution.
- Like n and p were the parameters for the binomial distribution, m and s were the parameters for the normal distribution, and df was the parameter for chi-squared, the F statistic also has parameters that define its shape.
- For F, the parameters are the degrees of freedom for the numerator and the degrees of freedom for the denominator of the F equation.
- These values are $df_{numerator}$ = k - 1 and $df_{denominator}$ = n - k, where k is the number of groups and n is the sample size.

Visualizing the F distribution

- The F-distribution has some similarity to the chi-squared distribution since they are both right-skewed and do not go below zero.

The F statistic for the ANOVA of technology use by degree

- The ANOVA comparing mean time using technology across categories of degree (`techuse.by.deg` object) had an F-statistic of 43.30 with 4 and 1404 degrees of freedom.

Interpreting the F statistic

- The F-statistic of 43.30 was far to the right in the tail of the distribution.
- The probability of an F-statistic this large or larger, if there was no difference among the means, was reported in the output as `< 2.2e-16`, which is < .001.
- With a p-value this tiny, the F-statistic would be considered statistically significant.

NHST Step 1: Write the null and alternate hypotheses

HO: The mean time spent on technology use is equal across degree groups.

HA: The mean time spent on technology use is not equal across degree groups.

NHST Step 2: Compute the test statistic

The F-satistic is 43.3.

```
# print the results of the ANOVA
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
```

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

The p-value is `< 2.2e-16`, which is very small. The value of an F-statistic being at least this large happens a tiny percentage of the time when the null hypothesis is true.

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

With a p-value < .001, the ANOVA indicates that there is likely a difference among the means of time spent using technology based on degree.

Interpretation:

- The mean time spent on technology use was significantly different across degree groups [$F(4,1404) = 43.3$; $p < .05$] indicating these groups likely came from a population with different mean time spent on technology use by educational attainment. The highest mean was grad school% of time used for technology for those with graduate degrees. The lowest mean was < high school% of the time for those with less than a high school diploma.