ANOVA

Zachary Himmelberger

Introduction

In this tutorial, we will examine how to calculate an ANOVA by hand and using R. We will focus on the interpretation. The initial example will use a between-subjects one-way ANOVA, so called because it involve one independent variable. However, I will also show R code for a between-subjects factorial ANOVA, which involves more than one indpendent variable.

Problem Statement

A researcher is interested in the effects of mindfulness affects test anxiety. Prior to an exam, students were grouped into one of three experimental conditions: mindfulness, conversation, and control. The mindfulness group spent 15-minutes focusing on their breathing and the conversation group spent 15-minutes having a friendly conversation with a researcher. Following the 15-minute period, the participants rated their anxiety on a 10-point scale about the upcoming exam on a 10-point scale. The control rated their anxiety immediately upon getting to the exam. Here are the results:

Control	Mindfulness	Conversation	
10	4	5	
8	4	6	
6	2	5	
8	2	6	
$\bar{x}_{ctrl} = 8$	$\bar{x}_{mind} = 3$	$\bar{x}_{conv}$ =5.5	$\bar{x} = 5.5$

Note that we have a sample mean for each group and an overall mean (denoted with $\bar{x}$ here), which we call the grand mean. We want to know whether the population means differ between the three conditions.

t-tests

One way we could analyze this experiment is to conduct multiple *t*-tests. We could contrast Control v. Mindfulness, Control v. Conversation, and Mindfulness v. Conversation. This strategy is known as making all-pairwise comparisons. However, this method would increase our type I error because each test has a specified alpha. If we use an alpha of .05, then our type I error rate across all three contrasts (called the experiment-wise error rate) is $1 - (.95 * .95 * .95) \approx .143$.

One solution that will help keep our type I error rate at .05 is apply a correction. Conceptually, the easiest choice is to make a Bonferroni correction, which involves dividing the experiment-wise alpha by the number of tests conducted and using that value for each test. In our case, this would mean running each *t*-test with an alpha of $\frac{.05}{3} \approx .017$. We will make use of this approach later. For now, just know that it significantly reduces our statistical power (i.e., our ability to reject the null hypothesis), especially when we are making many comparisons. ANOVA offers an alternative solution.

One-way ANOVA

Null Hypothesis

The (typical) null hypothesis of an ANOVA is that the population means are equal. In symbols, this is given by

$$\mu_1 = \mu_2 = \mu_3 = \mu_n$$

where n is any number of means. The alternative hypothesis is that there is at least one population mean that is not equal. This is harder to put in symbols. Note that rejecting the null hypothesis is equivalent to saying that the populations means are not equal, but we do not know which mean differs from which other mean. I will come back to this point below.

Sum of Squares

The result is an F-test. To get an F-statistic, we will need to calculate sum of squares and degrees of freedom. Remember,

$$SS = \sum_{i=1}^{N}(x_i - \bar{x})^2$$

For each group, we will calculate SS in three different ways.

1. $SS_{total} = \sum_{i=1}^{N}(x_i - \bar{x})^2$,
2. $SS_{between} = \sum_{i=1}^{N}(\bar{x}_g - \bar{x})^2$,
3. $SS_{within} = \sum_{i=1}^{N}(x_i - \bar{x}_g)^2$,

where $\bar{x}$ is the grand mean and $\bar{x}_g$ is the group mean.

We can think of SS_{total} as representing the total variation, $SS_{between}$ as the variation between the group and the grand mean, and SS_{within} as the variation within each group. Note that $SS_{total} = SS_{between} + SS_{within}$. If we think of the independent variable (i.e., the grouping variable) as the explanatory variable, then $SS_{between}$ represents the variation we can explain, and SS_{within} represents the variation that we cannot explain.

Control

x	$x - \bar{x}$	$(x - \bar{x})^2$	$\bar{x}_{ctrl} - \bar{x}$	$(\bar{x}_{ctrl} - \bar{x})^2$	$x - \bar{x}_{ctrl}$	$(x - \bar{x}_{ctrl})^2$
10	4.5	20.25	2.5	6.25	2	4
8	2.5	6.25	2.5	6.25	0	0
6	0.5	0.25	2.5	6.25	-2	4
8	2.5	6.25	2.5	6.25	0	0
		$SS_{total} = 33$		$SS_{between} = 25$		$SS_{within} = 8$

Mindfulness

x	$x - \bar{x}$	$(x - \bar{x})^2$	$\bar{x}_{mind} - \bar{x}$	$(\bar{x}_{mind} - \bar{x})^2$	$x - \bar{x}_{mind}$	$(x - \bar{x}_{mind})^2$
4	-1.5	2.25	-2.5	6.25	1	1
4	-1.5	2.25	-2.5	6.25	1	1
2	-3.5	12.25	-2.5	6.25	1	1
2	-3.5	12.25	-2.5	6.25	1	1
		$SS_{total} = 29$		$SS_{between} = 25$		$SS_{within} = 4$

Conversation

x	$x - \bar{x}$	$(x - \bar{x})^2$	$\bar{x}_{conv} - \bar{x}$	$(\bar{x}_{conv} - \bar{x})^2$	$x - \bar{x}_{conv}$	$(x - \bar{x}_{conv})^2$
5	-0.5	0.25	0	0	-0.5	0.25
6	0.5	0.25	0	0	0.5	0.25
5	-0.5	0.25	0	0	-0.5	0.25
6	0.5	0.25	0	0	0.5	0.25
		$SS_{total} = 1$		$SS_{between} = 0$		$SS_{within} = 1$

Now we can sum across the groups.

$$\begin{aligned} SS_{total} &= 33 + 29 + 1 = 63 \\ SS_{between} &= 25 + 25 + 0 = 50 \\ SS_{within} &= 8 + 4 + 1 = 13 \end{aligned}$$

Degrees of Freedom

For each group, we will calculate df in three different ways.

1. $df_{total} = N - 1$,
2. $df_{between} = k - 1$,
3. $df_{within} = N - k$,

where N is the sample size and k is the number of groups.

Note that $df_{total} = df_{between} + df_{within}$. For our problem, we have

$$\begin{aligned} df_{total} &= 12 - 1 = 11 \\ df_{between} &= 3 - 1 = 2 \\ df_{within} &= 12 - 3 = 9. \end{aligned}$$

Mean Square and the F-statistic

Mean square is a *correction* to sum of squares that takes into account the degrees of freedom.

$$MS = \frac{SS}{df}$$

As in sum of squares and degrees of freedom, we can calculate the mean square separately between and within groups.

1. $MS_{between} = \frac{SS_{between}}{df_{between}}$
2. $MS_{within} = \frac{SS_{within}}{df_{within}}$

The F-statistic is the ratio of mean square between to mean square within,

$$F = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/df_{between}}{SS_{within}/df_{within}}.$$

In our example, we have

$$
\begin{aligned}
MS_{between} &= 50/2 = 25 \\
MS_{within} &= 13/9 = 1.44 \\
F &= 25/1.44 = 17.36.
\end{aligned}
$$

Using our above conceptual definition of sum of squares, we can think of this as being the ratio of what we can explain (using the independent variable) to what we cannot explain. It is helpful to put all of this information in a chart.

Source	SS	df	MS	F
between	50	2	25	17.36
within	13	9	1.44	
total	63	11		

p-value

As with other inferential statistics, a given value of the F-statistic has an associated p-value.

In our example, we have $p = .0008$. Thus, we can reject the null hypothesis. This means that we have evidence that the population means are not equal. However, we still cannot say which population mean is different from which other population mean.

Statistical Assumptions

There are three main assumptions for a between-subjects one-way ANOVA. These assumptions are a generalization of the same assumptions for the t-test.

1. homogeneity of variance (i.e., the population variances are equal)
2. the errors are normally distributed
3. the errors are independent

ANOVA Using R

First, I will put the data in R.

```
df <- data.frame(
  condition = c(rep("control", 4), rep("mindfulness", 4), rep("conversation", 4)),
  anxiety = c(10, 8, 6, 8, 4, 4, 2, 2, 5, 6, 5, 6)
)
```

```
print(df)
```

```
##       condition anxiety
## 1       control      10
## 2       control       8
## 3       control       6
## 4       control       8
## 5   mindfulness       4
## 6   mindfulness       4
## 7   mindfulness       2
## 8   mindfulness       2
## 9  conversation       5
## 10 conversation       6
## 11 conversation       5
```

```
## 12 conversation       6
```

For completeness, I will get descriptive statistics and create a plot, though I will not comment further on either.

```
# import packages
library(psych)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

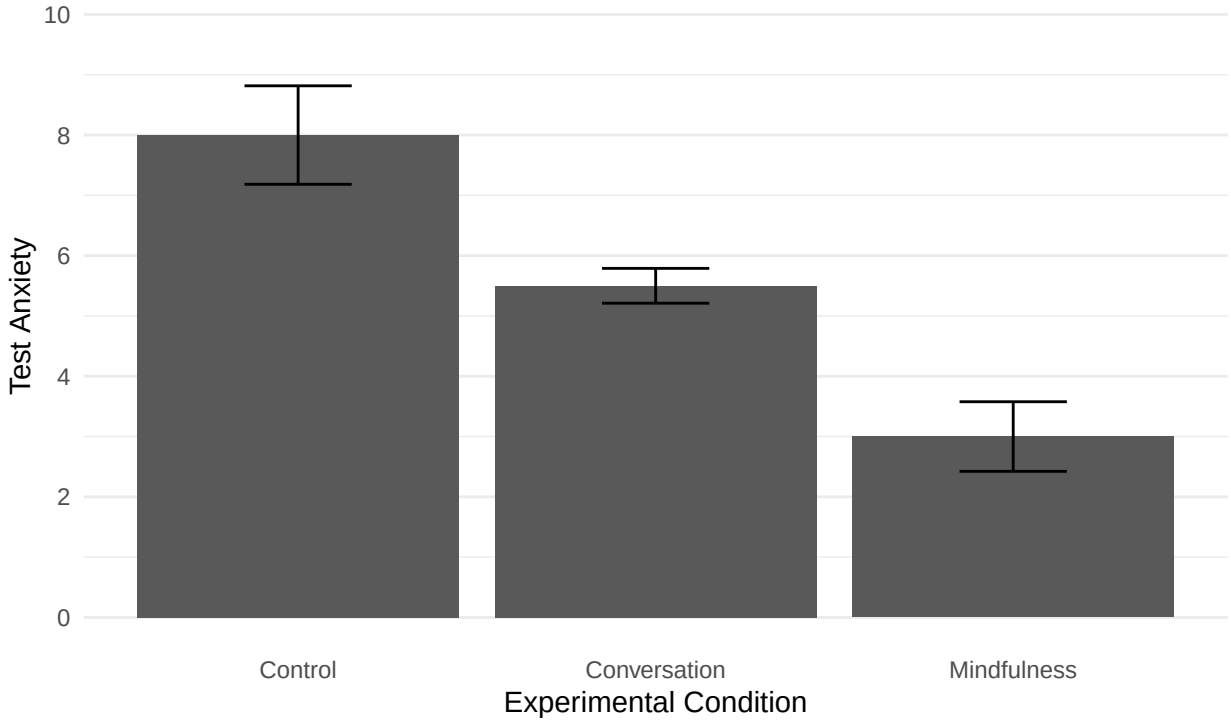
```
# descriptive statistics
describeBy(df$anxiety, group=df$condition)
```

```
##
##  Descriptive statistics by group
## group: control
##    vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 4    8 1.63      8       8 1.48   6  10     4    0    -1.88 0.82
## ------------------------------------------------------------
## group: conversation
##    vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 4  5.5 0.58    5.5     5.5 0.74   5   6     1    0    -2.44 0.29
## ------------------------------------------------------------
## group: mindfulness
##    vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 4    3 1.15      3       3 1.48   2   4     2    0    -2.44 0.58
```

```
# plot with additional customizations for future reference
ggplot(data = df, aes(x = condition, y = anxiety)) +
  stat_summary(fun = "mean", geom = "bar") +
  stat_summary(fun.data = "mean_se", geom = "errorbar", width = .3) +
  scale_x_discrete(name = "Experimental Condition",
                   labels = c("Control", "Conversation", "Mindfulness")) +
  scale_y_continuous(name = "Test Anxiety",
                     breaks = seq(0, 10, by=2),
                     limits = c(0,10)) +
  labs(title = expression(bold("Figure 1.")~"Comparison of Means"),
       subtitle = expression(italic("Note:")~"error bars represent 1 SE.")) +
  theme_minimal() +
  theme(plot.title = element_text(family = "Times", size = 12),
        plot.subtitle = element_text(family = "Times", size = 12),
        panel.grid.major.x = element_blank())
```

Figure 1. Comparison of Means

Note: error bars represent 1 SE.




```
# you can use theme(plot.title = element_text(hjust = 0.5)) to center the title
```

The ANOVA will be calculated in two steps. First, we will create the model. Second, we will show the results in a typical ANOVA table.

```
anova_model <- aov(anxiety ~ condition, data=df)
anova(anova_model)
```

```
## Analysis of Variance Table
##
## Response: anxiety
##           Df Sum Sq Mean Sq F value    Pr(>F)
## condition  2     50 25.0000  17.308 0.0008236 ***
## Residuals  9     13  1.4444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note a few things. First, there is some rounding error in the F-statistic, but the numbers otherwise match exactly. Second, instead of the language between and within, R uses the name of the variable to indicate between-group variability and residual to indicate the within-group variability.

Post hoc Tests

Now that we have a significant ANOVA, we must conduct follow-up tests to determine which groups are different from which groups. Typically, we are interested in all-pairwise comparisons and we use a correction. This is the same thing as conducting a separate t-test for each comparison, though the denominator will be different than a traditional t-test. If we didn't have a significant ANOVA, then it would be inappropriate to conduct post hoc testing.

I will show how to conduct post hoc tests using Bonferroni and Tukey corrections. I will also contrast this approach with conducting separate traditional t-tests with the same correction applied.

In R, we can get the Bonferroni follow-up tests using the following command.

```
pairwise.t.test(df$anxiety, df$condition,
                p.adjust.method="bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$anxiety and df$condition
##
##              control conversation
## conversation 0.0493  -
## mindfulness  0.0007  0.0493
##
## P value adjustment method: bonferroni
```

The resulting table shows three p-values, one for each comparison. As stated above, a Bonferroni correction typically involves dividing alpha. However, the same thing can be accomplished by multiplying the p-values by the number of tests conducted. To see that, let's run it again with no correction.

```
pairwise.t.test(df$anxiety, df$condition,
                p.adjust.method="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$anxiety and df$condition
##
##              control conversation
## conversation 0.01644 -
## mindfulness  0.00023 0.01644
##
## P value adjustment method: none
```

If we look at the conversation v. control comparison, we can see that the Bonferroni corrected p-value is $0.01644 \times 3 = 0.0493$ (with a small rounding error). This means that it is identical if we (1) run the comparisons with no correction and use an alpha of $.05/3 = .016$ or (2) use the Bonferroni correction above with an alpha of .05. In practice, you should always choose option 1 because it is conventional in our field.

Many researchers don't like using a Bonferroni correction because it is overly strict in controlling the experiment-wise alpha. There are many other options for post hoc tests, though the most common is the Tukey's HSD. This can be accessed in R using the following.

```
TukeyHSD(anova_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = anxiety ~ condition, data = df)
##
## $condition
##                           diff       lwr        upr     p adj
## conversation-control     -2.5 -4.872749 -0.1272515 0.0395847
## mindfulness-control      -5.0 -7.372749 -2.6272515 0.0006122
## mindfulness-conversation -2.5 -4.872749 -0.1272515 0.0395847
```

We would use an alpha of .05 for all Tukey comparisons. It is important to note that all p-values are smaller when using the Tukey HSD instead of a Bonferroni correction. In general, we would say that Bonferroni is a more conservative test, which means that it is less likely to lead to a type I error but will also be less likely to show a true difference in population means.

At this point, you should wonder what is different between conducting multiple t-tests v. conducting an ANOVA and then conducting the t-tests. In short, the significant ANOVA and the homogeneity of variance assumption allows us to use the pooled denominator from all three groups; whereas, a traditional t-test would use the pooled variance from only the two groups in that comparison. This is easier to see if we look at the formula.

A traditional t-test uses a pooled standard error from the two groups.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The post hoc testing uses the mean square within from the ANOVA (corrected by the sample size for the two groups in the comparison). Remember that this is pooled from all groups.

$$\sqrt{MS_{within}(\frac{1}{n_1} + \frac{1}{n_2})}$$

While this detail is technical and not necessarily important to memorize, I do think it is important to understand it broadly.

In our example, regardless of which correction we use, we can say that there is evidence that all population means are different. Thus, the students who performed mindfulness training experienced less test anxiety than the students who had a conversation or were in the control group. Further, the students in the conversation group were less anxious about the test than those in the control group.

Writing Results in APA Format

An one-way analysis of variance was conducted to test the effect of mindfulness on test anxiety. The test revealed a significant difference between the 3 groups, F (2, 9) = 17.31, $p < .001$. Post hoc testing revealed that participants in the mindfulness condition (M = 3.0, SD = 1.2) had less test anxiety than participants in the conversation condition (M = 5.5, SD = 0.6) and the control condition (M = 8.0, SD = 1.6). Participants in the conversation condition had significantly less test anxiety than those in the control condition.