Section 12
Analysis of Variance (ANOVA)

12.1 – Comparing Three or More Groups

Introduction

In the activity you completed this week, we investigated how we might make a comparison of three or more groups at once. We examined data on various oil supplements to determine if they had any significantly different impacts on reducing blood pressure. The goal of this was to come up with a way to test all of these groups simultaneously rather than look at all possible pairs. Why don't we just do many pairwise comparisons, that is, compute *t*-tests for every possible pair of groups? Two reasons:

- This can become very difficult to do quickly when the number of groups grows large. If there are 6 groups, that already means you have 15 comparisons to make, and this grows quickly!
- With the number of tests you would perform, this may result in a higher chance of rejecting incorrectly: with a 0.05 significance level, you would expect 1 in 20 independent tests to be rejected even if the null were true. While these aren't necessarily independent since the same groups are being re-used across pairs, it still becomes more likely to erroneously reject the null for some pair if you are comparing many groups or populations. (e.g., 5 groups = 10 possible pairs)

We derived our own statistic to measure overall differences each group had from the mean of all observations across all the groups. This likely involved looking at the sum of absolute value or squared differences from the overall mean. Generally, we will prefer looking at the sum of squared differences, as this is more calculus-friendly, and it also has the nice property of not weighing minor differences that could have been a product of random chance too heavily, relative to larger differences that get proportionally much larger when squaring.

The method we used in-class worked out well because we were comparing groups of equal sizes. While this is hopefully how we would design a study, you are likely to get groups of unequal sizes for many reasons. For example, in an experiment, subjects may drop out by the end of the study. In an observational study, some populations many be naturally larger than others, meaning it would be difficult to get groups of equal size to compare.

If we think about the implications of comparing groups of unequal sample sizes, there are a few things to consider: first, groups with larger sample sizes tend to have means that are more consistent (remember the central limit theorem!). Additionally, if there were a group mean based on only a few observations, it would get equal consideration as a group with a much larger sample size if we were to just take each group mean's difference from the overall mean.

To generalize the methods we derived in class earlier this week, we will derive the method of analysis of variance (ANOVA) in this section. ANOVA generalizes group comparisons when you have more than 2 groups. That is, it tests the following hypotheses:

H_0:

H_a:

Like the informal test we computed in TinkerPlots this week, ANOVA will be decent at identifying if there are differences that exist between groups, but not necessarily what groups differ. We will also examine methods that will illuminate specific between-group differences within the ANOVA test.

The one-way ANOVA model

Say we've collected some data from several groups. That is, we have the following data from k populations or groups:

$$\text{Population 1: } X_{11}, X_{12}, \ldots, X_{1n_1}$$

$$\text{Population 2: } X_{21}, X_{22}, \ldots, X_{2n_2}$$

$$\vdots$$

$$\text{Population } k\text{: } X_{k1}, X_{k2}, \ldots, X_{1n_k}$$

Note that the groups don't need to have equal sample sizes, hence the use of different subscripts for the last data point. Our goal is to determine if the means of the populations or groups are significantly different or not. To determine this, we can conceive of a model similar to how we built the regression model:

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Here, we define μ to be the population mean across all groups combined, α_i to be the _____ of group i, and similarly to regression, we define ε_{ij} as the _____ for data value X_{ij}. Note that the way we define α_i gives us that μ_i, the population mean of group i is defined to be $\mu + \alpha_i$. Thus, we have defined a model that gives us data in terms of their population's respective mean, plus some error term.

Calculating an ANOVA test

So far, there has been no reason to believe that this method should be called "Analysis of Variance" given that the main focus of this method is via examining means. However, we can do a population mean comparison by looking at the variability now that we have more than two population means to compare. We can examine the variability of those means themselves and see how their variability compares to the variability of the data itself within each group. First, we define the total variability of our data, that is, the total sum of the squared error terms. Here, we define this in terms of the data rather than the model:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

We define $\bar{X}$ above as the overall sample mean, combining our data from all groups together:

$$\bar{X} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}}{N}, \text{where } N = \sum_{i=1}^{n} n_i$$

Now, let's get at the two sources of variability that we want to compare: the variability among the means themselves, and the variability of the data within each group. We define these each as the _____ (SSG) and the _____, or due to error (SSE):

$$SSG = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 \qquad SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

Here, we define $\bar{X}_i$ to be the sample mean for group *i*. Through carefully deconstructing the sums above, it can be shown that SSG + SSE = SST, which illustrates that we have decomposed the total variability into two sources. SSG is very similar to the statistics we collected on in TinkerPlots earlier this week, but now it is scaled by the sample size to weigh larger groups more heavily.

Going back to our initial intuition, we cannot simply compare the two values here, as it's not a fair comparison given that we can have data sets of many sizes. We might have 1000 total data values from 3 populations, or just 150 total data values from 6 populations; the former would have a relatively high SSE by just having more data values, where the latter may have a relatively high SSG due to having more groups. Thus, we should standardize these sums to these sizes.

The way we do this is via dividing by the degrees of the freedom for the test that we will be conducting, which gives us the mean squares between groups, and mean squares due to error.

$$MSG = \frac{SSG}{k-1} \qquad MSE = \frac{SSE}{N-k}$$

Now that we've standardized these measures of variability, we can compute a test-statistic to see if the variability between the group means (MSG) is too large relative to how the data varies within their groups (MSE). A relatively large MSG would be evidence that there do in fact exist differences between the groups relative to their means. Thus, to get a single value that carries this comparative idea, we can take the ratio of these two measures of variability. This is known as an F-test statistic:

$$F = \frac{MSG}{MSE}$$

This may seem a bit predictable, but the F-statistic follows an F distribution. The F distribution is only defined on positive values and is skewed to the right, which makes sense since F is a ratio of measures of variability, which must be positive.

The shape of an F distribution can vary wildly depending on its degrees of freedom:

In the context of ANOVA, the F distribution we will be concerned with will be F(k – 1, N – k), that is, the F distribution on k – *1* and N – k degrees of freedom, as this is the specific F distribution that the ANOVA F-test statistic follows. To compute probabilities on the F distribution in R, you can use the following function:

```
pf(F, df1, df2)
```

As the F-test statistic will always get larger as the means get more different, you will always want the complement of this probability as the *p*-value.

This applet is helpful in visualizing the relationship between the variation within groups relative to the variations between the groups themselves, and how that impacts the F-statistic. You can adjust the MSE by the "standard deviation" slider, and you can adjust the MSG by moving the black dots to adjust the group means to be more similar or more different.

12.2 – One-way ANOVA

The ANOVA table

It's common for results of ANOVA analyses to be summarized in a table. For the basic type of ANOVA we've described, the table is typically formatted as shown below:

	DF	SS	MS	F	P-value
Groups	$k-1$	SSG	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	`1 - pf(F, k-1, N-k)`
Error	$N-k$	SSE	$MSE = \frac{SSE}{N-k}$		
Total	$N-1$	SST			

Example: The critical flicker frequency (cff) is the highest frequency (in cycles/second) at which a person can detect the flicker in a flickering light source. At frequencies above the cff, the light appears to be continuous even though it is flickering. A study in the Journal of General Psychology aimed to assess if the color of a person's iris (Brown, Green, or Blue) has an effect on their cff. Researchers took independent random samples of people with each eye color: 8 with brown eyes, 5 with green eyes, and 6 with blue eyes. Assume that the necessary assumptions for ANOVA are met. A partial ANOVA table is given below – complete the table and conduct this test at level $\alpha = 0.05$.

	DF	SS	MS	F	P-value
Groups		22.99			
Error		38.31			
Total		61.30			

In that example, we assumed that the assumptions of conducting ANOVA were met, but did not yet discuss those assumptions. Those four assumptions required to conduct ANOVA are:

1. Each sample of data obtained is a random sample taken from its respective population.
2. Each sample of data we have was sampled independently of the other samples.
3. Each sample of data comes from a population that is normally distributed (can be overridden by $n_i \geq 25$ for each i)
4. Each sample of data comes from a population with the same standard deviation.

All of these assumptions are not too surprising, maybe except for the last one. The reason for this is that we calculated MSE in a way that computes the squared distance of each data point from its group's mean, and combined these squared terms together to measure variability. This implicitly assumes that the groups have equal variability, since this is the single measure used as a baseline for how much the data normally varies within groups – we don't use a different measure of variability within each group.

To check this assumption, there are many methods that can be used to verify this. One such method is called Levene's test, which tests the following hypotheses:

H_0:

H_a:

Thus, in a weird twist of fate, coming to the decision of _____ means that this assumption is valid. The typical level of significance used for this test is $\alpha = 0.10$. We will see how to conduct Levene's test in R later.

Post-hoc analysis

As mentioned previously, ANOVA can only tell you that differences exist, and does not provide information about which groups differ. After conducting an ANOVA test and finding significant differences between groups, a post-hoc analysis can be conducted to determine which groups actually differ. Many post-hoc methods exist, with one example being Tukey's multiple comparisons. This does two-sample, two-sided tests on every possible pair of groups to determine possible differences.

Important note: these two-sample tests use adjusted *p*-values to account for the fact that you are testing multiple hypotheses simultaneously. Remember: the more things you test, the higher the chance that you incorrectly reject the null hypothesis. The Tukey method raises the *p*-values appropriately to account for this, and it is noted in the table with the column label "`p adj`" for adjusted *p*-values.

Example: For the cff example above, the following Tukey comparison results were given from R. Summarize the results of this analysis using level $\alpha = 0.05$.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = cff ~ eye, data = cff)

$`eye`
              diff   lwr   upr   p adj
Brown-Green  -1.33 -3.61  0.94 0.07944
Brown-Blue   -2.58 -4.74 -0.42 0.00485
Green-Blue   -1.25 -3.66  1.17 0.10815
```

Conducting one-way ANOVA in R

Example: In a large lecture, 200-level class, there are 115 students in class on a particular day. The instructor noted each student's general location in the classroom as one of four categories: Front, Mid-Front, Mid-Back, and Back. The instructor also looked up each student's current class standing and GPA. This data is stored in the file **seating.csv**. Carry out an ANOVA test to see if there are differences in the mean GPA for students that sit in differing regions of the lecture hall using level $\alpha = 0.05$, and determine which pairs of seating groups differ.

Check assumptions:

```
table(seating$Seat)
front = subset(seating, Seat=="Front")
qqnorm(front$GPA)
qqline(front$GPA)
```

```
library(car)
leveneTest(GPA~Seat, data=seating, center="mean")
```

Create model:

```
model = aov(GPA~Seat, data=seating)
anova(model)
```

Post-hoc:

```
TukeyHSD(model)
```

12.3 – Two-way ANOVA

Theoretical model

In the previous example, we didn't use the information about students class level at all. We could though investigate if there are differences according to the students' seat location and their class level simultaneously. This introduces two factors into our model, hence we are doing two-way ANOVA and making a comparison of our response variable across two factors simultaneously.

We can write out the statistical model for two-way ANOVA as follows:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

Like before, we define μ to be the population mean across all groups combined. We call α_i and β_j the _____ for group i of the first factor and group j for the second factor, respectively. We add a new _____ term to this model γ_{ij}, and once again, we define ε_{ijk} as the _____ for data value $X_{ijk.}$

Through this setup, we can test hypotheses across the main effects and the interaction effects. This gives us three different null hypotheses to test:

H_0^A:

H_0^B:

H_0^{AB}:

These effects can similarly be summarized via an ANOVA table, which has more rows to consider the three tests being conducted as part of a two-way ANOVA. For the table below, say there are a different levels of Factor A, and b different levels of Factor B.

	DF	SS	MS	F	P-value
Factor A	$a - 1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$	`1 - pf(Fa, a-1, N-ab)`
Factor B	$b - 1$	SSB	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$	`1 - pf(Fb, b-1, N-ab)`
Interaction AB	$(a - 1)(b - 1)$	SSAB	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$	`1 - pf(Fab, (a-1)*(b-1), N-ab)`
Error	$N - ab$	SSE	$MSE = \frac{SSE}{N-ab}$		
Total	$N - 1$	SST			

Building a two-way ANOVA model

Since there are now multiple effects and factors to tease out, we need to use variable selection as we did for regression. One general procedure used in two-way (or n-way) ANOVA is as follows:

1. Run model with interaction term (or all interaction terms, if going beyond two-way)
2. If the (highest order) interaction term is significant, keep it in the model. Otherwise remove.
3. If removed, re-run without that interaction and repeat step 2 until a significant interaction is present.

Example: Re-run the example regarding GPA and seat location from before, adding in class standing as a second factor of a two-way ANOVA mode. Are there any significant main effects from the class standing or from the interaction effects at level $\alpha = 0.05$?

```
model1 = aov(GPA~Seat*Class, data=seating)
anova(model1)
```

```
model2 = aov(GPA~Seat+Class, data=seating)
anova(model2)
```

It's important to note that if an interaction term is significant in the model despite the main effects being insignificant, we can still conclude that the main effects have an impact on the response variable. Main effects describe only the average or overall effect, but the interaction between the two can highlight that they still have an impact. The following is an example of such an effect:

Example: In an experiment to determine ways to treat people with Seasonal Affective Disorder (SAD), researchers randomly assigned 150 participants with similar levels of depression to one of two physical programs: aerobics and meditation, as well as replaced all home lighting with one of three types of lights: fluorescent, LED, and a natural lighting replica. Each subject was given the Beck Depression Instrument after the study, a questionnaire that measures the depression levels of patients. Data for this study can be found in the **SAD.csv** file. Determine if these two factors significantly impact the patients' average depression level.

```
smodel = aov(BDI~Lighting*Condition, data=SAD)
```

```
aggregate(BDI~Condition, data=SAD, mean)
aggregate(BDI~Lighting, data=SAD, mean)
aggregate(BDI~Lighting+Condition, data=SAD, mean)
```

Means	Fluorescent	LED	Natural	Overall
Meditation				
Aerobics				
Overall				

12.4 – Additional Practice

Example: Students who have recently completed their first college course in statistics were recruited for a teaching experiment on building statistical learning models. These students were assigned to one of two lessons: one that was lecture-based, and one that involved an interactive activity. The students were then asked to complete an assessment after the lesson, which was scored based on their correct answers. They were also asked about their motivation for learning statistics (low, mid, or high). The researchers conducting this experiment are interested in determining what factors are important in determining the students scores on the assessment. The data for this study can be found in the file **teaching.csv.**

We will first determine if the motivation level impacted students' scores on the assessment. Check the assumptions for conducting an ANOVA test on this data.

Conduct a one-way ANOVA test to determine if there are differences in the assessment scores based on students' motivation.

Conduct a post-hoc analysis to determine which pairs of groups are different at a 5% level. Use a connected lines visualization to illustrate which groups are determined to be similar/different.

Conduct a two-way ANOVA to determine how both the teaching style and motivation affect the assessment score. Should an interaction term be included?

Find the average score for each pair of teaching style and motivation. How do these means explain a possible interaction effect?