Introduction to ANOVA

Lamb Weight Gain Example from Text
The following table contains fictitious data on the weight gain of lambs on three different diets over a 2 week period.

<table>
<thead>
<tr>
<th>Weight Gain (lbs.)</th>
<th>Diet 1</th>
<th>Diet 2</th>
<th>Diet 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What is a question of interest?

How do we analyze this data? We have independent samples so why not use independent samples t-tests? Answer: Using the $t$-distribution to make more than one comparison of a pair of independent samples drives up the chance for error.

Recall: The Type I error rate (the probability of going with the alternative when we shouldn’t) of a $t$-test is the significance level, $\alpha$.

The lamb weight data is comprised of three independent samples. How many pair-wise comparisons can we make?

Then, we can make a Type I Error in any or all of these comparisons. Let’s look at what happens to the probability of at least one Type I Error when making multiple comparisons:

More on what to do about this error rate problem later...for now, we introduce the ANOVA.
A one-way analysis of variance, or just “ANOVA”, that we’ll be learning is a hypothesis testing procedure that uses the following hypotheses:

\[ H_0: \]

\[ H_A: \]

The term “one-way” refers to the fact that there is only one variable defining the groups (in our example this is “Diet”).

Notation:

- \( I \) = number of groups
- \( i \) denotes the \( i^{th} \) group and \( j \) denotes the \( j^{th} \) observation
  - \( y_{ij} = y_{12} \) denotes the 2\(^{nd} \) observation in the first group
- \( n_i \) = sample size for the \( i^{th} \) group
- \( \bar{y}_i \) = sample mean for the \( i^{th} \) group
- \( n_* = \sum_{i=1}^{I} n_i \) (the total sample size across all groups)
- \( \bar{y} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}}{n_*} \) (sample mean combining data across all groups)

**Sum of Squares (SS), Degrees of Freedom (df), and Mean Squares (MS)**

\[ SS(\text{within}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{I} (n_i - 1) s_i^2 \quad \text{MS(\text{within})} = \frac{SS(\text{within})}{df(\text{within})} \]

\[ df(\text{within}) = n_* - I \]

\[ SS(\text{between}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2 \quad \text{MS(\text{between})} = \frac{SS(\text{between})}{df(\text{between})} \]

\[ df(\text{between}) = I - 1 \]

\[ SS(\text{total}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \quad \text{MS(\text{total})} = \frac{SS(\text{total})}{df(\text{total})} \]

\[ df(\text{total}) = n_* - 1 \]
Consider the deviation from an observation to the overall mean written in the following way:

\[ y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) \]

Notice that the left side is at the heart of SS(total), and the right side has the analogous pieces of SS(within) and SS(between). It actually works out (with a bit of math) that

\[
\sum_{i=1}^{l} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{l} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{l} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2
\]

\[ SS(\text{total}) = SS(\text{within}) + SS(\text{between}) \]

The analysis of variance is centered around this idea of breaking down the total variation of the observations from their grand mean into its two pieces: the variation within groups (error variation) and the variation between groups (treatment variation).

**The F test for ANOVA**

The test statistic for the ANOVA is

\[ F_s = \frac{MS(\text{between})}{MS(\text{within})} \]

If the data indicates large differences in group means compared to each other relative to the variability within groups around group means, F_s will be large. Big values of F_s indicate evidence against H_0 (if the group means are the same, the variability between them should not be larger than the natural variation of the data within each group).

The F distribution has degrees of freedom for the numerator (between) and the denominator (within). We say F_s ~ F(v_1, v_2). The following picture from Wikipedia illustrates a few F pdf’s.

P-values for the F test in ANOVA are tail area quantities that will be calculated for you.
Software packages performing the $F$ test for ANOVA return something called an “ANOVA table”. The following is the ANOVA output from Minitab 16 for the lamb weight data. Notice where the numbers in the table come from.

**One-way ANOVA: Diet 1, Diet 2, Diet 3**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>2</td>
<td>36.0</td>
<td>18.0</td>
<td>0.77</td>
<td>0.491</td>
</tr>
<tr>
<td>Error</td>
<td>9</td>
<td>210.0</td>
<td>23.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>246.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$S = 4.830$ $R$-Sq = 14.63% $R$-Sq(adj) = 0.00%

Individual 95% CIs For Mean Based on Pooled StDev

| Level | N   | Mean | StDev | --------+---------+---------+---------+--|
|-------|-----|------|-------|--------+---------+---------+---------+--|
| Diet 1| 3   | 11.0 | 4.359 | (---------------*--------------)
| Diet 2| 5   | 15.0 | 4.950 | (---------------*--------------)
| Diet 3| 4   | 12.0 | 4.967 | (---------------*--------------)
|       |     | 8.0  | 12.0  | 16.0  | 20.0  |

Pooled StDev = 4.830

**ANOVA in R**

```r
weightgain<-c(8,16,9,9,16,21,11,18,15,10,17,6)
diet<-c(1,1,1,2,2,2,2,2,3,3,3,3)
diet<-factor(diet)
> summary(aov(weightgain~diet))

Df   Sum Sq Mean Sq   F value   Pr(>F)
diet     2   36.0  18.00     0.771     0.491
Residuals 9 210.0  23.33
```

Conduct an ANOVA for the lamb data.
A few final notes on the one way ANOVA:

It assumes a common standard deviation for the populations.

Notice that \( MS(\text{within}) = MS(\text{error}) = \text{MSE} = \frac{\sum_{i=1}^{I}(n_i-1)s_i^2}{n-1} \) is a pooled (weighted or average) variance for the groups. We view this MS(error) as an estimate of the population common variance across the treatments. Then, the estimate of the common population standard deviation is

And this is often referred to as

We check the assumption that the variation is the same for all groups before carrying out an ANOVA, possibly by looking at side-by-side boxplots (among other ways).

There are other assumptions (and methods of checking them) we will discuss in the Regression / Correlation lecture notes.

So, what do we do if the ANOVA F test rejects \( H_0 \) and we conclude there is at least one population mean that is different? Then, we can turn to the methods under a blanket of topics called “multiple comparisons” where we make pairwise comparisons to determine which treatments are significantly larger (or smaller) than the others. Healthy study is given to this set of theories, as careful attention must be paid to the error rates. Section 11.9 of your text gives an introduction to this topic.