1-Way ANOVA

MATH 143

Department of Mathematics and Statistics
Calvin College
An example ANOVA situation

Example (Treating Blisters)

- Subjects: 25 patients with blisters
- Treatments: Treatment A, Treatment B, Placebo (P)
- Measurement: # of days until blisters heal
- Data [and means]:
  - A: 5,6,6,7,7,8,9,10 [7.25]
  - B: 7,7,8,9,9,10,10,11 [8.875]
  - P: 7,9,9,10,10,10,11,12,13 [10.11]

Question: Are these differences significant?
(Or would we expect differences this large by random chance?)
Are these differences significant?

Whether differences between the groups are significant depends on

- the difference in the means
- the amount of variation within each group
- the sample sizes
Some notation for ANOVA

Entire data set:

- $n =$ number of individuals all together
- $k =$ number of groups
- $y_{ij} =$ value for individual $i$ in group $j$
- $\bar{y} =$ sample mean of quant. variable for entire data set
- $s =$ sample s.d. of quant. variable for entire data set

Group $j$:

- $n_j =$ # of individuals in group $j$
- $\mu_j =$ population mean for group $j$
- $\bar{y}_j =$ sample mean for group $j$
- $\sigma_j =$ population standard deviation for group $j$
- $s_j =$ sample standard deviation for group $j$
The ANOVA model

\[ y_{ij} = \mu_j + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma) \]

DATA \quad FIT \quad ERROR

Relationship to Regression:

- FIT gives mean for each level of categorical variable, but the means need not fall in a pattern as in regression.
- ANOVA can be thought of as regression with a categorical predictor.

Assumptions of this model: Each group . . .

- is normally distributed about its group mean (\( \mu_j \))
- has the same standard deviation
The ANOVA model (take 2)

\[ y_{ij} = \mu + \tau_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma) \]

- The time we express FIT as an overall average (\( \mu \)) plus a treatment effect (\( \tau_j \)) for each group.

Assumptions of this model still the same: Each group . . .
- is **normally distributed** about its group mean (\( \mu + \tau_j \))
- has the **same standard deviation**
Checking the assumptions

Once again we look at residuals:

\[ \text{residual}_{ij} = e_{ij} = y_{ij} - \bar{y}_j \]

- **Normality Check:**
  - normal quantile plots of residuals (boxplots, histograms)
  - ANOVA can handle some skewness, but is quite sensitive to outliers

- **Equal Variance Check:**
  - \( s_j \)'s should be approximately equal
  - Rule of Thumb: \( \leq 2:1 \) ratio between largest and smallest of group sample standard deviations
    - 4:1 ratio between variances is equivalent

- Remember that the assumptions are about the population, not the sample.
  - These are hard to check when the data sets are small.
What does ANOVA do?

At its simplest (there are extensions) ANOVA tests the following hypotheses:

- $H_0$: The means of all the groups are equal.
  - $\mu_1 = \mu_2 = \cdots = \mu_k$
  - $\tau_1 = \tau_2 = \cdots = \tau_k = 0$

- $H_a$: Not all the means are equal.
  - doesn’t say how or which ones differ
  - can follow up with “multiple comparisons” if we reject $H_0$
A quick look at the ANOVA table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>SS</th>
<th>MS</th>
<th>F-stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>2</td>
<td>34.74</td>
<td>17.37</td>
<td>6.45</td>
<td>0.0063</td>
</tr>
<tr>
<td>Residuals</td>
<td>22</td>
<td>59.26</td>
<td>2.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>94.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The ANOVA test statistic (ingredients)

ANOVA measures two sources of variation in the data and compares their relative sizes

- variation BETWEEN groups
  - for each data value look at the difference between its group mean and the overall mean

\[
SS_{Tr} = \sum (\bar{y}_j - \bar{y})^2 = \sum (\text{response} - \text{group mean})^2
\]

\[
MS_{Tr} = \frac{SS_{Tr}}{k - 1}
\]

- variation WITHIN groups
  - for each data value we look at the difference between that value and the mean of its group

\[
SS_E = \sum (y_{ij} - \bar{y}_j)^2 = \sum (\text{residual})^2
\]

\[
MS_E = \frac{SS_E}{n - k}
\]
The ANOVA test statistic (example)

We want to measure and compare the amount of variation due to BETWEEN group variation and WITHIN group variation:

So for each data value, we calculate its contribution to:

- BETWEEN group variation: \((\bar{y}_j - \bar{y})^2\)
- WITHIN group variation: \((y_{ij} - \bar{y}_j)^2\)

Example (An Even Smaller Dataset)

Suppose we have three groups

- Group 1: 5.3, 6.0, 6.7 \[\bar{y}_1 = 6.00\]
- Group 2: 5.5, 6.2, 6.4, 5.7 \[\bar{y}_2 = 5.95\]
- Group 3: 7.5, 7.2, 7.9 \[\bar{y}_3 = 7.53\]
Computing $F$

overall mean: 6.44; $F = \frac{2.5528}{0.25025} = 10.21575$
Connections between $SS_{Tot}$, $MS_{Tot}$, and standard deviation

Ignoring groups for a moment:

$$s_y^2 = \frac{\sum(y_{ij} - \bar{y})^2}{n - 1}$$

$$= \frac{SS_{Tot}}{DF_{Tot}} = MS_{Tot}$$

So $MS_{Tot} = s_y^2$ and $SS_{Tot} = (n - 1)s_y^2$.

These measure the TOTAL variation (of $y$).
Connections between $SS_E$, $MS_E$, and standard deviation

If we just look at the data from one of the groups:

\[ s_j^2 = \frac{\sum_j (y_{ij} - \bar{y}_j)^2}{n_j - 1} = \frac{SS[\text{within group } i]}{DF_j} \]

So $SS[\text{within Group } i] = (s_j^2)(DF_j)$

This means

\[
SS_E = SS[\text{within}] = \sum_j SS[\text{within group } i]
\]

\[
= \sum_j s_j^2 (n_j - 1)
\]

\[
= \sum_j s_j^2 \cdot DF_j
\]
Pooled estimate for $\sigma$

One of the ANOVA assumptions is that all groups have the same standard deviation. We can estimate this with a weighted average:

$$s_p^2 = \frac{DF_1 \cdot s_1^2 + DF_2 \cdot s_2^2 + \cdots + DF_k \cdot s_k^2}{DF_1 + DF_2 + \cdots + DF_k}$$

$$s_p^2 = \frac{SS_E}{DF_E} = MS_E$$

So $MS_E$ is the pooled estimate of variance parameter $\sigma^2$.

- If we do this with only two groups, we get the 2-sample $t$-test with pooled estimate of variance (the check box in StatCrunch that we have been turning off).
- A similar thing can be done to estimate $\sigma$ in the simple linear regression model.
Tukey’s Honest Significant Differences

One method of adjusting the simultaneous confidence intervals for all the pairs is called Tukey’s method.

Tukey 95% Simultaneous Confidence Intervals

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>A subtracted from B</td>
<td>0.4365</td>
<td>3.6865</td>
</tr>
<tr>
<td>A subtracted from P</td>
<td>0.8577</td>
<td>4.8645</td>
</tr>
<tr>
<td>B subtracted from P</td>
<td>0.7673</td>
<td>3.2395</td>
</tr>
</tbody>
</table>

- This corresponds to 98.01% CIs for each pairwise difference.
- Only P vs A is significant in our example (lower and upper limits have the same sign).
Other views of Multiple Comparisons

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

diff lwr upr
B-A 1.6250 -0.43650 3.6865
P-A 2.8611 0.85769 4.8645
P-B 1.2361 -0.76731 3.2395

Individual 95% CIs For Mean
Based on Pooled StDev

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>7.250</td>
<td>1.669</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>8.875</td>
<td>1.458</td>
</tr>
<tr>
<td>P</td>
<td>9</td>
<td>10.111</td>
<td>1.764</td>
</tr>
</tbody>
</table>

Pooled StDev = 1.641

7.5 9.0 10.5