3.0 ANOVA - ANALYSIS OF VARIANCE
(updated Spring 2001)

ANOVA is extensively used in the science and engineering.

Study the coagulation time for samples of blood drawn from 24 animals receiving 4 different diets. Does diet affect blood coagulation time? Twenty four (24) animals were randomly allocated to four diets (A,B,C, and D). The blood samples were taken and tested in random order.

The analysis that follows is exactly justified if the data are random samples from four normal populations, and is approximately justified based on the randomization dist\textsuperscript{b} idea presented previously.

We will assume that the four populations have equal variance, \( \sigma^2 \). Is there evidence to indicate real differences between the diets/treatments?

\[ H_0: \mu_A = \mu_B = \mu_C = \mu_D \]
\[ H_A: \text{At least one of the means differs from the others.} \]

In the following table, test order is listed in the parenthesis.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>63 (12)</td>
<td>62 (20)</td>
<td>68 (16)</td>
<td>56 (23)</td>
</tr>
<tr>
<td>67 (9)</td>
<td>60 (2)</td>
<td>66 (7)</td>
<td>62 (3)</td>
</tr>
<tr>
<td>71 (15)</td>
<td>63 (11)</td>
<td>71 (1)</td>
<td>60 (6)</td>
</tr>
<tr>
<td>64 (14)</td>
<td>59 (10)</td>
<td>67 (17)</td>
<td>61 (18)</td>
</tr>
<tr>
<td>65 (4)</td>
<td>68 (13)</td>
<td>63 (22)</td>
<td></td>
</tr>
<tr>
<td>66 (8)</td>
<td>68 (21)</td>
<td></td>
<td>64 (19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>63 (5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>59 (24)</td>
<td></td>
</tr>
</tbody>
</table>

\[ n_A = 6 \quad n_B = 4 \quad n_C = 6 \quad n_D = 8 \]

Treat. Avg. \( \bar{Y}_A = 66 \quad \bar{Y}_B = 61 \quad \bar{Y}_C = 68 \quad \bar{Y}_D = 61 \)

Grand Avg. 64

Are the differences between the treatment averages greater than expected given the variation within treatments?
• Variation within treatments

\[
s^2_A = \frac{1}{n_a - 1} \sum_{i=1}^{N_A} (Y_{Aj} - \bar{Y}_A)^2 = \frac{S_A}{v_A} = \text{Sum of squares within treatment } A \\
\text{Degree of freedom for treatment } A
\]

\[
s^2_A = \frac{(63 - 66)^2 + (67 - 66)^2 + \ldots + (66 - 66)^2}{6 - 1} = \frac{S_A}{v_A} = \frac{40}{5} = 8
\]

Similarly,

\[
s^2_B = \frac{S_B}{v_B} = \frac{10}{3} = 3.33
\]

\[
s^2_C = \frac{S_C}{v_C} = \frac{14}{5} = 2.8
\]

\[
s^2_D = \frac{S_D}{v_D} = \frac{48}{7} = 6.857
\]

Since \( s^2_A, s^2_B, s^2_C, \) and \( s^2_D \) all provide estimates of the unknown variance, \( \sigma^2_y \). They may be combined to provide a pooled estimate.

\[
s^2_p = \frac{v_A s^2_A + v_B s^2_B + v_C s^2_C + v_D s^2_D}{v_A + v_B + v_C + v_D} = \frac{S_A + S_B + S_C + S_D}{v_A + v_B + v_C + v_D}
\]

We’ll refer to this sample variance as the within treatment sample variance.

\[
s^2_w = \frac{S_w}{v_w} = \frac{40 + 10 + 14 + 48}{5 + 3 + 5 + 7} = \frac{112}{20} = \frac{\text{Within Treatment Sum of Squares}}{\text{Within Treatment Degree of Freedom}}
\]

\( s^2_w = 5.6 \). It is the within treatment mean square estimate of the true variance, \( \sigma^2_y \).

**One-Way ANOVA**

In general,
Variation Between Treatments

If there is no difference between treatment means, a second estimate of $\sigma_y^2$ can be obtained based on the variation of the treatment means about $\bar{y}$.

If we calculate

$$\sum_{OverAllTreatments} \frac{(\bar{Y}_{treatment} - \bar{Y})^2}{#Treatments - 1},$$

it can be used as an estimate of $\sigma_y^2$.

But we want an estimate of $\sigma_y^2$. Since $\sigma_y^2 = n\sigma_F^2$, let’s scale

$$(\bar{Y}_{treatment} - \bar{Y})^2$$

by $n_{treat}$.

So, in general,

$$\sum_{i=1}^{k} n_i (\bar{Y}_{treatment} - \bar{Y})^2 = \frac{s_{BT}^2}{\nu_B} = \frac{S_B}{\nu_B}$$

where $k$ is the number of treatments. It is the Between Treatment estimate of $\sigma_y^2$, or, the Between Treatment Mean Square.

For our Problem,

<table>
<thead>
<tr>
<th>Treatment</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{Y}_i$</td>
<td>66</td>
<td>61</td>
<td>68</td>
<td>61</td>
</tr>
<tr>
<td>$\bar{Y}$</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>$(\bar{Y}_i - \bar{Y})$</td>
<td>2</td>
<td>-3</td>
<td>4</td>
<td>-3</td>
</tr>
<tr>
<td>$n_i$</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>
$S_B = 6(2)^2 + 4(-3)^2 + 6(4)^2 + 8(-3)^2 = 228, \nu_B = 3,$
and $s^2_B = \frac{S}{\nu_B} = \frac{228}{3} = 76.0$

**Summary:**

Within Treatments:  
**Sum of Squares:** $S_W = 112$  
**Degrees of Freedom:** $\nu_W = 20$  
**Mean Square:** $s^2_W = 5.6$

Between Treatments:  
**Sum of Square** $S_B = 228$  
**Degrees of Freedom** $\nu_B = 3$  
**Mean Square** $s^2_B = 76.0$

Under $H_0$, both $s^2_W$ and $s^2_B$ are estimates of $\sigma^2_y$

If there is a difference between treatment means, the between treatment variance will be inflated. If a sample of size $n$ is drawn from a normal population with variance $\sigma^2$, The quantity,

$$\frac{s^2(n-1)}{\sigma^2}$$

follows the $\chi^2$ distribution. If $\sigma^2$ is postulated, $\chi^2_{calc} = \frac{s^2(n-1)}{\sigma^2}$ can be computed and compared to the existing values. If $\sigma^2$ is not postulated, a $100(1-\alpha)\%$ confidence interval for $\sigma^2$ can be calculated.

$$\frac{s^2(n-1)}{\chi^2_{\nu,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{\nu,\frac{\alpha}{2}}}$$
Sampling distribution of Two Variances

If sample of size $n_1$ is drawn from normal distribution with a variance of $\sigma_1^2$, and sample of size $n_2$ is drawn from normal distribution with a variance of $\sigma_2^2$, estimates of the population variances may be calculated: $s_1^2$ and $s_2^2$.

The quantity $\frac{s_1^2}{\sigma_1^2}$ is distributed and $\frac{s_2^2}{\sigma_2^2}$ is distributed. The ratio

$$\frac{\chi^2_{v_1}}{v_1} \sim F_{v_1, v_2}$$

follows the F distribution with $v_1, v_2$ degrees of freedom.

Therefore,

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{v_1, v_2}$$

or

$$\frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} \sim F_{v_1, v_2}$$

If $\sigma_1^2 = \sigma_2^2$, then

$$\frac{s_1^2}{s_2^2} \sim F_{v_1, v_2}$$

For the blood coagulation effect problem, $s_w^2 = 5.6$ ($v_w=20$) and $s_B^2 = 76.0$ ($v_B=3$).

Under $H_0$, both are estimates of $\sigma_y^2$, and $F_{calc} = \frac{s_B^2}{s_w^2} = 13.57$. Note
that this ratio should be distributed according to \( F_{v_1, v_2} = 3, v_2 = 20 \) under \( H_0 \).

From a \( F \)-distribution table, it is found that \( F(3, 20, .95) = 3.10 \) and \( F(3, 20, .99) = 4.94 \). Since both numbers are less than \( F_{\text{calc}} \), strong evidence shows that \( F_{\text{calc}} \) is not a typical value. This means that \( s_B^2 \) is inflated by differences between \( \mu_A, \mu_B, \mu_C, \) and \( \mu_D \). Therefore, reject \( H_0 \). At least one of the means is different.

To confirm ANOVA results, we would like to display \( \bar{Y} \)'s on their reference distribution. It may be difficult, since each \( \bar{Y} \) is based on different Degrees of Freedom (\( \sigma^2_y \) differs from each other).

\[
\bar{Y} = 64, \; s_y^2 = 5.6, \; s_y = 2.366.
\]

We need to calculate a \( t \) for each \( \bar{Y} \).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>66</td>
<td>61</td>
<td>68</td>
<td>61</td>
</tr>
<tr>
<td>( n )</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>( s_{\bar{Y}} )</td>
<td>0.966</td>
<td>1.183</td>
<td>0.966</td>
<td>0.837</td>
</tr>
<tr>
<td>( t_{\text{calc}} )</td>
<td>2.070</td>
<td>-2.536</td>
<td>4.141</td>
<td>-3.584</td>
</tr>
</tbody>
</table>

The four calculated \( t \) points are plotted on a \( t_{20} \) distribution plot. From the plot, we can ask if these four \( t \) values be drawn randomly from a \( t_{20} \) distribution.

**Decomposition of the variance**

The coagulation time for the \( i \)th treatment and \( j \)th trial within the treatment is \( Y_{ij} \) where \( i \) goes from 1 to \( k \) (\( k \) is the number of treatments/diets) and \( j \) goes from 1 to \( n_i \) (\( n_i \) is the number of trials for the \( i \)th treatment).

The coagulation time, \( Y_{ij} \), may be expressed as
\[ Y_{ij} = \bar{Y} + (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i) \]

Squaring both sides and summing over all trials and treatments, and after simplification,
\[
\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \bar{Y}^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2
\]

\[
S_{TOT} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2, \text{ is the Total Sum of Squares}
\]
\[
S_{AVG} = N \bar{Y}^2, \text{ is the Sum of Squares Due to the Mean}
\]
\[
S_B = \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y})^2, \text{ is the Between Treatment Sum of Squares.}
\]
\[
S_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2, \text{ is the Within Treatment Sum of Squares.}
\]

For our blood coagulation time example, \( S_{TOT} = 98644, S_{AVG} = 98304, S_B = 228, S_W = 112. \)

We observe that in fact \( S_{AVG} + S_B + S_W = S_{TOT} \)

Additionally, \( \nu_{TOT} = \nu_{AVG} + \nu_B + \nu_W \)
\( (24) = (1) + (3) + (20) \)

What we know can be summarized in an ANOVA Table

<table>
<thead>
<tr>
<th>Source of Var.</th>
<th>SS</th>
<th>D.of F</th>
<th>Mean Sqr.</th>
<th>F_{calc}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>98304</td>
<td>1</td>
<td>98304</td>
<td>17554.3</td>
</tr>
<tr>
<td>Between Treat.</td>
<td>228</td>
<td>3</td>
<td>76</td>
<td>13.57</td>
</tr>
<tr>
<td>Within Treat.</td>
<td>112</td>
<td>20</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>98644</td>
<td>24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the above table, the average mean square error (98304) is an estimate of \( \sigma_y^2 \) under the assumption that \( \mu_y = 0 \). The Between Treatment Mean Square Error (76) is an estimate of \( \sigma_y^2 \) under the assumption that
\[ \mu_A = \mu_B = \mu_C = \mu_D. \]

To test the hypothesis that \( \mu_A = \mu_B = \mu_C = \mu_D \), compare Between/Within Treatment Mean Square

\[ \frac{s_B^2}{s_W^2} = 13.57 = F_{\text{calc}} \]

to \( F_{v_1 = 3, v_2 = 20} = 3.10 \) and \( F(3,20,.99) = 4.94 \).

To test the hypothesis that \( \mu_y = 0 \), compare average and within treatment mean squares

\[ \frac{s_{AVG}^2}{s_W^2} = 17554.3 \] to \( F_{v_1 = 3, v_2 = 20} = 4.35 \) and \( F(1,20,.99) = 8.10 \).

So, the average is significant, so as are the treatments.

Earlier, we expressed the response as:

\[ Y_{ij} = \bar{Y} + (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i) \]

where \( \bar{Y} \) is the grand mean, \( (\bar{Y}_i - \bar{Y}) \) is treatment effect, and \( (Y_{ij} - \bar{Y}_i) \) is experimental error.

\[ y = \eta + \tau_i + \epsilon \]

\( \bar{y} \) is an estimate of \( \eta \). A model for the response is:

\[ \hat{y} = \hat{\eta} + \hat{\tau}_i \]

In blood coagulation experiment, we studied one variable, diet. We will now look at two variables (1 blocking variables and 1 treatment variable or 2 blocking variables).