One-way ANOVA Summary with Contrasts

```sas
data line;
    * SAS code;
    infile 'line.dat';
    input type strength;
    title 'FISHING LINE';
    if type = 1 then brand = 'Trilene XL';
    else if type = 2 then brand = 'Trilene XT';
    else if type = 3 then brand = 'Stren';
    run;

proc tabulate data=line;
    class brand;
    var strength;
    table brand, (n mean*strength var*strength*f=10.4 std*strength*f=10.4);
    run;

proc glm data=line;
    class brand;
    model strength = brand;
    run;
```

<table>
<thead>
<tr>
<th>brand</th>
<th>N</th>
<th>strength</th>
<th>variance</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stren</td>
<td>10.00</td>
<td>11.10</td>
<td>0.0133</td>
<td>0.1155</td>
</tr>
<tr>
<td>Trilene XL</td>
<td>10.00</td>
<td>11.52</td>
<td>0.0173</td>
<td>0.1317</td>
</tr>
<tr>
<td>Trilene XT</td>
<td>10.00</td>
<td>11.64</td>
<td>0.0160</td>
<td>0.1265</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>2</td>
<td>1.60800000</td>
<td>0.80400000</td>
<td>51.69</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Within</td>
<td>27</td>
<td>0.42000000</td>
<td>0.01555556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>2.02800000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square  Coeff Var  Root MSE  strength Mean
0.792899  1.092136  0.124722  11.42000  Overall Mean

Pooled SD/Mean % Pooled SD
Pooled SD = 1.09% of Mean $\sqrt{\text{MSE}}$
\[ y_{ij} = j^{th} \text{ value in } i^{th} \text{ treatment group.} \]
\[ i = 1, \ldots, t, \quad t = \# \text{ of treatment groups} \]
\[ j = 1, \ldots, n_i \quad n_i = \# \text{ of values for } i^{th} \text{ treatment} \]

### Treatment Group

For \( t = 3 \) treatment groups
\[
\begin{array}{c|ccc}
  j = 1 & i = 1 & i = 2 & i = 3 \\
  y_{11} & y_{21} & y_{31} \\
  y_{12} & y_{22} & y_{32} \\
  \vdots & \vdots & \vdots \\
  y_{1,n_i} & y_{2,n_i} & y_{3,n_i} \\
\end{array}
\]

Mean
\[
\begin{array}{c}
  \bar{y}_{1}, \quad \bar{y}_{2}, \quad \bar{y}_{3}, \\
  11.41 \quad 11.52 \quad 11.64 \quad \text{for fishing line} \\
\end{array}
\]

A dot (·) means sum or average over that index
\[
\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \quad \text{Mean of } i^{th} \text{ group.}
\]

\[ \bar{y}_i = \text{overall mean} = \text{average of all } y \text{ values} = \frac{\sum_i \sum_j y_{ij}}{\sum n_i} = \frac{\sum_i n_i \bar{y}_i}{\sum n_i} \]

\[ = \text{weighted average of } \bar{y}_i \text{ values weighted by } n_i \]

We want to test \( H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_t \) where \( t = \# \text{ treatment groups} \)

- Assuming
  - Independent data or residuals
  - Normal populations or residuals
  - Equal variances

### ANOVA Table Notation

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS = SS/df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>( t - 1 )</td>
<td>SSB</td>
<td>MSB</td>
<td>MSB/MSW</td>
</tr>
<tr>
<td>Within</td>
<td>( \sum(n_i - 1) )</td>
<td>SSW</td>
<td>MSW = ( S_w^2 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( \sum n_i - 1 )</td>
<td>TSS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Equivalent names used by SAS:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS = SS/df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>( t - 1 )</td>
<td>SSB</td>
<td>MSB</td>
<td>MSB/MSW</td>
</tr>
<tr>
<td>Error</td>
<td>( \sum(n_i - 1) )</td>
<td>SSW</td>
<td>MSW = ( S_w^2 = S_p^2 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( \sum n_i - 1 )</td>
<td>TSS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

df = degrees of freedom \quad SS = \text{Sum of Squares} \quad MS = \text{Mean Square}

To estimate the common \( \sigma^2 \), we pool sample variances.

\[
MSW = MS \text{ Within} = S_w^2 = S_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+(n_3-1)s_3^2}{(n_1-1)+(n_2-1)+(n_3-1)} = \frac{9(0.0133)+9(0.0173)+9(0.0160)}{27}
\]

A weighted average of \( S_i^2 \) sample variances weighted by \( df_i = n_i - 1 \).
SS Within = SS E = df_w * \( S_w^2 \) is also the sum of squared errors.
- Error = residual = \( y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i \).
- \( SS Within = SS Error = \sum (error_{ij})^2 = \sum (residual_{ij})^2 = \sum (y_{ij} - \bar{y}_i)^2 \)
- An important perspective, but harder way to calculate \( MSW = S_w^2 \)
- SS Within called SS Error in SAS ANOVA
- SS Within called SS Residual in regression

**SS Between = SS Treatments = \( \sum n_i (\bar{y}_i - \bar{y}_{..})^2 \)**

\[
= 10(11.1 - 11.42)^2 + 10(11.52 - 11.42)^2 + 10(11.64 - 11.42)^2
\]
- How much do the treatment means vary?
- How much do the different treatment groups' means differ from overall mean?
- SSB = 0 \( \Rightarrow \) All sample means the same.
- SSB is also the improvement in sum of squared errors using the treatment groups for fitted values compared to not knowing treatment groups.
- \( TSS = \) Total SS = sum of squared errors *without* using treatment group info for fitting.
  - Using \( \bar{y}_{..} \) for all fitted values.
  - Error = \( y_{i} - \bar{y}_{..} \).
- SSB = improvement in error SS by using treatment information
- \( SSB= SS \ error \ without \ treatment \ Info - SS \ error \ with \ treatment \ Info = TSS - SSW \)
- \( R^2 = \% \) or fraction improvement in error SS using treatment group info

\[
R^2 = \frac{SS \ error \ without \ treatment \ Info - SS \ error \ with \ treatment \ Info}{SS \ error \ without \ treatment \ Info}
\]

\[
= \frac{TSS - SSW}{TSS} = \frac{SSB}{TSS}
\]
- Fraction of Variability Explained by Treatments

Fraction of Variability Explained by Within Treatment Variability = \( \frac{SSW}{TSS} \)

- To test \( H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_t \)
  - \( F = \frac{MSB}{MS \ Treatments} = \frac{MSB}{MS \ Error} \)
  - Reject \( H_0 \) if \( F \) is too big.
    - Reject \( H_0 \) if the variation amongst the sample means is big relative to the within group variability.
  - Compare to \( F \) table with \( df_1 = \) numerator \( df = df_{Between} = df_{Treat} \) and \( df_2 = df_{Within} = df_{Error} \)
    - \( p-value = P(F > F_{Calculated}) \)

- To find ANOVA table from summary statistics
  - First find df
    - \( df \) Between group = \( df \) Treatments
      - We have \( t - 1 \) \( df \) for looking at variation between \( t \) treatment groups
    - \( df \) Within groups = \( df \) Error
      - We have \( (n_1 - 1) + (n_2 - 1) + \cdots \) \( df \) for pooling variances within each group
      - Pooling \( S_1^2, S_2^2, \ldots \)
Next find MS Within = MS Error
- Pooled variance
- Our pooled estimate of $\sigma^2$
  - Expected Value of MS Within = $E(\text{MS Within}) = \sigma^2$
  - $\text{MSW} = s^2_W = \frac{(n_1-1)s^2_1 + (n_2-1)s^2_2 + (n_3-1)s^2_3}{(n_1-1)+(n_2-1)+(n_3-1)}$

Next find SS Between group = SS Treatments
- $\text{SS Between} = \text{SS Treatments} = \sum n_i(\bar{y}_{i.} - \bar{y}_{..})^2$
- $\text{MS Between} = \frac{\text{SS Between}}{\text{df Between}} = \frac{\sum n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{t-1}$
- Since $\text{Var}(\bar{y}_{i.}) = \frac{\sigma^2}{n_i}$, if all groups have the same population means
  - Multiplying by $n_i$ makes $\text{MS Between}$ come out to about $\sigma^2$ if all groups have the same population means.
  - Expected Value of MS Between = $E(\text{MS Between}) = \sigma^2$ if $\mu_1 = \mu_2 = \ldots = \mu_t$

$F = \frac{\text{MS Between}}{\text{MS Within}} = \frac{\text{MS Treatments}}{\text{MS Error}}$
- If all groups have the same population means, then $F = \frac{\text{MS Between}}{\text{MS Within}} \approx \frac{\sigma^2}{\sigma^2} = 1$
- If F is too much bigger than 1.0, larger variation between treatment group means than expected if all groups have the same population means, then reject $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_t$

- Confidence Intervals and T-tests
  - For a linear combination, $l$, of means
    
    \[
    l = \sum_{i=1}^t a_i \mu_i = a_1 \mu_1 + a_2 \mu_2 + a_3 \mu_3
    \]
    
    \[
    \text{Estimate} = \hat{l} = \sum a_i \bar{y}_{i.} = a_1 \bar{y}_1. + a_2 \bar{y}_2. + a_3 \bar{y}_3.
    \]
    
    \[
    \text{Var(Estimate)} = a_1^2 \text{Var}(\bar{y}_1.) + a_2^2 \text{Var}(\bar{y}_2.) + a_3^2 \text{Var}(\bar{y}_3.) = a_1^2 \frac{\sigma^2}{n_1} + a_2^2 \frac{\sigma^2}{n_2} + a_3^2 \frac{\sigma^2}{n_3} = \sigma^2 \sum_{i=1}^t a_i^2
    \]
    
    \[
    \text{Estimate of } \sigma^2 = \text{MS Error} = \text{MS Within} = S^2_W = S^2_p
    \]
    
    \[
    \text{SE}_{\text{Estimate}} = \sqrt{\frac{a_1^2 \text{MS Error}}{n_1} + \frac{a_2^2 \text{MS Error}}{n_2} + \frac{a_3^2 \text{MS Error}}{n_3}}
    \]
    
    \[
    df \text{ for estimate of } \sigma = \frac{df_{\text{Within}} = df_{\text{Error}} = \sum_{i=1}^t (n_i - 1)}{t-1}
    \]
    
    \[
    \text{Two-sided Interval} = \text{Estimate} \pm t_{0.025, SE_{\text{Estimate}}}
    \]
    
    \[
    t-test \quad H_0: \text{Parameter} = \# \quad t = \frac{\text{Estimate} - \#}{\text{SE}_{\text{Estimate}}}
    \]
• Contrasts

\[ \frac{t^2_{df}}{\frac{t^2}{df}} = F_{1,df} \]

\[ t^2 = \frac{Estimate^2}{(SE_{Estimate})^2} = \frac{Estimate^2}{MS_{Error}} \]
\[ = \frac{Estimate^2}{\sum(a_i^2/n_i)} = \frac{MS_{Contrast}}{MS_{Error}} = \frac{SS_{Contrast}/1}{MS_{Error}} \]

The df for a single contrast is 1.

The SS for a contrast is \( t^2 \) except not divided by \( MS_{Error} \) yet.

\[ SS_{Contrast} = \frac{Estimate^2}{\sum(a_i^2/n_i)} = \frac{Estimate^2}{\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + \frac{a_3^2}{n_3} + \ldots} \]

• For a single contrast, the contrast tells us nothing more and even less that a confidence interval and p-value for the estimate.

• Since the contrast just gives us an F test with its p-value, not a confidence interval, we can leave fractions out of the contrasts.
  - \( H_0: \mu_{STREN} - \frac{\mu_{XL} - \mu_{XT}}{2} = 0 \) is the same as \( H_0: 2 \cdot \mu_{STREN} - \mu_{XL} - \mu_{XT} = 0 \)
  - In a SAS "estimate" statement, keep fractions.
  - A SAS "contrast" statement gives the same results with or without equivalent fractions.

• For SAS
  - For a single contrast, an estimate statement does everything a contrast statement does, plus more.
  - The one functionality a contrast statement has beyond an estimate statement is that we can test an \( H_0 \) that includes more than one contrast.
    - For example \( H_0: \mu_{XL} - \mu_{XT} = 0 \) and \( H_0: \mu_{STREN} - \mu_{XL} = 0 \).
    - In the F test, \( df_1 \) = the numerator df = number of linearly independent contrasts.
      - \( H_0: \mu_{XL} - \mu_{XT} = 0 \) and \( H_0: \mu_{STREN} - \mu_{XL} = 0 \)
      - contrast 'two contrasts' brand 0 1 -1, brand 1 -1 0;
      - With two linearly independent contrasts, the numerator df for this test is 2.
      - This is the same as the F test of \( H_0: \mu_{XL} = \mu_{XT} = \mu_{STREN} \).