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></th>
<th>Mean</th>
<th>Variance</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>strength</td>
<td>strength</td>
</tr>
<tr>
<td>brand</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stren</td>
<td>10.00</td>
<td>11.10</td>
<td>0.0133</td>
</tr>
<tr>
<td>Trilene XL</td>
<td>10.00</td>
<td>11.52</td>
<td>0.0173</td>
</tr>
<tr>
<td>Trilene XT</td>
<td>10.00</td>
<td>11.64</td>
<td>0.0160</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: 0.792899  Coeff Var: 1.092136  Root MSE: 0.124722  strength Mean: 11.42000  Overall Mean: 11.42000

Pooled SD/Mean %  Pooled SD
Pooled SD = 1.09% of Mean \(\sqrt{MSE}\)
$y_{ij} = \text{j}^{th} \text{ value in i}^{th} \text{ treatment group.}$

$i = 1, \ldots, t. \quad t = \# \text{ of treatment groups}$

$j = 1, \ldots, n_i \quad n_i = \# \text{ of values for } i^{th} \text{ treatment}$

<table>
<thead>
<tr>
<th>Treatment Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i = 1$</td>
</tr>
<tr>
<td>$y_{11}$</td>
</tr>
<tr>
<td>$y_{12}$</td>
</tr>
<tr>
<td>$\vdots$</td>
</tr>
<tr>
<td>$y_{1n_1}$</td>
</tr>
</tbody>
</table>

Mean

$\bar{y}_1, \bar{y}_2, \bar{y}_3.$

11.41 11.52 11.64 for fishing line

A dot (·) means sum or average over that index

$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}. \text{ Mean of } i^{th} \text{ group.}$

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

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 = $S^2_w$</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 = $S^2_w = S^2_p$</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 = Sum of Squares \quad MS = Mean Square

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

$$MSW = MS \text{ Within} = S^2_w = S^2_p = \frac{(n_1-1)s^2_1+(n_2-1)s^2_2+(n_3-1)s^2_3}{(n_1-1)+(n_2-1)+(n_3-1)} = \frac{9(0.0133)+9(0.0173)+9(0.0160)}{27}$$

A weighted average of $S^2_i$ sample variances weighted by df$_i = n_i - 1$. 
SS Within = SS $E = df_w \cdot S_w^2$ is also the sum of squared errors.

- Error = residual $= y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i \cdot}$.
- $SSWithin = SS Error = \sum (error_{ij})^2 = \sum (residual_{ij})^2 = \sum (y_{ij} - \bar{y}_{i \cdot})^2$
- 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 = $\sum n_i (\bar{y}_{i \cdot} - \bar{y}.)^2$**

- 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.
  - $TSS = \text{Total SS} = \text{sum of squared errors without using treatment group info for fitting.}$
    - Using $\bar{y}_{i \cdot}$ for all fitted values.
  - $SSB = \text{improvement in error SS by using treatment information}$
  - $SSB= SS error without treatment Info - SS error with treatment Info = TSS - SSW$
- $R^2 = \%$ or fraction improvement in error SS using treatment group info
  
  $R^2 = \frac{SS error without treatment Info - SS error with treatment Info}{SS error without treatment Info}$

  - $= \frac{TSS - SSW}{TSS} = \frac{SSB}{TSS}$
  - $\text{Fraction of Variability Explained by Treatments}$

  $\text{Fraction of Variability Explained by Within Treatment Variability} = \frac{SSW}{TSS}$

- To test $H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_t$
  - $MSB = \frac{MS Treatments}{df_{Between}}$
  - $MSW = \frac{MS Error}{df_{Within}}$
  - $F = \frac{MSB}{MSW}$
  - $\text{Reject } H_0 \text{ if } F \text{ is too big.}$
    - $\text{Reject } H_0 \text{ if the variation amongst the sample means is big relative to the within group variability.}$
  - $p-value = P(F > F_{Calculated})$

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

Next find SS Between group = SS Treatments
- \[ SS \text{ Between } = SS \text{ Treatments } = \sum n_i (\bar{y}_i - \bar{y}_{..})^2 \]
- MS Between = $\frac{SS \text{ Between}}{df \text{ Between}} = \frac{\sum n_i (\bar{y}_i - \bar{y}_{..})^2}{t-1}$
  - Since $Var(\bar{y}_i) = \frac{\sigma^2}{n_i}$, if all groups have the same population means
    - Multiplying by $n_i$ makes $MS \text{ Between}$ 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 $\mu_1 = \mu_2 = \ldots = \mu_t$

\[ F = \frac{MS \text{ Between}}{MS \text{ Within}} = \frac{MS \text{ Treatments}}{MS \text{ Error}} \]
- If all groups have the same population means, then $F = \frac{MS \text{ Between}}{MS \text{ Within}} \approx \frac{\sigma^2}{\sigma^2} = 1$
- 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 $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_t$