Concepts and Definitions

Objective: Identify factors responsible for variability in observed data

Specify one or more factors that could account for variability (e.g. location, time, etc.). Each factor is associated with a particular set of populations or treatments (e.g. particular sampling stations, sampling days, etc.). One-way analysis of variance (ANOVA) considers only a single factor.

Suppose a random sample \([x_{i1}, x_{i2}, ..., x_{ij}]\) is obtained for treatment \(i\). There are \(i = 1, ..., I\) treatments (e.g. each treatment may correspond to a different sampling location).

Arrange data in a table/array -- rows are treatments, columns are replicates:

\[
\begin{align*}
[x_{11}, x_{12}, ..., x_{1j}] \\
[x_{21}, x_{22}, ..., x_{2j}] \\
& \quad \vdots \\
[x_{I1}, x_{I2}, ..., x_{IJ}]
\end{align*}
\]

Here we assume each treatment has same number of replicates \(J\). The ANOVA procedure may be generalized to allow different number of replicates for each treatment.

Each random sample has a CDF \(F_{xi}(x_i)\). The different \(F_{xi}(x_i)\) are assumed identical except for their means, which may differ. Classical ANOVA also assumes that all data are normally distributed.

Each random variable \(x_{ij}\) is decomposed into several parts, as specified by the following one-factor model:

\[
x_{ij} = \mu_i + e_{ij} = \mu + a_i + e_{ij}
\]

- \(\mu_i = E[x_{ij}]\) is unknown mean of \(x_i\) (for all \(j\)).
- \(\mu = \) unknown grand mean (average of \(\mu_i\)).
- \(a_i = \mu_i - \mu\) = unknown deviation of treatment mean from grand mean (often called an effect).
- \(e_{ij}\) = random residual for treatment \(i\), replicate \(j\).

\(E[e_{ij}] = 0, \ Var[e_{ij}] = \sigma^2\), for all \(i, j\)
Objective is to estimate/test values of $a_i$'s, which are the unknown distributional parameters of the $F_{x_i}(x_i)$'s.

Formulating the Problem as a Hypothesis Test

If the factor does not affect variability in the data then all $a_i$'s = 0. Use hypothesis test:

$$ H_0: a_1 = a_2 = \ldots = a_I = 0 $$

It is better to test all $a_i$ simultaneously than individually or in pairs. Test that sum-of-squared $a_i$'s = 0.

$$ H_0: \sum_{i=1}^{I} a_i^2 = 0 $$

Derive a test statistic based on sums-of-squares of data.

Sums-of-Squares Computations

Define the sample treatment and grand means:

$$ m_{xi} = \frac{1}{J} \sum_{j=1}^{J} x_{ij} = \bar{x}_i. $$

$$ m_x = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij} = \bar{x}. $$

The total sum-of-squares $SST$ measures variability of $x_{ij}$ around $m_x$:

$$ SST = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - m_x)^2 $$

$$ = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - m_{xi})^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (m_{xi} - m_x)^2 $$

$$ = SSE + SSTr $$

$SST$ can be divided into error sum-of-squares $SSE$ and treatment sum-of-squares $SSTr$.

$SSE$ measures variability of $x_{ij}$ around $m_{xi}$, within treatments:
\[ \text{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - m_{xi})^2 \]

\[ \text{SSTr} \text{ measures variability of } m_{xi} \text{ around } m_x, \text{ across treatments:} \]

\[ \text{SSTr} = \sum_{i=1}^{I} \sum_{j=1}^{J} (m_{xi} - m_x)^2 \]

Error and treatment mean squared values:

\[ \text{MSE} = \frac{\text{SSE}}{I(J-1)} \]

\[ \text{MSTr} = \frac{\text{SSTr}}{I-1} \]

\[ E[\text{MSE}] = \sigma^2 \]

\[ E[\text{MSTr}] = \sigma^2 + \frac{J}{I-1} \sum_{i=1}^{I} a_i^2 \]

\[ \text{MSE} \text{ is an unbiased estimate of } \sigma^2, \text{ even if } a_i's \text{ are not zero.} \]
\[ \text{MSTr} \text{ is an unbiased estimate of } \sigma^2, \text{ only if all } a_i's \text{ are zero.} \]

Test Statistic

Use ratio \( \text{MSTr} / \text{MSE} \) as a test statistic:

\[ \mathcal{F} (\text{MSE}, \text{MSTr}) = \frac{\text{MSTr}}{\text{MSE}} \]

When \( H_0 \) is true and \( x_{ij} \)'s are normally distributed this statistic follows \( F \) distribution with \( \nu_T = I - 1 \) and \( \nu_E = I(J-1) \) degrees of freedom. Check normality by plotting \( (x_{ij} - m_{xi}) \) with \text{normplot}.

One-sided rejection region (rejects only if \( \text{MSTr} \) is large):

\[ R_0 : \mathcal{F} (\text{MSE}, \text{MSTr}) \geq F^{-1}_{\mathcal{F}, \nu_T, \nu_E} [\alpha] \]
One-sided \( p \)-value:

\[
p = 1 - F_{\nu_T,\nu_E} \left( F(MSE, MSTr) \right)
\]

Unbalanced ANOVA problems with different sample sizes for different treatments can be handled by modifying formulas slightly (see Devore, Section 10.3).

Single Factor ANOVA Tables

Above calculations are typically summarized in an ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>( F )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>( SSTr )</td>
<td>( \nu_T = I-1 )</td>
<td>( MSTr = SSTr / \nu_T )</td>
<td>( F = MSTr / MSE )</td>
<td>( p = 1 - F_{\nu_T,\nu_E}(F) )</td>
</tr>
<tr>
<td>Error</td>
<td>( SSE )</td>
<td>( \nu_E = I(J-1) )</td>
<td>( MSE = SSE / \nu_E )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( SST )</td>
<td>( \nu_T = IJ-1 )</td>
<td>( MST = SST / \nu_T )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example -- Effect of Season on Oxygen Level

Consider following set of dissolved oxygen concentration data \((x_{ij})\) obtained in 4 different seasons/treatments (rows), 6 replicates per season (columns):

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.62</td>
<td>6.12</td>
<td>6.62</td>
<td>6.21</td>
<td>7.08</td>
<td>5.36</td>
</tr>
<tr>
<td></td>
<td>7.70</td>
<td>8.31</td>
<td>8.80</td>
<td>8.24</td>
<td>7.87</td>
<td>7.44</td>
</tr>
<tr>
<td></td>
<td>2.52</td>
<td>5.44</td>
<td>4.94</td>
<td>2.99</td>
<td>4.39</td>
<td>4.44</td>
</tr>
<tr>
<td></td>
<td>6.77</td>
<td>6.65</td>
<td>6.01</td>
<td>6.26</td>
<td>7.09</td>
<td>6.05</td>
</tr>
</tbody>
</table>

Use a single factor ANOVA to determine if season has a significant impact on oxygen variability.

The MATLAB `anova1` function derives the error and treatment sums of squares and computes \( p \) value. **When using anova1 be sure to transpose the data array** (MATLAB requires treatments in columns and replicates in rows).

Results are presented in this standard single factor ANOVA table:
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS=SS/df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>47.1642</td>
<td>3</td>
<td>15.7214</td>
<td>29.8</td>
<td>1.4E-7</td>
</tr>
<tr>
<td>Error</td>
<td>10.5518</td>
<td>20</td>
<td>0.5276</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>57.716</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The very low $p$ value indicates that seasonality is **highly significant** in this case. Note that $MSTr$, which depends on the $a_i$’s, is much larger than $MSE$.

$$F \text{ CDF, } \nu_T = 3, \nu_E = 5$$