The random effects balanced 1-way ANOVA model

\[ y_{ij} = \mu + \tau_j + \epsilon_{ij} \]

with the error terms satisfying

\[ \epsilon_{ij} \text{ i.i.d. } N(0, \sigma^2) \]

for \( j = 1, 2, \ldots, k \) and \( i = 1, 2, \ldots, n_j \).

**Strategy for checking model assumptions:**

The following should be done in the given order:

1. Check the form of the model; i.e., is \( E(y_{ij}) = \mu + \tau_j \) correct?
2. Check for outliers
3. Check for independence
4. Check for constant variance
5. Check for normality

For all we will use the residuals

\[ \hat{\epsilon}_{ij} = y_{ij} - \bar{y}_j \]

or in effect, the standardized residuals defined by

\[ z_{ij} = \frac{\hat{\epsilon}_{ij}}{\sqrt{\frac{SSE}{N-1}}} \]
Residual plots are plots of the standardized residuals versus the levels of another variable (usually the factors).

How to check whether the model assumptions hold (see above for the numeric correspondence):

1. Plot $z_{ij}$ versus the levels of each independent variable. If the residuals exhibit a non-random pattern about zero in any such plot, the lack of fit is indicated.

2. If **all** model assumptions hold, then approximately 68% of $z_{ij}$ should be between -1 and 1; 95% should be between -2 and 2; 99.7% should be between -3 and 3. If not, run the analysis with and without the outlier and see how it affects the conclusions. If it does, the point is influential and should be reported.

3. The most likely cause of non-independence in the error terms is the similarity of experimental units close together in time or space. Plot $z_{ij}$ against the order in which they were collected or against any spatial arrangement of the corresponding experimental units. If there is no discernible pattern, the independence assumption holds. Also, you can carry out a formal test for independence for data collected over time using the Durbin-Watson test statistic:

$$DW = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_{ij} - \hat{\epsilon}_{i-1,j})^2}{\sum_{i=1}^{n} \hat{\epsilon}_{ij}^2}$$

It tests for lag 1 autocorrelation. The rule of thumb is as follows: If $DW \geq 1.7$, there is no lack of independence; if $DW \leq 1$, then there is.

In the case of lack of independence and when there is a clear trend in the residual plot (such as a linear trend), it may be advisable to add terms into the model to represent the time or space effect. For example, one can use the model

$$y_{ij} = \mu + \tau_j + \gamma t_{ij} + \epsilon_{ij}$$

where the errors are i.i.d. $N(0, \sigma^2)$ and $t_{ij}$ is the time at which the $i$th observation was recorded. The above is an analysis of covariance model.
4. If \( n_1 = n_2 = \ldots = n_k = n \), then the violation of the equal variance assumption does not have serious effects. But if \( n_l \neq n_j \) then the inference is off. To detect it, plot \( z_{ij} \) versus \( \hat{y}_{ij} \). The rule of thumb is: If

\[
\frac{s^2_{\text{max}}}{s^2_{\text{min}}} \leq 3
\]

the assumption holds. \( s^2_{\text{max}} = \max(s^2_1, s^2_2, \ldots, s^2_k) \) and \( s^2_{\text{min}} = \min(s^2_1, s^2_2, \ldots, s^2_k) \). When there is heteroskedasticity present, you can apply variance stabilizing transformations:

\( y^2, \sqrt{y}, \log(y), -1/\sqrt{y}, -1/y \). \( \sqrt{y}, \log(y) \) are normality inducing transformations. For Poisson \( y \)'s use \( \sqrt{y} \), and for binomial \( y \)'s use arcsin \( \sqrt{y} \).

5. Normal probability plots of the standardized residuals can be used to check the normality assumption.

For a SAS example, see handout 2.